Low-rank methods for heterogeneous and multi-source data

Geneviève Robin

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

NNT : 2019SACLX026

ÉCOLE POLYTECHNIQUE UNIVERSITÉ PARIS-SACLAY

THÈSE DE DOCTORAT

de

L'UNIVERSITÉ PARIS-SACLAY

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : Ecole polytechnique

Laboratoire d'accueil : Centre de mathématiques appliquées de Polytechnique, UMR 7641 CNRS

Spécialité de doctorat : Mathématiques appliquées

Geneviève ROBIN

Méthodes de rang faible pour l'analyse de données multi-sources, hétérogènes et incomplètes

Date de soutenance : 11 Juin 2019

Lieu de soutenance : Palaiseau

Après avis des rapporteurs :
PR. JÉRÉMIE BIGOT (Institut de Mathématiques de Bordeaux, Université de Bordeaux)
PR. TREVOR HASTIE (Statistics department, Stanford University)

Jury de soutenance :

PR. FRANCIS BACH	(SIERRA project-team, INRIA - DI ENS) Président
PR. GÉRARD BIAU	(LPSM, Sorbonne Université) Examinateur
PR. JULIE JOSSE	(CMAP, École Polytechnique) Codirectrice de thèse
PR. OLGA KLOPP	(ESSEC Business School) Examinatrice
PR. KARIM LOUNICI	(CMAP, École Polytechnique) Examinateur
PR. ÉRIC MOULINES	(CMAP, École Polytechnique) Codirecteur de thèse
PR. STÉPHANE ROBIN	(MIA, AgroParisTech - INRA) Examinateur

Fondation mathématique FMJH Jacques Hadamard

CMAP

Résumé

Dans les applications modernes des statistiques et de l'apprentissage machine, les praticiens sont encouragés à produire de plus en plus de données, ce qui conduit souvent à assouplir les techniques d'acquisition et à agréger des sources d'information diverses. En conséquence, les analystes sont confrontés à de nombreuses imperfections dans les données. En particulier, les données sont souvent *hétérogènes*, c'est-à-dire qu'elles combinent des informations quantitatives et qualitatives, *incomplètes*, avec des valeurs manquantes dues à des pannes de machines ou au phénomène de non-réponse, et *multi-sources*, lorsque les données résultent de l'agrégation de plusieurs sources de données. Cela remet souvent en question les cadres classiques des statistiques et de l'apprentissage machine, où la plupart des résultats théoriques sont généralement spécifiques à des données numériques et complètes.

L'objet de cette thèse est de développer de nouvelles méthodes d'analyse de telles données multi-sources, hétérogènes et incomplètes. En particulier, nous cherchons à prédire les données manquantes. Pour ce faire, nous développons un cadre complet basé sur des modèles bas rang dans des familles exponentielles hétérogènes. Par rapport aux travaux antérieurs, l'originalité de cette thèse est qu'elle aborde simultanément les phénomènes de données multi-sources, hétérogènes et incomplètes, tandis que la plupart des travaux existants ne traitent ces imperfections qu'une à la fois.

Les contributions de ce manuscrit consistent en deux grandes catégories de résultats. Dans les trois premiers chapitres, nous nous concentrons sur l'imputation des données de comptage, en présence d'informations secondaires constituées de caractéristiques quantitatives et qualitatives. Nous proposons un cadre combinant les modèles linéaires généralisés et la complétion de matrice dans la famille exponentielle, afin de tirer parti de l'information secondaire dans le processus d'imputation. Nous obtenons des garanties statistiques pour notre méthode, fournissons un logiciel libre avec un tutoriel, et des évaluations empiriques. Nous appliquons également la méthode à un problème d'écologie, et estimons la tendance temporelle de la taille de la population de trois espèces d'oiseaux d'eau.

Dans les chapitres 6, 7 et 8, nous nous concentrons sur l'imputation de données multi-sources, hétérogènes et incomplètes. Nous introduisons un cadre très général qui incorpore comme cas particuliers plusieurs exemples d'intérêt dans les applications. Encore une fois, nous fournissons une étude théorique approfondie, un logiciel libre ainsi qu'un tutoriel et des évaluations empiriques des méthodes proposées. Enfin, nous les utilisons pour imputer un sous-échantillon d'un registre médical concernant les traumatisés sévères traités dans plusieurs hôpitaux français.

Abstract

In modern applications of statistics and machine learning, the urge of producing more data often leads to relaxing acquisition techniques, and compounding diverse sources. As a results, analysts are often confronted to many data imperfections. In particular, data are often *heterogeneous*, i.e. combine quantitative and qualitative information, *incomplete*, with missing values caused by machine failures or by the nonresponse phenomenon, and *multi-source*, when the data result from the aggregation of several data sets. One of the most important characteristics of these data imperfections is probably that they often occur all together. This challenges the classical frameworks in statistics and machine learning, where most theoretical results are usually specific to numeric and complete data.

The subject of this dissertation is to develop new methods to analyze such multi-source, heterogeneous and incomplete data. In particular, we seek to predict the missing data. To do so, we develop a complete framework based on heterogeneous exponential family low-rank models. Compared to prior work, the originality of this dissertation is that it tackles multi-source, heterogeneous and incomplete data phenomenons simultaneously, while most existing work only handle these imperfections one at a time.

The contributions of this manuscript consist in two main classes of methods. In the first three chapters, we focus on the imputation of count data, in the presence of side information consisting of quantitative and qualitative features. We propose a framework combining generalized linear models and exponential family matrix completion, to take advantage of the side information in the imputation process. We derive statistical guarantees for our method, provide an open-source software with a tutorial, and empirical evaluations. We also apply the method to a problem in ecology, and estimate the temporal trend of the population size of three waterbird species

In Chapters 6, 7 and 8, we focus on the imputation of multi-source, heterogeneous and incomplete data. We introduce a very general framework which incorporates as special cases several examples of interest in applications. Again, we provide a thorough theoretical study, an open-source software along with a tutorial, and empirical evaluations of the proposed methods. Finally, we use them to impute a subsample of a severe trauma registry from French hospitals.

Remerciements

Éric et Julie, voici l'occasion de vous remercier du fond du cœur pour tout ce que vous m'avez transmis, scientifiquement et humainement, pendant ces années de thèse. Éric, merci de m'avoir appris la rigueur avec bienveillance, et de m'avoir fait entrevoir la richesse des mathématiques, merci également pour ta bonne humeur communicative et pour tes traits d'esprit. Julie, merci de m'avoir proposé des sujets passionants, accompagnée sur tous les plans, et pour la confiance que tu m'as accordée en me permettant de participer à ton enseignement et à tes collaborations, merci aussi pour tous les bons moments passés en dehors du travail, en Californie et ailleurs.

Je suis extrêmement reconnaissante envers Jérémie Bigot et Trevor Hastie d'avoir accepté de rapporter ce travail, ce dont je suis honorée. Trevor, thank you for accepting to review this manuscript, I am honored to count you in my jury; thank you also for your warm welcome in Stanford. Un grand merci également à Francis Bach, Gérard Biau, Olga Klopp, Karim Lounici et Stéphane Robin d'avoir accepté de faire partie de mon jury de thèse. Olga, merci d'avoir collaboré avec moi, de m'avoir fait profiter de ton expertise en complétion de matrices, et d'avoir été aussi impliquée : sans toi, une bonne partie de ce manuscrit n'existerait tout simplement pas. Stéphane, merci à toi et à Christophe de m'avoir initiée à la recherche pendant mon stage de master, et pour l'accueil chaleureux que je reçois toujours, chaque fois que je reviens au MIA.

Merci à tous ceux avec qui j'ai eu la chance de collaborer durant ma thèse: François Husson, Balasubramanian Narasimhan, Sylvain Sardy, Rob Tibshirani, Hoi To Wai. Naras, thank you for giving me the opportunity to visit you in Stanford, and for collaborating with me. Rob, thank you for sharing your ideas and for your support, I am honored to have worked with you. To, I am lucky to have worked and spent time with you in Paris, Boston and Montréal. Merci également à l'association Traumabase, et en particulier à Tobias Gauss et Sophie Hamada pour leur aide. Je souhaite enfin adresser un merci tout particulier à l'équipe de la Tour du Valat: Laura Dami, Clémence Deschamps, Pierre Defos du Rau, Elie Gaget, Jean-Yves Mondain-Monval et Marie Suet. Merci pour cette collaboration passionante, pour votre accueil plus que chaleureux, et de m'avoir permis d'échapper régulièrement à la grisaille parisienne pendant cette dernière année. Merci également d'avoir pris le temps de relire ce manuscrit et de m'avoir aidée à l'améliorer.

Merci à mes amis qui ont courageusement relu et corrigé des parties de ce manuscrit: Frédéric, Imke, Nicolas, Pierre, et Victor. Fred, merci pour ton amitié, les séances d'escalades, les bières, et d'avoir égayé mes journées au CMAP. Imke, merci pour ta gentillesse et pour ta compagnie apaisante. Nicolas, merci pour ta motivation au travail inspirante et pour ton soutien. Pierre, merci pour les nombreux moments musicaux, scientifiques et festifs passés ensemble, merci aussi de m'avoir accueillie chaleureusement à Telecom quand je n'avais pas le courage d'aller jusqu'à Saclay. Victor, merci pour une amitié totale, à travers la boue, les sciences, les questionnements existentiels, les

moments de joie et de tristesse.

Merci également à tous mes camarades du CMAP. Merci Aude pour avoir été un rayon de soleil dans le bureau et pour ton amitié, Belhal pour ta bonne humeur à toute épreuve, Florian de m'avoir tenu compagnie quand nous n'avions pas d'élèves en tutorat, Heythem pour ta motivation pour toutes les sorties doctorants et d'avoir corrigé mon arabe approximatif, Jaouad pour les bons moments passés à Montréal, Kevish pour ton enthousiasme pour les concerts et les soirées, Mathilde pour les séances d'escalade et de foot, Rémi d'avoir été le meilleur numéro 10 dans ma team, Ruben d'avoir animé les pauses café, et Wei d'avoir été une camarade de combat dans la bataille pour les bureaux. Merci à Antoine, Céline, Corentin et Corentin, Florian, Juliette, Léa, Matthieu, Martin, Nicolas, Paul, Paulin et Quentin pour avoir partagé avec moi de nombreux moments au CMAP. Merci aussi à Batiste et Kevin, qui m'ont fait une petite place dans leur bureau pour m'éviter de prendre le RER et permis passer de meilleures journées en leur compagnie. I would also like to thank all the people who welcomed me in Stanford, and with whom I spent happy moments in California: Augustin, Agatha, Claire, Håvard, Jonathan, Léo, Lorenzo, Mona, Nina and Theodor.

Je souhaite également remercier du fond du cœur mes amis avec qui j'ai la chance de passer mon temps libre. Zhor, merci d'être arrivée soudainement dans ma vie pour ne plus en partir, d'être une vraie partenaire, pour tes mots et marques d'encouragement tout au long de cette thèse. Un merci infini à Alexandre, Corentin, Emma, Guillaume, Hadrien, Lætitia, Louise, Noé, Radia, Raphaël, Sophie et Yanis pour leur façon de rendre la vie douce. Merci à Camille, Camille, Margaux et Natalie d'être des amies inconditionnelles.

أريد أيضا أن أشكر زملاء و معلمي دروس العربية على الوقت الأسبوعي الممتع و الجميل الذي سمح لي بنسيان الرياضيات.

Enfin, je remercie infiniment toute ma famille, pour qui je ne saurais exprimer convenablement ma gratitude. Merci à mes parents de m'avoir transmis leur curiosité scientifique et de m'avoir poussée et soutenue dans mes études, merci à ma maman, mon frère Paul et mes sœurs Bertille et Justine pour leur présence et pour leur soutien constant. Merci à Caligula de m'avoir apporté la tranquillité d'esprit nécessaire pour étudier. Pour finir, merci Mansour de m'avoir accompagnée au long de cette thèse, et de m'avoir encouragée par un savant mélange d'humour et de gentillesse.

Contents

List of Figures

List of Tables

Essential nomenclature

m_1	number of rows in a data frame
m_2	number of columns in a data frame
$[\![m_1]\!]$	set of integers $\{1, \ldots, m_1\}$
$\boldsymbol{Y}$	data frame in $\mathbb{R}^{m_1 \times m_2}$
$\boldsymbol{X}$	matrix in $\mathbb{R}^{m_1 \times m_2}$
$\boldsymbol{X}_{i,j}$	(i,j)-th entry of $\boldsymbol{X}$
$\boldsymbol{X}_{i,\cdot}$	i-th row of $\boldsymbol{X}$
$\boldsymbol{X}_{\cdot,j}$	j-th column of $\boldsymbol{X}$
$\boldsymbol{X}_{I,J}$	submatrix $\{(\boldsymbol{X}_{i,j}), i \in I, j \in J\}$
$\boldsymbol{X}^\top$	transpose of $\boldsymbol{X}$
$\langle \boldsymbol{X}, \boldsymbol{X}' \rangle$	scalar product of $\mathbb{R}^{m_1 \times m_2}$
$\|\boldsymbol{X}\|_F$	Frobenius norm of $\boldsymbol{X}$
$\|\boldsymbol{X}\|_*$	nuclear norm of $\boldsymbol{X}$ (the sum of singular values)
$\|\boldsymbol{X}\|$	operator norm of $\boldsymbol{X}$ (the largest singular value)
$\|\boldsymbol{X}\|_\infty$	infinity norm of $\boldsymbol{X}$ (the largest entry in absolute value)
$\|\boldsymbol{X}\|_0$	ℓ_0 norm of $\boldsymbol{X}$ (the number of nonzero entries)
$\|\boldsymbol{X}\|_1$	ℓ_1 norm of $\boldsymbol{X}$ (the sum of entries in absolute value)
$\mathrm{rank}(\boldsymbol{X})$	rank of $\boldsymbol{X}$ (the number of nonzero singular values)
α	vector
α_k	k-th entry of α
$\langle \alpha, \alpha' \rangle$	scalar product
$\|\alpha\|_2$	Euclidean norm of α
$\|\alpha\|_\infty$	infinity norm of α (the largest entry in absolute value)
$\|\alpha\|_0$	ℓ_0 norm of α (the number of nonzero entries)
$\|\alpha\|_1$	ℓ_1 norm of α (the sum of entries in absolute value)
i.i.d.	independent and identically distributed
Ω	set of observed entries $\Omega := \{(i,j) \in [\![m_1]\!] \times [\![m_2]\!]; \boldsymbol{Y}_{i,j}$ observed$\}$
$\Omega_{i,j}$	indicator of observed entries $\Omega_{i,j} = 1$ if $\boldsymbol{Y}_{i,j}$ and 0 otherwise
$\mathbb{1}_m$	vector of ones of length m
$\mathbb{1}_I$	indicator vector of set $I \subset [\![m]\!]$, $(\mathbb{1}_I)_k = 1$ if $k \in I$ and 0 otherwise

Chapter 1

Résumé à l'intention des non mathématiciens et mathématiciennes

Une donnée, dans son sens le plus général, est ce qui est connu, et peut être utilisé afin d'élaborer un raisonnement, ou de déterminer la solution à un problème (définition du dictionnaire Larousse). Les données sont donc en quelque sorte la matière première à partir de laquelle on peut produire de l'information. D'ailleurs, on associe souvent aux données l'adjectif *brutes*, pour indiquer qu'elles ne peuvent pas être utilisées telles quelles, mais doivent être explorées, analysées, découpées, réduites, pour en extraire une information compréhensible. Cet exercice de transformation de données brutes en information, nous le pratiquons tous, tous les jours, lorsque nous reconnaissons des visages, des objets ou des sons familiers à partir des données sensorielles qui parviennent à notre cerveau. L'activité professionnelle de certains d'entre nous est également fondée sur une expertise à extraire un certain type d'information d'un certain type de données. Ainsi, un médecin produit un diagnostic, c'est-à-dire une information compréhensible, à partir de résultats d'examens variés, qui représentent des données incompréhensibles pour qui n'a pas reçu de formation médicale.

Exemple introductif

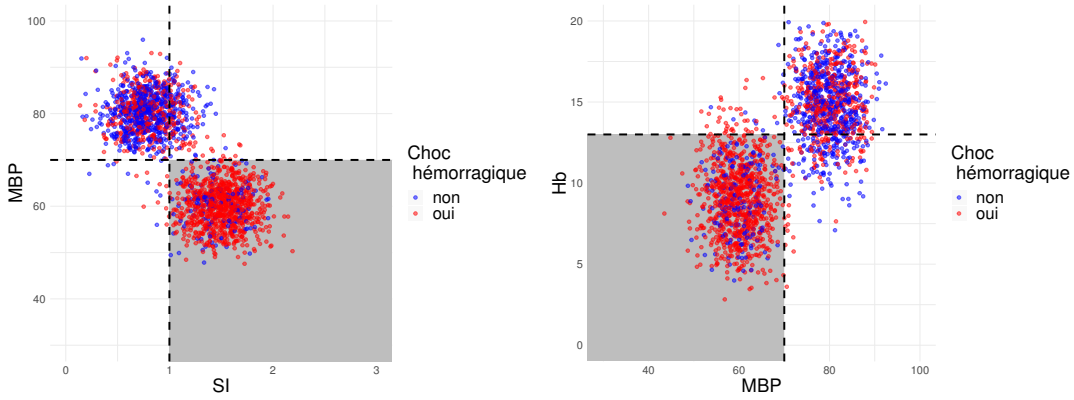
Prenons un exemple concret. Lorsqu'une personne subit un traumatisme grave, suite à une chute, un accident de voiture, etc., la principale cause de mortalité et morbidité est l'apparition d'un choc hémorragique (Hamada et al., 2018). Un choc hémorragique correspond à une forte diminution de la masse sanguine en circulation, qui entraîne à son tour une diminution de la distribution d'oxygène et une chute du débit cardiaque, et donc de potentielles séquelles, voire la mort (Cannon, 2018). La détection précoce du choc hémorragique permet donc de mieux traiter les patients concernés et, dans certains cas, d'éviter leur décès. Pour cela, des médecins ont mis en place un protocole qui consiste à mesurer sur les patients cinq facteurs associés au choc hémorragique. Ensuite, selon les valeurs observées, ils lancent, ou non, une alerte indiquant que le patient court un risque élevé de connaître un choc hémorragique. Les mesures effectuées concernent trois facteurs quantitatifs, c'est-à-dire des nombres dont une valeur particulièrement faible (ou particulièrement élevée selon le facteur regardé) correspond à un risque de choc hémorragique : l'indice de choc noté SI (le rapport du rythme cardiaque sur la pression

artérielle systolique), le taux d'hémoglobine mesuré au moment de l'accident (Hb), et la pression artérielle moyenne (calculée à partir des pressions artérielles systolique et diastolique), notée MBP. Ils mesurent également deux facteurs binaires, c'est-à-dire des questions auxquelles la réponse est soit "oui" soit "non" : le patient a-t-il été intubé ? Le patient souffre-t-il d'une fracture instable du bassin ?

En statistique, il s'agit d'un problème de *prédiction*, qui consiste à prédire un trait non observé, ici le choc hémorragique, à partir d'autres traits observés, ici l'indice de choc, le taux d'hémoglobine, la pression artérielle moyenne, les fractures du bassin et intubations éventuelles. Pour déterminer la procédure de prédiction, les médecins se basent sur des données déjà existantes. En effet, dans le passé, de nombreux patients ont été traités pour des traumatismes graves, leurs caractéristiques ont été mesurée, et l'on sait si oui ou non ils ont subi un choc hémorragique. Donc, les médecins vont chercher à déterminer si, parmi les patients déjà traités, ceux qui ont souffert d'un choc hémorragique se différencient des autres par leurs caractéristiques. Et en effet, pour chacun des cinq facteurs étudiés, ils ont identifié des valeurs associées à un plus grand risque de choc hémorragique : lorsque l'indice de choc SI est plus grand que 1, le taux d'hémoglobine Hb plus petit que 13g/dl (grammes par décilitre), la pression artérielle moyenne MBP est plus petite que 70mmHg (millimètres de mercure), lorsque le patient a été intubé, et lorsqu'il ou elle souffre d'une fracture instable du bassin. De plus, ils ont identifié que le fait de lancer une alerte lorsqu'au moins deux de ces conditions étaient vérifiées en même temps permettait de détecter 85% des patients souffrant finalement d'un choc hémorragique. En somme, à partir des résultats de nombreux examens effectués sur les patients souffrant de traumatismes sévères, ils produisent une information simple et pertinente : une alarme qui se déclenche lorsque le patient a un risque élevé de choc hémorragique.

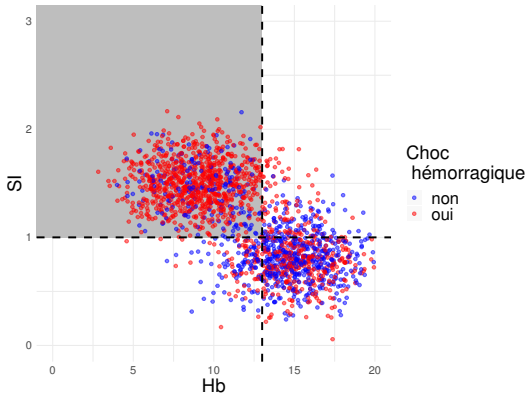
Ainsi, lorsqu'un nouveau patient arrive, et pour lequel on ne sait pas encore si un choc hémorragique va se déclencher, il suffit de vérifier si ses caractéristiques satisfont les conditions pour le déclenchement de l'alerte. Pour cela, on peut en particulier utiliser des outils visuels, qui consistent à représenter chaque individu sur des cartes, où les coordonnées correspondent aux facteurs quantitatifs mesurés (Hb, MBP, SI). Selon la position des individus dans ces espaces, l'alerte sera ou non déclenchée. De telles cartes sont représentées en Figure 1.1 pour des données synthétiques, c'est-à-dire qui ne correspondent pas à des patients réels. Dans la Figure 1.1, chaque point correspond à un (faux) individu. La couleur bleue indique que l'individu n'a pas eu de choc hémorragique, la couleur rouge indique la présence d'un choc hémorragique. Les quadrants gris indiquent les zones dans lesquelles les individus satisfont les conditions pour le déclenchement d'une alerte, c'est-à-dire où deux des critères détaillés plus haut (SI≥ 1, Hb≤ 13, MBP≤ 70) sont vérifiés.

De manière similaire, il existe des outils visuels permettant de représenter les informations binaires (fracture du bassin instable et intubation), comme le montre la Figure 1.2. Dans ces deux graphiques, chaque point correspond à un individu. La position le long de l'axe des abscisses indique la valeur prise par un facteur quantitatif (SI ou MBP) et la forme (triangle ou rond) indique la valeur prise par une variable binaire (fracture du bassin instable et intubation). La couleur bleue indique que l'individu n'a pas eu de choc hémorragique, la couleur rouge indique la présence d'un choc hémorragique. Les demi-espaces gris indiquent les zones dans lesquelles les individus satisfont le critère concernant la variable quantitative. L'alerte est déclenchée pour les individus correspondant aux triangles situés dans les zones grises. Comme indiqué dans l'étude originale (Hamada et al., 2018), en déclenchant l'alerte pour tout nouveau patient qui se trouve



(a) Pression artérielle moyenne (MBP) en fonction de l'indice de choc (SI).

(b) Taux d'hémoglobine (Hb) en fonction de la pression artérielle moyenne (MBP).

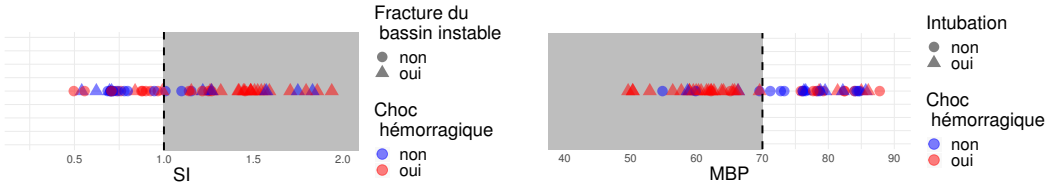


(c) Indice de choc (SI) en fonction du taux d'hémoglobine (Hb).

Figure 1.1: Carte en deux dimensions des individus (points) selon les valeurs de leurs trait quantitatifs. Les points bleus indiquent les individus n'ayant pas eu de choc hémorragique, les points rouges les individus ayant eu choc hémorragique. Les quadrants gris indiquent les zones dans lesquelles les individus satisfont les conditions pour le déclenchement d'une alerte.

dans l'un des quadrants gris de la Figure 1.1 ou correspondant à un triangle dans une zone grise de la Figure 1.2 (toutes les configurations ne sont pas représentées), permet de détecter 85% des chocs hémorragiques.

En réalité, l'exemple présenté ci-dessus repose sur trois hypothèses extrêmement importantes, qui ne sont pas vérifiées dans tous les problèmes. Premièrement, il suppose que l'on dispose d'information a priori sur le phénomène que l'on cherche à décrire : les cinq facteurs considérés ne sont pas pris au hasard, ils correspondent à des données dont les médecins savent qu'elles sont liées au choc hémorragique. Dans de nombreux exemples, en médecine et dans d'autres champs scientifiques, on ne dispose pas de telles informations a priori. Ainsi, pour certaines maladies moins bien connues que les traumatismes sévères, on ne connaît pas forcément de facteurs, ou indicateurs, qui puissent donner des informations sur l'apparition ou la progression de la maladie. Deuxièmement, il suppose que l'on dispose de données sur des individus ayant été traités pour des traumatismes graves et des chocs hémorragiques dans le passé. C'est la disponibilité de ces données qui permet de déterminer les critères précis (Hb≤ 13, MBP≤ 70, SI≥ 1) à regarder pour prédire le choc hémorragique, en fonction des valeurs prises dans le passé par ces traits pour les patients souffrant de choc hémorragique d'une part, et les patients



(a) Indice de choc (SI) et fracture du bassin instable (triangle : oui, ronds : non).

(b) Pression artérielle moyenne (MBP) et intubation (triangle : oui, ronds : non).

Figure 1.2: Représentation des individus selon un facteur quantitatif (SI ou MBP) et un facteur qualitatif (fracture du bassin instable ou intubation). Les individus correspondant aux triangles situés dans les zones grises satisfont les conditions de déclenchement de l'alerte.

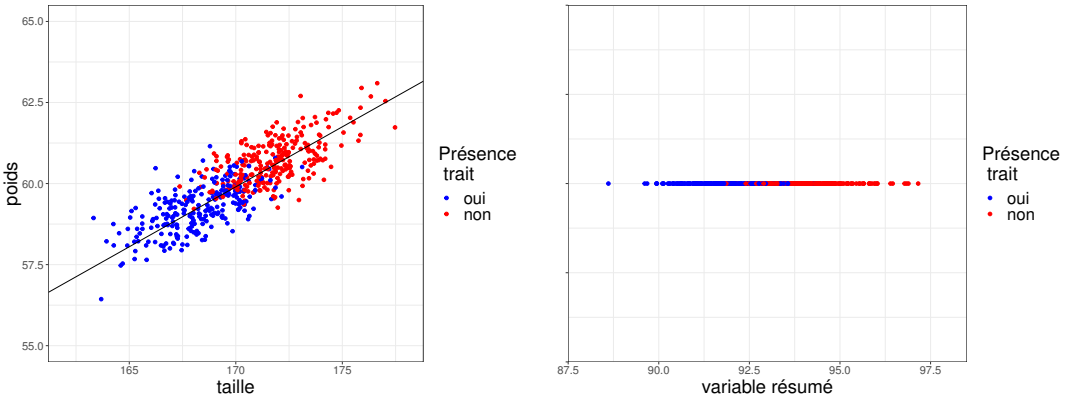
ne souffrant pas de choc hémorragique d'autre part. Troisièmement, lorsqu'un nouveau patient arrive, afin de déterminer si la procédure doit être déclenchée ou non, il est nécessaire de disposer d'un *modèle*, ou d'un outil d'interprétation, tel que les représentations graphiques des Figures 1.1 et 1.2. Ce modèle permet de synthétiser les données récoltées, et de savoir *où* se situe le nouveau patient dans l'espace caractérisé par les traits concernés (Hb, MBP, SI, fracture du bassin, intubation).

Réduction de la dimensionnalité

Cependant, dans la plupart des problèmes modernes en analyse de données, ces trois hypothèses ne sont pas, ou seulement partiellement, vérifiées. D'abord, souvent, les scientifiques ou analystes s'intéressent à un phénomène, tel que le développement d'une maladie, l'évolution temporelle de la taille de population d'espèces, etc. pour lesquels ils ne connaissent pas d'indicateur, variable, trait ou facteur, qui puissent les informer sur la question posée : cela remet en question la première hypothèse. Par exemple, en étude génomique du cancer, où l'on cherche à identifier des gènes potentiellement reliés au développement d'un cancer, on dispose en général de mesures faites pour des milliers, voire des millions de gènes, mais on ne sait pas lesquels sont pertinents pour le type de cancer étudié. Il est évidemment impossible à un médecin d'émettre un avis à partir de millions de mesures dont il ne sait pas lesquelles sont importantes. De même, en écologie, lorsque l'on cherche à surveiller une espèce (menacées, par exemple), on dispose parfois d'une grande quantité d'information géographiques et météorologiques sur son habitat et les activités anthropiques qui l'entoure, mais on ne sait pas toujours lesquelles ont véritablement une influence sur la taille des populations. Cela conduit au problème que l'on appelle en statistique la *réduction de la dimensionnalité* : à partir d'un grand nombre de variables (informations physiologiques sur un patient, météorologique sur un certain site environnemental, etc.), peut-on produire un petit nombre d'entre elles qui soient associées au phénomène auquel on s'intéresse (maladie, surveillance d'espèce, etc.) ?

Ce problème est illustré en Figure 1.3 à l'aide d'un exemple fictif. Dans cet exemple,

on s'intéresse à trois caractéristiques dans une population d'individus: la taille, le poids, et à un trait qualitatif dont on cherche à distinguer les individus qui le possèdent de ceux qui ne le possèdent pas. Ce trait peut correspondre, par exemple, à la présence d'une maladie, ou d'un symptôme particulier. Dans la figure Figure 1.3a, on représente les individus (un point par individu) dans un plan où les coordonnées correspondent à la taille pour l'abscisse et au poids pour l'ordonnée. La couleur du point indique la présence du trait, avec du bleu lorsque l'individu possède le trait particulier (maladie, etc.), et du rouge sinon. Cette représentation est en deux dimensions géométriques, puisque chaque individu est caractérisé par sa taille d'une part, et son poids d'autre part. Cependant, il est très clair visuellement que le nuage de point suit une structure particulière, avec une relation linéaire entre le poids et la taille. La droite noire représentée en Figure 1.3a représente cette relation linéaire, et indique la direction selon laquelle on observe le plus de *variabilité* entre les individus : le nuage de points est le plus "étalé" le long de cette direction. De plus, le long de cette droite, on observe que les individus possédant le trait considéré (les points bleus), se trouvent en amont, c'est-à-dire plus proches de l'origine correspondant à un poids et une taille de zero, que les individus ne possédant pas ce trait (les points rouges). Cela conduit à la réflexion suivante : il n'est pas nécessaire de renseigner le poids et la taille de chaque individu, mais seulement leur position le long de cette droite. Cela revient à créer une "variable résumé", qui rassemble à l'information importante contenue dans le poids, et l'information importante contenue dans la taille. Ce type de transformation est dans le même esprit que la constructions de l'indice de masse corporelle (IMC) par exemple. Plutôt que de conserver deux nombres (poids et taille) pour chaque individu, on construit une troisième variable qui les résume. Ainsi, on obtient le graphe *unidimensionnel* de la Figure 1.3b, où chaque individu est placé le long d'une droite correspondant à la variable résumé. Il s'agit bien d'une réduction de la dimensionnalité, puisque l'on est passé d'une représentation en deux dimensions, à une représentation en une seule dimension.



(a) Représentation des individus en fonction de leur poids et leur taille.

(b) Représentation des individus selon leur positionnement sur la droite noire de la Figure 1.3a.

Figure 1.3: Réduction de la dimensionnalité pour une collection de données concernant le poids, la taille, et la présence d'un trait d'intérêt (maladie, etc.) dans un échantillon d'individus. Chaque point représente un individu, et la couleur indique la présence du trait étudié (maladie, etc.). Les individus bleus possèdent le trait, les individus bleus ne possèdent pas le trait. La droite en noir sur la Figure 1.3a indique la direction du plan selon laquelle on observe le plus de variabilité.

Dans cet exemple simple, l'utilité d'une telle réduction de la dimensionnalité n'est

pas évidente, puisque le graphique 1.3a est déjà relativement interprétable tel quel. Cependant, elle apparaît clairement lorsque l'on analyse des données où de nombreuses variables ont été mesurées sur chaque patient. Considérons par exemple un jeu de données publiques provenant d'une étude médicale sur le diabète chez les indiennes Pima (Newman et al., 1998) disponible dans le package R `mlbench` (Leisch and Dimitriadou, 2010). Un extrait du jeu de données est présenté en Table 1.1. Les variables mesurées sont le nombre de fois où la personne est tombée enceinte, le taux de glucose, la pression artérielle diastolique, l'épaisseur du pli cutané du triceps, l'insuline sérique, l'indice de masse corporelle, la fonction pédigrée du diabète (information sur les antécédents familiaux concernant le diabète), l'âge, et enfin, la présence du diabète. Une question possible est de savoir si ces variables mesurées sont liées au diabète, et donc permettrait par exemple de prédire, pour une nouvelle personne, ses chances d'être diabétique. L'un des problèmes rencontrés dans l'analyse de ces données est qu'en raison du nombre de variables (sept caractéristiques quantitatives et un trait binaire, le diabète), il est difficile de visualiser les individus dans l'espace comme dans les exemples précédents.

	enceinte	glucose	pression	triceps	insuline	masse	pedigree	age	diabetes
1	6	148	72	35	0	33.6	0.63	50	pos
2	1	85	66	29	0	26.6	0.35	31	neg
3	8	183	64	0	0	23.3	0.67	32	pos
4	1	89	66	23	94	28.1	0.17	21	neg
5	0	137	40	35	168	43.1	2.29	33	pos
6	5	116	74	0	0	25.6	0.2	30	neg

Table 1.1: Extrait d'un jeu de données concernant le diabète chez les indiennes Pima.

Cependant, à l'aide de méthodes de réduction de dimensionnalité, on peut produire, à partir des septs variables quantitatives mesurées, un petit nombre de variables résumé, par exemple deux ou trois, qui peuvent à leur tour être utilisées pour visualiser les données en deux ou trois dimensions. La méthode de réduction de dimensionnalité la plus ancienne et la plus répandue est l'analyse en composante principales, ou ACP, (Pearson, 1901; Hotelling, 1933), qui consiste à chercher des directions orthogonales selon lesquelles les points sont les plus variables (comme la droite en Figure 1.3a). Une telle représentation est montrée en Figure 1.4, où chaque femme présente dans le jeu de données est indiquée par un point dont les coordonnées sont les valeurs que prend chaque personne pour de nouvelles "variables résumé" calculée en utilisant l'ACP à partir des sept variables quantitatives données dans la Table 1.1.

Une telle représentation présente plusieurs intérêts. D'abord, elle facilite l'analyse et l'interprétation en permettant de visualiser les données dans l'espace plutôt que sous la forme d'un tableau (cf. Table 1.1) : elle permet de confirmer que les variables mesurées sont bien liées au diabète, puisque l'on observe clairement que les femmes atteintes de diabètes sont séparées de celles ne souffrant pas de diabète dans cet espace. Ensuite, elle a un intérêt computationnel, puisqu'il suffit ensuite de garder en mémoires les coordonnées des individus selon trois axes, au lieu des sept variables initiales; cela allège également le coût des calculs. Ce dernier point est encore plus important dans les cas où des milliers voire des millions de caractéristiques sont mesurées sur chaque personne, comme en génomique par exemple. Notons que l'analyste a la possibilité de choisir le nombre de dimensions qu'il veut conserver, deux, trois ou plus selon ses besoins, et tant que ce nombre ne dépasse pas le nombre de variables initial. Le nombre

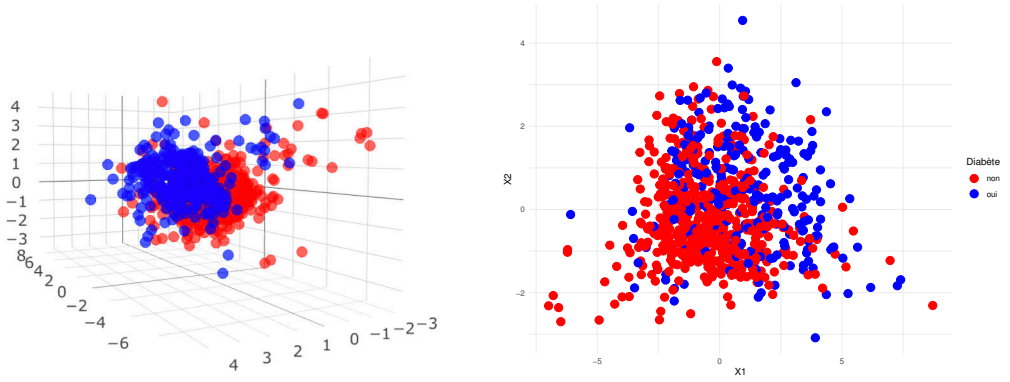


Figure 1.4: Représentation en deux (droite) et trois (gauche) dimensions du jeu de données sur le diabète des indiennes Pima à l'aide de l'ACP.

de dimensions conservées s'appelle le *rang*. Dans cette thèse, nous développons des méthodes dites de *rang faible*, c'est-à-dire des méthodes qui permettent de produire un petit nombre de nouvelles variables importantes à partir d'un grand nombre de variables initiales, afin de visualiser et analyser les données plus facilement.

Données multi-sources

Depuis environ une dizaine d'années, et dans de nombreux champs scientifiques, la quantité et la complexité des données disponibles, entraînée par le développement de techniques d'acquisitions de plus en plus efficaces et de moins en moins coûteuses, a explosé. En parallèle, les méthodes d'apprentissage automatique, qui apprennent à résoudre des tâches à partir de données, par exemple à prédire la présence d'une maladie à partir de caractéristiques des individus, se sont développées rapidement. En particulier, l'efficacité de ces méthodes d'apprentissage grandit avec le nombre de données disponibles. Par exemple, supposons que l'on cherche à prédire si des individus souffrent d'une maladie, à partir de certaines de leurs caractéristiques, et à partir de données concernant des patients déjà diagnostiqués. Dans ce cas, l'erreur de prédiction, c'est-à-dire le nombre de patients pour lequel on prédit qu'ils sont malades alors qu'ils ne le sont pas, ou l'inverse, va avoir tendance a diminuer avec le nombre de patients déjà diagnostiqués inclus dans l'étude. Ainsi, la Figure 1.5 représente l'erreur de prédiction d'une méthode d'apprentissage en fonction du nombre d'individus pour lesquels le diagnostic est déjà connu : cette erreur diminue lorsque le nombre d'individus augmente, pour se stabiliser à une erreur minimale se trouvant autour de 10%.

Pour cette raison, les analystes sont poussés à chercher à augmenter la quantité de données disponibles, afin de fournir de meilleures résultats. Cela passe, par exemple, par la mise en commun de jeux de données. Ainsi, il est fréquent dans les études médicales ou sociales, que les individus proviennent de plusieurs hôpitaux, plusieurs écoles, plusieurs villes, etc. C'est ce que l'on appelle des données *multi-sources*. De façon générale, la mise en commun des données doit permettre d'améliorer les performances des méthodes d'apprentissage, en accroissant les bases de données. Cependant, il peut arriver que cette mise en commun détériore en réalité les résultats obtenus, en particulier lorsque les différentes sources (hôpitaux, écoles, etc.) présente une forte hétérogénéité. Par exemple, en raison de disparités géographiques ou démographiques, il arrive que les individus provenant de deux hôpitaux distincts aient des caractéristiques très différentes.

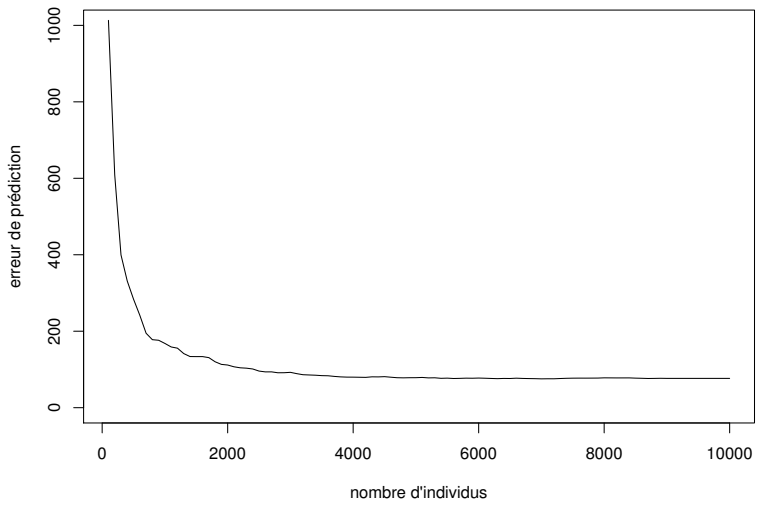


Figure 1.5: Erreur de prédiction d'une méthode d'apprentissage en fonction du nombre d'individus inclus dans la base de données d'entraînement.

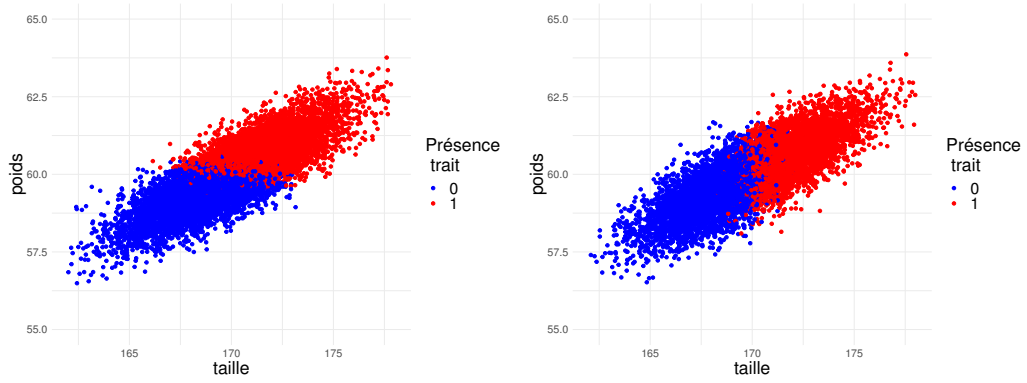


Figure 1.6: Représentations en deux dimensions d'individus en fonction de leur taille et leur poids, pour des individus issus de deux groupes différents (deux sources). La couleur indique la présence d'un trait binaire d'intérêt (la présence d'une maladie par exemple).

Dans ce cas, chercher à décrire les deux populations dans une seule et même analyse peut mener à l'échec. Pour le voir, reprenons l'exemple du jeu de données contenant le poids et la taille d'individus, ainsi qu'un trait binaire (maladie par exemple). Cependant, cette fois, nous considérons deux jeux de données correspondant à deux hôpitaux différents, et nous introduisons de l'hétérogénéité entre les individus des deux hôpitaux, comme présenté en Figure 1.6, où la présence du trait binaire ne se répartit pas de la même façon entre les individus du premier hôpital (graphique de gauche), et ceux du deuxième hôpital (graphique de droite).

À partir de ces deux jeux de données, on peut comparer les résultats d'une méthode d'apprentissage (une régression logistique) entraînée sur chacun des jeux de données séparément, ou sur la combinaison des deux. On obtient les résultats suivants : en entraînant la méthode séparément sur les deux jeux de données, le bon diagnostic est effectué dans 99.2% des cas pour le premier jeu de données (figure de gauche), et dans 96.8% des cas pour le deuxième jeu de données (figure de droite); en revanche lorsque la méthode est entraînée sur la mise en commun des jeux de données, on obtient le

bon diagnostic dans seulement 92.4% des cas. Ce résultat très simple montre que, si l'aggrégation de données permet en théorie d'obtenir des méthodes plus performantes, il est crucial de prendre en compte l'hétérogénéité des différentes sources de données. Les méthodes de rang faible développées dans cette thèse sont conçues pour prendre en compte ces différences possibles entre plusieurs sources de données, afin de pouvoir les agréger et bénéficier de davantage de données, sans risquer de détériorer les performances si les sources sont trop différentes.

Données hétérogènes

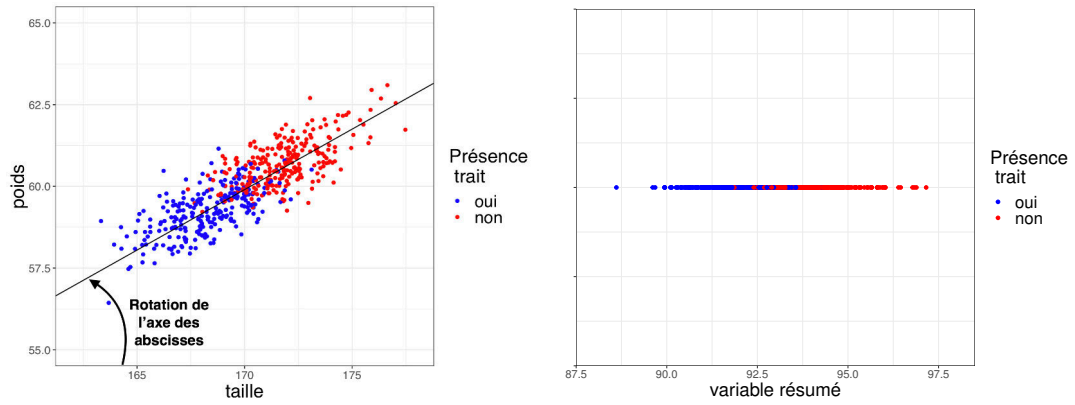
Jusqu'ici nous avons présenté des exemples où les variables utilisées dans la réduction de dimensionnalité étaient des variables *quantitatives*. En effet, nous avons regardé des traits binaires tels que la présence d'un choc hémorragique ou de diabète, mais nous cherchions à prédire ces traits, et les "variables résumé" étaient calculées uniquement à partir des variables quantitatives. Or, dans plusieurs applications d'intérêt pour cette thèse, les données contiennent des variables quantitatives *et* qualitatives, que l'on cherche à analyser simultanément. Par exemple, le jeu de données Traumabase (`http://www.traumabase.eu/fr_FR`), contient des informations concernant des patients polytraumatisés, dont des informations quantitatives telles que le temps passé en réanimation, mais aussi des informations qualitatives, telles que le type d'accident. Par ailleurs, les données sont *multi-sources*, puisque les patients proviennent de plusieurs hôpitaux français (Bicêtre, Pitié Salpétrière, etc.). Le jeu de données contient aussi de nombreuses données manquantes, indiquées par "NA" : ce point sera détaillé dans la section suivante.

Centre	Radio poumons	Radio bassin	Accident	Temps en réa (h)
Bicêtre	NA	Normal	Chute d'une hauteur	NA
HEGP	NA	NA	Chute d'une hauteur	2
Pitié Salpétrière	NA	NA	Accident piéton-voiture	NA
Lille	Normal	NA	Chute d'une hauteur	2
Beaujon	NA	NA	Chute de sa hauteur	NA
Lille	NA	NA	Chute de sa hauteur	NA

Table 1.2: Extrait de jeu de données Traumabase.

Ce type de données contenant un mélange d'informations quantitatives et qualitatives est qualifié d'*hétérogènes*, ou de *mixtes* (Pagès, 2004). Le caractère hétérogène des données est essentiel, car la plupart des méthodes visant à réduire la dimension des données sont *géométriques*, et ne peuvent pas être appliquées directement à des données qualitatives. Considérons l'exemple de la Figure 1.7, déjà analysée plus haut.

L'analyse en composante principales permet de trouver une "variable résumé"–la droite noire de la Figure 1.7a–à partir de rotations des variables initiales. C'est-à-dire que, partant de la Figure 1.7a, on cherche à tourner les axes des abscisses et des ordonnées, jusqu'à trouver la direction selon laquelle les points sont les plus variables (les plus étalés). Cela conduit au graphe Figure 1.7b, on l'on n'a plus qu'une seule dimension quantitative (l'axe des abscisses de la Figure 1.7b), et une information qualitative : la présence d'un certain trait, indiquée par la couleur des points. Supposons qu'on veuille encore réduire la dimensionnalité de nos données et ne plus garder, pour chaque point, qu'une seule caractéristique qui résume à la fois la variable quantitative (qui résume déjà le poids et la taille), et la couleur des points. On ne peut pas appliquer



(a) Représentation des individus en fonction de leur poids et leur taille.

(b) Représentation des individus selon leur positionnement sur la droite noire de la Figure 1.3a.

Figure 1.7: Réduction de la dimensionnalité pour une collection de données concernant le poids, la taille, et la présence d'un trait d'intérêt (maladie, etc.) dans un échantillon d'individus. Chaque point représente un individu, et la couleur indique la présence du trait étudié (maladie, etc.). Les individus bleus possèdent le trait, les individus bleus ne possèdent pas le trait. La droite en noir sur la figure de gauche indique la direction du plan selon laquelle on observe le plus de variabilité.

directement la même méthode que pour passer de Figure 1.7a à Figure 1.7b. En effet, la notion de rotation est fondamentalement géométrique, et ne s'applique pas à un trait binaire tel que la couleur bleue ou rouge. Ainsi, pour pouvoir appliquer des méthodes de réduction de dimensionnalité à des données hétérogènes, il est nécessaire de développer des alternatives à celles qui existent déjà pour les données quantitatives. De telles méthodes ont été proposé dans le passé, mais elles souffrent de certaines limites, soit car elles ne prennent pas en compte le caractère *multi-source* des données, soit parce qu'elle ne disposent pas de garanties théoriques. Dans cette thèse, nous proposons des méthodes de réduction de dimension, qui s'adaptent à des données multi-sources et hétérogènes, et nous prouvons plusieurs résultats théoriques, qui garantissent l'efficacité des méthodes, sous des conditions qui sont vérifiées dans la plupart des cas qui nous intéressent.

Données incomplètes

Pour définir les contours du sujet de cette thèse, il reste à discuter le phénomène des données *incomplètes*, ou *manquantes*. En effet, comme on l'a déjà entrevu avec l'exemple de la Traumabase en Table 1.2, il est courant, dans les applications modernes de l'apprentissage statistique, qu'une partie des données nécessaires pour l'application des méthodes classiques ne soit pas disponible. Par exemple, dans la Table 1.2, on voit que certaines caractéristiques n'ont pas été mesurées pour certains patients; c'est ce qu'indiquent les cases contenant "NA" (Non Applicable). Ces cases non remplies sont appelées *données manquantes*, et les jeux de données contenant de telles données manquantes, des *données incomplètes*. Ce phénomène apparaît dans la quasi-totalité des applications des statistiques et de l'apprentissage. Par exemple, en médecine, et particulièrement dans la médecine d'urgence telle que le traitement des traumatisés sévères, il est rares que toutes les mêmes mesures puissent être effectuées sur tous les patients. En effet, le manque de temps ou la gravité de l'état du patient conduit souvent à ne pas ren-

Figure 1.8: Deux canards milouin (*Aythya ferina*).

seigner l'intégralité des caractéristiques supposées être contenur dans la base de données. En statistiques sociales, lorsque l'on interroge des individus à l'aide de questionnaires, il arrive très régulièrement que certains d'entre eux ne souhaitent pas répondre à toutes les questions, en particulier lorsque le questionnaire contient des sujets sensibles comme les revenus, la consommation de drogues ou les pratiques sexuelles. Enfin, il arrive que certaines données ne soient tout simplement pas accessibles.

Par exemple, en écologie, afin de surveiller la taille des populations d'espèces, des écologues et des personnes volontaires se déplacent sur différents sites, pour compter les individus présents. Ainsi, dans le cadre de la surveillance des oiseaux d'eau migrateurs en Afrique du Nord, plusieurs instituts dont le Mediterranean Waterbirds Network (MNW), Groupe de Recherche pour la Protection des Oiseaux au Maroc/BirdLife Morocco, la Direction Générale des Forêts (Algérie), l'Association "les Amis des Oiseaux"/BirdLife (Tunisie), la Libyan Society for Birds, l'Egyptian Environment Affairs Agency, l'Office National de la Chasse et de la Faune Sauvage (ONCFS, France), et l'institut de la Tour du Valat (France), organisent annuellement le comptage de plusieurs espèces d'oiseaux d'eau dans 785 sites répartis en Afrique du Nord. Le but est, pour chaque espèce, d'estimer le nombre d'oiseaux présents dans toute la région, afin de déterminer quelles espèces sont stables, et lesquelles sont en déclin. Pour compter l'intégralité des populations, il faut en théorie visiter chaque site tous les ans : si un site majeur est manqué, on risque de grandement sous-évaluer ou sur-évaluer la taille de la population. Cependant, pour des raisons financières et politques, il est impossible de visiter chaque site chaque année. Ainsi, le jeu de données disponible est en réalité grandement incomplet. Ce phénomène est illustré en Table 1.3, où est montré un extrait du jeu de données résultant du comptage des oiseaux d'eau en Afrique du Nord pour une espèce particulière : le canard milouin (*Aythya ferina*), dont une photo est montrée en Figure 1.8.

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Site 1	0	0	0	0	0	NA	5	0	NA	0	0	0	0	0	0	0	0	0
Site 2	0	0	0	0	0	0	23	8	4	50	25	126	0	0	0	12	4	2
Site 3	NA	NA	NA	NA	NA	NA	NA	0	NA	0	0	9	0	0	0	0	0	0
Site 4	NA	NA	NA	NA	NA	NA	NA	NA		NA	NA	NA	NA	NA	NA	0	0	0
Site 5	NA	NA	NA	NA	NA	NA	NA	12	NA	NA	NA	NA	NA	NA	NA	0	20	102
Site 6	782	0	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 1.3: Extrait du jeu de données de comptage du canard milouin en Afrique du Nord.

Dans la Table 1.3, on voit que, dans de nombreux cas, des sites n'ont pas pu être visités : ils sont indiqués par "NA". En théorie le but est, à partir du tableau d'abondances

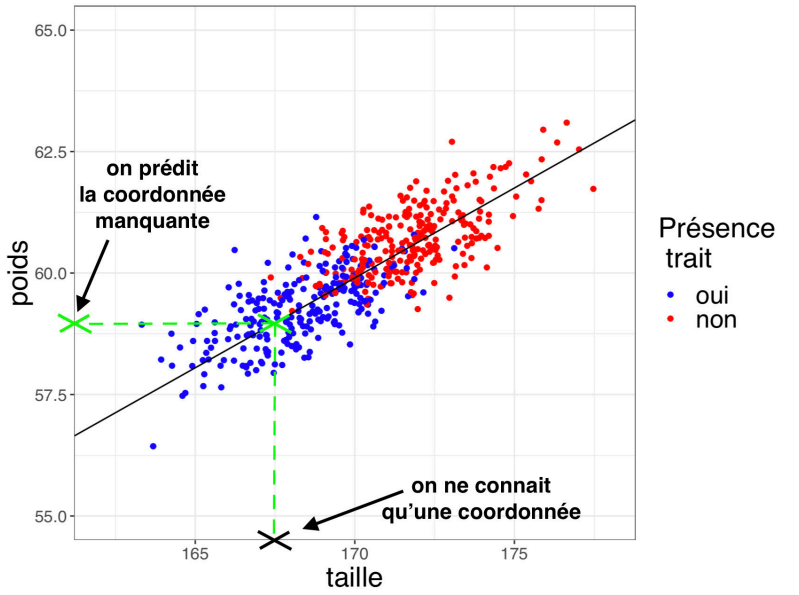


Figure 1.9: Prédiction d'une donnée manquante en utilisant la réduction de dimension.

dont un extrait est montré en Table 1.3, de calculer la somme des comptes observés dans tous les sites, par année, pour voir l'évolution temporelle de l'abondance totale du canard milouin. Malheureusement, ce n'est pas possible directement en raison des données manquantes.

Cependant, encore une fois, les méthodes de réduction de dimensionnalité permettent de pallier ce problème. Prenons l'exemple de la Figure 1.9. Ici, les deux coordonées (taille et poids) sont observés pour tous les individus sauf un, indiqué par une croix noire, pour lequel on ne dispose que de la taille. Cependant, grâce aux individus dont on connaît toutes les caractéristiques, on peut calculer la position de la droite noire, qui correspond toujours à la même variable, résumant la taille et le poids. Ainsi, pour l'individu dont on ne connaît pas le poids, on peut le prédire, à l'aide de sa taille et de la direction calculée, comme indiqué par les segments et les croix vertes. Le fait de prédire les données manquantes est également appelé, en statistique, l'*imputation* de données manquantes.

Ici également, la méthode utilisée repose sur le caractère quantitatif des données. Cependant, dans les données d'abondance du canard milouin, les données sont *discrètes*, puisqu'il s'agit de nombre entiers. Dans le cas des données médicales de la Traumabase, les données sont hétérogènes. Dans cette thèse, nous développons des méthodes d'imputation de données manquantes, qui permettent d'imputer des données quantitatives, qualitatives et discrètes, qui peuvent également être multi-sources, comme discuté dans la section précédente. Ces méthodes sont utilisées pour imputer le jeu de données contenant les abondances d'oiseaux d'eau, et déterminer, pour trois espèces différentes, l'évolution temporelle de la taille des populations. Par ailleurs, elles sont également appliquées à l'imputation d'un extrait de la Traumabase contenant des informations hétérogènes sur des patients souffrant de traumatisme crânien.

Résultats de la thèse

Après avoir situé le travail effectué dans cette thèse dans son contexte statistique et applicatif, nous pouvons à présent résumer les résultats obtenus. Ce manuscrit contient quatre types de contributions. Premièrement, nous proposons des méthodes de bas rang

pour l'imputation de données multi-sources, hétèrogènes et incomplètes, qui permettent de prédire les données manquantes dans tous les cas de figures discutés précédemment. Deuxièmement, nous prouvons, sous la forme de plusieurs théorèmes, que les méthodes proposées sont efficaces sous des conditions réalistes. En particulier, nous prouvons des résultats de la forme suivante. Supposons qu'une donnée n'a pas été observée, par exemple le nombre de canards dans un certain site. Supposons que, si elle avait été observée, elle prendrait la valeur y, c'est-à-dire si quelqu'un avait pu se déplacer pour compter le nombre de canards, il aurait compté y canards. Alors, sous certaines conditions techniques réalistes, en utilisant nos méthodes d'imputation de données manquantes, on prédit pour cette donnée non observée une valeur $\hat{y}$ qui satisfait:

$$(y - \hat{y}) \leq \delta, \text{ et } (\hat{y} - y) \leq \delta$$

c'est-à-dire que la donnée prédite est proche de la vraie donnée, et l'erreur est d'au plus une quantité δ, dont on prouve également qu'elle est optimale, c'est-à-dire qu'il est impossible de faire mieux. Troisièmement, nous distribuons les méthodes proposées sous la forme de deux logiciels publics disponibles en ligne. Le premier logiciel, appelé *lori* (`https://CRAN.R-project.org/package=lori`), permet d'imputer des données de comptes, telles que les données d'abondance d'oiseaux présentées en Table 1.3. Le deuxième logiciel, appelé *mimi* (`https://CRAN.R-project.org/package=mimi`), permet d'imputer des données hétérogènes, telles que les données de la Traumabase présentées en Table 1.2. Enfin, nous présentons un certains nombre d'expériences empiriques, qui évaluent les performances pratiques des logiciels *lori* et *mimi*. Nous montrons qu'ils se comparent globalement de faç favorable aux techniques existentes, et mettons en valeurs les cas de figure où ces techniques sont supérieures aux méthodes existantes. Ainsi, nous proposons un cadre complet, avec des résultats théoriques et empiriques, ainsi que des logiciels réutilisables, pour l'imputation de données multi-sources, hétèrogènes et incomplètes, à l'aide de méthodes de bas rang.

Chapter 2

Introduction

Contents

2.1 Low-rank data tables

In statistics and machine learning, the most natural way to store data is usually to arrange them in tables, where rows correspond to examples or individuals, and columns to attributes. Organizing data in such a way is convenient but, in practice, confronts analysts to data imperfections. Indeed, through the data collection process, it is common that some information is missed, because of machine failures, lack of resources, or simply because individuals do not wish to provide it. For this reason, as data tables are filled, some of their entries remain empty. In ecology for instance, monitored species are counted at regular time intervals in multiple ecological sites (Choler, 2005; Peres-Neto et al., 2016; Sayoud et al., 2017). This process is costly, and requires to visit

remote locations. Thus, at time points when the economical or political situations are not favorable–which happens regularly–some of the sites are not sampled. In species abundance tables, such events result in *missing values*. In addition, data tables are flat by definition, and put on the same level attributes of different types which cannot be straightforwardly compared. For example, in social surveys (Heeringa et al., 2010, Chapters 5 and 6), individuals are asked to provide quantitative information (such as income, rent, etc.), as well as qualitative attributes (employment and marital status, etc.). These *heterogenous* data tables are difficult to analyze as such. Missingness and heterogeneity may also result from the compounding of several data sources, as data aggregation is often seen as an opportunity to increase statistical power. Recently, the promise of personalized medicine has encouraged hospitals to centralize electronic health records (EHR), in order to increase the chance of finding patients with similar profiles. Thus, an essential characteristic of EHR is that they are often *multi-source*, as registries contain patients coming from multiple hospitals. When practices are not standardized, aggregating hospital data may produce heterogeneity and missing values. For example, it is common that two different hospitals resort to different examinations to obtain the same information.

The most important aspect of such data imperfections is probably that they often happen all together. In ecology, incomplete abundance tables are usually supplemented by geographical and meteorological information about the sites and time points where species were counted; this side information is often retrospectively scrapped from the web, and may contain quantitative and qualitative attributes. For instance, the average rainfall, a numerical variable, and the country where the sites are located, a categorical feature. Similarly, in recommendation systems, categorical ratings are collected together with users and items attributes, which may be of different types (Agarwal et al., 2011). Furthermore, in this application, only a small proportion of ratings are observed, and the goal is usually to predict the missing entries. In the medical field, in addition to being multi-source, EHR contain quantitative clinical features like blood pressure and physiological measurements, as well as categorical information like gender, disease stage, type of accident, etc. (Murdoch and Detsky, 2013); EHR are also often incomplete, since not all measurements are made on all patients.

In all these applications, an essential challenge is to compute effective data table summaries, with the objective in mind to use them to assess relationships between variables of different types, and to predict missing values. To do so, statisticians have exploited the mathematical counterpart of data tables: *matrices*. In particular, data tables may be approximated by *low-rank matrices*, whose rows and columns lie in low-dimensional vector spaces. Low-rank approximation methods have been extensively used for tasks such as visualization, clustering, and missing values imputation. The most famous example is probably Principal Component Analysis (PCA), invented by Pearson (1901) and formalized by Hotelling (1933). Since then, a jungle of extensions were proposed, to adapt PCA–which is designed for quantitative numeric data–to more general data types and structures (Greenacre, 1984; Kiers, 1991; Collins et al., 2001; de Leeuw, 2006; Zou et al., 2006; Mohamed et al., 2009; Xu et al., 2010; Candès et al., 2011; Li and Tao, 2013; Mardani et al., 2013; Kateri, 2014; Hastie et al., 2015; Pagès, 2015; Liu et al., 2016; Udell et al., 2016).

In this dissertation, we are concerned with the analysis of multi-source, heterogeneous and incomplete data tables. In particular, through two practical problems: the analysis of a waterbird abundance data set and of a severe trauma registry. To do so, we develop new tools based on low-rank models, and adapted to multi-source, heterogeneous and

incomplete data. Before introducing some background on low-rank methods, we present two motivating examples inspired by collaborations with the Tour du Valat institute and the Traumabase group.

2.1.1 Waterbird monitoring

Waterbirds are defined as the families of bird species who depend upon wetland sites for at least part of their life cycle, for instance through food, habitat, or breeding. The monitoring of waterbirds has been a global concern for at least 50 years, when Wetlands International (`www.wetlands.org`), a global organization dedicated to the conservation and restoration of wetlands, launched the first International Waterbird Census (IWC) in 1967. In these censuses, birds are counted synchronously in over 25,000 wetland sites in more than 100 countries, to monitor the population sizes and the changes in the number and distribution of waterbirds. One of the objectives is to provide information to global conservation organizations; this may have an impact on regulatory measures at national and international levels.

In this respect, a major challenge is to analyze such bird census data at flyway or regional scales, including in areas where there may be gaps in the temporal or site sampling schemes. North Africa in particular is a region of major importance, as it acts as a last stopover for migratory waterbirds before they cross the Sahara or the Mediterranean Sea. In this region, before 2013, abundance data have been collected partially, without any coordination across countries. Since 2013, waterbirds monitoring in North-Africa is conducted through coordinated region-wide censuses (Sayoud et al., 2017). In particual, Sayoud et al. (2017) revealed how to reduce census cost by focusing on important sites, assessed the effectiveness of conservation policies, and detected environmental predictors related to waterbird distributional ecology. However, waterbird monitoring data at this regional scale were not yet analyzed across more than one year. Because of the irregular spatial coverage, imputation of nonexistent count data is a challenge of major importance in this application.

The first contributions presented in this dissertation are motivated by the analysis of a waterbird abundance data set, which gathers waterbird counts across 785 wetland sites (across the 5 countries in North Africa), between 1990 and 2017. For political and financial reasons, not all wetland sites could be visited every year, and thus there are many missing values in this data set (between 40% and 60% depending on the species). The final goal is to estimate the temporal trends over all sites, in order to identify declining species. We approach this task with a missing values imputation perspective: we seek to predict the unobserved counts before computing estimated yearly totals. The originality of the approach compared to existing imputation methods, is that we use side information concerning the sites and years (such as meteorological anomalies, latitude and longitude, distance to the closest urban center) to improve the predictions. As a result, in the process, we also estimate the effect of geographical and meteorological covariates on the bird counts. This is also interesting in itself, for example to detect environmental factors which may be adverse to some of the species. The general model and procedure are described in Chapter 3, the analysis of the waterbird data set in Chapter 4, and an R (R Core Team, 2017) tutorial in Chapter 5.

2.1.2 Multi-center major trauma registry

Major trauma is defined as any injury that can cause prolonged disability or death. It has been pointed out by the World Health Organization (`https://www.who.int/home`) as a major cause of handicap and mortality on a global scale (Hay, 2017). It has also been shown that management of major trauma based on standardized and protocol-based care improves prognosis of patients (David JS, 2018; Rossaint et al., 2016). But before evaluating the care process, a first inevitable step is to collect reliable data that describe it. To do so, in France, the Traumabase Group (`http://www.traumabase.eu/en_US`) maintains a national observatory of major trauma. Since 2012, fifteen French trauma centers have contributed to the data base, which now contains information on more than 20,000 patients from admission until discharge from critical care.

The resulting data set can be seen as an aggregation of smaller data sets coming from multiple centers. In addition, a broad heterogeneity of trauma care processes across French hospitals has been reported in existing studies (Hamada et al., 2015; Jouffroy et al., 2018). Thus, it is crucial to take into account the multi-center, or *multilevel* structure of the Traumabase. Furthermore, another distinctive characteristic of the Traumabase is that it contains mixed data, i.e. both qualitative and quantitative variables. For example, the systolic and diastolic blood pressures are numeric variables (measured in millimeters of mercury mmHg). The variable indicating the type of accident (fall, car accident, gun shot, etc.), on the other hand, is qualitative. The time passed between the accident and arrival at the hospital, measured in hours, is a discrete variable. Finally, many entries are missing in the data set.

From a clinical point of view, there are several important questions such as finding predictors of morbidity and mortality, to describe the strategies deployed in major trauma care, and to develop decision support tools. From a statistical point of view, these challenges involve, e.g., performing predictive models, exploratory data analysis, and causal inference, all of this with missing values. In this problem too, missing values imputation is a flexible and attractive option, as one may apply any statistical method once the data set is complete (although care must be taken if the imputation is single), instead of adapting every model to an incomplete framework. Thus, the second part of our contributions is focused on providing methods to impute multilevel mixed data. The first method, presented in Chapter 8, is based on multilevel singular value decomposition (SVD). The second method, presented in Chapter 6, is an alternative based on a general probabilistic model. This second method is implemented in an R package for which we provide a tutorial in Chapter 7.

2.1.3 Why are low-rank models relevant?

To see it, we need some general background on low-rank matrices. The rank of a matrix $\boldsymbol{X} \in \mathbb{R}^{m_1 \times m_2}$ is defined as the dimension of the vector space generated by its columns, or rows. Indeed, it can be proved that the row and column spaces are in fact of the same dimension. Along this dissertation, we will denote the rank of $\boldsymbol{X}$ by $\mathrm{rank}(\boldsymbol{X})$. The matrix $\boldsymbol{X}$ is said to be of low-rank if $\mathrm{rank}(\boldsymbol{X})$ is small compared to the dimensions m_1 and m_2. Usually, small compared to m_1 and m_2 means smaller than a predefined integer $r_{\mathsf{max}} < \min(m_1, m_2)$.

If $\boldsymbol{X} \in \mathbb{R}^{m_1 \times m_2}$ is of rank r, then, it can be factorized as follows:

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top, \tag{2.1}$$

where $\boldsymbol{U} \in \mathbb{R}^{m_1 \times r}$ and $\boldsymbol{V} \in \mathbb{R}^{m_2 \times r}$ are orthonormal matrices, and $\boldsymbol{D} \in \mathbb{R}^{r \times r}$ is diagonal with nonnegative entries. Decomposition (2.1) is called the *singular value decomposition* (SVD) of $\boldsymbol{X}$. The matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ contain *singular vectors* of $\boldsymbol{Y}$, and the diagonal matrix $\boldsymbol{D}$ contains the *singular values* of $\boldsymbol{Y}$. The SVD of $\boldsymbol{Y}$, which generalizes eigen-decompositions to rectangular matrices, is given by the triplet of matrices $(\boldsymbol{U}, \boldsymbol{D}, \boldsymbol{V})$. From factorization (2.1), it results that $\boldsymbol{X}$ can be represented using $r(m_1 + m_2 - r)$ parameters. Indeed, the free parameters consist in r nonnegative values, r orthonormal directions in $\mathbb{R}^{m_1}$ $(\boldsymbol{U})$, and r orthonormal directions in $\mathbb{R}^{m_2}$ $(\boldsymbol{V})$. The number of free parameters in $\boldsymbol{U}$ is $(m_1 - 1) + \ldots + (m_1 - r) = rm_1 - r^2/2 - r/2$, since for $k \in [\![r]\!]$, the column $\boldsymbol{U}_{.,k}$ is constrained to have a unitary norm, and to be orthogonal to the $(k-1)$ previous ones. Similarly, the number of free parameters in $\boldsymbol{V}$ is $(m_2 - 1) + \ldots + (m_2 - r) = rm_2 - r^2 - r/2$. Finally, the number of free parameters in $\boldsymbol{D}$ is r.

Thus, in parametric models with matrix parameters, the rank controls the complexity of a model, through the number of free parameters involved. Consider a data matrix $\boldsymbol{Y}$ with m_1 rows and m_2 columns. Low-rank methods rely on the assumption that $\boldsymbol{Y}$ can be well approximated by a matrix $\boldsymbol{X}^0$ of low-rank. For example, in this dissertation, we will consider models of the form:

$$\boldsymbol{Y} = \mathcal{F}(\boldsymbol{X}^0), \tag{2.2}$$

where $\mathcal{F} : \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^{m_1 \times m_2}$ is a matrix-valued function and may have a stochastic component. In this case, a low-rank method aims at estimating the underlying matrix $\boldsymbol{X}^0$, from the noisy and/or incomplete observations $\boldsymbol{Y}$, subject to a rank constraint. If r is small enough, approximating a full-rank data matrix $\boldsymbol{Y}$ by a rank-r matrix can reduce dramatically the storage and computational costs. Furthermore, in matrix estimation problems, the number of parameters to estimate is $m_1 m_2$, and the number of observed values is equal to $m_1 m_2$, or smaller when the data matrix has missing entries. In high-dimensional statistics jargon, this corresponds to a "$p \geq n$" problem, where constraints on the parameter space are usually made. In this respect, the low-rank assumption can be seen as an equivalent of the *bet on sparsity principle* (Hastie et al., 2015, Chapter 1)–which is usually endorsed for vector parameters–for matrix parameters. To back up these popular low-rank models, recent works provided evidence that a vast class of matrices are well approximated by a low-rank counterpart (Alon et al., 2013; Chatterjee, 2015; Udell and Townsend, 2018). Finally, factorization (2.1) also has the advantage of easily providing interpretation tools. Indeed, $\boldsymbol{U}$ and $\boldsymbol{V}$ are in fact r-dimensional approximations of the row and column vector spaces of $\boldsymbol{Y}$. These low-dimensional vector spaces can be used, for instance, to visualize the data points.

2.1.4 Principal component analysis

The next question is: how to compute a low-rank approximation $\boldsymbol{X}$ of $\boldsymbol{Y}$? In model (2.2), this implies choosing a link function $\mathcal{F}$, and a cost function to measure the distance between the data $\boldsymbol{Y}$ and a candidate low-rank matrix. A simple example is the following model:

$$\boldsymbol{Y} = \boldsymbol{X}^0 + \boldsymbol{E}, \tag{2.3}$$

where $\boldsymbol{Y}$ is a realization of the low-rank matrix $\boldsymbol{X}^0$ perturbed by i.i.d. additive noise $\boldsymbol{E} = (\boldsymbol{E}_{i,j})$. In this case, a natural distance measure is the usual Euclidean distance,

given in $\mathbb{R}^{m_1 \times m_2}$ by the Frobenius norm. In this example, the underlying low-rank matrix $\boldsymbol{X}^0$ may be estimated by solving the following optimization problem:

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{Y} - \boldsymbol{X}\|_F^2 \\ \text{subject to} & \operatorname{rank}(X) \leq r. \end{array} \tag{2.4}$$

Program (2.4) admits a closed-form solution given by the rank-r *truncated* SVD of $\boldsymbol{Y}$. Assume without loss of generality that $m_1 \leq m_2$, and note that we will keep this assumption along the whole dissertation. If the data matrix $\boldsymbol{Y}$ has full rank, i.e. $\operatorname{rank}(\boldsymbol{Y}) = m_1$, its SVD is given by:

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top, \tag{2.5}$$

with $\boldsymbol{U} \in \mathbb{R}^{m_1 \times m_1}$, $\boldsymbol{V} \in \mathbb{R}^{m_2 \times m_1}$ two orthonormal matrices and $\boldsymbol{D} \in \mathbb{R}^{m_1 \times m_1}$ a diagonal matrix with nonnegative entries. The rank-r *truncated* SVD of $\boldsymbol{Y}$ consists, from equation (2.5), in keeping only the first r columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, and the first $r \times r$ block in $\boldsymbol{D}$:

$$\operatorname{SVD}_r(\boldsymbol{Y}) := \boldsymbol{U}_{\cdot,[\![r]\!]}\boldsymbol{D}_{[\![r]\!],[\![r]\!]}(\boldsymbol{V}_{\cdot,[\![r]\!]})^\top, \tag{2.6}$$

where $\boldsymbol{U}_{\cdot,[\![r]\!]}$is the submatrix of $\boldsymbol{U}$ defined by its r first columns $\boldsymbol{U}_{\cdot,[\![r]\!]} = (\boldsymbol{U}_{i,j}), i \in [\![m_1]\!], j \in [\![r]\!]$. Similarly, $\boldsymbol{V}_{\cdot,[\![r]\!]}(\boldsymbol{V}_{i,j}), i \in [\![m_2]\!], j \in [\![r]\!]$, and $\boldsymbol{D}$ is the submatrix of $\boldsymbol{D}$ defined by the intersection of the first r rows and columns: $\boldsymbol{D}_{[\![r]\!],[\![r]\!]} = (\boldsymbol{D}_{i,j}), i \in [\![r]\!], j \in [\![r]\!]$.

This simple example is of course very well-known under the name of principal component analysis (PCA), when the columns (or rows) of $\boldsymbol{Y}$ have been centered. It can be shown that the solution to the rank-r PCA problem (2.4) is exactly the rank r truncated SVD (2.6), where the *principal components* are defined by the columns of $\boldsymbol{U}_{\cdot,[\![r]\!]}$. From a statistical point of view, the principal components are orthonormal directions along which the variance of the samples $\boldsymbol{X}_{i,\cdot}$ is maximized. In data analysis, they are often used as a new basis of reduced dimension where samples can be represented.

The majority of low-rank methods can be seen as extensions or variants of PCA, for example adapted to non numeric data, with general link functions and distance measures. Another important extension is the processing of missing values. In the context of incomplete data, low-rank methods have been used for two purposes. The first is to perform PCA (or alternatives) *in spite of the missing values*; this led to extensions of PCA to incomplete data settings, using EM-type algorithms for instance (Josse and Husson, 2012). The second standpoint is to exploit a low-rank structure *to recover the missing values*; this is the starting point of a vast field of the statistics literature called *matrix completion* (Candès and Recht, 2009; Recht et al., 2010).

2.2 Nuclear norm heuristics

Contrary to PCA, most rank constrained problems have no closed-form solution, and are NP-hard in general, for instance, if some entries of $\boldsymbol{Y}$ are unobserved. Since NP-hardness comes from the rank constraint $\operatorname{rank}(\boldsymbol{X}) \leq r$, it is often replaced in practice by tractable proxies. Equivalently to the definition given in Section 2.1.3, the rank of a matrix $\boldsymbol{X}$ is also defined as its number of nonzero singular values. Denote, for $k \in [\![m_1]\!]$, $\sigma_k(\boldsymbol{X})$ the k-th largest singular value of $\boldsymbol{X}$. If $\boldsymbol{X}$ has rank r, then

$$\sigma_1(\boldsymbol{X}) \geq \ldots \geq \sigma_r(\boldsymbol{X}) > 0, \text{ and } \sigma_{r+1}(\boldsymbol{X}) = \ldots = \sigma_{m_1}(\boldsymbol{X}) = 0. \tag{2.7}$$

Thus, the rank of $\boldsymbol{X}$ is in fact the ℓ_0 norm of the vector containing its singular values:

$$\operatorname{rank}(\boldsymbol{X}) = \sum_{k=1}^{m_1} \mathbb{1}_{\{\sigma_k(\boldsymbol{X})>0\}}. \tag{2.8}$$

Similarly to what was done in sparse vector estimation problems, an option is to replace the ℓ_0 norm by a convex relaxation, such as the ℓ_1 norm. This is precisely the definition of the nuclear norm:

$$\|\boldsymbol{X}\|_* = \sum_{k=1}^{m_1} \sigma_k(\boldsymbol{X}). \tag{2.9}$$

It has been shown that, in many cases, nuclear norm heuristics provide accurate (even exact) solutions, and can be computed with tractable algorithms, for example in matrix completion.

Matrix completion addresses the problem of recovering a matrix from partial observation of its entries. The most famous practical example is probably the Netflix problem (see, e.g., Bell and Koren (2007) and the references therein). In this problem, items (columns) are rated by users (rows). However, each user typically only rates a few items, so that only very few entries of the data matrix are actually observed. The goal is to complete the missing ratings, in order to recommend relevant items to users. Formally, let $\Omega \subseteq [\![m_1]\!] \times [\![m_2]\!]$ denote the indices of the observed entries in $\boldsymbol{Y}$, and denote by $n = |\Omega|$ the cardinality of Ω, i.e., the number of observed entries. Of course, matrix completion is impossible in general if $n < m_1 m_2$ and without additional constraints. But, if the data matrix $\boldsymbol{Y}$ has rank r, then it is intrinsically characterized by $r(m_1 + m_2 - r)$ parameters, and thus there is hope to recover the unobserved entries. The so-called low-rank matrix completion problem has been extensively studied.

2.2.1 Exact matrix completion

The question asked in exact low-rank matrix completion is: can we find a matrix $\boldsymbol{X}$ of minimal rank which matches exactly $\boldsymbol{Y}$ on the set of observed entries Ω? This boils down to the following optimization problem:

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(\boldsymbol{X}) \\ \text{subject to} & \boldsymbol{X}_{i,j} = \boldsymbol{Y}_{i,j} \text{ for all } (i,j) \in \Omega. \end{array} \tag{2.10}$$

Problem (2.10) is NP-hard (Candès and Recht, 2009), and was replaced in seminal works (Recht et al., 2010; Candès and Recht, 2009; Candès and Tao, 2010) by the following nuclear norm heuristic:

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{X}\|_* \\ \text{subject to} & \boldsymbol{X}_{i,j} = \boldsymbol{Y}_{i,j} \text{ for all } (i,j) \in \Omega. \end{array} \tag{2.11}$$

Note that problem (2.11) is stochastic whenever the set of observed entries Ω are sampled through a random process. It was shown in Recht et al. (2010); Candès and Recht (2009); Candès and Tao (2010); Recht (2011); Gross (2011) that, under an incoherence assumption discussed later on, and when the number of observed entries is large enough, the incomplete matrix $\boldsymbol{Y}$ can be recovered exactly by solving problem (2.11). In particular, Recht (2011) prove the following result.

Definition 1 (Recht (2011), Definition 1.1). *Let U be a subspace of $\mathbb{R}^m$ of dimension r, and $\mathbf{P}_U$ the orthogonal projection on U. Then, the coherence of U with respect to the standard basis (e_i) is defined to be*

$$\mu(U) = \frac{m}{r} \max_{1 \leq i \leq m} \|\mathbf{P}_U e_i\|_2^2. \tag{2.12}$$

Intuitively, the coherence $\mu(U)$ of the subspace U measures the correlation between the subspace and the canonical basis. Indeed, denote $(U_1, \ldots, U_r)$ an orthonormal basis of U, with $U_k \in \mathbb{R}^m$ for all $k \in [\![r]\!]$. Then, the coherence of U may be written simply as follows:

$$\mu(U) = \frac{m}{r} \max_{1 \leq i \leq m} \left(\sum_{k=1}^{r} \langle e_i, U_k \rangle^2 \right). \tag{2.13}$$

The coherence $\mu(U)$ satisfies $1 \leq \mu(U) \leq m/r$. The lower bound 1 is achieved, for example, if U is of dimension 1 and $U_1 = (1/\sqrt{m}, \ldots, 1/\sqrt{m})$. The upper bound m/r is achieved if $(U_1, \ldots, U_r)$ contains a standard basis vector e_i, $i \in [\![m]\!]$. Let $\boldsymbol{Y}$ be an $m_1 \times m_2$ matrix of rank r with SVD $(\boldsymbol{U}, \boldsymbol{D}, \boldsymbol{V})$. Assume without loss of generality that $\boldsymbol{D}$ is of size $r \times r$, $\boldsymbol{U}$ is of size $m_1 \times r$ and $\boldsymbol{V}$ is of size $m_2 \times r$. Then, using (2.13), the coherence of the row and column vector spaces are given by:

$$\mu(\boldsymbol{U}) = \frac{m_1}{r} \max_{1 \leq i \leq r} \left(\sum_{k=1}^{r} \langle e_i, \boldsymbol{U}_{\cdot,k} \rangle^2 \right), \quad \mu(\boldsymbol{V}) = \frac{m_2}{r} \max_{1 \leq j \leq r} \left(\sum_{k=1}^{r} \langle e_j, \boldsymbol{V}_{\cdot,k} \rangle^2 \right). \tag{2.14}$$

In Recht (2011), the authors prove the following theorem.

Theorem 1 (Recht (2011), Theorem 1.1). *Assume that:*

- *The row and column vector spaces of $\boldsymbol{Y}$ have coherence bounded above by some positive μ_0.*
- *The matrix $\boldsymbol{U}\boldsymbol{V}^\top$ has a maximum entry bounded by $\mu_1 \sqrt{r/(m_1 m_2)}$ in absolute value for some μ_1 positive.*
- *The entries of $\boldsymbol{Y}$ are observed with locations sampled uniformly at random.*

Assume that the number of observed entries n satisfies

$$n \geq 32 \max(\mu_1^2, \mu_0) r (m_1 + m_2) \beta \log^2(2m_2), \tag{2.15}$$

with $\beta > 1$. Then, program (2.11) *has a unique minimizer which is equal to $\boldsymbol{Y}$ with probability at least $1 - 6\log(m_2)(m_1 + m_2)^{2-2\beta} - m_2^{2-2\beta^{1/2}}$.*

The incoherence condition in Theorem 1, $\max(\mu(\boldsymbol{U}), \mu(\boldsymbol{V})) \leq \mu_0$, controls the correlation between the singular vector spaces $\boldsymbol{U}$ and $\boldsymbol{V}$ and the orthonormal bases of $\mathbb{R}^{m_1}$ and $\mathbb{R}^{m_2}$. Intuitively, it is related to the amount of information provided by each entry of $\boldsymbol{Y}$ about the other entries of $\boldsymbol{Y}$. For example, it is impossible to recover a matrix which is equal to zero everywhere except in one entry, unless the nonzero entry is observed. If the incoherence condition is satisfied, each entry of $\boldsymbol{Y}$ contains a sufficient amount of information about the singular vector spaces $\boldsymbol{U}$ and $\boldsymbol{V}$, so that they can be recovered exactly by sampling the entries of $\boldsymbol{Y}$ uniformly at random. The result (2.15) essentially shows that the nuclear norm heuristic (2.11) yields an exact solution, provided that the number of observed entries is of the order of $\mathcal{O}(r(m_1 + m_2)\log(m_2)^2)$. Note that the result holds with high probability, in reference to the sampling of the observed entries, which are assumed to be sampled uniformly at random.

2.2.2 Noisy matrix completion

In many cases, it is unlikely that the observations are exact, and more plausible to assume the data matrix $\boldsymbol{Y}$ corrupted by noise. For now, we assume for simplicity that the noise is additive; we will consider more general cases in Section 2.3. In other words, we observe:

$$\boldsymbol{Y}_{i,j} = \boldsymbol{X}^0_{i,j} + \boldsymbol{E}_{i,j}, \quad (i,j) \in \Omega, \tag{2.16}$$

where $\{\boldsymbol{E}_{i,j}, (i,j) \in \Omega\}$ are independent noise terms. In this case, fitting the observations exactly is pointless, and the matrix completion problem (2.11) is reformulated as:

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{X}\|_* \\ \text{subject to} & \|\mathcal{P}_\Omega(\boldsymbol{X} - \boldsymbol{Y})\|_F^2 \leq \delta, \end{array} \tag{2.17}$$

where $\mathcal{P}_\Omega$ is the projection on the set of observed entries. For $\boldsymbol{M} \in \mathbb{R}^{m_1 \times m_2}$:

$$\mathcal{P}_\Omega(\boldsymbol{M}) = \sum_{(i,j)\in\Omega} \boldsymbol{M}_{i,j} e_i f_j^\top, \tag{2.18}$$

where (e_i) and (f_j) are the standard bases of $\mathbb{R}^{m_1}$ and $\mathbb{R}^{m_2}$ respectively. This problem was introduced by Candes and Plan (2010), who showed that the estimation error of program (2.17) is proportional to the noise level, under conditions which are similar to those required in the noiseless setting. Noisy matrix completion was later also studied in Keshavan et al. (2010); Foygel and Srebro (2011); Gaiffas and Lecué (2011); Koltchinskii (2011b); Koltchinskii et al. (2011); Rohde and Tsybakov (2011); Agarwal et al. (2012); Klopp (2014). In particular, Klopp (2014) shows optimal convergence rates for noisy matrix completion, without requiring the incoherence condition described in Theorem 1, and used in Candes and Plan (2010). Klopp (2014) considers the following nuclear norm penalized estimation problem:

$$\begin{array}{ll} \text{minimize} & \|\mathcal{P}_\Omega(\boldsymbol{X} - \boldsymbol{Y})\|_F^2 + \lambda\|\boldsymbol{X}\|_* \\ \text{subject to} & \|\boldsymbol{X}\|_\infty \leq a, \end{array} \tag{2.19}$$

where $\lambda > 0$ is a regularization parameter controlling the trade-off between fitting the data and enforcing low-rank solutions. Klopp (2014) shows that (2.19) has minimax optimal estimation error. They only assume an upper bound a on the absolute value of the matrix $\boldsymbol{X}$, which is a much weaker condition than the incoherence condition of Candes and Plan (2010); Keshavan et al. (2010). The result of Klopp (2014) may be summarized as follows.

Theorem 2. *Assume that:*

- *The noise $\boldsymbol{E}$ and the sampling Ω are independent.*
- *The probability of observing an entry in column j (resp. row i) is bounded above by $L/\min(m_1, m_2)$, with $L \geq 1$.*
- *Every entry is observed with probability at least $(\mu m_1 m_2)^{-1}$, with $\mu \geq 1$.*
- *The noise $(\boldsymbol{E}_{i,j})$ is subexponential: there exists $K > 0$ such that for all $(i,j) \in \Omega$, $\mathbb{E}\left[\exp(|\boldsymbol{E}_{i,j}|/K)\right] \leq \mathrm{e}$.*

Then, there exists a value of λ (which we omit here for sake of clarity), such that, for $n \geq 2\log^2(m_1+m_2)\min(m_1,m_2)/L$, and with probability at least $1-3/(m_1+m_2)$, any solution $\hat{\boldsymbol{X}}$ of (2.19) *satisfies:*

$$\frac{\|\hat{\boldsymbol{X}}-\boldsymbol{X}^0\|_F^2}{m_1 m_2} \lesssim \mu^2 L \frac{\mathrm{rank}(\boldsymbol{X}^0)m_2}{n}, \tag{2.20}$$

where $\lesssim$ denotes the inequality up to constant and logarithmic factors. The rate (2.20) is parametric; furthermore, Klopp (2014) also shows that it is minimax optimal, up to constant and logarithmic factors. Theorem 2 provides a high probability upper bound, this time in reference to the sampling of the observed entries, *and* to the distribution of the noise.

2.3 General data types

The low-rank methods reviewed so far are all based on a least squares loss function, which may be interpreted as an implicit additive noise model, and is not adapted to all types of data. Several extensions of PCA and matrix completion have thus been proposed to accommodate multiplicative noise, such as Poisson noise for instance. Other works proposed generalizations to heterogeneous noise, with quantitative and qualitative data.

2.3.1 Generalized principal component analysis

Extending PCA to general loss functions boils down to replacing problem (2.4) by:

$$\begin{array}{ll} \text{minimize} & \mathcal{L}(\boldsymbol{Y},\boldsymbol{X}) \\ \text{subject to} & \mathrm{rank}(\boldsymbol{X}) \leq r, \end{array} \tag{2.21}$$

where $\mathcal{L}$ is a function measuring how well $\boldsymbol{X}$ fits the data $\boldsymbol{Y}$. For example, in parameteric models, $\mathcal{L}(\boldsymbol{Y},\boldsymbol{X})$ may be the negative log-likelihood of the distribution of $\boldsymbol{Y}$ parametrized by $\boldsymbol{X}$. Because (2.21) does not have a closed-form solution in general, in practice, a factorized version is often solved instead:

$$\begin{array}{ll} \text{minimize} & \mathcal{L}(\boldsymbol{Y},\boldsymbol{U}\boldsymbol{V}^\top) \\ \text{subject to} & \boldsymbol{U} \in \mathbb{R}^{m_1\times r}, \text{ and } \boldsymbol{V} \in \mathbb{R}^{m_2\times r}. \end{array} \tag{2.22}$$

Problem (2.22) was introduced in Collins et al. (2001), where $\mathcal{L}$ is derived from probabilistic models of the exponential family. This includes the Gaussian model $\boldsymbol{Y}_{i,j} \sim \mathcal{N}(\boldsymbol{X}^0_{i,j},\sigma^2)$, $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, in which case $\mathcal{L}$ is simply the least squares loss. If $\boldsymbol{Y}$ contains counts, then one may use a Poisson model with an exponential link $\boldsymbol{Y}_{i,j} \sim \mathcal{P}(\exp(\boldsymbol{X}^0_{i,j}))$, in which case $\mathcal{L}(\boldsymbol{Y},\boldsymbol{X}) = \sum_{i=1}^{m_1}\sum_{i=1}^{m_2}(-\boldsymbol{Y}_{i,j}\boldsymbol{X}_{i,j} + \exp(\boldsymbol{X}_{i,j}))$. Gordon (2002) further extended the framework to loss functions based on the generalized Bregman divergence of convex functions. Many other works have considered problem (2.22), with Bayesian (Mohamed et al., 2009) and probabilistic (Chiquet et al., 2018) versions of exponential family PCA, as well as scalable methods for high-dimensional data sets (Liu et al., 2016). Other works studied even more general losses, and additional regularization terms (Srebro, 2004; de Leeuw, 2006; Singh and Gordon, 2008; Udell et al., 2016). In particular, Udell et al. (2016) consider *heterogeneous* loss functions which are allowed to depend on the columns of $\boldsymbol{Y}$, so that heterogeneous data can be modeled: $\mathcal{L}(\boldsymbol{Y},\boldsymbol{X}) = \sum_{j=1}^{m_2}\mathcal{L}_j(\boldsymbol{Y}_{\cdot,j},\boldsymbol{X}_{\cdot,j})$.

With a different perspective, other types of generalizations for non numeric data consist in applying transformations to $\boldsymbol{Y}$ before computing the classical PCA solution:

$$\begin{array}{ll} \text{minimize} & \|\tilde{\boldsymbol{Y}} - \boldsymbol{X}\|_F^2 \\ \text{subject to} & \operatorname{rank}(\boldsymbol{X}) \leq r, \end{array} \tag{2.23}$$

where $\tilde{\boldsymbol{Y}}$ is the transformed data set. This is the case of Correspondence Analysis (CA) (Greenacre, 1984) for count data, Multiple Correspondence Analysis (MCA) (Greenacre and Blasius, 2006) for categorical data, as well Factorial Analysis of Mixed Data (FAMD) (Pagès, 2015) and PCAMIX (Kiers, 1991), both for mixed data. These geometric methods are designed without any reference to probabilistic models. However, often times, such transformations are in fact first-order approximations of non Gaussian probabilistic models. For example, it has been shown that CA approximates a Poisson log-bilinear model for small entry-wise values of $\boldsymbol{Y}$ (Escoufier, 1982). There are also similar results for MCA and a multi-logit bilinear model (Fithian and Josse, 2017).

2.3.2 Exponential family matrix completion

In the same way, matrix completion was extended to more general data types, mainly by replacing (2.16) and (2.19) by similar problems where non Gaussian data fitting terms are minimized, with additional constraints or regularizations. One of the first extensions of matrix completion to non numeric data was done in Davenport et al. (2012). In this paper, the authors introduced one-bit matrix completion, which is similar to the original matrix completion problem (2.11), with one crucial difference: only the sign of the observations $\boldsymbol{Y}_{i,j} = \boldsymbol{X}^0_{i,j} + \boldsymbol{E}_{i,j}$ are observed ($+1$ if $\boldsymbol{Y}_{i,j} \geq 0$ and -1 if $\boldsymbol{Y}_{i,j} < 0$). In Davenport et al. (2012), the authors consider the optimization problem

$$\begin{array}{ll} \text{minimize} & -\sum_{(i,j)\in\Omega} \left(\mathbb{1}_{\boldsymbol{Y}_{i,j}=1} \log(f(\boldsymbol{X}_{i,j})) + \mathbb{1}_{\boldsymbol{Y}_{i,j}=-1} \log(1 - f(\boldsymbol{X}_{i,j})) \right) \\ \text{subject to} & \|\boldsymbol{X}\|_* \leq a\sqrt{r m_1 m_2} \text{ and } \|X\|_\infty \leq a, \end{array} \tag{2.24}$$

with f a link function. In (2.24), a logistic loss is penalized instead of a least squares loss. The nuclear norm and infinity norm constraints are used as proxies for a low-rank constraint. Indeed, any matrix $\boldsymbol{X}$ of rank r satisfies $\|\boldsymbol{X}\|_* \leq \sqrt{r}\|\boldsymbol{X}\|_F$. If, in addition, $\|\boldsymbol{X}\|_\infty \leq a$, then one obtains $\|\boldsymbol{X}\|_* \leq a\sqrt{r m_1 m_2}$. Davenport et al. (2012) obtain the first theoretical guarantees for (2.24). In particular, they show that, when the number of observations n is large enough, the rank-r signal matrix $\boldsymbol{X}^0$ can be estimated with an error of the order

$$\frac{\|\hat{\boldsymbol{X}} - \boldsymbol{X}^0\|_F^2}{m_1 m_2} \lesssim C_{f,a} \sqrt{\frac{r(m_1 + m_2)}{n}}, \tag{2.25}$$

where $\lesssim$ denotes the inequality up to constant and log factors, and $C_{f,a}$ is a constant which depends on the regularity of the link function f on $[-a, a]$. The bound reported in (2.25) is a high probability upper bound, as a result of the noise and the sampling of entries. One-bit matrix completion was also studied in Cai and Zhou (2013) for general (non uniform) sampling distributions, and with a constraint on the max-norm of $\boldsymbol{X}$ rather than the nuclear norm; they obtain a convergence rate similar to (2.25). Cao and Xie (2016) further extended the approach of Davenport et al. (2012) to Poisson matrix completion, with the following matrix recovery problem:

$$\begin{array}{ll} \text{minimize} & \sum_{(i,j)\in\Omega} (-\boldsymbol{Y}_{i,j}\boldsymbol{X}_{i,j} + \exp(\boldsymbol{X}_{i,j})) \\ \text{subject to} & \|\boldsymbol{X}\|_* \leq a\sqrt{r m_1 m_2}, \\ & b \leq \boldsymbol{X}_{i,j} \leq a \text{ for all } (i,j) \in [\![m_1]\!] \times [\![m_2]\!]. \end{array} \tag{2.26}$$

Other works extended the nuclear norm penalized version of matrix completion (2.19) by solving problems of the following form:

$$\begin{array}{ll} \text{minimize} & \sum_{(i,j)\in\Omega} \mathcal{L}(\boldsymbol{Y}_{i,j}, \boldsymbol{X}_{i,j}) + \lambda\|\boldsymbol{X}\|_* \\ \text{subject to} & \|\boldsymbol{X}\|_\infty \leq a. \end{array} \tag{2.27}$$

For example, in Klopp et al. (2015), the authors introduce multinomial matrix completion, where the observations are allowed to take more than two values. As a special case, they study the one-bit matrix completion problem in its penalized form, where

$$\mathcal{L}(\boldsymbol{Y}_{i,j}, \boldsymbol{X}_{i,j}) := -\mathbb{1}_{\boldsymbol{Y}_{i,j}=1} \log(f(\boldsymbol{X}_{i,j})) - \mathbb{1}_{\boldsymbol{Y}_{i,j}=-1} \log(1 - f(\boldsymbol{X}_{i,j})). \tag{2.28}$$

In this case, they provide a minimax optimal estimator with a convergence rate of the order of:

$$\frac{\|\hat{\boldsymbol{X}} - \boldsymbol{X}^0\|_F^2}{m_1 m_2} \lesssim C_{f,a}^2 \frac{r(m_1 + m_2)}{n}. \tag{2.29}$$

Similarly to (2.25), the bound (2.29) holds with high probability. Note that, compared to (2.25), the above rate is faster, in the sense that it converges to 0 with a dependence on the number of observations n of the order of $1/n$ instead of $1/\sqrt{n}$. For other (unbounded) data types, Lafond (2015) studied exponential family matrix completion in the same framework as Klopp et al. (2015). In Lafond (2015), problem (2.27) is solved for loss functions $\mathcal{L}$ derived from exponential family models. This includes (but is not restricted to) the Gaussian, binomial, Poisson and exponential models. Lafond (2015) provides minimax convergence rates of the same order as (2.29).

2.4 Hybrid low-rank structures

In this dissertation, we are interested in incorporating several sources of data in the same low-rank model. For example, a count table resulting from species censuses together with geographical and meteorological information scrapped from the web. Or, in our application to the Traumabase registry, several data sets from multiple hospitals. Our approach at doing this will be to consider hybrid low-rank structures. In particular, the parameter matrix $\boldsymbol{X}$ may be decomposed into two components, one of them being low-rank. This type of composite structure has been studied before. We review here some models which will be useful to contextualize our contributions.

2.4.1 Main effects and interactions

The concept of main effects and interactions first emerged in the design of experiments and analysis of variance. In these tasks, one tries to describe how an outcome variable y varies with conditions, or variables $(x_1, \ldots, x_q)$. Qualitatively, a *main effect* is the effect of a variable x_i, $i \in [\![q]\!]$ on y *independent* of other variables $(x_j)_{j\neq i}$. On the contrary, an *interaction* refers to the effect of a variable x_i, $i \in [\![q]\!]$ on y *dependent* on another variable x_j, $j \neq i$. For example, in waterbird monitoring, the outcome y is a count variable, and $(x_1, \ldots, x_q)$ are geographical and meteorological information about the place and time where the birds were counted. Suppose that x_i is a measure of the rainfall. A main effect may be: the rainfall has a positive impact on the bird counts across all sites and time points. An interaction may be: the rainfall has a positive impact on the bird counts in sites located far from urban centers, and a negative impact on the sites located close to urban centers.

Log-linear models In count data analysis, log-linear models (Kateri, 2014, Chapter 4) are often used to estimate main effects and interactions. Suppose that $\boldsymbol{Y}$ is a count table with independent entries. Equivalently, each observation $\boldsymbol{Y}_{i,j}$ can be represented as a triplet (i_k, j_k, y_k), $k \in [\![m_1 m_2]\!]$, where (i_k, j_k) are two categorical variables indicating the row and column where the count y_k is located in the table $\boldsymbol{Y}$. In its simplest form, the log-linear model assumes:

$$\log \mathbb{E}\left[\boldsymbol{Y}_{i,j}\right] = \mu + \alpha_i + \beta_j. \tag{2.30}$$

In (2.30), μ is an offset, and α_i and β_j are the main effects of the row i and column j, respectively. For example, in waterbird monitoring, if i is a site of major importance, α_i may be large, indicating that the bird counts are large in site i *across all years*. Respectively, if year j was favorable (resp. unfavorable), β_j may be large (resp. small), indicating that the bird counts are large (resp. small) during year j *across all sites*. Model (2.30) is simplistic, and often times does not reflect reality, as sites and years are known to *interact*. Especially at a large regional scale, a specific year j may be favorable in some sites, and unfavorable in others. For instance, the weather conditions may vary across North Africa, so that year j was a favourable year for birds in Morocco, but a defavourable year for birds in Libya. To incorporate such interactions, model (2.30) may be generalized into:

$$\log \mathbb{E}\left[\boldsymbol{Y}_{i,j}\right] = \mu + \alpha_i + \beta_j + \boldsymbol{\Theta}_{i,j}, \tag{2.31}$$

where $\boldsymbol{\Theta}_{i,j}$ is an interaction term between the i-th row and the j-th column. Model (2.31) above is overparametrized, or saturated. To reduce the number of parameters, a low-rank structure has often been imposed to the interaction matrix $\boldsymbol{\Theta}$ (see, e.g., Goodman (1985); de Falguerolles (1998)), thus obtaining the following model:

$$\log \mathbb{E}\left[\boldsymbol{Y}_{i,j}\right] = \mu + \alpha_i + \beta_j + \boldsymbol{\Theta}_{i,j},\ \operatorname{rank}(\boldsymbol{\Theta}) \leq r, \tag{2.32}$$

where r is a predefined maximal rank. Thus, the parameter matrix $\boldsymbol{X}$, dfined by $\boldsymbol{X} = (\mu + \alpha_i + \beta_i + \boldsymbol{\Theta}_{i,j})$ consists of the superimposition of a two-way linear regression term and a low-rank component.

Multilevel regression Other examples of models where main effects and interactions are estimated are multilevel regression models (Gelman and Hill, 2007). Multilevel regression aims to analyze hierarchically structured data, where examples (patients, students, etc.) are nested within groups (hospitals, schools, etc.). The idea behind it is that several regression models may be fitted in each group, and that the regression coefficients may depend on the groups themselves.

In the Traumabase example, denote $\boldsymbol{Y} = (\boldsymbol{Y}_{i,j})$ the data frame containing the patients in rows and the attributes in columns. Consider the j-th variable, and assume it is quantitative (time spent in critical care for instance). If $\boldsymbol{U}_i \in \mathbb{R}^K$ is a vector of patients characteristics, a regression model may be:

$$\boldsymbol{Y}_{i,j} = \mu_j^0 + \langle \alpha_j^0, \boldsymbol{U}_i \rangle + \boldsymbol{\Theta}_{i,j}^0. \tag{2.33}$$

In (2.33), μ_j^0 is an intercept, α_j^0 is a vector of regression coefficients, and $\boldsymbol{\Theta}_{i,j}^0$ is a residual. Note that, in (2.33), the hospital center where individual i was treated is not taken into account, and the regression coefficients are the same for every trauma center. However, (2.33) may not reflect reality. For instance, some hospitals are known to treat patients with more severe injuries, which may lead to the average time in critical care

being larger in these centers, compared to the overall average. To model this variability across hospitals, model (2.33) may be generalized into

$$\boldsymbol{Y}_{i,j} = \mu^0_{c(i),j} + \langle \alpha^0_j, \boldsymbol{U}_i \rangle + \boldsymbol{\Theta}^0_{i,j}, \tag{2.34}$$

where $c(i)$ indicates the group to which individual i belongs. This corresponds to the *varying intercept* multilevel regression framework (Gelman and Hill, 2007), where the intercept $\mu^0_{c(i),j}$ depends on the group, but the regression coefficients α^0_j are constant across hospitals. Model (2.34) may be further generalized to account for variability of the regression coefficients:

$$\boldsymbol{Y}_{i,j} = \mu^0_{c(i),j} + \langle \alpha^0_{c(i),j}, \boldsymbol{U}_i \rangle + \boldsymbol{\Theta}^0_{i,j}. \tag{2.35}$$

Such models are also referred to as *random effects* models, when the variations of the coefficients (intercept, slope, or both) across groups, are modeled as random.

In addition, the residuals $\boldsymbol{\Theta}^0_{i,j}$ are sometimes also estimated, in which case some assumptions need to be made, to constrain the parameter space. In Chapter 6, we study such a framework, and constrain the matrix of residuals (or interactions) $\boldsymbol{\Theta}^0_{i,j}$ to be of low-rank. Intuitively, this may be interpreted as assuming that a few archetypical individuals and a few summary variables are sufficient to characterize the interactions.

2.4.2 Low-rank plus sparse matrix decomposition

Low-rank plus sparse decomposition consists in the following problem. Suppose we (partially) observe a data matrix $\boldsymbol{Y}$, which is the superimposition of a low-rank component and a sparse component. Is it possible to recover both components? This problem has in fact many practical applications, for instance in graphical modeling (Chandrasekaran et al., 2012) and robust PCA (Candès et al., 2011). A large body of work has tackled the problem of reconstructing a sparse and a low-rank term exactly from the observation of their sum. Formally, assume the data matrix can be decomposed into:

$$\boldsymbol{Y} = \boldsymbol{\Theta}^0 + \boldsymbol{A}^0, \tag{2.36}$$

where $\boldsymbol{\Theta}^0$ is a low-rank matrix, and $\boldsymbol{A}^0$ is entry-wise sparse. For example, the matrix $\boldsymbol{A}^0$ may contain gross errors or corruptions. Chandrasekaran et al. (2011) derived sufficient conditions under which such a model is identifiable, based on a *rank-sparsity incoherence* principle, detailed later on. Under this condition, they prove that the two components $\boldsymbol{\Theta}^0$ and $\boldsymbol{A}^0$ may be recovered *exactly* with the following convex program:

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{\Theta}\|_* + \lambda \|\boldsymbol{A}\|_1 \\ \text{subject to} & \boldsymbol{\Theta} + \boldsymbol{A} = \boldsymbol{Y}, \end{array} \tag{2.37}$$

where $\|\boldsymbol{A}\|_1$ denotes the ℓ_1 norm of the matrix $\boldsymbol{A}$ (the sum of entries in absolute value). The result of Chandrasekaran et al. (2011) may be summarized as follows. For $\boldsymbol{M} \in \mathbb{R}^{m_1 \times m_2}$, let $R(\boldsymbol{M})$ be the tangent space at $\boldsymbol{M}$ with respect to the variety of all matrices with rank less than or equal to $\operatorname{rank}(\boldsymbol{M})$. Define the quantity:

$$\xi(\boldsymbol{M}) := \max_{\boldsymbol{N} \in R(\boldsymbol{M}), \|\boldsymbol{N}\| \leq 1} \|\boldsymbol{N}\|_\infty, \tag{2.38}$$

with $\|\boldsymbol{N}\|_\infty$ denoting the largest entry of $\boldsymbol{N}$ in absolute value. Intuitively, if $\xi(\boldsymbol{M})$ is small, then $\boldsymbol{M}$ cannot be too sparse. Similarly, let $S(\boldsymbol{M})$ be the tangent space at $\boldsymbol{M}$

with respect to the variety of all matrices with number of nonzero entries less than or equal to $\|\boldsymbol{M}\|_0$. Define the quantity:

$$\mu(\boldsymbol{M}) := \max_{\boldsymbol{N} \in S(\boldsymbol{M}), \|\boldsymbol{N}\|_\infty \leq 1} \|\boldsymbol{N}\|. \tag{2.39}$$

In (2.38), $\|\boldsymbol{N}\|$ denotes the operator norm (the largest singular value). Intuitively, if $\mu(\boldsymbol{M})$ is small, then the singular values $\boldsymbol{M}$ cannot be too large.

Theorem 3 (Theorem 2, Chandrasekaran et al. (2011)). *Let $\boldsymbol{Y} = \boldsymbol{\Theta}^0 + \boldsymbol{A}^0$ with $\xi(\boldsymbol{\Theta}^0)\mu(\boldsymbol{A}^0) < 1/6$. Then, for $\lambda = \sqrt{3\xi(\boldsymbol{\Theta}^0)/\mu(\boldsymbol{A}^0)}$, the unique minimum of* (2.37) *is $(\boldsymbol{\Theta}^0, \boldsymbol{A}^0)$.*

Hsu et al. (2011) studied a similar model, and also provided recovery guarantees. In Candès et al. (2011), the authors studied decomposition (2.36), under a probabilistic model where the locations of the nonzero entries in $\boldsymbol{A}^0$ are chosen uniformly at random. Under this model, they show that, if $\boldsymbol{\Theta}^0$ satisfies the same incoherence condition as in Theorem 1, then $\boldsymbol{\Theta}^0$ and $\boldsymbol{A}^0$ can be recovered exactly without requiring the rank-sparsity incoherence condition. In particular, this allows to recover matrices with larger ranks and sparsity patterns. Xu et al. (2010) extended the model to study *column-wise sparsity*. Mardani et al. (2013) studied an even broader framework with general sparsity pattern, where $\boldsymbol{Y} = \boldsymbol{\Theta}^0 + \boldsymbol{R}\boldsymbol{A}^0$, with $\boldsymbol{R} \in \mathbb{R}^{m_1 \times p}$ is a compression matrix and $\boldsymbol{A}^0 \in \mathbb{R}^{p \times m_2}$ is a sparse matrix. In this model, Mardani et al. (2013) determined conditions under which exact recovery of both $\boldsymbol{\Theta}^0$ and $\boldsymbol{A}^0$ is possible.

In many instances, it is unrealistic to assume that the data matrix $\boldsymbol{Y}$ is observed exactly, and noisy models are more plausible. In Klopp et al. (2017), the authors study noisy matrix completion in the presence of entry-wise or column-wise corruptions:

$$\boldsymbol{Y} = \boldsymbol{\Theta}^0 + \boldsymbol{A}^0 + \boldsymbol{E}, \tag{2.40}$$

where $\boldsymbol{E} = (\boldsymbol{E}_{i,j})$ is a matrix of additive noise. For the entry-wise sparse case, they consider the following estimation procedure:

$$\begin{array}{ll} \text{minimize} & \frac{1}{n}\sum_{(i,j)\in\Omega}(\boldsymbol{Y}_{i,j} - \boldsymbol{\Theta}_{i,j} - \boldsymbol{A}_{i,j})^2 + \lambda_1\|\boldsymbol{\Theta}\|_* + \lambda_2\|\boldsymbol{A}\|_1 \\ \text{subject to} & \|\boldsymbol{\Theta}\|_\infty \leq a, \|\boldsymbol{A}\|_\infty \leq a. \end{array} \tag{2.41}$$

Similar to noisy matrix completion (Theorem 2), they only assume an upper bound a on the largest entries in absolute value of the matrices $\boldsymbol{\Theta}^0$ and $\boldsymbol{A}^0$ which, here also, is a much weaker assumption than the incoherence condition of Hsu et al. (2011); Candès et al. (2011); Mardani et al. (2013). Klopp et al. (2017) shows that the estimation procedure (2.41) is minimax optimal. Denote by $\|\boldsymbol{A}\|_0$ the number of nonzero entries in the matrix $\boldsymbol{A}$. The result of Klopp et al. (2017) may be summarized as follows.

Theorem 4 (Corollary 11, Klopp et al. (2017)). *Assume that:*

- *The probability of observing an entry in column j (resp. row i) is bounded above by $L/\min(m_1, m_2)$, with $L \geq 1$.*
- *Every entry is observed with probability at least $(\mu m_1 m_2)^{-1}$, with $\mu \geq 1$.*
- *The noise $(\boldsymbol{E}_{i,j})$ is subexponential: there exists $K > 0$ such that for all $(i,j) \in \Omega$, $\mathbb{E}\left[|\boldsymbol{E}_{i,j}|/K\right] \leq \mathrm{e}$.*

Then, there exist values of λ_1 and λ_2 (which we omit here for sake of clarity), such that, for n large enough, any solution of (2.41) *$(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{A}})$ satisfies:*

$$\frac{\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^0\|_F^2}{m_1 m_2} + \frac{\|\hat{\boldsymbol{A}} - \boldsymbol{A}^0\|_F^2}{m_1 m_2} \lesssim \left(\frac{\mathrm{rank}(\boldsymbol{\Theta}^0)m_2 + \|\boldsymbol{A}^0\|_0}{n} + \frac{\|\boldsymbol{A}^0\|_0}{m_1 m_2}\right). \tag{2.42}$$

2.4.3 Multilevel PCA

Multilevel simultaneous component analysis (MLSCA, Timmerman (2006)), or multilevel PCA (MLPCA), is an extension of PCA designed to model *numeric data with group structures*. Along this dissertation, we will always refer to it as MLPCA, for simplicity. MLPCA is based on multilevel SVD, which consists in decomposing the variability of the data into two components, the between and within groups variability, and performing an SVD on both parts. Assume that the numeric data $\boldsymbol{Y}$ is naturally the row concatenation of K smaller data sets $\boldsymbol{Y}_k \in \mathbb{R}^{n_k \times m_2}$, $k \in [\![K]\!]$, such that the k-th group contains n_k individuals and $\sum_{k=1}^{K} n_k = m_1$:

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \hline \boldsymbol{Y}_2 \\ \hline \vdots \\ \hline \boldsymbol{Y}_K \end{pmatrix} \begin{matrix} \updownarrow n_1 \\ \updownarrow n_2 \\ \vdots \\ \updownarrow n_K \end{matrix} . \tag{2.43}$$

For a group $k \in [\![K]\!]$, an individual of the k-th group $i_k \in [\![n_k]\!]$ and a variable $j \in [\![p]\!]$, we denote by $y_{k,i_k,j}$ the value of variable j taken by individual i_k in group k. The entries of $\boldsymbol{Y}$ can be decomposed, for a group $k \in [\![K]\!]$, an individual $i_k \in [\![n_k]\!]$ in the k-th group and a variable $j \in [\![p]\!]$, as

$$y_{k,i_k,j} = \mathbf{m}_j + (\mathbf{m}_j^k - \mathbf{m}_j) + (y_{k,i_k,j} - \mathbf{m}_j^k), \tag{2.44}$$

where $\mathbf{m}_j = m_1^{-1} \sum_{i=1}^{m_1} \boldsymbol{Y}_{i,j}$ is the mean of the j-th variable, and $\mathbf{m}_j^k = m_1^{k-1} \sum_{i=1}^{m_1^k} y_{k,i_k,j}$ is the mean of the j-th variable in group k. The deviation of group k to the overall mean of variable j, $\mathbf{m}_j^k - \mathbf{m}_j$, is usually referred to as the *between* groups variability, and the deviation of individual i_k to the mean of variable j in group k, $y_{k,i_k,j} - \mathbf{m}_j^k$, as the *within* groups variability. Written in matrix form, we obtain:

$$\boldsymbol{Y} = \mathbb{1}_{m_1} \mathbf{m}^\top + \boldsymbol{Y}_b + \boldsymbol{Y}_w,$$

where $\mathbf{m} = (\mathbf{m}_1, \ldots, \mathbf{m}_K)$, $(\boldsymbol{Y}_b)_{i,j} = \mathbf{m}_j^k - \mathbf{m}_j$, for i in group k and $j \in [\![m_2]\!]$, and $(\boldsymbol{Y}_w)_{i,j} = y_{k,i_k,j} - \mathbf{m}_j^k$, for i in group k and $j \in [\![m_2]\!]$. The multilevel extension of PCA, MLPCA, consists in assuming two low-rank models, for the between matrix $\boldsymbol{Y}_b$, that we approximate by a matrix of rank r_b, and for the within matrix $\boldsymbol{Y}_w$, that we approximate by a matrix of rank r_w. This yields the following decomposition:

$$\boldsymbol{Y} = \mathbb{1}_{m_1} \mathbf{m}^\top + \mathbf{U}_b \mathbf{V}_b^\top + \mathbf{U}_w \mathbf{V}_w^\top + \boldsymbol{E}. \tag{2.45}$$

Model (2.45) implies that there are two target low-rank matrices, which we seek to recover from noisy observations of their sum. In Chapter 8, we will introduce extensions of MLPCA to heterogeneous and incomplete data.

2.5 Summary of contributions

Despite the abundant literature on low-rank matrix approximations for data analysis, a number of shortcomings still limit their application. In particular, although a number of methods address multi-source, heterogeneous or incomplete data, to the best of our knowledge, none of them address these three problems simultaneously. As a result, in relation with the applications we have in mind, existing works suffer from either

model-related limitations, or theoretical gaps. The objective of this dissertation is to provide general models, which adapt to multi-source, heterogeneous and incomplete data simultaneously, and in particular to the ecological and medical applications described in Sections 2.1.1 and 2.1.2. For these models, we will be committed to providing complete methodologies, from estimation methods, to theoretical guarantees and ready-to-use implementations.

2.5.1 Low-rank model with covariates for count data analysis

In Chapter 3, we address the problem of analyzing and imputing incomplete count data using side information. Consider a count table $\boldsymbol{Y}$ of size $m_1 \times m_2$, from which some values are missing. For example, $\boldsymbol{Y}$ might contain abundance data about a particular species, measured across ecological sites (rows) and time points (columns). Assume that, in addition to $\boldsymbol{Y}$, two covariate matrices $\boldsymbol{R} \in \mathbb{R}^{m_1 \times K_1}$ and $\boldsymbol{C} \in \mathbb{R}^{m_2 \times K_2}$ are available. $\boldsymbol{R}$ contains side information about the rows of $\boldsymbol{Y}$, such as geographical information about the sampling sites, and $\boldsymbol{C}$ contains side information about its columns, for example meteorological characteristics of the different time points. Both covariate matrices may contain quantitative and qualitative row and column features. Our objective is dual. On the one hand, we seek to take advantage of available side information $\boldsymbol{R}$ and $\boldsymbol{C}$ to better impute the missing counts: if the covariates are good predictors of the counts $\boldsymbol{Y}$, using them may improve the imputation. On the other hand, we seek to estimate the relation between $\boldsymbol{R}$ and $\boldsymbol{C}$ and the counts $\boldsymbol{Y}$ *in spite of the missing observations*. In particular, to detect factors which may be associated to larger or smaller counts. We provide a complete methodology including a general model, an estimation procedure for which we derive statistical guarantees, and an optimization algorithm. In addition, we evaluate our method empirically on synthetic and ecological data.

We consider a parametric Poisson probabilistic model, and assume the counts to follow distributions of the form:

$$\boldsymbol{Y}_{i,j} \sim \mathcal{P}(\exp(\boldsymbol{X}^0_{i,j})), \tag{2.46}$$

where $\mathcal{P}(\lambda)$ denotes the Poisson distribution of intensity λ. We model the effects of the covariates on the counts through the following log-linear model (2.32):

$$\boldsymbol{X}^0_{i,j} = \mu^0 + \boldsymbol{R}_{i,\cdot}\alpha^0 + \boldsymbol{C}_{j,\cdot}\beta^0 + \boldsymbol{\Theta}^0_{i,j}. \tag{2.47}$$

In (2.47), $\mu^0 \in \mathbb{R}$ is an offset, $\alpha^0 \in \mathbb{R}^{K_1}$ is a vector modeling the effects of the row covariates, and $\beta^0 \in \mathbb{R}^{K_2}$ is a vector modeling the effects of the column covariates. Finally, $\boldsymbol{\Theta}^0$ is a row-column interaction matrix, which we assume to be low-rank. Intuitively, this low-rank assumption may be interpreted as modeling a few archetypical rows and columns, which interact in a multiplicative manner. Models related to (2.47) have been considered before in statistical ecology applications, for instance in Brown et al. (2014); ter Braak et al. (2017). However, to the best of our knowledge, their theoretical and empirical properties have not been thoroughly studied. On the other hand, the literature on convex low-rank matrix completion is abundant and benefits from a substantial theoretical background. However, as far as we know, solutions which incorporate side information consider only numeric data (Mao et al., 2017). Some works on generalized low-rank models allow to incorporate covariates, but they have no statistical guarantees (Fithian and Mazumder, 2018; Chiquet et al., 2018). In this sense, the scope of this paper is to develop a complete methodology for the inference of model (2.47),

with theoretical guarantees and ready-to-use implementation, bridging the gap between convex low-rank matrix completion and model-based count data analysis. We estimate the parameters of model (2.47) through a convex program, where a Poisson data fitting term penalized by the nuclear norm of $\boldsymbol{\Theta}$ is minimized:

$$\begin{array}{ll} \text{minimize} & \mathcal{L}(\boldsymbol{Y}, \mu, \alpha, \beta, \boldsymbol{\Theta}) + \lambda \|\boldsymbol{\Theta}\|_* \\ \text{subject to} & |\mu + \boldsymbol{R}_{i,\cdot}\alpha + \boldsymbol{C}_{j,\cdot}\beta + \boldsymbol{\Theta}_{i,j}| \leq a,\ (i,j) \in [\![m_1]\!] \times [\![m_2]\!], \end{array} \tag{2.48}$$

where $\mathcal{L}$ is the Poisson loss:

$$\mathcal{L}(\boldsymbol{Y}, \mu, \alpha, \beta, \boldsymbol{\Theta}) = \sum_{(i,j)\in\Omega} [-\boldsymbol{Y}_{i,j}(\mu + \boldsymbol{R}_{i,\cdot}\alpha + \boldsymbol{C}_{j,\cdot}\beta + \boldsymbol{\Theta}_{i,j}) + \exp(\mu + \boldsymbol{R}_{i,\cdot}\alpha + \boldsymbol{C}_{j,\cdot}\beta + \boldsymbol{\Theta}_{i,j})].$$

On the theoretical side, our main contribution is to show that the estimation procedure (2.48) guarantees an estimation error of the same order of magnitude as the minimax optimal rates of Klopp (2014) and Lafond (2015). In particular, we show that (under assumptions detailed in Chapter 3):

$$\|\hat{\boldsymbol{X}} - \boldsymbol{X}^0\|_F^2 \lesssim \frac{\operatorname{rank}(\boldsymbol{\Theta}^0) m_2}{p}, \tag{2.49}$$

where $\lesssim$ denotes the inequality up to constant ant logarithmic factors, and every entry is observed with probability at least $p > 0$. On the practical side, our main contribution is to propose an optimization algorithm, and to evaluate the method on synthetic and ecological data. We demonstrate empirically that our method outperforms state-of-the-art count data imputation procedures, in particular when the proportion of missing values is large and the main effects and interactions are of similar orders of magnitude. We also provide interpretation tools through visual displays, and illustrate the method with the analysis of a well-known plant abundance data set. In particular, we show that the arising interpretation is consistent with known results from the original study. The method is implemented in the R package `lori`, available on the CRAN, for which we provide a tutorial in Chapter 5.

2.5.2 Estimation of waterbird population trends with multiple imputations

Chapter 3 provided a single imputation procedure for count data with missing values, when side information is available. Although the numerical results were promising, we identified two directions of improvement.

First, in settings where the proportion of observed values is very small, and especially when the number of covariates is simultaneously large, the model developed in Chapter 3, which does not constrain the covariate coefficients, may be statistically and computationally limiting. In Chapter 4, we extend the methodology introduced in Chapter 3, by adding a LASSO-type penalty to regularize the vector of main effects. We thus obtain an estimation problem with a hybrid LASSO and nuclear norm penalty. We introduce a mixed coordinate gradient descent algorithm (MCGD), which efficiently solves the resulting optimization problem in large dimensions. This new estimation procedure also defines a single imputation method.

Second, single imputation is useful when one seeks to predict the missing entries as well as possible. However, if data analyses are performed after imputation, multiple imputation (Rubin, 1987)–which consists in predicting several plausible values for

the missing entries in order to reflect uncertainty in the imputation–is the statistically sound approach. We develop a bootstrap-based multiple imputation procedure, based on our doubly penalized single imputation method and a resampling method which combines nonparametric and parametric bootstrap. Note that, because of the two penalties, our imputation procedure is biased, and thus we cannot hope to obtain valid confidence regions for our prediction even with multiple imputation. However, the resampling procedure allows to derive intervals of variability which reflect the variability of our imputations with respect to the missing values and the noise of the observations. We evaluate the methods in terms of imputation and coverage in realistic settings, and show that it improves on state-of-the art count data imputation methods which are currently used in ecology. The complete methods, implemented in the R package `lori`, is finally used to impute the waterbird data set.

Let $\boldsymbol{Y} \in \mathbb{R}^{m_1 \times m_2}$ denote the data set containing the waterbird abundances: the rows of $\boldsymbol{Y}$ correspond to different sites, and the columns to different time points (around the 15$^{\text{th}}$ of January, during the years 1990 to 2017). The entry $\boldsymbol{Y}_{i,j}$ contains an integer number corresponding to the number of birds counted at the i-th site, in the j-th year. As discussed in Section 2.1.1, the data set $\boldsymbol{Y}$ has many missing values, and we seek to impute them in order to compute yearly total abundances, i.e. to compute the column-wise sums of $\boldsymbol{Y}$. More precisely, we will perform multiple imputations, which allow us to compute intervals of variability for the predicted values.

To do so, we start by introducing a new single imputation procedure, which take advantage of available side information. Let $\boldsymbol{U} \in \mathbb{R}^{m_1 m_2 \times K}$ denote a matrix of covariates about the rows and columns of $\boldsymbol{Y}$. For every entry $\boldsymbol{Y}_{i,j}$ corresponds a row of $\boldsymbol{U}$, denoted $\boldsymbol{U}(i,j) := \boldsymbol{U}_{(j-1)m_1+i,\cdot} \in \mathbb{R}^K$, which contains geographical information about the i-th site (latitude, longitude, etc.), meteorological information about the j-th year (temperature abnormalities, etc.), as well as information about the pair (i-th site, j-th year), such as the yearly rainfall in the site's area, and yearly economical indices of the site's country. We use a parametric Poisson model similar to (2.46) and (2.47):

$$\boldsymbol{Y}_{i,j} \sim \mathcal{P}(\mu^0_{i,j}),\ \log(\mu^0_{i,j}) = \alpha^0_i + \beta^0_j + \sum_{k=1}^{K} \epsilon^0_k \boldsymbol{U}(i,j)_k + \boldsymbol{\Theta}^0_{i,j}. \tag{2.50}$$

In (2.50), the vectors $\alpha^0 \in \mathbb{R}^{m_1}$ and $\beta^0 \in \mathbb{R}^{m_2}$ contain main row and column effects, $\epsilon^0 \in \mathbb{R}^K$ contains main effects of covariates, and $\boldsymbol{\Theta}^0$ is a matrix of interactions. We estimate the parameters of model (2.50) by minimizing the Poisson negative log-likelihood, with two additional penalties:

$$(\hat{\alpha}, \hat{\beta}, \hat{\epsilon}, \hat{\boldsymbol{\Theta}}) \in \operatorname{argmin} \mathcal{L}(\boldsymbol{Y}; \alpha, \beta, \epsilon, \boldsymbol{\Theta}) + \lambda_1 \|\boldsymbol{\Theta}\|_* + \lambda_2 \left(\|\alpha\|_1 + \|\beta\|_1 + \|\epsilon\|_1\right). \tag{2.51}$$

In (2.51), $\mathcal{L}(\boldsymbol{Y}; \alpha, \beta, \epsilon, \boldsymbol{\Theta})$ is the Poisson negative log-likelihood:

$$\sum_{(i,j)} \Omega_{i,j}[-\boldsymbol{Y}_{i,j}(\alpha_i + \beta_j + \sum_{k=1}^{K} \epsilon_k \boldsymbol{U}(i,j)_k + \boldsymbol{\Theta}_{i,j}) + \exp(\alpha_i + \beta_j + \sum_{k=1}^{K} \epsilon_k \boldsymbol{U}(i,j)_k + \boldsymbol{\Theta}_{i,j})].$$

In comparison to the method described in Chapter 3, we introduce an additional LASSO-type regularization term $\lambda_2 \left(\|\alpha\|_1 + \|\beta\|_1 + \|\epsilon\|_1\right)$. Indeed, in practice, this improved the imputation, reduced the computational time, and had the advantage of selecting important covariates, which is useful for interpretation purposes. We will provide statistical

guarantees for the estimation problem (2.51) in Chapter 6, as a special case of our general result. Finally, the single imputation procedure consists in predicting, for every $(i, j) \in [\![m_1]\!] \times [\![m_2]\!] \setminus \Omega$:

$$\hat{\boldsymbol{Y}}_{i,j} = \exp(\hat{\alpha}_i + \hat{\beta}_j + \sum_{k=1}^{K} \hat{\epsilon}_k \boldsymbol{U}(i,j)_k + \hat{\boldsymbol{\Theta}}_{i,j}). \tag{2.52}$$

We are not the first to consider imputation of count data using side information. In particular, TRIM (Trends and Indices in Monitoring data, Pannekoek and van Strien (2001)), is a popular count data imputation method used by ecologists, and is also based on a Poisson log-linear model. Compared to TRIM, our model allows to include quantitative and categorical covariates, while TRIM requires to categorize the quantitative traits. Second, TRIM does not model row-column interactions, which we believe to be important in the waterbird application. On the other hand, low-rank models for incomplete count data such as Correspondence Analysis (CA, Greenacre (1984)) and Poisson matrix completion (Cao and Xie, 2016) may be interpreted as low-rank interaction models, but do not incorporate covariates.

We produce multiple imputations by combining a resampling procedure to the single imputation procedure (2.52). First, we produce M new data set using nonparametric bootstrap. To generate them, we interpret the count data matrix $\boldsymbol{Y}$ as a contingency table. We assume that the location of birds in time and space are independent and identically distributed: each bird is observed in the site i and at the year j with probability $\pi_{i,j}$, where the probability $\pi_{i,j}$ is the same for all birds, and the birds are observed independently. Based on this assumption, we generate new samples $\tilde{\boldsymbol{Y}}^1, \ldots, \tilde{\boldsymbol{Y}}^M$ using nonparametric bootstrap, and fit our Poisson model to each of them. We thus obtain M imputation model, and produce M imputed data sets $\hat{\boldsymbol{Y}}^1, \ldots, \hat{\boldsymbol{Y}}^M$. Then, to account for the uncertainty related to the missingness pattern, we generate new missing values in the completed data sets $\hat{\boldsymbol{Y}}^1, \ldots, \hat{\boldsymbol{Y}}^M$, and re-estimate M imputation models based on new data with different observed values and different missingness patterns. This first process accounts for uncertainty about the imputation model.

Then, we must also assess the variability of each imputation model. To do so, we use parametric bootstrap to estimate the variability of each model separately. In other word, for each of the M imputation models, we generate D completed data set, using the same stochastic imputation procedure. Finally, we obtain MD imputed data sets.

On the practical side, we demonstrate empirically that, in the regime where the waterbirds data set is located, with many missing values and some large interactions, the proposed imputation method outperforms existing techniques which are currently used by ecologists. We also evaluate the coverage of the multiple imputation procedure, and the robustness of the method to model misspecification, with an experiment on zero-inflated and overdispersed count data. We finally apply the method to the estimation of the temporal trends three waterbird species, and provide estimates of the total yearly bird counts as well as intervals of variability for these estimates.

2.5.3 Main effects and interactions in mixed and incomplete data frames

In Chapter 6, we develop a general framework for main effects and interactions in mixed and incomplete data frames, based on a heterogeneous exponential family probabilistic model. The main contribution with respect to prior work, is to generalize low-rank models to accommodate side information and heterogeneous noise *simultaneously*. In particular, we will consider hybrid low-rank structures, where the parameter matrix $\boldsymbol{X}^0$ is decomposed into two components, one of them being low-rank, and the other a regression term on an arbitrary, fixed, dictionary of matrices. Our noise model is also very general, through the use of column-wise heterogeneous loss functions, in the spirit of Udell et al. (2016). After introducing our general model, and motivating its main features through several examples of interest in applications, we propose an estimation procedure. The estimation is based on the minimization of a heterogeneous loss functions, with a hybrid penalty inducing low-rank solutions for the matrix of interactions and sparse solution for the vector of main effects. We derive a block coordinate gradient descent algorithm (BCGD), and provide a convergence result. Then, we derive statistical guarantees for our estimates, in the form of upper and lower bounds on the estimation error of the main effects and interactions *simultaneously*. We then evaluate the method in terms of estimation, and demonstrate that it compares favorably to state-of-the art methods for mixed data imputation in terms of prediction of the missing entries. Finally, we illustrate the applicability of the method with a short analysis of a subsample of the Traumabase data set.

Formally, we model mixed data types using a data-fitting term based on heterogeneous exponential family quasi-likelihoods. Let h, and g be functions, and denote by $\mathrm{Exp}^{(h,g)} = \{f_x^{(h,g)} : x \in \mathbb{R}\}$ the canonical exponential family with base function h and link function g. We denote by $f_x^{(h,g)}$ the density given by:

$$f_x^{(h,g)}(y) = h(y) \exp\left(yx - g(x)\right), \tag{2.53}$$

The exponential family is a flexible framework for different data types. For example, for numerical data, we set $g(x) = x^2\sigma^2/2$ and $h(y) = (2\pi\sigma^2)^{-1/2}\exp(-y^2/\sigma^2)$. In this case, $\mathrm{Exp}^{(h,g)}$ is the family of Gaussian distributions with mean $\sigma^2 x$ and variance σ^2. For count data, we set $g(x) = \exp(ax)$ and $h(y) = 1/y!$, where $a \in \mathbb{R}$. In this case, $\mathrm{Exp}^{(h,g)}$ is the family of Poisson distributions with intensity $\exp(ax)$. For binary data, $g(x) = \log(1 + \exp(x))$ and $h(y) = 1$. Here, $\mathrm{Exp}^{(h,g)}$ is the family of Bernoulli distributions with success probability $1/(1+\exp(-x))$. To model mixed data, we choose a collection $\{(g_j, h_j),\ j \in [\![m_2]\!]\}$ of link functions and base functions corresponding to the types of each column in $\boldsymbol{Y}$ (numeric, binary, etc.). For each $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, we denote by $\boldsymbol{X}^0_{i,j}$ the value of the parameter minimizing the divergence between the distribution of $\boldsymbol{Y}_{i,j}$ and the exponential family $\mathrm{Exp}^{(h_j,g_j)}$, $j \in [\![m_2]\!]$. Finally, we consider a setting with missing values, so that $\boldsymbol{Y}_{i,j}$ is observe only for a subset of entries. We use the following data-fitting term defined by the heterogeneous exponential family negative quasi log-likelihood

$$\mathcal{L}(\boldsymbol{X}; \boldsymbol{Y}, \Omega) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \Omega_{i,j}\left\{-\boldsymbol{Y}_{i,j}\boldsymbol{X}_{i,j} + g_j(\boldsymbol{X}_{i,j})\right\}. \tag{2.54}$$

We model main effects and interactions in the parameter space, through decompo-

sitions of the form:

$$\boldsymbol{X}^0 = \sum_{k=1}^{N} \alpha_k^0 \boldsymbol{U}^k + \boldsymbol{\Theta}^0, \tag{2.55}$$

where $\mathcal{U} := (\boldsymbol{U}^1, \ldots, \boldsymbol{U}^N)$ is a dictionary of matrices of $\mathbb{R}^{m_1 \times m_2}$. Model (2.55) is closest to the decomposition studied in Mardani et al. (2013). In this work, the authors consider the problem of separating a low-rank matrix and the product of a compression matrix by a sparse matrix $\boldsymbol{X}^0 = \boldsymbol{\Theta}^0 + \boldsymbol{RA}$. Mardani et al. (2013) study the *exact* decomposition problem, when $\boldsymbol{X}^0$ is observed directly and without errors. The main difference with our contribution in Chapter 6, is that we consider a noisy setting. Furthermore, we consider a general and heterogeneous noise framework. Finally, we estimate α^0 and $\boldsymbol{\Theta}^0$ we the following program:

$$\begin{array}{rll} (\hat{\alpha}, \hat{\boldsymbol{\Theta}}) \in & \underset{\text{subject to}}{\text{argmin}} & \mathcal{L}(\mathsf{f}_U(\alpha) + \boldsymbol{\Theta}; Y, \Omega) + \lambda_1 \|\boldsymbol{\Theta}\|_* + \lambda_2 \|\alpha\|_1 \\ & & \|\alpha\|_\infty \leq a, \|\boldsymbol{\Theta}\|_\infty \leq a, \end{array} \tag{2.56}$$

with $\lambda_1 > 0$ and $\lambda_2 > 0$. We discuss the statistical guarantees of our procedure, with two simultaneous upper bounds on the estimation errors of the sparse and low-rank components. We show that the estimation errors of (2.56) are of the order of:

$$\begin{array}{l} \|\mathsf{f}_U(\alpha^0) - \mathsf{f}_U(\hat{\alpha})\|_2^2 \leq \frac{a\|\alpha^0\|_0}{p} \phi(\mathcal{U}), \\ \|\boldsymbol{\Theta}^0 - \hat{\boldsymbol{\Theta}}\|_F^2 \leq \frac{\mathrm{rank}(\boldsymbol{\Theta}^0) m_2}{p} \frac{a\|\alpha^0\|_0}{p} \phi(\mathcal{U}), \end{array} \tag{2.57}$$

where $\phi(\mathcal{U})$ is a factor which depends on the geometry of the dictionary, and accounts for interplay between main effects and interactions. To assess the tightness of our convergence rates, we derive lower bounds, and show that in a number of situations, our upper bounds are near optimal. We also propose a block coordinate gradient descent algorithm to compute our estimator, and a convergence result. Numerical results are presented to support our theoretical claims. We also show that our method performs comparably to state-of the art mixed data imputation methods in terms of prediction of the missing values. The method is available in the R (R Core Team, 2017) package `mimi`, for which we propose a tutorial in Chapter 7.

2.5.4 Imputation of mixed data with multilevel SVD

In Chapter 8, we introduce an extension of multilevel PCA (MLPCA), which is designed for data sets with categorical and quantitative information, and can be used to impute missing values in multilevel mixed data frames. Recall that in the multilevel setting, the data set $\boldsymbol{Y} \in \mathbb{R}^{m_1 \times m_2}$ is naturally the row concatenation of K smaller data sets $\boldsymbol{Y}_k \in \mathbb{R}^{n_k \times m_2}$, $k \in [\![K]\!]$. $\boldsymbol{Y}$ collects the measurements of m_2 variables across a population of m_1 individuals categorized in K groups, such that the k-th group contains n_k individuals and $\sum_{k=1}^K n_k = m_1$:

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \hline \boldsymbol{Y}_2 \\ \hline \vdots \\ \hline \boldsymbol{Y}_K \end{pmatrix} \begin{array}{l} \updownarrow n_1 \\ \updownarrow n_2 \\ \vdots \\ \updownarrow n_K \end{array}. \tag{2.58}$$

Famous examples include pupils nested within schools or patients within hospitals. Throughout this chapter, we focus on this latter example with a running application

to the Traumabase data set (see Section 2.1.2). In this work, we focus on the imputation of missing values in the multilevel data set $\boldsymbol{Y}$. There are several methodological contributions in this chapter. First, we introduce two multilevel component methods to analyze qualitative and mixed data respectively. Second, we extend these methods to accommodate missing values, and to impute categorical and mixed variables with multilevel structures. We demonstrate on synthetic data that our methods have smaller prediction errors than competitors when the data are generated with a multilevel model. Finally, we illustrate the methods with the imputation of the Traumabase register. We also discuss how the computations may be distributed, as an incentive for hospitals to participate in the program. The methods are implemented in the R (R Core Team, 2017) package missMDA (Josse and Husson, 2016).

We start by proposing a counterpart of MLPCA to analyse categorical variables. Our method is based on multiple correspondence analysis (MCA, ?Husson et al. (2010)), that we extend to handle multilevel structures. More precisely, assume that categorical data are coded as a complete disjunctive table $\boldsymbol{Z}$ where all categories of all variables are represented as indicator vectors. In other words $z_{ic} = 1$ if individual i takes the category c and 0 otherwise. For example, if there are $m_2 = 2$ variables with 2 and 3 levels respectively, we have the following equivalent codings:

$$Y = \begin{pmatrix} 1 & 1 \\ 2 & 3 \\ 1 & 2 \\ 2 & 3 \\ 2 & 2 \\ 2 & 2 \end{pmatrix} \quad \Longleftrightarrow \quad \boldsymbol{Z} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix}.$$

For $1 \leq j \leq m_2$ we denote by C_j the number of categories of variable j, and $C = \sum_{j=1}^{m_2} C_j$ the total number of categories. For $1 \leq c \leq C$, $\boldsymbol{Z}_{.,c}$ is the c-th column of $\boldsymbol{Z}$ corresponding to the indicator of category c. We define $\pi_c = m_1^{-1} \mathbb{1}_{m_1}^\top \boldsymbol{Z}_{.,c}$ the proportion of observations in category c, $\pi = (\pi_1, \ldots, \pi_C)^\top$ and $\boldsymbol{D}_\pi$ the $C \times C$ diagonal matrix with π on its diagonal. Multiple correspondence analysis (MCA) is defined as the SVD of the matrix

$$\boldsymbol{A} = \frac{1}{m_1 m_2} \left(\boldsymbol{Z} - \mathbb{1}_{m_1} \pi^\top \right) {\boldsymbol{D}_\pi}^{-1/2}. \tag{2.59}$$

We introduce the following strategy for multilevel MCA (MLMCA). From the indicator matrix of dummy variables $\boldsymbol{Z}$, we start by defining a between part and a within part. MCA, in the sense of the SVD of a transformed matrix (2.59), will then be applied on each part. For $k \in [\![K]\!]$, define $\boldsymbol{Z}_k$ the sub-matrix of $\boldsymbol{Z}$ containing all categories and the rows corresponding to individuals of group k. The between part is defined block-wise as the mean of the indicator matrix per group k with the following $n_k \times p$ matrices, stacked below one another:

$$\boldsymbol{Z}_{b,k} = \frac{1}{n_k} \mathbb{1}_{n_k} \mathbb{1}_{n_k}^\top \boldsymbol{Z}_k.$$

The entries of $\boldsymbol{Z}_{b,k}$ contain the proportion of observations taking each category in group k (for instance the proportion of individuals carrying some disease in a particular hospital).

Finally

$$\boldsymbol{Z}_b = \left(\begin{array}{c} \boldsymbol{Z}_{b,1} \\ \hline \boldsymbol{Z}_{b,2} \\ \hline \vdots \\ \hline \boldsymbol{Z}_{b,K} \end{array} \right).$$

MCA (2.59) is afterwards applied to the fuzzy indicator matrix $\boldsymbol{Z}_b$, *i.e.* SVD is applied to

$$(\boldsymbol{Z}_b - \mathbb{1}_{m_1}\pi^\top)\boldsymbol{D}_\pi^{-1/2}.$$

This results in obtaining between component scores $\boldsymbol{F}_b \in \mathbb{R}^{m_1 \times Q_b}$ and between loadings $\boldsymbol{V}_b \in \mathbb{R}^{m_1 \times Q_b}$. The estimated between matrix is then $\hat{\boldsymbol{Z}}_b = \boldsymbol{F}_b \boldsymbol{V}_b^\top \boldsymbol{D}_\pi^{1/2} + \mathbb{1}_{m_1}\pi^\top$. As for the within part, MCA is applied to the data where the between part has been swept out, *i.e.* SVD is applied to the following matrix:

$$(\boldsymbol{Z} - \boldsymbol{Z}_b)\, \boldsymbol{D}_\pi^{-1/2}. \tag{2.60}$$

Chapter 3

Low-rank model for count data with covariates

Contents

3.1 Introduction

Let $\boldsymbol{Y}$ be an $m_1 \times m_2$ observation matrix of counts, $\boldsymbol{R} \in \mathbb{R}^{m_1 \times K_1}$ and $\boldsymbol{C} \in \mathbb{R}^{p \times K_2}$ be two matrices containing row and column covariates, respectively. In the waterbird monitoring application introduced Section 2.1.1, the rows of the count table represent ecological sites, and columns represent years. For $(i, j) \in [\![m_1]\!] \times [\![m_2]\!]$, $\boldsymbol{Y}_{i,j}$ counts the abundance of waterbirds measured in site i during the year j. The row feature $\boldsymbol{R}_{i\ell}$, $\ell \in \{1, \ldots, K_1\}$ embeds geographical information about the site i (latitude, longitude, distance to coast, etc.) while the column feature $\boldsymbol{C}_{j\ell}$, $\ell \in \{1, \ldots, K_2\}$ codes meteorological characteristics of the year j (precipitation, etc.). In addition, some entries of $\boldsymbol{Y}$ are missing. For example ecological sites are sometimes inaccessible because of meteorological or political conditions, and therefore cannot be counted. In this chapter, we develop a complete methodology to impute the missing entries, and to analyze the

relationship between the covariates and the counts.

To do so, we assume a probabilistic framework with independent entries $\boldsymbol{Y}_{i,j}$ following a Poisson model

$$\boldsymbol{Y}_{i,j} \sim \mathcal{P}(\mathrm{e}^{\boldsymbol{X}^0_{i,j}}), (i,j) \in [\![m_1]\!] \times [\![m_2]\!], \tag{3.1}$$

and approach the problem through the estimation of the underlying parameter matrix $\boldsymbol{X}^0$. To estimate it, we rely on a hybrid model combining regression on the covariates $\boldsymbol{R}$ and $\boldsymbol{C}$ and on a low-rank assumption. In particular, we build upon existing low-rank models for count data. For instance, the *generalized additive main effects and multiplicative interaction* model, or *row-column* model (see, e.g., Goodman (1985); de Falguerolles (1998)), assumes

$$\boldsymbol{X}^0_{i,j} = \mu^0 + \alpha^0_i + \beta^0_j + \boldsymbol{\Theta}^0_{ij}, \quad \operatorname{rank}(\boldsymbol{\Theta}^0) \leq \min(m_1 - 1, m_2 - 1). \tag{3.2}$$

In this model, μ^0 is an offset, the terms which only depend on the index of the row or column (α^0_i and β^0_j) are called *main effects*, and the terms which depend on both (here $\boldsymbol{\Theta}^0_{i,j}$) are called *interactions* (Kateri, 2014, Section 4.1.2, p.87).

To incorporate side information in this framework, we express the row and column effects α^0_i and β^0_j as regression terms on the covariates. In other words, for $\mu^0 \in \mathbb{R}$, $\alpha^0 \in \mathbb{R}^{K_1}$, $\beta^0 \in \mathbb{R}^{K_2}$ and $\boldsymbol{\Theta}^0 \in \mathbb{R}^{m_1 \times m_2}$,

$$\boldsymbol{X}^0_{i,j} = \mu^0 + \underbrace{\sum_{k=1}^{K_1} \boldsymbol{R}_{i,k}\alpha^0_k}_{\text{row effect}} + \underbrace{\sum_{l=1}^{K_2} \boldsymbol{C}_{j,l}\beta^0_l}_{\text{column effect}} + \boldsymbol{\Theta}^0_{i,j}, \quad \operatorname{rank}(\boldsymbol{\Theta}^0) \leq \min(m_1 - 1, m_2 - 1). \tag{3.3}$$

This extension is relevant in practice for two purposes. First, estimated covariates coefficients (and in particular their signs) can be used to determine whether the studied covariates have positive or negative effects on the counts; this is useful in ecology, to check whether meteorological, geographical or political conditions are favorable or adverse to species. Second, when the proportion of missing values is large, which is often the case in bird monitoring, incorporating (relevant) covariates may improve the imputation significantly. Thus, the estimation of model (3.3) serves the dual objective of count data analysis and imputation.

Models related to (3.3) have been considered for statistical ecology applications in (Brown et al., 2014; ter Braak et al., 2017). However, to the best of our knowledge, their theoretical properties have not been thoroughly studied. On the other hand, the literature on convex low-rank matrix estimation is abundant and benefits from a substantial theoretical background, but few software with ready to use solution are available for practitioners, and applications for count data outside image analysis (Luisier et al., 2011; Salmon et al., 2014; Cao and Xie, 2016) and recommendation systems (Gopalan et al., 2014) have not been attempted. The scope of this paper is to develop a complete methodology for the inference of model (3.3), bridging the gap between convex low-rank matrix completion and model-based count data analysis.

After detailing related work, we introduce in Section 3.2 a general model which includes (3.3) as a special case; we propose an estimation procedure through the minimization of a data fitting term penalized by the nuclear norm of the interaction matrix,

which acts as a convex relaxation of the rank constraint. In the same section, building upon existing results on nuclear norm regularized loss functions, we derive statistical guarantees for our estimation procedure. In particular, we provide an upper bound for the Frobenius norm of the estimation error. In Section 3.3, we propose an optimization algorithm, and two methods to choose the regularization parameter automatically. We provide a simulation study in Section 3.4 revealing that the method outperforms state-of-the-art methods when the proportion of missing values is large and the interactions are of significant order compared to the main effects. In Section 3.5, we show on plant abundance data with side information, how the results of our procedure can be interpreted through visual displays. In particular, the arising interpretation is consistent with known results from the original study (Choler, 2005). In Section 3.6, we use our method to analyze a waterbirds abundance data set from the French national agency for wildlife and hunting management (ONCFS). The proofs of the statistical guarantees are postponed to Section 3.8, and the method is available as an R package (R Core Team, 2017) called `lori` (LOw-Rank Interaction) on the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=lori`.

Model (3.3) is closely related to other models previously suggested in the statistical ecology literature to analyze count tables with row and column covariates. For instance, Brown et al. (2014) and ter Braak et al. (2017) suggested the following model:

$$\boldsymbol{X}^0_{i,j} = \mu^0 + \alpha^0_i + \beta^0_j + \epsilon^0_{RC} R_i C_j, \tag{3.4}$$

with R_i, $1 \leq i \leq n$ a row trait and C_j, $1 \leq j \leq p$ a column trait. The interaction between covariates is modeled by $\epsilon^0_{RC} R_i C_j$, where ϵ^0_{RC} is an unknown parameter measuring the strength of the interaction between the two traits. The main difference with model (3.3) is that we incorporate the covariates in the main effects rather than the interactions, which leads to different interpretations. In terms of estimation properties, the main advantage of (3.3) is that, as long as $K_1 \leq m_1$ and $K_2 \leq m_2$, we estimate less parameters. This is an important point for us since in many applications we are interested in (see e.g. Section 2.1.1), a large proportion of entries is missing, limiting the amount of available data. Finally, model (3.4) was developed with the aim of testing significant associations between covariates, rather than estimating the parameters or imputing missing values. In particular, its theoretical properties, as far as we know, were not studied.

In the low-rank matrix completion literature, related approaches for count matrix recovery and dimensionality reduction can be embedded within the framework of low-rank exponential family estimation (Collins et al., 2001; de Leeuw, 2006; Li and Tao, 2013; Josse and Wager, 2016; Liu et al., 2016) as well as its Bayesian counterpart (Mohamed et al., 2009; Gopalan et al., 2014). In terms of statistical guarantees, the theoretical performance of nuclear norm penalized estimators for Poisson denoising has been studied in Cao and Xie (2016), where the authors prove uniform bounds on the empirical error risk. Estimation rates are also given in Lafond (2015), where optimal bounds are proved for matrix completion in the exponential family. These two papers do not account for available covariates.

More recently, Chiquet et al. (2018) developped a probabilistic PCA framework for the exponential family, where covariates can be included in the parameter space. Fithian and Mazumder (2018) present a variety of low-rank problems including the generalized

nuclear norm penalty (Angst et al., 2011), that can be used to include row and column covariates. Similar estimation problems were also considered, e.g., in Agarwal and Chen (2009); Abernethy et al. (2009). However, to the best of our knowledge, these papers did not provide statistical guarantees and the practical advantages of such extensions compared to classical low-rank methods have not been thoroughly studied.

3.2 The low-rank interactions (LORI) model

3.2.1 General model

We now introduce a general version of the model (3.3) described in the introduction. Indeed, assuming a Poisson probabilistic model may be restrictive; thus, we relax it and rely on a pseudo-likelihood data fitting term instead. Consider the following assumption on the distribution of $\boldsymbol{Y}_{i,j}$, $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$.

H1. *The random variables* $\boldsymbol{Y} = \{\boldsymbol{Y}_{i,j}\}_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]}$ *are independent and there exist* $\gamma > 0$, $\sigma_- > 0$ *and* $\sigma_+ < \infty$ *such that for all* $i \in [\![m_1]\!]$ *and* $j \in [\![m_2]\!]$

$$\mathrm{e}^{-\gamma} \leq \mathbb{E}[\boldsymbol{Y}_{i,j}] \leq \mathrm{e}^{\gamma} \text{ and } \sigma_-^2 \leq \mathrm{var}[\boldsymbol{Y}_{i,j}] \leq \sigma_+^2.$$

Assumption **H** 1 means that the random variables $\boldsymbol{Y}_{i,j}$ have bounded expectations and variances. In particular, their expectations satisfy $\mathbb{E}[\boldsymbol{Y}_{i,j}] > 0$, for $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$. We do not assume the entries follow Poisson distributions, but we define our target parameter $\boldsymbol{X}^0_{i,j}$ as:

$$\boldsymbol{X}^0_{i,j} = \mathrm{argmin}_{x\in\mathbb{R}}\{-\mathbb{E}[\boldsymbol{Y}_{i,j}]x + \exp(x)\}\,. \tag{3.5}$$

In other words, the target parameter $\boldsymbol{X}^0_{i,j}$ minimizes the Kullback-Leibler divergence between the distribution of $\boldsymbol{Y}_{i,j}$ and a Poisson distribution. Thus, it defines the best Poisson approximation of the distribution of $\boldsymbol{Y}_{i,j}$ (in the sense of the Kullback-Leibler divergence). Note that the assumption $\mathbb{E}[\boldsymbol{Y}_{i,j}] > 0$ implies that (3.5) is always well-defined.

Similarly, we generalize the decomposition introduced in (3.3), to incorporate the broadest family of models possible. The main feature of model (3.3) is that it decomposes the parameter matrix $\boldsymbol{X}^0$ into two components, one of them being low-rank. Furthermore, the low-rank component is not arbitrary: it corresponds to the "residual term" after modeling main row and column effects. Thus, from a given parameter matrix $\boldsymbol{X}^0$, $\boldsymbol{\Theta}^0$ is obtained by centering the rows and columns of $\boldsymbol{X}^0$. This corresponds to the following matrix computation:

$$\boldsymbol{\Theta}^0 = \boldsymbol{X}^0 - \mathbb{1}_{m_1}\mathbb{1}_{m_1}^\top \boldsymbol{X}^0 - \boldsymbol{X}^0 \mathbb{1}_{m_2}\mathbb{1}_{m_2}^\top + \mathbb{1}_{m_1}\mathbb{1}_{m_1}^\top \boldsymbol{X}^0 \mathbb{1}_{m_2}\mathbb{1}_{m_2}^\top. \tag{3.6}$$

In other words, the rows and columns of $\boldsymbol{X}^0$ are projected onto the vector spaces orthogonal to $\mathbb{1}_{m_1}$ and $\mathbb{1}_{m_2}$ respectively. Starting from (3.6), the decomposition may be generalized by considering projections on arbitrary row and column vector spaces. Let S_1 and S_2 be fixed linear subspaces of $\mathbb{R}^{m_1}$ and $\mathbb{R}^{m_2}$ respectively. Let P_1 and P_2 be the orthogonal projection matrices on S_1 and S_2, $\mathcal{P}^\perp : \boldsymbol{X} \in \mathbb{R}^{m_1\times m_2} \mapsto P_1 \boldsymbol{X} P_2^\top$, $\mathcal{P} : \boldsymbol{X} \in \mathbb{R}^{m_1\times m_2} \mapsto \boldsymbol{X} - \mathcal{P}^\perp(\boldsymbol{X})$, $\mathcal{X}_0 \subset \{\boldsymbol{X} \in \mathbb{R}^{m_1\times m_2}; \mathcal{P}^\perp(\boldsymbol{X}) = 0\}$ and $\mathcal{T} = \{\boldsymbol{X} \in \mathbb{R}^{m_1\times m_2}; \mathcal{P}(\boldsymbol{X}) = 0\}$. For example, if S_1 is the span of the constant vector

$(1, \ldots, 1)$ of size m_1, and S_2 is the span of the constant vector $(1, \ldots, 1)$ of size m_2, the orthogonal projection operator $\mathcal{P}^\top$ consists in substracting the row and column means, with operation (3.6). We consider general decompositions of the form:

$$\boldsymbol{X}^0 = \boldsymbol{A}^0 + \boldsymbol{\Theta}^0, \quad \boldsymbol{A}^0 \in \mathcal{X}_0, \boldsymbol{\Theta}^0 \in \mathcal{T}. \tag{3.7}$$

Such general decompositions may be used to model particular types of interactions explicitly. For example, interactions between covariates, similarly to what is done in (3.4). Note that the original model, (3.3), is included in (3.7) by setting $S_1 = \{u \in \mathbb{R}^{m_1}; \mathbb{1}_{m_1}^\top u = 0\}$, $S_2 = \{v \in \mathbb{R}^{m_2}; \mathbb{1}_{m_2}^\top v = 0\}$, and

$$\mathcal{X}_0 = \left\{ \left(\mu + \sum_{k=1}^{K_1} \boldsymbol{R}_{ik}\alpha_k + \sum_{k=2}^{K_2} \boldsymbol{C}_{ik}\beta_k \right)_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} ; \mu \in \mathbb{R}, \alpha \in \mathbb{R}^{K_1}, \beta \in \mathbb{R}^{K_2} \right\}.$$

The dimension of this subspace is at most $1 + K_1 + K_2$ and the rank of a matrix in $\mathcal{X}_0$ is less than 3. In the general case, we denote:

$$r = \max\left(\{\mathrm{rank}(\boldsymbol{A}) \,:\, \boldsymbol{A} \in \mathcal{X}_0\}\right). \tag{3.8}$$

We finally consider a setting with missing observations. Denote by $\Omega \subset [\![m_1]\!] \times [\![m_2]\!]$ the set of observed entries: $(i, j) \in \Omega$ if and only if $\boldsymbol{Y}_{i,j}$ is observed. Define also the random variables (ω_{ij}) such that $\omega_{ij} = 1$ if $\boldsymbol{Y}_{i,j}$ is observed and $\omega_{ij} = 0$ otherwise. We assume that (ω_{ij}) and $\boldsymbol{Y}$ are independent, and a Missing Completely At Random (MCAR) scenario (Little and Rubin, 2002), where (ω_{ij}) are independent Bernoulli random variables. For $(i, j) \in \{1, \ldots, n\} \times \{1, \ldots, p\}$, we denote $\pi_{ij} = \mathbb{P}(\omega_{ij} = 1)$. We assume the probability of observing any entry is positive, i.e. there exists $\pi > 0$ such that

$$\min\left\{\pi_{ij} \,:\, (i, j) \in [\![m_1]\!] \times [\![m_2]\!]\right\} = \pi > 0\,. \tag{3.9}$$

For $j \in [\![m_2]\!]$, denote by $\pi_{\cdot j} = \sum_{i=1}^n \pi_{ij}$ the probability of observing an element in the j-th column. Similarly, for $i \in [\![m_1]\!]$, denote by $\pi_{i\cdot} = \sum_{j=1}^p \pi_{ij}$ the probability of observing an element in the i-th row. We define the following upper bound:

$$\max\left(\{\pi_{i\cdot} \,:\, i \in [\![m_1]\!]\} \cup \{\pi_{\cdot j} \,:\, j \in [\![m_2]\!]\}\right) \leq \beta\,. \tag{3.10}$$

3.2.2 Estimation and main results

We define a data-fitting term based on the Poisson pseudo-likelihood:

$$\mathcal{L}(\boldsymbol{X}) = \sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} \omega_{ij}\left\{-\boldsymbol{Y}_{i,j}\boldsymbol{X}_{i,j} + \exp(\boldsymbol{X}_{i,j})\right\}. \tag{3.11}$$

Denote $\|\cdot\|$ the operator norm (the largest singular value), $\|\cdot\|_\infty$ the infinity norm (the largest entry in absolute value) and $\|\cdot\|_*$ the nuclear norm (the sum of singular values). Our estimator of model (3.3), for a given regularization parameter $\lambda > 0$, is the minimizer of the data-fitting term (3.11) penalized by the nuclear norm of $\boldsymbol{\Theta}$:

$$(\hat{\boldsymbol{A}}, \hat{\boldsymbol{\Theta}}) \in \begin{array}{cc} \text{argmin} & \mathcal{L}(\boldsymbol{A} + \boldsymbol{\Theta}) + \lambda\|\boldsymbol{\Theta}\|_*, \\ \text{subject to} & \|\boldsymbol{A} + \boldsymbol{\Theta}\|_\infty \leq \gamma, (i, j) \in [\![m_1]\!] \times [\![m_2]\!], \\ & \boldsymbol{A} \in \mathcal{X}_0, \text{ and } \boldsymbol{\Theta} \in \mathcal{T}. \end{array} \tag{3.12}$$

Denote $\hat{\boldsymbol{X}} = \hat{\boldsymbol{A}} - \hat{\boldsymbol{\Theta}}$. After solving the estimation problem (6.11), we compute an imputed data set $\hat{\boldsymbol{Y}}$ as follows:

$$\begin{array}{ll} \hat{\boldsymbol{Y}}_{i,j} & = \exp(\hat{\boldsymbol{X}}_{ij}), \quad (i,j) \in [\![m_1]\!] \times [\![m_2]\!] \setminus \Omega, \\ \hat{\boldsymbol{Y}}_{i,j} & = \boldsymbol{Y}_{ij}, \qquad\qquad (i,j) \in \Omega. \end{array} \tag{3.13}$$

We evaluate the statistical properties of the estimation procedure (3.12) in terms of the estimation error $\hat{\boldsymbol{X}} - \boldsymbol{X}^0$, where $\hat{\boldsymbol{X}} = \hat{\boldsymbol{A}} - \hat{\boldsymbol{\Theta}}$. Indeed, for $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, the distance $(\hat{\boldsymbol{X}}_{ij} - \boldsymbol{X}^0_{ij})^2$ is related to the Kullback-Leibler divergence between the Poisson distributions parameterized by $\exp(\hat{\boldsymbol{X}}_{ij})$ and $\exp(\boldsymbol{X}^0_{ij})$, respectively. Thus, it measures how well the distribution of $\boldsymbol{Y}_{i,j}$ is approximated by a Poisson distribution of intensity $\exp(\hat{\boldsymbol{X}}_{ij})$ and, by extension, the quality of the imputation procedure (3.13).

The following statistical guarantees essentially show that the Frobenius norm of the estimation error $\hat{\boldsymbol{X}} - \boldsymbol{X}^0$ is of the order of

$$\|\hat{\boldsymbol{X}} - \boldsymbol{X}^0\|_F^2 \lesssim \frac{\operatorname{rank}(\boldsymbol{\Theta}^0) m_2}{\pi} + \frac{r m_2}{\pi},$$

where $\lesssim$ denotes the inequality up to constant and logarithmic factors. The first term corresponds to the usual bound in low-rank matrix estimation and completion (Klopp, 2014; Lafond, 2015). The additional term rm_2/π accounts for explicit modeling of the covariates in the main effects. To derive this upper bound, we need additional assumptions. Let (E_{ij}) be the matrices of the canonical basis of $\mathbb{R}^{m_1 \times m_2}$. Let

$$\nabla \mathcal{L}(\boldsymbol{X}) = \sum_{(i,j) \in [\![m_1]\!] \times [\![m_2]\!]} \omega_{ij} \left\{ -\boldsymbol{Y}_{i,j} + \exp(\boldsymbol{X}_{i,j}) \right\} E_{ij}$$

be the gradient of $\mathcal{L}$ evaluated at $\boldsymbol{X}$. Denote also $\partial^2 \mathcal{L} / \partial x_{ij}^2$ the second derivative of $\mathcal{L}$ with respect to the (i,j)-th coordinate.

H2. *The function $\mathcal{L}$ is strongly convex and smooth on $[-\gamma - \varepsilon, \gamma + \varepsilon]^{m_1 \times m_2}$ for some $\varepsilon > 0$. There exist $\sigma_- > 0$ and $\sigma_+ < \infty$ such that for all $\boldsymbol{X} \in [-\gamma - \varepsilon, \gamma + \varepsilon]^{m_1 \times m_2}$ and $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, $\sigma_-^2 \leq \partial^2 \mathcal{L}(\boldsymbol{X}) / \partial x_{ij}^2 \leq \sigma_+^2$.*

We also introduce the following random matrix related to the distribution of the missing entries. Let (ϵ_{ij}), $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$ be i.i.d. Rademacher random variables independent of $\boldsymbol{Y}$ and Ω and define

$$\Sigma_R = \sum_{(i,j) \in [\![m_1]\!] \times [\![m_2]\!]} \epsilon_{ij} \omega_{ij} E_{ij}. \tag{3.14}$$

The derivation of our statistical guarantees follow two steps. First, we derive in Theorem 5 a deterministic upper bound, which depends on the two random matrices Σ_R (explicitly) and $\nabla \mathcal{L}(\boldsymbol{X}^0)$ (through an assumption of the theorem).

Theorem 5. *Assume **H** 1-2, and $\lambda \geq 2\|\nabla \mathcal{L}(\boldsymbol{X}^0)\|$. Then for all $m_1, m_2 \geq 1$, with probability at least $1 - 8(m_1 + m_2)^{-1}$,*

$$\left\| \boldsymbol{X}^0 - \hat{\boldsymbol{X}} \right\|_F^2 \leq \frac{C}{\pi^2} \left(\left[\frac{\lambda^2}{\sigma_-^4} + (\mathbb{E}\|\Sigma_R\|)^2 \gamma^2 \right] (\operatorname{rank}(\boldsymbol{\Theta}^0) + r) + \log(n+p) \right), \tag{3.15}$$

where r, γ are defined in (3.8) *and* (3.12), *C is a numerical constant whose value can be found in the proof and which is independent of m_1, m_2 and $\boldsymbol{X}^0$.*

Proof. The proof is postponed to Section 3.8.1. □

Then, we use tail bounds on the norms of sums of random matrices, to control the quantities $\mathbb{E}\|\Sigma_R\|$ and $\|\nabla\mathcal{L}(\boldsymbol{X}^0)\|$. Doing so, we obtain an upper bound on $\mathbb{E}\|\Sigma_R\|$, and compute a value of λ such that the condition $\lambda \geq 2\|\nabla\mathcal{L}(\boldsymbol{X}^0)\|$ holds with high probability. Finally, we obtain a high probability upper bound which only depends on the parameters of the problem (in Theorem 6). We will need the following additional assumption on the distribution of the counts:

H3. *There exists $\delta > 0$ such that for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$,*

$$\mathbb{E}\left[\exp(|\boldsymbol{Y}_{i,j}|/\delta)\right] < +\infty\,.$$

Assumption (3) means that the entries $\boldsymbol{Y}_{i,j}$, and thus the entries of $\nabla\mathcal{L}(\boldsymbol{X}^0)$, are subexponential. Equipped with Theorem 5, we now derive our main theorem. Define the following quantities, with C^* a numerical constant defined in Lemma 5 and r, β and γ defined in (3.8), (3.10) and (3.12) respectively:

$$\begin{aligned}
\Phi_1 &= 48\sigma_+^2\beta\log(m_1+m_2),\\
\Phi_2 &= 36\delta^2(\mathrm{e}-1)^2\log^2\left(1+8\delta^2\tfrac{m_1m_2}{\beta\sigma_-^2}\right)\log^2(m_1+m_2)\\
\Phi_3 &= 4C^{*2}\max(\beta,\log\{\min(m_1,m_2)\}).
\end{aligned} \tag{3.16}$$

Considering only the parameters which depend on the size of the problem, i.e. on the dimensions m_1 and m_2, or on the proportion of observed entries through the parameter β, the orders of magnitude of these three quantities are:

$$\begin{aligned}
\Phi_1:&\quad \beta\log(m_1+m_2),\\
\Phi_2:&\quad \log^2(m_1m_2/\beta)\log^2(m_1+m_2),\\
\Phi_3:&\quad \max(\beta,\log\{\min(m_1,m_2)\}).
\end{aligned} \tag{3.17}$$

Theorem 6. *Assume **H** 1-**H** 3 and set*

$$\lambda = \max\left\{4\sigma_+\sqrt{3\beta\log(m_1+m_2)}, 12\delta(e-1)\log\left(1+8\delta^2\frac{m_1m_2}{\beta\sigma_-^2}\right)\log(m_1+m_2)\right\}.$$

Then with probability at least $1-10(m_1+m_2)^{-1}$,

$$\|\hat{\boldsymbol{X}}-\boldsymbol{X}^0\|_F^2 \leq \frac{C}{\pi^2}\left\{(\max(\Phi_1,\Phi_2)+\Phi_3)\left(\mathrm{rank}(\boldsymbol{\Theta}^0)+r\right)+\log(m_1+m_2)\right\}, \tag{3.18}$$

where C is a numerical constant independent of m_1, m_2 and $\boldsymbol{X}^0$.

Proof. The proof is postponed to Section 3.8.2. □

Denoting $\lesssim$ the inequality up to constant and logarithmic factors, and using the orders of magnitude reported in (3.17), we recover an upper bound of the order of:

$$\|\hat{\boldsymbol{X}}-\boldsymbol{X}^0\|_F^2 \lesssim \frac{\mathrm{rank}(\boldsymbol{\Theta}^0)\beta}{\pi^2}+\frac{r\beta}{\pi^2}.$$

The first term correspond to the usual bound in low-rank matrix estimation and completion (Klopp, 2014; Lafond, 2015), and is equal to $\mathrm{rank}(\boldsymbol{\Theta}^0)\max(m_1,m_2)/\pi$ when the sampling is almost uniform ($c_1\pi \leq \pi_{ij} \leq c_2\pi$). The additional term $r\beta/\pi^2$ accounts for explicit modeling of the covariates in the main effects. The constant term appearing in bound (3.18) grows linearly with the upper bound σ_+^2 and quadratically with the inverse of σ_-^2. This means that by relaxing Assumption 1 to allow $\mathrm{var}(\boldsymbol{Y}_{i,j})$ to grow as fast as $\log(m_1+m_2)$ or decrease as fast as $1/\log(m_1+m_2)$, we only lose a log-polynomial factor in bound (3.18).

3.3 Implementation

3.3.1 Optimization algorithm

In this section, we propose an algorithm to solve the estimation problem (3.12) for the initial model

$$\boldsymbol{X}^0_{i,j} = \mu^0 + \sum_{k=1}^{K_1} \boldsymbol{R}_{i,k}\alpha^0_k + \sum_{l=1}^{K_2} \boldsymbol{C}_{j,l}\beta^0_l + \boldsymbol{\Theta}^0_{ij}.$$

We use an *alternating minimization* (Csiszár and Tusnády, 1984) algorithm, which consists in updating μ, α, β and $\boldsymbol{\Theta}$ alternatively, each time along a descent direction. Note that, in the algorithm and the entire numerical section, we relax the constraint $|\mu + \boldsymbol{R}_{i,.}\alpha + \boldsymbol{C}_{j,.}\beta + \boldsymbol{\Theta}_{i,j}| \leq \gamma$, which is mainly required to obtain statistical guarantees. Denote

$$\mathcal{F}(\mu, \alpha, \beta, \boldsymbol{\Theta}) = \mathcal{L}(\mu + \boldsymbol{R}_{i,.}\alpha + \boldsymbol{C}_{j,.}\beta + \boldsymbol{\Theta}),$$

and $\nabla_{\boldsymbol{\Theta}}\mathcal{F}$ the gradient of $\mathcal{F}$ with respect to $\boldsymbol{\Theta}$ defined by $(\nabla_{\boldsymbol{\Theta}}\mathcal{F}(\mu, \alpha, \beta, \boldsymbol{\Theta}))_{ij} = -\boldsymbol{Y}_{i,j} + \exp(\mu + R_{i,.}\alpha + C_{j,.}\beta + \boldsymbol{\Theta}_{i,j})$ if $\omega_{ij} = 1$ and $(\nabla_{\boldsymbol{\Theta}}\mathcal{F}(\mu, \alpha, \beta, \boldsymbol{\Theta}))_{ij} = 0$ otherwise. Denote $(\mu^{[0]}, \alpha^{[0]}, \beta^{[0]}, \boldsymbol{\Theta}^{[0]})$ the initialized parameters, and $(\mu^{[t]}, \alpha^{[t]}, \beta^{[t]}, \boldsymbol{\Theta}^{[t]})$ the value of the parameters at iteration t, $t \geq 1$. At every iteration we solve three sub-problems in μ, α, β and $\boldsymbol{\Theta}$ in order to update $\mu^{[t]}$,$\alpha^{[t]}$, $\beta^{[t]}$ and $\boldsymbol{\Theta}^{[t]}$. First, the sub-problem in μ can be solved in closed form:

$$\mu^{[t]} \in \operatorname{argmin} \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \omega_{ij}\{-Y_{i,j}\mu + \exp(\mu + \sum_{k=1}^{K_1} \boldsymbol{R}_{i,k}\alpha_k^{[t-1]} + \sum_{l=1}^{K_2} \boldsymbol{C}_{j,l}\beta_l^{[t-1]} + \boldsymbol{\Theta}_{i,j}^{[t-1]})\},$$

which yields:

$$\mu^{[t]} = \log\left\{ \frac{\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \omega_{ij}\boldsymbol{Y}_{i,j}}{\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \omega_{ij}\exp(\sum_{k=1}^{K_1} \boldsymbol{R}_{i,k}\alpha_k^{[t-1]} + \sum_{l=1}^{K_2} \boldsymbol{C}_{j,l}\beta_l^{[t-1]} + \boldsymbol{\Theta}_{i,j}^{[t-1]})} \right\},$$

as long as $\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \omega_{ij}\boldsymbol{Y}_{i,j} > 0$, that is, there is at least one positive count observed in $\boldsymbol{Y}$. Then, the updates in α and β may be done simultaneously by estimating a Poisson generalized linear model with offsets:

$$(\alpha^{[t]}, \beta^{[t]}) \in \operatorname{argmin} \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \omega_{ij}\Bigg\{ -Y_{i,j}\left(\sum_{k=1}^{K_1} \boldsymbol{R}_{i,k}\alpha_k + \sum_{l=1}^{K_2} \boldsymbol{C}_{j,l}\beta_l\right) + \exp(\mu^{[t]} + \sum_{k=1}^{K_1} \boldsymbol{R}_{i,k}\alpha_k + \sum_{l=1}^{K_2} \boldsymbol{C}_{j,l}\beta_l + \boldsymbol{\Theta}_{i,j}^{[t-1]}) \Bigg\},$$

which can be done for instance with standard algorithms implemented in available libraries. Finally, we perform the update in $\boldsymbol{\Theta}$ along the proximal gradient direction. Denote by $\mathcal{D}_\lambda$ the soft-thresholding operator of singular values at level λ (Cai et al., 2010, Section 2). We update $\boldsymbol{\Theta}$ by soft-thresholding the singular values

$$\boldsymbol{\Theta}^{[t]} = \mathcal{D}_\lambda[\boldsymbol{\Theta}^{[t-1]} - \tau\nabla_{\boldsymbol{\Theta}}\mathcal{F}(\mu^{[t]}, \alpha^{[t]}, \beta^{[t]}, \boldsymbol{\Theta}^{[t-1]})],$$

where the step size τ is tuned using backtracking line search. The complete procedure is sketched in Algorithm 1.

Algorithm 1 Alternating minimization for problem (6.11)

Initialize $\mu^{[0]}$, $\alpha^{[0]}$, $\beta^{[0]}$, $\mathbf{\Theta}^{[0]}$
for $t = 0, 1, \ldots, T-1$ **do**
  $\mu^{[t+1]} \in \operatorname{argmin} \mathcal{F}(\mu, \alpha^{[t]}, \beta^{[t]}, \mathbf{\Theta}^{[t]})$,
  $(\alpha^{[t+1]}, \beta^{[t+1]}) \in \operatorname{argmin} \mathcal{F}(\mu^{[t]}, \alpha, \beta, \mathbf{\Theta}^{[t]})$,
  $\tau = 1$,
  $\mathbf{\Theta}^{[t+1]} = \mathcal{D}_\lambda[\mathbf{\Theta}^{[t]} - \tau \nabla_{\mathbf{\Theta}} \mathcal{F}(\mu^{[t]}, \alpha^{[t]}, \beta^{[t]}, \mathbf{\Theta}^{[t]})]$
  while $\mathcal{F}(\mu^{[t+1]}, \alpha^{[t+1]}, \beta^{[t+1]}, \mathbf{\Theta}^{[t+1]}) + \lambda \|\mathbf{\Theta}^{[t+1]}\|_* > \mathcal{F}(\mu^{[t+1]}, \alpha^{[t+1]}, \beta^{[t+1]}, \mathbf{\Theta}^{[t]}) + \lambda \|\mathbf{\Theta}^{[t]}\|_*$ **do**
    $\tau = \tau / 2$
    $\mathbf{\Theta}^{[t+1]} = \mathcal{D}_\lambda[\mathbf{\Theta}^{[t]} - \tau \nabla_{\mathbf{\Theta}} \mathcal{F}(\mu^{[t]}, \alpha^{[t]}, \beta^{[t]}, \mathbf{\Theta}^{[t]})]$.
  end while
end for
Return $\mu^{[T]}$, $\alpha^{[T]}$, $\beta^{[T]}$, $\mathbf{\Theta}^{[T]}$

Note that if $K_1 + K_2 > |\Omega|$, with $|\Omega|$ denoting the cardinality of Ω, the update in α and β does not have a unique solution. However in our targeted applications, typically $K_1 + K_2 \ll |\Omega|$. Note that, even though our theoretical guarantees require a MCAR mechanism, the estimation method still holds when entries are *missing at random* (Little and Rubin (2002), Section 1.3).

3.3.2 Automatic selection of λ

A common way to select the regularization parameter is cross-validation, which consists in erasing a fraction of the observed cells in $\boldsymbol{Y}$, estimating a complete parameter matrix $\hat{\boldsymbol{X}}$ for a range of λ values, and choosing the parameter λ that minimizes the imputation error. This can be performed directly using our method (LORI) without modifying the code, simply by modifying the weights ω_{ij}. However, this procedure is computationally costly, as it implies estimating the LORI model many times. We suggest an alternative to cross-validation, inspired by Donoho and Johnstone (1994) and the work of Giacobino et al. (2016) on *quantile universal threshold*. In Theorem 7 below, we define the so-called *null-thresholding statistic* of estimator (6.11), a function of the data $\lambda_0(\boldsymbol{Y})$ for which the estimated interaction matrix $\hat{\mathbf{\Theta}}^{\lambda_0(\boldsymbol{Y})}$ is null, and the same estimate $\hat{\mathbf{\Theta}}^{\lambda} = 0$ is obtained for any $\lambda \geq \lambda_0(\boldsymbol{Y})$.

Theorem 7 (Null-thresholding statistic). *The estimated interaction matrix $\hat{\mathbf{\Theta}}^{\lambda}$ for a regularization parameter λ is null if and only if $\lambda \geq \lambda_0(\boldsymbol{Y})$, where $\lambda_0(\boldsymbol{Y})$ is the null-thresholding statistic*

$$\lambda_0(\boldsymbol{Y}) = \|\nabla \mathcal{L}(\hat{\boldsymbol{A}})\|, \quad \text{where} \quad \hat{\boldsymbol{A}} \in \operatorname{argmin}_{\boldsymbol{A} \in \mathcal{X}_0} \quad \mathcal{L}(\boldsymbol{A}). \tag{3.19}$$

Proof. The proof is postponed to Section 3.8.3. □

Here, $\|.\|$ denotes the operator norm (the largest singular value). We propose a heuristic selection of λ based on this null-thresholding statistic $\lambda_0(\boldsymbol{Y})$. To explain further the procedure, we first need to define the following test:

$$\mathbf{H}_0 : \mathbf{\Theta}^0 = 0 \quad \text{against the alternative} \quad \mathbf{H}_1 : \mathbf{\Theta}^0 \neq 0 \tag{3.20}$$

which tests whether the parameter matrix $\boldsymbol{X}^0$ can be explained only in terms of linear combinations of the measured covariates. For a probability $\varepsilon \in (0, 1)$, consider the upper ε-quantile λ_ε of the null-thresholding statistics, namely that satisfies $\mathbb{P}_{\mathbf{H}_0}(\lambda_0(\boldsymbol{Y}) > \lambda_\varepsilon) <$

ε. The test which consists in comparing the statistics $\lambda_0(\boldsymbol{Y})$ to λ_ε is of level $1-\varepsilon$ for (3.20). This can be seen as an alternative to the χ^2 test for independence which handles covariates. In practice we do not have access to the distribution under the null $\mathbb{P}_{\mathbf{H}_0}(\lambda_0(\boldsymbol{Y}) < \lambda)$, but perform parametric bootstrap (Efron, 1979) to compute a proxy $\check{\lambda}_\varepsilon$. In practice we use $\varepsilon = 0.05$ and $\lambda_{\mathsf{QUT}} := \check{\lambda}_{.05}$; we refer to it in what follows as *quantile universal threshold* (QUT). This selection of the regularization parameter is essentially the universal threshold of Donoho and Johnstone (1994) extended to our setting.

3.4 Simulation study

3.4.1 Simulation scheme

In this experiment section we simulate count data under the LORI model (3.3). First, we generate covariate matrices $\boldsymbol{R} \in \mathbb{R}^{300\times 3}$ and $\boldsymbol{C} \in \mathbb{R}^{30\times 4}$ drawn from independent multivariate Gaussian distributions with mean 0 and diagonal covariance matrices. Then, we set $\mu^0 = 1$, $\alpha^0 = (0.5, 0.5, 0)$ and $\beta^0 = (0.5, 0.5, 0, 0)$. Then, an interaction matrix $\boldsymbol{\Theta}^0$ of rank 5 is generated by sampling orthonormal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, and fixing five decreasing singular values. We also "plant" some values in $\boldsymbol{\Theta}^0$: 5 entries are fixed to 3, to model outlier profiles. Denote by $\boldsymbol{A}^0$ the matrix corresponding to all the main effects, defined by $(\boldsymbol{A}^0)_{i,j} = \mu^0 + \sum_{k=1}^{K_1} \boldsymbol{R}_{i,k}\alpha_k^0 + \sum_{l=1}^{K_2} \boldsymbol{C}_{j,l}\beta_l^0$. The Frobenius norm of $\boldsymbol{\Theta}^0$ is then controlled through a parameter $\tau = \|\boldsymbol{\Theta}^0\|_F / \|\boldsymbol{A}^0\|_F$. We obtain a parameter matrix $\boldsymbol{X}^0$: $\boldsymbol{X}^0_{i,j} = (\mu^0 + \boldsymbol{R}_{i,.}\alpha^0 + \boldsymbol{C}_{j,.}\beta^0)_{i,j} + \boldsymbol{\Theta}^0_{i,j}$. Finally, we simulate $\boldsymbol{Y} \in \mathbb{N}^{300\times 30}$ under model (3.1): $\boldsymbol{Y}_{i,j} \sim \mathcal{P}(\exp(\boldsymbol{X}^0_{i,j}))$.

3.4.2 Estimation

We compare the performance of LORI in terms of estimation of the regression coefficients α^0 and β^0, and compare it to a standard Poisson Generalized Linear Model (GLM) estimated with the `glm` function in R. We repeat the experiment 100 times for decreasing values of the ratio $\tau = \|\boldsymbol{\Theta}^0\|_F / \|\boldsymbol{A}^0\|_F$, where $\boldsymbol{A}^0 = ((\mu^0 + \boldsymbol{R}_{i,.}\alpha^0 + \boldsymbol{C}_{j,.}\beta^0)_{i,j})$ is fixed. We look at the Root Mean Square Error (RMSE) for the estimation of $\boldsymbol{A}^0$; the results are given in Table 3.1, where we observe that LORI and the Poisson GLM are equivalent for $\tau = 0$, and that LORI outperforms the GLM for non-zero interactions, with a gap widening as τ increases.

Second, we compare LORI to a convex low-rank matrix estimatiob procedure with a Poisson loss function and where covariates are not modeled (e.g. Lafond (2015)), in terms of the relative estimation error $\|\hat{\boldsymbol{X}} - \boldsymbol{X}^0\|_F / \|\boldsymbol{X}^0\|_F$ (because $\|\boldsymbol{X}^0\|_F$ varies with τ). We refer to this competitor as "Poisson LRM". Again, we reproduce the experiment 100 times for decreasing values of the ratio $\tau = \|\boldsymbol{\Theta}^0\|_F / \|\boldsymbol{A}^0\|_F$. On Table 3.2, we observe that LORI achieves lower errors than Poisson LRM, which is expected as we simulated under the LORI model. As τ decreases–i.e. the size of the main effects increases relative to the interactions–both errors decrease as well, and the gap between LORI and the Poisson LRM widens, indicating that modeling covariates explicitly improves the estimation.

τ		Mean of RMSE$*100$	Standard deviation of RMSE$*100$
1	LORI	23	1.6
	GLM	76	15.5
0.5	LORI	8.3	1.2
	GLM	9.0	1.2
0.25	LORI	3.1	0.9
	GLM	3.2	1.0
0.1	LORI	2.4	0.8
	GLM	2.4	0.8
0	LORI	2.3	0.7
	GLM	2.4	0.7

Table 3.1: Estimation error (RMSE) of regression coefficients $\sqrt{\|\hat{\alpha} - \alpha^0\|_2^2 + \|\hat{\beta} - \beta^0\|_2^2}$ of LORI and a Poisson GLM, for decreasing values of $\tau = \|\boldsymbol{\Theta}\|_F / \|\boldsymbol{A}^0\|_F$. The results are aggregated across 100 replications of the experiment.

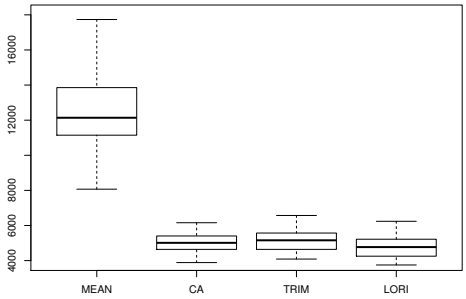
τ		Mean of Relative RMSE$*100$	Std dev. of Relative RMSE$*100$
1	LORI	63	0.5
	Poisson LRM	93	0.4
0.5	LORI	45	0.16
	Poisson LRM	99	0.14
0.25	LORI	24	0.1
	Poisson LRM	100	0.06
0.1	LORI	10	0.1
	Poisson LRM	100	0.06
0	LORI	2.1	0.6
	Poisson LRM	100	0.06

Table 3.2: Estimation error (Relative RMSE) of parameter matrix $\|\hat{\boldsymbol{X}} - \boldsymbol{X}^0\|_F / \|\boldsymbol{X}^0\|_F$ of LORI and a Poisson GLM, for decreasing values of $\tau = \|\boldsymbol{\Theta}^0\|_F / \|\boldsymbol{A}^0\|_F$.

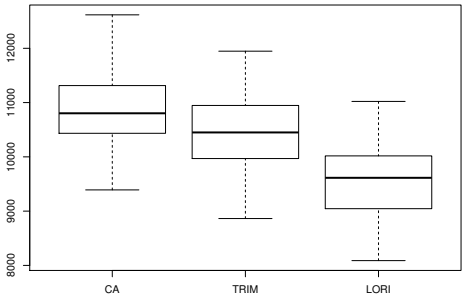
3.4.3 Imputation

Using the same simulation scheme, we now compare LORI in terms of missing values imputation to Correspondence Analysis (CA) and Trends & Indices for Monitoring data Pannekoek and van Strien (2001) (TRIM), a method based on a Poisson log-linear model used to impute bird abundance data. To do so we erase an increasing proportion of entries in the data and impute them using LORI, CA and TRIM, replicating the experiment 100 times. In the first experiment, we also include imputation of the missing values using the column means, as a baseline referred to as "MEAN"; we remove it in subsequent experiments to improve visibility. We observe on Figure 3.1 that LORI performs best, which is expected as we simulate under the LORI model. Moreover, the gap widens as the percentage of missing values increases. In particular, as the proportion of missing values increases, TRIM is not able to impute all the rows, and removes some of them; thus the imputation error of TRIM–which is already larger than LORI–is computed on a subsample of the rows (with the least missing entries).

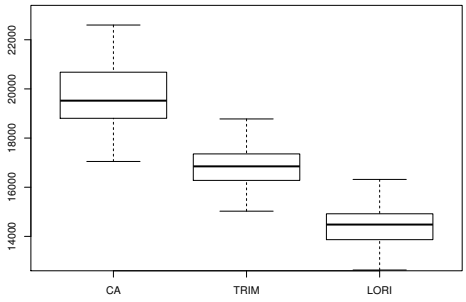
Then, we evaluate the imputation performances of LORI on the Aravo data set



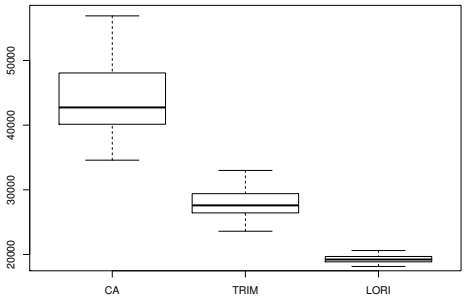
(a) 20% of missing entries



(b) 40% of missing entries



(c) 60% of missing entries



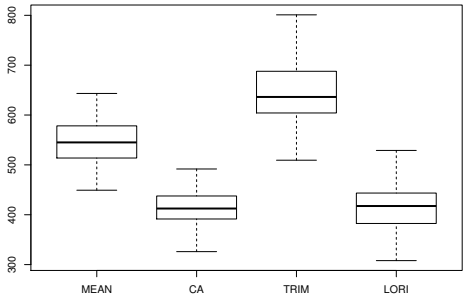
(d) 80% of missing entries

Figure 3.1: Imputation error $\sum_{(i,j)\in\Omega}(\boldsymbol{Y}_{i,j} - \hat{\boldsymbol{Y}}_{i,j})^2$, aggregated across 100 replications. The compared methods are, from left to right in the boxplots, imputation by the column means (MEAN), correspondence analysis (CA), trends and indices in monitoring data (TRIM), and low-rank interactions (LORI). The results are given for increasing proportions of missing values: 20% (top left), 40% (top right), 60% (bottom left) and 80% (bottom right).

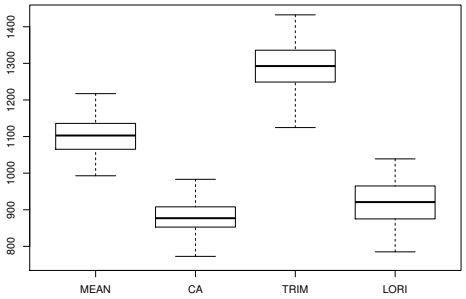
(described in more details in the next section). We introduce an increasing amount of (completely at random) missing values, and compute the prediction errors of CA, TRIM and LORI. The results are displayed in Figure 3.2, which essentially shows that LORI performs similarly as CA on this data set.

3.5 Analysis of the Aravo data

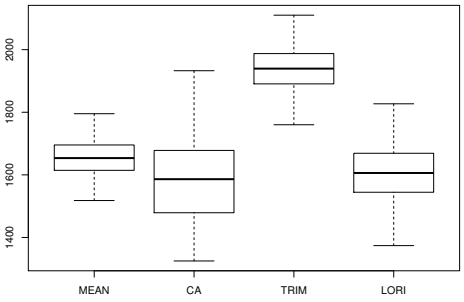
The Aravo data set (Choler, 2005) counts the abundance of 82 species of alpine plants in 75 sites in France; covariates about the environments and species are also available. We focus on 8 species traits providing physical information about plants (height, spread, etc.), and 4 environmental variables giving geographical and meteorological information about sites. We apply our method LORI after scaling the covariates and tuning the regularization parameter with the QUT method. This results in estimates for the main effects of the environment characteristics α and of the species traits β, as well as an estimate of the interaction matrix $\boldsymbol{\Theta}$.



(a) 20% of missing entries



(b) 40% of missing entries



(c) 60% of missing entries

Figure 3.2: Imputation error $\sum_{(i,j)\in\Omega}(\boldsymbol{Y}_{i,j} - \hat{\boldsymbol{Y}}_{i,j})^2$, for the Aravo data set, aggregated across 100 replications. The compared methods are, from left to right in the boxplots, imputation by the column means (MEAN), correspondence analysis (CA), trends and indices in monitoring data (TRIM), and low-rank interactions (LORI). The results are given for increasing proportions of missing values: 20% (top left), 40% (top right), 60% (bottom left) and 80% (bottom right).

Aspect	Slope	PhysD	Snow
0.01	0.02	-0.01	-0.02

Table 3.3: Main effect of the Aravo environment characteristics estimated with LORI. The regularization parameter is tuned using QUT.

Height	Spread	Angle	Area	Thick	SLA	Nmass	Seed
0.02	-0.06	-0.05	-0.05	-0.03	-0.04	0.05	-0.03

Table 3.4: Main effect of the Aravo species traits estimated with LORI. The regularization parameter is tuned using QUT.

The main effects of environment characteristics are given in Table 3.3 and the main effects of the species traits in Table 3.4. First we observe that overall, species traits have larger effects than environment characteristics on the observed abundances. In particular, the mass-based leaf nitrogen content (Nmass) has a large positive effect, which seems to indicate that plants with a large Nmass tend to be more abundant across all environments. On the other hand, the maximum lateral spread of clonal plants (Spread), area of single leaf (Area) leaf elevation angle estimated at the middle of the lamina (An-

gle) and specific leaf area (SLA) have large negative effects on the abundances.

The estimated rank of the interaction matrix $\hat{\Theta}$ (number of singular values above 10^{-6}) is 2. The environments (rows) and species (columns) can be visualized on a biplot (de Rooij and Heiser, 2005, Section 2.5), where rows and columns are represented simultaneously in a normalized Euclidean space. In such plots, the dimensions of the Euclidean space are given by the principal directions of $\hat{\Theta}$, scaled by the square root of the singular values of $\hat{\Theta}$. Figure 3.3 shows such a display, which can be interpreted in terms of distance between points: a species and an environment that are close interact highly, and two species or two environments that are close have similar profiles. Justifications for such a distance interpretation can be found in (de Rooij and Heiser, 2005, Section 2.5) or (Fithian and Josse, 2017, Section 2). Note that, in both plots, the signs of the directions may be flipped simultaneously. We can then look at the relations between the

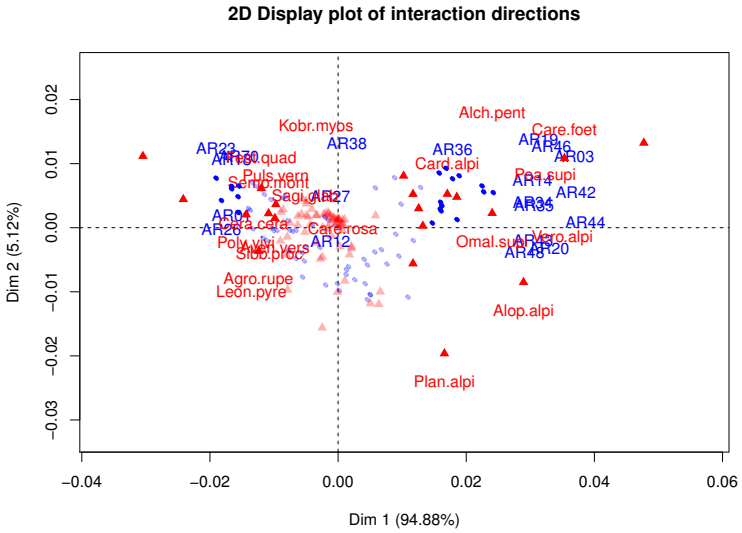
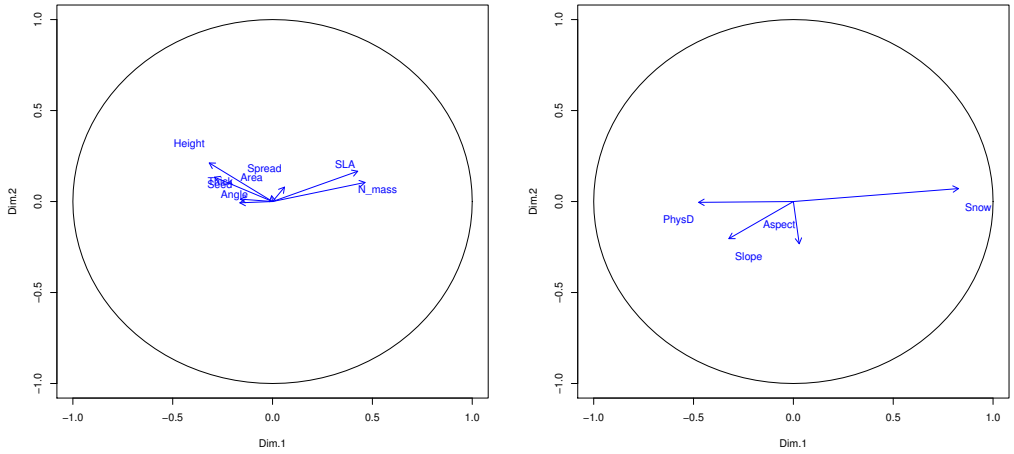


Figure 3.3: Display of the two first dimensions of interaction estimated with LORI. Environments are represented with blue points and species with red triangles.

known traits and the interaction directions of $\hat{\Theta}$. Figure 3.4a shows that the two first directions of interaction are correlated with the species covariates; the correlation is particularly high for the Nmass and SLA variables. Thus, on Figure 3.3, the two directions separate the plants with large SLA and Nmass (top right corner) from those with small SLA and Nmass (bottom left corner). Then, Figure 3.4a shows that the directions of interaction are also correlated with the environment covariates, and particularly with the mean snowmelt date (Snow). Thus, on Figure 3.3, the two directions separate the late melting environments (top right corner) from the early melting environments (bottom left corner). Combining the interpretation of Figure 3.3, Figure 3.4a and Figure 3.4b, we deduce that plants with large Nmass and SLA interact highly with late melting sites (large value of Snow). This was in fact the main result obtained in the original study Choler (2005) (see, e.g., the summary of findings in the abstract), which advocates the good properties of LORI in terms of interpretation.

Then, we evaluate the performance of LORI in terms of imputation of the missing values on the Aravo data: we introduce increasing proportions of missing values, and



(a) Environment covariates

(b) Species traits

Figure 3.4: Correlation between the two first dimensions of interaction and the covariates (the covariates are not used in the estimation).

compare LORI to mean imputation, CA and TRIM.

3.6 Using covariates to impute ecological data

The waterbirds data are constituted by counts of migratory waterbirds in 785 wetland sites (across the 5 countries in North Africa), between 1990 and 2017 (Sayoud et al., 2017). One of the objectives is to assess the effect of time on species abundances, to monitor the populations and assess wetlands conservation policies. Ornithologists have also recorded side information concerning the sites and years, which may influence the counts. For instance, meteorological anomalies, latitude and longitude. The count table contains a large amount of missing entries (70%), but the covariate matrices which contain respectively 6 covariates about the 785 sites and 8 covariates about the 18 years, are fully observed. Our method allows to take advantage of the available covariates to provide interpretation for spatio-temporal patterns. As a by-product, it produces an imputed contingency table.

Tables 3.5 and 3.6 show the estimated main effects of some of the sites and years characteristics. Sites with large altitudes are associated to smaller counts, as well as sites which are far from the coast. Sites which are located far from towns, and sites with large water surfaces are associated to larger counts. The four year covariates given in Table 3.6 concern meteorological anomalies. The associated coefficients are much smaller than those of the site covariates, which may indicate that there is more variability in the counts across sites than across years.

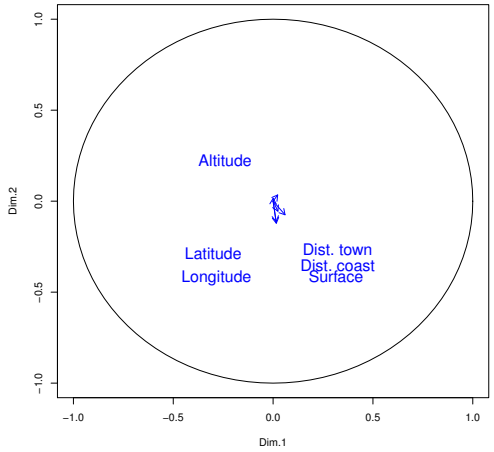
Latitude	Longitude	Altitude (m)	Dist. town(m)	Dist. coast (m)	Surface (km^2)
0.98	-0.70	-5.67	0.46	-6.17	0.71

Table 3.5: Main effect of the sites characteristics estimated with LORI. The regularization parameter is tuned using QUT.

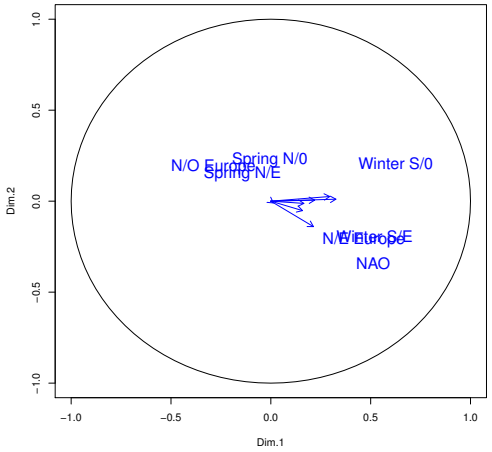
Spring N/O	Spring N/E	Winter S/O	Winter S/E	N/O Europe	N/E Europe	NAO
0.06	-0.01	0.07	0.09	0.05	0.00	0.08

Table 3.6: Main effect of the years characteristics estimated with LORI.

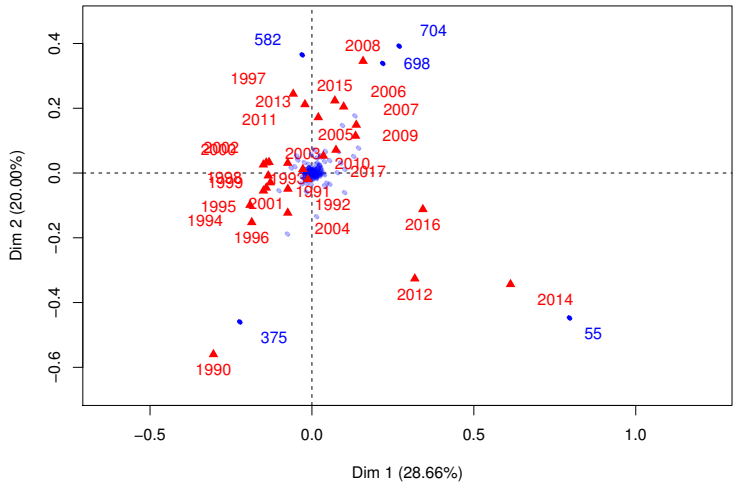
The sites and years can also be displayed using the same visual tools as described in Section 3.5. Figure 3.5a and 3.5b show the correlations between the covariates and the directions of interaction. On the two-dimensional display on Figure 3.5c, the first dimension is correlated with meteorological anomalies. Simultaneously, on Figure 3.5c, we observe a very clear temporal gradient along the first dimension, indicating that over time, meteorological abnormalities increase (in the sense of a summary anomaly variable embodied by the second direction). Several sites (55, 375, 582, 698, 704) lay out of the point cloud, and correspond to sites with very large surface.



(a) Correlation between sites characteristics and the two first directions of interaction.



(b) Correlation between year characteristics and the two first directions of interaction.



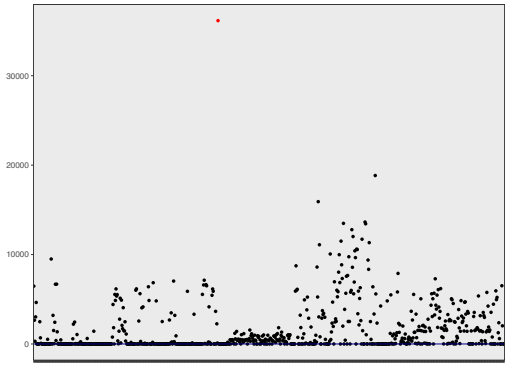
(c) Display of the two first dimensions of interaction estimated with LORI. Environments are represented with blue points and years with red triangles.

Figure 3.5: Visual display of LORI results for the waterbirds data.

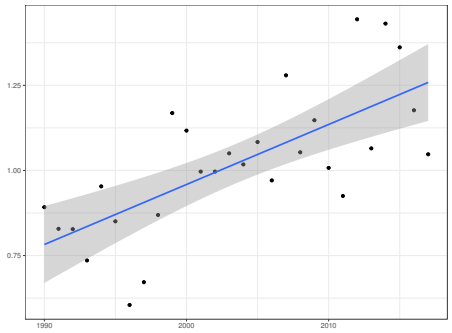
LORI also returns counts estimates, which can be used to compute an estimation of the total yearly abundances (i.e. counts estimates summed across sites). To better assess the temporal trend, one can decompose the estimated counts into three factors corresponding to the site effects, year effects and interactions respectively. Indeed, for $(i, j) \in \{1, \dots, n\} \times \{1, \dots, p\}$, one can write

$$\exp(\hat{X}_{ij}) = \exp(\hat{\mu}) \exp(R_{i,\cdot}\hat{\alpha}) \exp(C_{j,\cdot}\hat{\beta}) \exp(\hat{\mathbf{\Theta}}_{ij}).$$

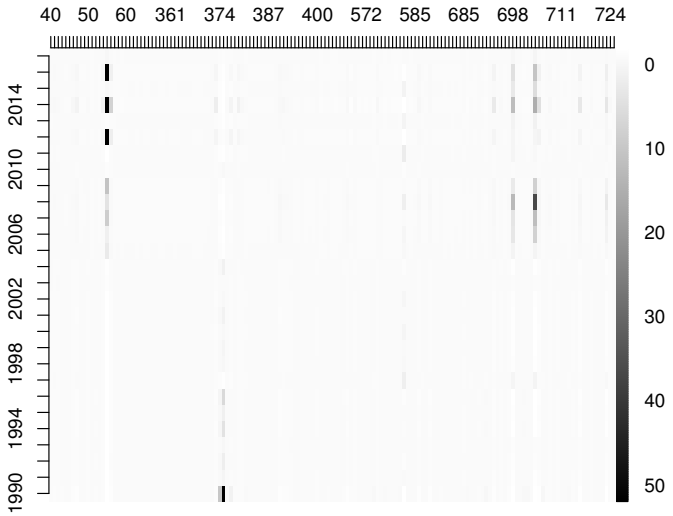
Figure 3.5 shows the last three factors of this decomposition separately.



(a) Total site effects in count scale ($\exp(R_{i,}\hat{\alpha})$, for $1 \leq i \leq n$). One site has $\exp(R_{i,}\hat{\alpha}) \geq \exp(5)$ (large red point). Horizontal line: $\exp(R_{i,}\alpha) = 1$.



(b) Total year effects in count scale ($\exp(C_{j,}\hat{\beta})$, for $1 \leq j \leq p$). Blue line: loess (standard deviation in gray).



(c) Interaction in count scale ($\exp(\hat{\mathbf{\Theta}}_{ij})$) for 150 sites (to improve the display).

Figure 3.6: Decomposition of the estimated counts into multiplicative site effects (top left), year effects (top right) and interactions (bottom).

On Figure 3.6a we see that most sites have multiplicative effects around 1 on count scale. One site (site 376, large red point) stands out; again, it corresponds to an extremely large site (6000km^2, 5 times larger than the second, 300 times larger than the

mean). In this respect, the row effects act as normalization factors accounting for surface. We also observe tenuous levels along the x axis, corresponding to sites of different countries. On Figure 3.6b we observe a slighlty increasing temporal trend. This means that, all other things being equal, later years tend to produce larger abundances. As illustrated in Figure 3.5b, this temporal trend can be associated with the effects of meteorological abnormalities. This may indicate that more birds migrated from Europe to North Africa in the recent years, due to increasing meteorological abnormalities in Europe. Note that the temporal effects (top right) are much smaller in amplitude than the spatial effects (top left). Indeed, more variability is observed in the counts between sites in a given year, than between years for a given site.

Finally, looking at the interaction matrix on Figure 3.6c, we see that the interactions are mainly driven by a few sites which interact more or less highly with every year. In particular the sites 55 and 375 present large interactions with part of the years. Site 55 corresponds to a site where the number of birds counted every year has extreme variations. For most of the years, the counts are around the median of the data set (482). On the contrary, in 2012, 2014 and 2016, the observed counts (respectively $113,990$, $239,069$, and $92,730$) are in the largest 1% of the entries: these years correspond to the large interactions observed for site 55. The site 375, on the other hand, has very few observed entries (5 across the 28 years). In particular, the count observed in 1990 ($254,749$) is much larger than the others (around $10,000$). This explains the large interaction observed for the pair (site 375, year 1990).

The site 704 also presents some large interactions. It corresponds to Ichkeul national park in Tunisia, which is a major site for most species; the abundances are very large in Ichkeul compared to other sites. However during several years including 2007, bad weather conditions prevented ornithologist to correctly count the birds, thus reported counts are significantly lower than expected. This explains the drop in the interaction in 2007 for Ichkeul, corresponding to an outlier behavior.

Such profiles, which correspond to outlying values, could not be highlighted without modeling interactions. This illustrates one of the advantages of LORI for such bird abundance data compared to state-of-the-art methods such as TRIM (Pannekoek and van Strien, 2001) which do not model interactions. In particular, in most cases the interaction terms absorb outliers (small or large), and indirectly account for the overdispersion which is known to occur in birds abundance data.

3.7 Conclusions and perspectives

In this chapter, we introduced a first low-rank method, with statistical guarantees and an open source implementation, to analyze count data with covariates. The method can either be seen as an imputation technique, which uses row and column covariates to better predict the missing values, or as a way to estimate main effects of covariates and interactions simultaneously. Promising numerical experiments suggest that it could be used successfully to impute the waterbird monitoring data set, and produce interpretable summaries. There are, however, several opportunities of improvement.

On the theoretical side, our estimation method only guarantees upper bounds on esti-

mation error of the parameter matrix $\boldsymbol{X}^0$. Although this already is a new result, it would be useful to control the error we make on the main effects and interactions separately. In particular, if we want to interpret the regression coefficients $\hat{\alpha}$ and $\hat{\beta}$. On the practical side, in settings where the number of covariates is large (or if we have qualitative covariates with many categories), and we observe only a small proportion of entries, this first model, which does not constrain the covariates coefficients, may be limiting. Indeed, as the number of covariates increases, two phenomenons occur. Firstly, the computational cost increases significantly; secondly, the number of observations may not be sufficient to estimate the covariate coefficients. To solve all these limitations simultaneously, an option would be to penalize the main effects, for example with a LASSO-type penalty.

Another limitation is that we have defined a single imputation procedure, which does not allow to estimate the variability of our parameters' point estimates. As our goal is to impute the data set prior to data analysis, single imputation is not a statistically sound approach. Thus, an interesting extension would be to build a multiple imputation method, based on this single imputation method. We investigate these two extensions in Chapter 4, by adding a regularization term to the LORI model, to constrain the covariates coefficients, and proposing a multiple imputation procedure.

3.8 Proofs

3.8.1 Proof of Theorem 5

We will first derive an upper bound for $\sum_{(i,j)\in\Omega}(\hat{\boldsymbol{X}}_{i,j}-\boldsymbol{X}^0_{i,j})^2$, then control $\|\hat{\boldsymbol{X}}-\boldsymbol{X}^0\|_F^2$ by $\|\hat{\boldsymbol{X}}-\boldsymbol{X}^0\|_F^2 \leq \sum_{(i,j)\in\Omega}(\hat{\boldsymbol{X}}_{i,j}-\boldsymbol{X}^0_{i,j})^2 + \mathrm{D}$, with D a residual term defined later on. By definition of $\hat{\boldsymbol{X}}=\hat{\boldsymbol{A}}+\hat{\boldsymbol{\Theta}}$, $\mathcal{L}(\hat{\boldsymbol{X}})+\lambda\|\hat{\boldsymbol{\Theta}}\|_* \leq \mathcal{L}(\boldsymbol{X}^0)+\lambda\|\boldsymbol{\Theta}^*\|_*$. Using the strong convexity of $\mathcal{L}$ and substracting $\langle\nabla\mathcal{L}(\boldsymbol{X}^0),\hat{X}-\boldsymbol{X}^0\rangle$ on both sides of this inequality, we obtain

$$\frac{\sigma_-^2\sum_{(i,j)\in\Omega}(\hat{\boldsymbol{X}}_{i,j}-\boldsymbol{X}^0_{i,j})^2}{2} \leq \underbrace{-\langle\nabla\mathcal{L}(\boldsymbol{X}^0),\hat{\boldsymbol{X}}-\boldsymbol{X}^0\rangle}_{I} + \underbrace{\lambda(\|\boldsymbol{\Theta}^*\|_*-\|\hat{\boldsymbol{\Theta}}\|_*)}_{II}. \quad (3.21)$$

We will bound separately the two terms on the right hand side of (3.21).

Given a matrix $\boldsymbol{X}\in\mathbb{R}^{m_1\times m_2}$, we denote $\mathcal{S}_1(\boldsymbol{X})$ (resp. $\mathcal{S}_2(\boldsymbol{X})$) the span of left (resp. right) singular vectors of $\boldsymbol{X}$. Let $P^{\perp}_{\mathcal{S}_1(\boldsymbol{X})}$ (resp. $P^{\perp}_{\mathcal{S}_2(\boldsymbol{X})}$) be the orthogonal projector in $\mathbb{R}^{m_1}$ on $\mathcal{S}_1(\boldsymbol{X})^{\perp}$ (resp. in $\mathbb{R}^{m_2}$ on $\mathcal{S}_2(\boldsymbol{X})^{\perp}$). We define the projection operator in $\mathbb{R}^{m_1\times m_2}$ $\mathcal{P}^{\perp}_{\boldsymbol{X}}:\tilde{X}\mapsto P^{\perp}_{\mathcal{S}_1(\boldsymbol{X})}\tilde{\boldsymbol{X}}P^{\perp}_{\mathcal{S}_2(\boldsymbol{X})}$, and $\mathcal{P}_{\boldsymbol{X}}:\tilde{\boldsymbol{X}}\mapsto\tilde{\boldsymbol{X}}-P^{\perp}_{\mathcal{S}_1(\boldsymbol{X})}\tilde{\boldsymbol{X}}P^{\perp}_{\mathcal{S}_2(\boldsymbol{X})}$. We use the following Lemma, proved in (Lafond, 2015, Lemma 16).

Lemma 1. *For all $\boldsymbol{M}$ and $\boldsymbol{M}'$ in $\mathbb{R}^{m_1\times m_2}$,*

(i) $\|\boldsymbol{M}+\mathcal{P}^{\perp}_{\boldsymbol{M}}(\boldsymbol{M})\|_* = \|\boldsymbol{M}\|_* + \|\mathcal{P}^{\perp}_{\boldsymbol{M}}(\boldsymbol{M})\|_*$,

(ii) $\|\boldsymbol{M}\|_* - \|\boldsymbol{M}'\|_* \leq \|\mathcal{P}_{\boldsymbol{M}}(\boldsymbol{M}-\boldsymbol{M}')\|_* - \|\mathcal{P}^{\perp}_{\boldsymbol{M}}(\boldsymbol{M}-\boldsymbol{M}')\|_*$,

(iii) $\|\mathcal{P}_M(\boldsymbol{M}-\boldsymbol{M}')\|_* \leq \sqrt{2\mathrm{rk}(\boldsymbol{M})}\|\boldsymbol{M}-\boldsymbol{M}'\|_F$.

Using $|\langle \nabla \mathcal{L}(\boldsymbol{X}^0), \hat{\boldsymbol{X}} - \boldsymbol{X}^0 \rangle| \leq \|\hat{\boldsymbol{X}} - \boldsymbol{X}^0\|_* \|\nabla \mathcal{L}(\boldsymbol{X}^0)\|$ and the triangular inequality gives that

$$I \leq \|\nabla \mathcal{L}(\boldsymbol{X}^0)\| \left(\|\mathcal{P}_{\boldsymbol{\Theta}^*}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*)\|_* + \|\mathcal{P}^{\perp}_{\boldsymbol{\Theta}^*}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*)\|_* + \|\hat{X}_0 - \boldsymbol{A}^0\|_* \right). \quad (3.22)$$

Then, Lemma 1 (ii) applied to $\hat{\boldsymbol{\Theta}}$ and $\boldsymbol{\Theta}^*$, results in

$$II \leq \lambda \left(\|\mathcal{P}_{\boldsymbol{\Theta}^*}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*)\|_* - \|\mathcal{P}^{\perp}_{\boldsymbol{\Theta}^*}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*)\|_* \right). \quad (3.23)$$

Plugging inequalities (3.22) and (3.23) in (3.21) we obtain

$$\begin{aligned} \sigma_-^2 \sum_{(i,j)\in\Omega} (\hat{\boldsymbol{X}}_{ij} - \boldsymbol{X}^0_{ij})^2 &\leq 2(\lambda + \|\nabla \mathcal{L}(\boldsymbol{X}^0)\|) \left\| \mathcal{P}_{\boldsymbol{\Theta}^*}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*) \right\|_* \\ &+ 2(\|\nabla \mathcal{L}(\boldsymbol{X}^0)\| - \lambda) \|\mathcal{P}^{\perp}_{\boldsymbol{\Theta}^*}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*)\|_* + 2\|\nabla \mathcal{L}(\boldsymbol{X}^0)\| \|\hat{\boldsymbol{A}} - \boldsymbol{A}^0\|_*. \quad (3.24) \end{aligned}$$

We now use the condition $\lambda \geq 2\|\nabla \mathcal{L}(\boldsymbol{X}^0)\|$ in (3.24):

$$\sigma_-^2 \sum_{(i,j)\in\Omega} (\hat{\boldsymbol{X}}_{ij} - \boldsymbol{X}^0_{ij})^2 \leq 3\lambda \|\mathcal{P}_{\boldsymbol{\Theta}^*}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*)\|_* + \lambda \|\hat{\boldsymbol{A}} - \boldsymbol{A}^0\|_*. \quad (3.25)$$

Then, $\text{rk}(\hat{\boldsymbol{A}} - \boldsymbol{A}^0) \leq r$ and $\|\hat{\boldsymbol{A}} - \boldsymbol{A}^0\|_F \leq \|\hat{\boldsymbol{X}} - \boldsymbol{X}^0\|_F$ imply that $\|\hat{\boldsymbol{A}} - \boldsymbol{A}^0\|_* \leq \sqrt{r}\|\boldsymbol{X}^0 - \hat{\boldsymbol{X}}\|_F$, which together with Lemma 1 (iii) and $\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*_0\|_F \leq \|\hat{\boldsymbol{X}} - \boldsymbol{X}^0\|_F$ yields

$$\sigma_-^2 \sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} \omega_{ij} (\hat{\boldsymbol{X}}_{ij} - \boldsymbol{X}^0_{ij})^2 \leq \lambda \left(3\sqrt{2\,\text{rank}(\boldsymbol{\Theta}^*)} + \sqrt{r} \right) \|\hat{\boldsymbol{X}} - \boldsymbol{X}^0\|_F. \quad (3.26)$$

We now derive the upper bound $\|\hat{\boldsymbol{X}} - \boldsymbol{X}^0\|_F^2 \leq \sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} \omega_{ij} (\hat{\boldsymbol{X}}_{ij} - \boldsymbol{X}^0_{ij})^2 + \mathsf{D}$. Define $\eta = 72\log(n+p)/(\pi \log(6/5))$,

$$\Sigma(\omega, \boldsymbol{X}) = \sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} \omega_{ij} \boldsymbol{X}^2_{ij} \quad (3.27)$$

and the set

$$\mathcal{C}(\eta, \rho) = \left\{ \boldsymbol{X} \in \mathbb{R}^{m_1 \times m_2}; \|\boldsymbol{X}\|_\infty \leq 1, \|\boldsymbol{X}\|_* \leq \sqrt{\rho}\|\boldsymbol{X}\|_F, \mathbb{E}\left[\Sigma(\omega, \boldsymbol{X})\right] > \eta \right\}. \quad (3.28)$$

We start by showing in the following Lemma that whenever $\hat{\boldsymbol{X}} - \boldsymbol{X}^0$ belongs to $\mathcal{C}(\eta, \rho)$ (for ρ and D defined later on), a restricted strong convexity property of the form $\|\hat{\boldsymbol{X}} - \boldsymbol{X}^0\|_F^2 \leq \sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} \omega_{ij} (\hat{\boldsymbol{X}}_{ij} - \boldsymbol{X}^*_{ij})^2 + \mathsf{D}$ holds. Define

$$\varsigma = 96\pi^{-1}[\rho(\mathbb{E}\|\Sigma_R\|)^2 + 8]. \quad (3.29)$$

Lemma 2. *Let $\eta = 72\log(n+p)/(\pi\log(6/5))$ and $\rho > 0$. With probability at least $1 - 8(n+p)^{-1}$, for all $X \in \mathcal{C}(\eta, \rho)$ we get*

$$|\Sigma(\omega, \boldsymbol{X}) - \mathbb{E}\left[\Sigma(\omega, \boldsymbol{X})\right]| \leq \frac{\mathbb{E}\left[\Sigma(\omega, \boldsymbol{X})\right]}{2} + \varsigma,$$

with Σ_R defined in (3.14).

Proof. Consider the event

$$\mathcal{B} = \left\{ \sup_{\boldsymbol{X} \in \mathcal{C}(\eta,\rho)} \left[|\Sigma(\omega, \boldsymbol{X}) - \mathbb{E}\left[\Sigma(\omega, \boldsymbol{X})\right]| - \frac{1}{2}\mathbb{E}\left[\Sigma(\omega, \boldsymbol{X})\right] \right] > \varsigma \right\}.$$

Define also for $l \in \mathbb{N}_*$

$$\mathcal{S}_l = \left\{ \boldsymbol{X} \in \mathcal{C}(\eta,\rho); \kappa^{l-1}\eta < \mathbb{E}\left[\Sigma(\omega, \boldsymbol{X})\right] < \kappa^l \eta \right\},$$

for $\kappa = 6/5$ and $\eta = 72\log(n+p)/(\pi\log(6/5))$. On $\mathcal{B}$, there exist $l \geq 1$ and $\boldsymbol{X} \in \mathcal{C}(\eta,\rho)$ such that $\boldsymbol{X} \in \mathcal{C}(\eta,\rho) \bigcap \mathcal{S}_l$, and

$$|\Sigma(\omega, \boldsymbol{X}) - \mathbb{E}\left[\Sigma(\omega, \boldsymbol{X})\right]| > \frac{1}{2}\mathbb{E}\left[\Sigma(\omega, \boldsymbol{X})\right] + \varsigma > \frac{1}{2}\kappa^{l-1}\eta + \varsigma = \frac{5}{12}\kappa^l\eta + \varsigma. \quad (3.30)$$

For $T > 0$, define the set

$$\mathcal{C}(\eta,\rho,T) = \{\boldsymbol{X} \in \mathcal{C}(\eta,\rho), \mathbb{E}\left[\Sigma(\omega, \boldsymbol{X})\right] \leq T\}$$

and the event

$$\mathcal{B}_l = \left\{ \sup_{\boldsymbol{X} \in \mathcal{C}(\eta,\rho,\kappa^l\eta)} |\Sigma(\omega, \boldsymbol{X}) - \mathbb{E}\left[\Sigma(\omega, \boldsymbol{X})\right]| > \frac{5}{12}\kappa^l\eta + \varsigma \right\}.$$

It follows from (3.30) that $\mathcal{B} \subset \bigcup_{l=1}^{+\infty} \mathcal{B}_l$; thus, it is enough to estimate the probability of the events $\mathcal{B}_l$, $l \in \mathbb{N}$, and then apply the union bound. Such an estimation is given in the following Lemma, adapted from Klopp (2015) (see Lemma 10). Define

$$Z_T = \sup_{\boldsymbol{X} \in \mathcal{C}(\eta,\rho,T)} |\Sigma(\omega, \boldsymbol{X}) - \mathbb{E}\left[\Sigma(\omega, \boldsymbol{X})\right]|. \quad (3.31)$$

Lemma 3. *Under the assumptions of Theorem 5,*

$$\mathbb{P}\left(Z_T \geq \frac{5}{12}T + \varsigma \right) \leq 4\mathrm{e}^{-\pi T/72}, \quad (3.32)$$

where ς is defined in (3.29).

Proof. We use the following Talagrand's concentration inequality and a symmetrization argument. Recall the statement of Talagrand's concentration inequality. Let $f : [-1,1]^m \mapsto \mathbb{R}$ a convex Lipschitz function with Lipschitz constant L, $\Xi_1, \ldots, \Xi_m$ be independent random variables taking values in $[-1,1]$, and $Z := f(\Xi_1, \ldots, \Xi_m)$. Then, for any $t \geq 0$, $\mathbb{P}(|Z - \mathbb{E}[Z]| \geq 16L + t) \leq 4\mathrm{e}^{-t^2/2L^2}$. For $\mathbf{x} = (x_{ij})$, $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, we apply this result to the function

$$f(\mathbf{x}) = \sup_{\boldsymbol{X} \in \mathcal{C}(\eta,\rho,T)} \left| \sum_{(i,j) \in [\![m_1]\!] \times [\![m_2]\!]} (x_{ij} - \pi_{ij}) \boldsymbol{X}_{ij}^2 \right|,$$

which is Lipschitz with Lipschitz constant $\sqrt{\pi^{-1}T}$:

$$
\begin{aligned}
&|f(x_{11},\dots,x_{np}) - f(z_{11},\dots,z_{np})| \\
&= \left| \sup_{\boldsymbol{X}\in\mathcal{C}(\eta,\rho,T)} \left| \sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} (x_{ij}-\pi_{ij})\boldsymbol{X}_{i,j}^2 \right| - \sup_{\boldsymbol{X}\in\mathcal{C}(\eta,\rho,T)} \left| \sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} (z_{ij}-\pi_{ij})\boldsymbol{X}_{i,j}^2 \right| \right| \\
&\le \sup_{\boldsymbol{X}\in\mathcal{C}(\eta,\rho,T)} \left| \left| \sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} (x_{ij}-\pi_{ij})\boldsymbol{X}_{i,j}^2 \right| - \left| \sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} (z_{ij}-\pi_{ij})\boldsymbol{X}_{i,j}^2 \right| \right| \\
&\le \sup_{X\in\mathcal{C}(\eta,\rho,T)} \left| \sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} (x_{ij}-\pi_{ij})\boldsymbol{X}_{i,j}^2 - \sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} (z_{ij}-\pi_{ij})\boldsymbol{X}_{i,j}^2 \right| \\
&\le \sup_{\boldsymbol{X}\in\mathcal{C}(\eta,\rho,T)} \left| \sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} (x_{ij}-z_{ij})\boldsymbol{X}_{i,j}^2 \right| \\
&\le \sup_{\boldsymbol{X}\in\mathcal{C}(\eta,\rho,T)} \sqrt{(i,j) \in \sum_{[\![m_1]\!]\times[\![m_2]\!]} \pi_{ij}^{-1}(x_{ij}-z_{ij})^2} \sqrt{\sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} \pi_{ij}\boldsymbol{X}_{i,j}^4} \\
&\le \sup_{\boldsymbol{X}\in\mathcal{C}(\eta,\rho,T)} \sqrt{\pi^{-1}} \sqrt{\sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} (x_{ij}-z_{ij})^2} \sqrt{\sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} \pi_{ij}\boldsymbol{X}_{i,j}^2} \\
&\le \sqrt{\pi^{-1}T} \sqrt{\sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} (x_{ij}-z_{ij})^2},
\end{aligned}
$$

where we have used $||a|-|b|| \le |a-b|, \|\boldsymbol{X}\|_\infty \le 1$ and $\mathbb{E}\left[\Sigma(\omega,\boldsymbol{X})\right] \le T$. Thus, Talagrand's inequality and the identity $\sqrt{\pi^{-1}T} \le T/(2\times 96) + 96/(2\pi)$ give

$$
\mathbb{P}\left(Z_T \ge \mathbb{E}(Z_T) + 768\pi^{-1} + \frac{1}{12}T + t \right) \le 4\mathrm{e}^{-t^2\pi/2T}.
$$

Taking $t = T/6$ we get

$$
\mathbb{P}\left(Z_T \ge \mathbb{E}(Z_T) + 768\pi^{-1} + \frac{3}{12}T \right) \le 4\mathrm{e}^{-\pi T/72}. \tag{3.33}
$$

Now we bound the expectation $\mathbb{E}[Z_T]$ using a symmetrization argument (Ledoux, 2001, Section 7.2). Let (ϵ_{ij}) be an i.i.d. Rademacher sequence. We have

$$
\mathbb{E}(Z_T) \le 2\mathbb{E}\left(\sup_{\boldsymbol{X}\in\mathcal{C}(\eta,\rho,T)} \left| \sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} \epsilon_{ij}\omega_{ij}\boldsymbol{X}_{i,j}^2 \right| \right), \tag{3.34}
$$

Then, the contraction inequality (see Koltchinskii (2011a), Theorem 2.2) yields

$$
\mathbb{E}(Z_T) \le 8\mathbb{E}\left(\sup_{\boldsymbol{X}\in\mathcal{C}(\eta,\rho,T)} \left| \sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} \epsilon_{ij}\omega_{ij}\boldsymbol{X}_{i,j} \right| \right) = 8\mathbb{E}\left(\sup_{\boldsymbol{X}\in\mathcal{C}(\eta,\rho,T)} |\langle \Sigma_R, \boldsymbol{X}\rangle| \right),
$$

where Σ_R is defined in (3.14). For $\boldsymbol{X} \in \mathcal{C}(\eta,\rho,T)$ we have that $\|\boldsymbol{X}\|_* \le \sqrt{\rho\pi^{-1}T}$. Then by duality between the nuclear and operator norms we obtain

$$
\mathbb{E}(Z_T) \le 8\mathbb{E}\left(\sup_{\|\boldsymbol{X}\|_*\le\sqrt{\rho\pi^{-1}T}} |\langle \Sigma_R, \boldsymbol{X}\rangle| \right) \le 8\sqrt{\rho\pi^{-1}T}\mathbb{E}\|\Sigma_R\|.
$$

Combined with (3.33) and using $8\sqrt{\rho\pi^{-1}T}\mathbb{E}\|\Sigma_R\| \leq \frac{T}{2\times 3} + \frac{3\times 8^2\rho\pi^{-1}}{2}(\mathbb{E}\|\Sigma_R\|)^2$ we finally obtain (3.32) using the definition of ς in (3.29). □

Lemma 3 implies that

$$\mathbb{P}(\mathcal{B}) \leq \sum_{l=1}^{+\infty} \mathbb{P}(\mathcal{B}_l) \leq 4\sum_{l=1}^{+\infty} \exp(-\pi\kappa^l\eta/72) \leq 8/(n+p),$$

which concludes the proof. □

Case 1 If $\sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} \pi_{ij}(\hat{\boldsymbol{X}}_{i,j} - \boldsymbol{X}^0_{i,j})^2 \leq \eta$, then $\|\hat{\boldsymbol{X}} - \boldsymbol{X}^0\|_2^2 \leq \eta/\pi$ and the result of Theorem 5 (3.18) is proved.

Case 2 If $\sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} \pi_{ij}(\hat{\boldsymbol{X}}_{i,j} - \boldsymbol{X}^0_{i,j})^2 > \eta$. Let us show that $(\hat{X} - \boldsymbol{X}^0)/2\gamma \in \mathcal{C}(\eta, 64\,\mathrm{rank}(\boldsymbol{X}^0))$. Using (3.24), $\sigma_-^2 \sum_{(i,j)\in\Omega}(\hat{\boldsymbol{X}}_{i,j} - \boldsymbol{X}^0_{i,j})^2 \geq 0$ and $\|\nabla\mathcal{L}(\boldsymbol{X}^0)\| \leq \lambda/2$, we obtain that

$$\|\mathcal{P}^{\perp}_{\boldsymbol{\Theta}^*}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*)\|_* \leq 3\|\mathcal{P}_{\boldsymbol{\Theta}^*}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*)\|_* + \|\hat{\boldsymbol{A}} - \boldsymbol{A}^*\|_*.$$

On the other hand,

$$\begin{aligned}
\|\hat{\boldsymbol{X}} - \boldsymbol{X}^0\|_* &\leq \|\mathcal{P}^{\perp}_{\boldsymbol{\Theta}^*}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*)\|_* + \|\mathcal{P}_{\boldsymbol{\Theta}^*}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*)\|_* + \|\hat{\boldsymbol{A}} - \boldsymbol{A}^*\|_* \\
&\leq 4\|\mathcal{P}_{\boldsymbol{\Theta}^*}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*)\|_* + 2\|\hat{\boldsymbol{A}} - \boldsymbol{A}^*\|_* \\
&\leq 2\sqrt{2\,\mathrm{rank}(\boldsymbol{\Theta}^*)}\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F + 2\sqrt{r}\|\hat{\boldsymbol{A}} - \boldsymbol{A}^*\|_F \\
&\leq \sqrt{64\,\mathrm{rank}(\boldsymbol{X}^0)}\|\hat{\boldsymbol{X}} - \boldsymbol{X}^0\|_F.
\end{aligned}$$

Thus, Lemma 2 implies that with probability at least $1 - 8(n+p)^{-1}$,

$$\begin{aligned}
\sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} \omega_{ij}(\hat{\boldsymbol{X}}_{i,j} - \boldsymbol{X}^0_{i,j})^2 \geq \frac{\mathbb{E}[\sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]} \omega_{ij}(\hat{\boldsymbol{X}}_{i,j} - \boldsymbol{X}^0_{i,j})^2]}{2} \\
- 384\gamma^2\pi^{-1}[64\,\mathrm{rank}(\boldsymbol{X}^0)(\mathbb{E}\|\Sigma_R\|)^2 + 8]. \quad (3.35)
\end{aligned}$$

Combining (3.35) and (3.26) we obtain

$$\begin{aligned}
\frac{\pi\|\hat{\boldsymbol{X}} - \boldsymbol{X}^0\|_F^2}{2} - \frac{384\gamma^2[64\,\mathrm{rank}(\boldsymbol{X}^0)(\mathbb{E}\|\Sigma_R\|)^2 + 8]}{\pi} \leq \\
\frac{\lambda}{\sigma_-^2}\left(3\sqrt{2\,\mathrm{rank}(\boldsymbol{\Theta}^*)} + \sqrt{r}\right)\|\hat{\boldsymbol{X}} - \boldsymbol{X}^0\|_F.
\end{aligned}$$

Finally, using the identity $ab \leq a^2 + b^2/4$ and $\mathrm{rank}(\boldsymbol{X}^0) \leq \mathrm{rank}(\boldsymbol{\Theta}^*) + r$ we obtain

$$\|\hat{\boldsymbol{X}} - \boldsymbol{X}^0\|_F^2 \leq \left(\frac{192\lambda^2}{\pi^2\sigma_-^4} + \frac{24576\gamma^2(\mathbb{E}\|\Sigma_R\|)^2}{\pi^2}\right)[\mathrm{rk}(\boldsymbol{\Theta}^*) + r] + \frac{6144}{\pi^2}. \quad (3.36)$$

3.8.2 Proof of Theorem 6

Theorem 6 derives from Theorem 5 and combining the two following steps: 1) computing a value of λ such that the condition $\lambda \geq 2\|\nabla \mathcal{L}(\boldsymbol{X}^0)\|$ holds with high probability and 2) controlling $\mathbb{E}\|\Sigma_R\|$. Let us start with 1). Define the random matrices $Z_{ij} = \omega_{ij}(-\boldsymbol{Y}_{i,j} + \exp(\boldsymbol{X}^0_{i,j}))E_{ij}$ and the quantity

$$\sigma_Z^2 = \max\left(\frac{1}{np}\left\|\sum_{i=1}^{n}\sum_{j=1}^{p}\mathbb{E}\left[Z_{ij}Z_{ij}^\top\right]\right\|, \quad \frac{1}{np}\left\|\sum_{i=1}^{n}\sum_{j=1}^{p}\mathbb{E}\left[Z_{ij}^\top Z_{ij}\right]\right\|\right). \tag{3.37}$$

Lemma 4. *Under the assumptions of Theorem 6,*

$$\frac{\sigma_-^2\beta}{np} \leq \sigma_Z^2 \leq \frac{\sigma_+^2\beta}{np}. \tag{3.38}$$

Proof. For all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, $Z_{ij}Z_{ij}^T = \omega_{ij}(-\boldsymbol{Y}_{i,j} + \exp(\boldsymbol{X}^0_{i,j}))^2 E_{ij}E_{ij}^\top$, and $\mathbb{E}[Z_{ij}Z_{ij}^\top] = \mathbb{E}[\omega_{ij}]\mathbb{E}[(-\boldsymbol{Y}_{i,j} + \exp(\boldsymbol{X}^0_{i,j}))^2]E_{ij}E_{ij}^\top$, which is a diagonal matrix with 0 everywhere except on the i-th element of its diagonal, where its value is $\mathbb{E}[\omega_{ij}]\mathbb{E}[(-\boldsymbol{Y}_{i,j} + \exp(\boldsymbol{X}^0_{i,j}))^2]$. Thus,

$$\sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]}\mathbb{E}[Z_{ij}Z_{ij}^\top]$$

is also a diagonal matrix, and the i-th element of its diagonal is $\sum_{j=1}^{p}\mathbb{E}[\omega_{ij}]\mathbb{E}[(-\boldsymbol{Y}_{i,j} + \exp(\boldsymbol{X}^0_{i,j}))^2]$. We obtain that

$$\frac{1}{np}\left\|\sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]}\mathbb{E}[Z_{ij}Z_{ij}^\top]\right\| = \frac{1}{np}\max_{i\in[\![m_1]\!]}\sum_{j=1}^{p}\mathbb{E}[\omega_{ij}]\mathbb{E}[(-\boldsymbol{Y}_{i,j} + \exp(\boldsymbol{X}^0_{i,j}))^2].$$

Using $\mathbb{E}[\boldsymbol{Y}_{i,j}] = \exp(\boldsymbol{X}^0_{i,j})$ and $\sigma_-^2 \leq \mathsf{var}(\boldsymbol{Y}_{i,j}) \leq \sigma_+^2$, we obtain:

$$\frac{\sigma_-^2}{np}\max_{i\in[\![m_1]\!]}\sum_{j=1}^{p}\mathbb{E}[\omega_{ij}] \leq \frac{1}{np}\left\|\sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]}\mathbb{E}[Z_{ij}Z_{ij}^\top]\right\| \leq \frac{\sigma_+^2}{np}\max_{i\in[\![m_1]\!]}\sum_{j=1}^{p}\mathbb{E}[\omega_{ij}]. \tag{3.39}$$

Using the same arguments, we also obtain

$$\frac{\sigma_-^2}{np}\max_{j\in[\![m_2]\!]}\sum_{i=1}^{n}\mathbb{E}[\omega_{ij}] \leq \frac{1}{np}\left\|\sum_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]}\mathbb{E}[Z_{ij}^\top Z_{ij}]\right\| \leq \frac{\sigma_+^2}{np}\max_{j\in[\![m_2]\!]}\sum_{i=1}^{n}\mathbb{E}[\omega_{ij}]. \tag{3.40}$$

Combining (3.39) and (3.39), we obtain that

$$\frac{\sigma_-^2}{np}\max\left\{\max_{i\in[\![m_1]\!]}\sum_{j=1}^{p}\mathbb{E}[\omega_{ij}], \max_{j\in[\![m_2]\!]}\sum_{i=1}^{n}\mathbb{E}[\omega_{ij}]\right\} \leq \sigma_Z^2 \leq \frac{\sigma_+^2}{np}\max\left\{\max_{i\in[\![m_1]\!]}\sum_{j=1}^{p}\mathbb{E}[\omega_{ij}], \max_{j\in[\![m_2]\!]}\sum_{i=1}^{n}\mathbb{E}[\omega_{ij}]\right\},$$

which concludes the proof. □

Note that $\mathbb{E}[Z_{ij}] = 0$ for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$ and $\nabla \mathcal{L}(\boldsymbol{X}^0) = \sum_{i=1}^n \sum_{j=1}^p Z_{ij}$. We use an extension of Theorem 4 in Koltchinskii (2013) to rectangular matrices via self-adjoint dilation (cf., for example, 2.6 in Tropp (2012)). Let $\Xi_1, \dots, \Xi_m$ be m independent $(m_1 \times m_2)$-matrices satisfying $\mathbb{E}[\Xi_i] = 0$ and

$$\inf\{K > 0 : \mathbb{E}[\exp(\|\Xi_i\|/K)] \leq e\} < M$$

for some constant M and for all $i \in \{1, \dots, m\}$. Define

$$\sigma^2 = \max\left(\frac{1}{m} \left\| \sum_{i=1}^m \mathbb{E}\left(\Xi_i \Xi_i^T\right) \right\|, \quad \frac{1}{m} \left\| \sum_{i=1}^m \mathbb{E}\left(\Xi_i^T \Xi_i\right) \right\| \right),$$

and $\bar{U} = M \log(1 + 2\frac{M^2}{\sigma^2})$. Then, for $t\bar{U} \leq 2(e-1)\sigma^2 m$,

$$\mathbb{P}\left\{ \left\| \frac{1}{m} \sum_{i=1}^m \Xi_i \right\| \geq t \right\} \leq 2(n+p) \exp\left\{ -\frac{t^2}{4m\sigma^2 + 2\bar{U}t/3} \right\}$$

and for $t\bar{U} > 2(e-1)\sigma^2 m$,

$$\mathbb{P}\left\{ \left\| \frac{1}{m} \sum_{i=1}^m \Xi_i \right\| \geq t \right\} \leq 2(n+p) \exp\left\{ -\frac{t}{(e-1)\bar{U}} \right\}.$$

Under Assumption 3 we may apply this result with $m = np$, $(\Xi_1, \dots, \Xi_m) = (Z_{11}, \dots, Z_{np})$, $M = 2\delta$, $\sigma^2 = \sigma_Z^2$ and $\bar{U} = 2\delta \log(1 + 8\delta^2/\sigma_Z^2)$. Taking

$$t \geq \max\left\{ 2\sigma_Z \sqrt{3np \log(n+p)}, 6\delta(e-1) \log(1 + 8\delta^2/\sigma_Z^2) \log(n+p) \right\}$$

and using Lemma 4, we get that with probability at least $1 - (n+p)^{-1}$,

$$\|\nabla \mathcal{L}(\boldsymbol{X}^0)\| \leq \max\left\{ 2\sigma_+ \left(3\beta \log(n+p)\right)^{1/2}, 6\delta(e-1) \log\{1 + 8\delta^2 np/(\beta\sigma_-^2)\} \log(n+p) \right\}.$$

Thus, taking λ as in Theorem 6 ensures that $\lambda \geq 2\|\nabla \mathcal{L}(\boldsymbol{X}^0)\|$ with probability at least $1 - (n+p)^{-1}$.

We now control $\mathbb{E}\|\Sigma_R\|$ with the following lemma.

Lemma 5. *There exists an absolute constant C^* such that the two following inequality holds*

$$\mathbb{E}[\|\Sigma_R\|] \leq C^* \left\{ \sqrt{\beta} + \sqrt{\log m} \right\}.$$

Proof. We use an extension to rectangular matrices via self-adjoint dilation of Corollary 3.3 in Bandeira and van Handel (2016).

Proposition 1. *Let $\boldsymbol{A}$ be an $m_1 \times m_2$ rectangular matrix with entries $\boldsymbol{A}_{i,j}$, $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, independent and centered bounded random variables. then, there exists a universal constant C^* such that*

$$\mathbb{E}[\|A\|] \leq C^* \left\{ \sigma_1 \vee \sigma_2 + \sigma_* \sqrt{\log(n \wedge p)} \right\},$$

$$\sigma_1 = \max_i \sqrt{\sum_j \mathbb{E}[\boldsymbol{A}_{i,j}^2]}, \quad \sigma_2 = \max_j \sqrt{\sum_i \mathbb{E}[\boldsymbol{A}_{i,j}^2]}, \quad \sigma_* = \max_{i,j} |\boldsymbol{A}_{i,j}|.$$

Applying Proposition 1 to Σ_R with $\sigma_1 \vee \sigma_2 \leq \sqrt{\beta}/|\Omega|$ and $\sigma_* \leq 1$ we obtain

$$\mathbb{E}[\|\Sigma_R\|] \leq C^* \left\{ \sqrt{\beta} + \sqrt{\log(n \wedge p)} \right\}.$$

□

Combining 1) and 2) with (3.36) and a union bound argument, we obtain the result of Theorem 6.

3.8.3 Proof of Theorem 7

In what follows we denote for $\boldsymbol{A} \in \mathcal{X}_0$ and $\boldsymbol{\Theta} \in \mathcal{T}$ $\mathcal{F}^\lambda(\boldsymbol{A}, \boldsymbol{\Theta}) = \mathcal{L}(\boldsymbol{A} + \boldsymbol{\Theta}) + \lambda \|\boldsymbol{\Theta}\|_*$. We establish below that $\lambda_0(\boldsymbol{Y})$ defined in (3.19) is equal to

$$\lambda_0(\boldsymbol{Y}) = \min_\lambda \quad 0 \in \partial_{\boldsymbol{\Theta}} \{ \mathcal{F}^\lambda(\hat{\boldsymbol{A}}, \boldsymbol{\Theta}) + \chi_{\mathcal{T}}(\boldsymbol{\Theta}) \} \mid_{\boldsymbol{\Theta}=0},$$

where for $\mathcal{K} \subset \mathbb{R}^{m_1 \times m_2}$,$\chi_{\mathcal{K}}(X)$ is the characteristic function of the set $\mathcal{K}$, equal to 0 on $\mathcal{K}$ and $+\infty$ elsewhere, and $\hat{\boldsymbol{A}} = \underset{\boldsymbol{A} \in \mathcal{X}_0}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{A})$ (see (3.19)). The subdifferential of the objective function $\mathcal{F}^\lambda$ with respect to $\boldsymbol{\Theta}$ is given by

$$\partial_{\boldsymbol{\Theta}} \mathcal{F}^\lambda(\hat{\boldsymbol{A}}, 0) = \nabla \mathcal{L}(\hat{\boldsymbol{A}} + \boldsymbol{\Theta}) \mid_{\boldsymbol{\Theta}=0} + \lambda \partial_{\boldsymbol{\Theta}} \|\boldsymbol{\Theta}\|_* \mid_{\boldsymbol{\Theta}=0} + \partial_{\boldsymbol{\Theta}} \chi_{\mathcal{T}}(\boldsymbol{\Theta}) \mid_{\boldsymbol{\Theta}=0}.$$

$0 \in \partial_{\boldsymbol{\Theta}} \chi_{\mathcal{T}}(\boldsymbol{\Theta}) \mid_{\boldsymbol{\Theta}=0}$. Lemma 6 ensures that $0 \in \partial \mathcal{F}^\lambda(\boldsymbol{\Theta}) \mid_{\boldsymbol{\Theta}=0}$ if and only if

$$0 \in \left\{ \nabla \mathcal{L}(\hat{\boldsymbol{A}}) + \lambda \boldsymbol{W};\ \|\mathcal{P}_{\mathcal{T}}(\boldsymbol{W})\| \leq 1 \right\}.$$

This is equivalent to $\lambda \geq \left\| \mathcal{P}_{\mathcal{T}}(\nabla \mathcal{L}(\hat{\boldsymbol{A}})) \right\|$. Additionally, at the optimum $\hat{\boldsymbol{A}}$, we have $\mathcal{P}_{\mathcal{T}}(\nabla \mathcal{L}(\hat{\boldsymbol{A}})) = \nabla \mathcal{L}(\hat{\boldsymbol{A}})$, which concludes the proof.

Lemma 6. *Let $g : \mathcal{T} \to \mathbb{R}_+$ be the function defined by $g(\boldsymbol{A}) = \|\boldsymbol{A}\|_*$ for $\boldsymbol{A} \in \mathcal{T}$. $\partial g(0) = \{\boldsymbol{W} \in \mathbb{R}^{m_1 \times m_2}, \|\mathcal{P}_{\mathcal{T}}(\boldsymbol{W})\| \leq 1\}$.*

Proof. By definition of the subdifferential we need to prove that for all $\boldsymbol{W} \in \mathbb{R}^{m_1 \times m_2}$, $\|\mathcal{P}_{\mathcal{T}}(\boldsymbol{W})\| < 1$, and for all $\boldsymbol{B} \in \mathcal{T}$, $g(\boldsymbol{B}) \geq g(0) + \langle \boldsymbol{W}, \boldsymbol{B} - 0 \rangle$. First $\boldsymbol{B} \in \mathcal{T}$ implies $\langle \boldsymbol{W}, \boldsymbol{B} \rangle = \langle \mathcal{P}_{\mathcal{T}}(\boldsymbol{W}), \boldsymbol{B} \rangle$, therefore $\|\mathcal{P}_{\mathcal{T}}(\boldsymbol{W})\| \leq 1$ is a sufficient condition for $\boldsymbol{W} \in \partial g(0)$. Now assume $\|\mathcal{P}_{\mathcal{T}}(W)\| > 1$ and let $\mathcal{P}_{\mathcal{T}}(W) = \boldsymbol{U} \Sigma \boldsymbol{V}^\top$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices of left and right singular vectors, and $\Sigma_{11} = \|\mathcal{P}_{\mathcal{T}}(\boldsymbol{W})\| > 1$. Let us define $\boldsymbol{B} = \boldsymbol{U} \tilde{\Sigma} \boldsymbol{V}^\top$, $\tilde{\Sigma}_{11} = 1$ and $\tilde{\Sigma}_{ij} = 0$ elsewhere; note that with this definition $\boldsymbol{B} \in \mathcal{T}$. We have $g(\boldsymbol{B}) = 1$ and $\langle \mathcal{P}_{\mathcal{T}}(\boldsymbol{W}), \boldsymbol{B} \rangle = \Sigma_{11} > g(\boldsymbol{B})$. Therefore $\|\mathcal{P}_{\mathcal{T}}(\boldsymbol{W})\| > 1 \Rightarrow \boldsymbol{W} \notin \partial g(0)$, from which we conclude

$$\partial g(0) = \left\{ \boldsymbol{W} \in \mathbb{R}^{m_1 \times m_2}, \|\mathcal{P}_{\mathcal{T}}(\boldsymbol{W})\| < 1 \right\}.$$

□

Chapter 4

Estimation of waterbird population trends with multiple imputations

Contents

4.1 Introduction

Birds are among the most studied species worldwide, and their monitoring is used as a surrogate to evaluate the state of biodiversity on a global scale. In particular, waterbirds are important ecosystem service providers, involved in the dispersion of seeds, acting as sentinels of epidemics and bioindicators of the condition of wetlands (Amat and Green, 2010). Waterbirds have been monitored since the 1960s, and are now counted yearly in more than 25,000 sites all over the world (Amano et al., 2017). These abundance data–and their reliability–are crucial to guide international conventions for the conservation of biodiversity. However, global studies have shown that increased surveillance efforts are required, in particular in the southern half of the Mediterranean basin, which is one of the most important biodiversity hotspots (Galewski et al., 2011). At this regional scale, the estimation of waterbird population trends is indeed often jeopardized by the lack of data. For example, the International Waterbird Census (IWC), initiated in 1967 by Wetlands International (`www.wetlands.org`) and conducted–in theory–every year, has been regular only since 1983 in Morocco, 1985 in Algeria, and 2002 in Tunisia. Furthermore, in some countries, the spatial coverage of the bird counts has remained

variable over the years, for financial and political reasons (Etayeb et al., 2015). From a data analysis perspective, this results in missing values.

For a concrete example, in this chapter, we analyze a North-African waterbird abundance data set, which contains bird counts across 785 sites in the five North-African countries (Algeria, Egypt, Libya, Morocco and Tunisia), between 1990 and 2017. The data consist of several count tables corresponding to different species, and where rows represent different sites and columns different years. In these count tables, after removing the sites where the species were never observed, there are between 40% and 60% of missing entries, depending on the species. The goal is to estimate total yearly counts, i.e. to compute column-wise sums, and to assess an uncertainty measure for these temporal trends. We approach the problem with a missing values imputation perspective: we predict plausible values for the missing entries, and then compute yearly totals. To provide an uncertainty measure, we resort to multiple imputation (Rubin, 1987). Multiple imputation consists in predicting, for each missing entry, several plausible values instead of a single prediction, in order to assess the uncertainty associated to the missing entries. An important feature of the waterbird data set, is that supplementary information about the rows and columns of the count table are also available. For example, geographical and meteorological information about where and when the birds were counted. In this chapter, we introduce a method which uses this side information with a dual goal: to improve the imputation quality and to select important predictors of waterbird abundances. Note that we analyze each species independently of the others.

Indeed, several studies have already shown that such factors are good predictors of waterbird abundances (Amano et al., 2017). However, as the supplementary data are often retrospectively scrapped from the web, and the available factors do not always correspond to actual biological hypotheses, it is unlikely that all variables have an effect on the observed counts. Thus, we expect to observe a sparsity phenomenon, where the effect of some of the variables included in the model take zero values. Furthermore, the spatial behavior of waterbirds is complex, and the available covariates may not be sufficient to explain the observations. In particular, one of the main hypotheses in this chapter is that sites and years *interact*. Indeed, we observed that bird counts are extremely variable across years for some of the sites, and quite stable for others. In light of these model assumptions, existing count imputation methods classically used by ecologists are not completely satisfactory, because they model either only effects of covariates, only interactions, or consist in *single imputation* methods without uncertainty assessment.

The most commonly used imputation method for waterbird count data is probably TRIM (trends and indices for monitoring data, Pannekoek and van Strien (2001)). TRIM is based on a Poisson log-linear model, and may include covariate effects. However, the implemented models are designed to incorporate *qualitative* covariates either about the rows *or* columns of the count table. In the present waterbird data analysis, the supplementary variables concern the rows (sites), columns (years), and also the row-column pairs (variables which depend on the sites and on the years); moreover most of the variables are quantitative. In addition, the TRIM model is not designed to incorporate row-column interactions. Another popular method is Correspondence Analysis (CA, Greenacre (1984)), a component method for count data which can be used for imputation purposes (Josse and Husson, 2016). CA imputes the missing values using a low-rank model which may be interpreted as interactions, but does not incorporate side information. Furthermore, CA and TRIM both correspond to *single imputation* methods, and do not evaluate the uncertainty of the imputed counts. Multiple imputation

methods for count data already exist. For example, the R package `countimp` (Kleinke and Reinecke, 2013), based on Multiple Imputation by Chained Equations (mice, van Buuren and Groothuis-Oudshoorn (2011)), implements multiple imputation methods for several count data models (Poisson, negative binomial, as well as zero-inflated and multilevel extensions), and may incorporate additional predictors. However, to the best of our knowledge, these methods do not model interactions.

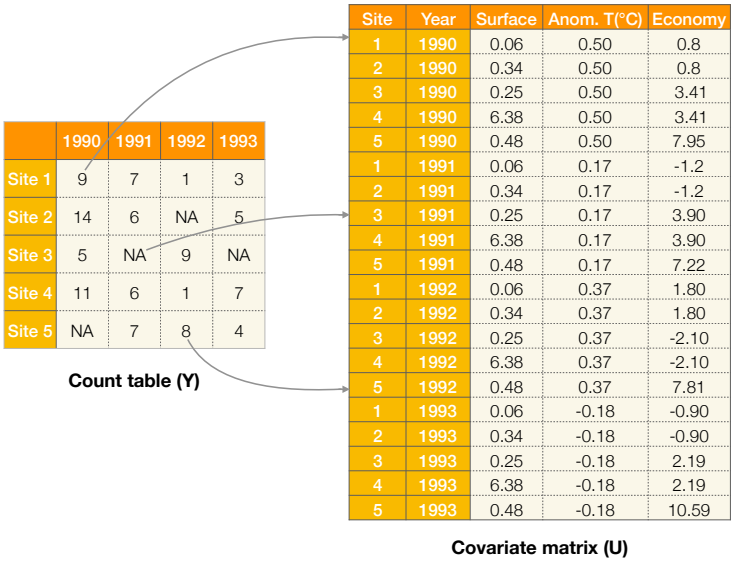
In this chapter, we introduce a new multiple imputation method for count data which incorporates effects of supplementary covariates (which may be quantitative or qualitative), models row-column interactions, and automatically selects important covariates. The method may be casted in the framework of bootstrap-based multiple imputation methods. However, note that, because our imputation procedure is biased, so we cannot hope to obtain valid confidence regions for our prediction even with multiple imputation. Nevertheless, the resampling procedure allows to derive intervals of variability which reflect the variability of our imputations with respect to the missing values and the noise of the observations. The rest of the chapter is organized as follows. In Section 4.2, we describe the complete procedure, namely the estimation problem and the single imputation procedure, as well as the resampling method and the resulting multiple imputation. Then, in Section 4.3, we evaluate the method in terms of imputation of the missing values and compare it to state-of-the-art count data imputation methods. We demonstrate that it yields smaller imputation errors when the data is generated according to a Poisson model, in particular when the proportion of missing values is large. We also evaluate the coverage of the method, and show that it is robust to model misspecification with an experiment on zero-inflated and overdispersed count data. To conclude Section 4.3, we evaluate the method and the same competitors on the imputation of a subsample of the waterbird data set. Finally, in Section 4.4, we apply the method to estimate the population trends of three waterbird species.

4.2 Multiple imputation

The new multiple imputation method developed in this chapter consists of two main building blocks: a single imputation method based on a Poisson log-linear model and a resampling procedure. In Section 4.2.1, we describe the distinctive features of our model, and try to provide intuitions about the underlying biological assumptions. Then, we propose an estimation procedure in Section 4.2.2, based on the minimization of a doubly penalized Poisson negative log-likelihood. We also describe the mixed coordinate gradient descent (MCGD) algorithm used to solve the optimization problem, and implemented in the R package `lori`. Finally, Section 4.2.3 describes the multiple imputation procedure.

4.2.1 Poisson log-linear model

Consider an abundance table $\boldsymbol{Y} \in \mathbb{N}^{m_1 \times m_2}$ containing missing values. Assume that a supplementary covariate matrix, $\boldsymbol{U} \in \mathbb{R}^{(m_1 m_2) \times q}$, contains side information about the rows and columns of the count table $\boldsymbol{Y}$. In the waterbird example, the rows of $\boldsymbol{Y}$ correspond to ecological sites, and the columns to years. The matrix $\boldsymbol{U}$ has $m_1 m_2$ rows, each corresponding to an entry of $\boldsymbol{Y}$, as represented on Figure 4.1. The columns of $\boldsymbol{U}$ correspond to variables describing either the rows, the columns, or the row-column pairs of $\boldsymbol{Y}$. For instance, in Figure 4.1, the first column of $\boldsymbol{U}$ indicates the surface of the sites



	1990	1991	1992	1993
Site 1	9	7	1	3
Site 2	14	6	NA	5
Site 3	5	NA	9	NA
Site 4	11	6	1	7
Site 5	NA	7	8	4

Count table (Y)

Site	Year	Surface	Anom. T(°C)	Economy
1	1990	0.06	0.50	0.8
2	1990	0.34	0.50	0.8
3	1990	0.25	0.50	3.41
4	1990	6.38	0.50	3.41
5	1990	0.48	0.50	7.95
1	1991	0.06	0.17	-1.2
2	1991	0.34	0.17	-1.2
3	1991	0.25	0.17	3.90
4	1991	6.38	0.17	3.90
5	1991	0.48	0.17	7.22
1	1992	0.06	0.37	1.80
2	1992	0.34	0.37	1.80
3	1992	0.25	0.37	-2.10
4	1992	6.38	0.37	-2.10
5	1992	0.48	0.37	7.81
1	1993	0.06	-0.18	-0.90
2	1993	0.34	-0.18	-0.90
3	1993	0.25	-0.18	2.19
4	1993	6.38	-0.18	2.19
5	1993	0.48	-0.18	10.59

Covariate matrix (U)

Figure 4.1: Incomplete count table and covariate matrix for one of the waterbird species.

(in km^2), a variable which depends only on the site index. Thus, the first column takes the same value in the rows that correspond to the same site (e.g. rows 1, 6, 11, 16). On the other hand, the second column of $\boldsymbol{U}$ indicates a global temperature anomaly, which only depends on the years. Thus, the second column takes the same value in the rows that correspond to the same year (e.g. rows 1-5, 6-10, etc.). Finally, the third column indicates a regional economical index, which depends on the location of the site *and* on the year. Note that the covariate vector associated to the (i,j)-th entry $\boldsymbol{Y}_{i,j}$ is the $(j-1)m_1+i$-th row of $\boldsymbol{U}$, $\boldsymbol{U}_{(j-1)m_1+i,\cdot} \in \mathbb{R}^q$. For simplicity, we denote this vector

$$\boldsymbol{U}^{i,j} := \boldsymbol{U}_{(j-1)m_1+i,\cdot}. \tag{4.1}$$

We also denote by $\boldsymbol{U}^{i,j}_k$ the k-th entry of $\boldsymbol{U}^{i,j}$.

We assume the entries of $\boldsymbol{Y}$ to be independent and to follow Poisson distributions. That is, we assume the following Poisson log-linear model:

$$\boldsymbol{Y}_{i,j} \sim \mathcal{P}(\exp(\boldsymbol{X}^0_{i,j})), \quad \boldsymbol{X}^0_{i,j} = \alpha^0_i + \beta^0_j + \textstyle\sum_{k=1}^{K_1+K_2} \boldsymbol{U}^{i,j}_k \epsilon^0_k + \boldsymbol{\Theta}^0_{i,j}, \tag{4.2}$$

where $\mathcal{P}(\lambda)$ denotes the Poisson distribution of intensity λ. In (4.2), α^0_i is a row effect, β^0_j a column effect, ϵ^0_k is the effect of the k-th covariate, and $\boldsymbol{\Theta}^0_{i,j}$ an interaction term between row i and column j. The model is saturated, and we make two main assumptions to constrain the parameter space. First, we assume that the vectors α^0, β^0 and ϵ^0 are sparse (contain zero values), meaning that not all sites, years or covariates have an effect on the counts. Second, we assume the interaction matrix $\boldsymbol{\Theta}^0$ has low-rank, meaning that the row-column interactions may be summarized by multiplicative interactions between a few latent row and column factors. To impose such structure to the parameters, we use an estimation procedure which includes sparsity and low-rank inducing regularization terms, as described in Section 4.2.2.

Denote by $\Omega \in \{0,1\}^{m_1 \times m_2}$ the observation mask, satisfying $\Omega_{i,j} = 1$ if $\boldsymbol{Y}_{i,j}$ is observed, and 0 otherwise. Denote by $\boldsymbol{Y}_{obs}$ the set of observed values, and $\boldsymbol{Y}_{mis}$ the set of unobserved values. Along this chapter, we assume a Missing At Random missing values mechanism. Denote the conditional probability of the observation mask

$$\mathbb{P}\left(\Omega | \boldsymbol{U} = \tilde{\boldsymbol{U}}, \boldsymbol{Y}_{obs} = \tilde{\boldsymbol{Y}}_{obs}, \boldsymbol{Y}_{mis} = \tilde{\boldsymbol{Y}}_{mis}\right), \tag{4.3}$$

where $\tilde{\boldsymbol{U}}$, $\tilde{\boldsymbol{Y}}_{obs}$ and $\tilde{\boldsymbol{Y}}_{mis}$ are possible values of the covariates and counts. The MAR assumption means that the probability given in (4.3) takes the same value for any $\tilde{\boldsymbol{Y}}_{mis}$, once $\tilde{\boldsymbol{U}}$ and $\tilde{\boldsymbol{Y}}_{obs}$ are fixed. Note that, in particular, the missingness pattern may depend on the covariate matrix $\boldsymbol{U}$, which we assume completely observed.

4.2.2 Estimation

To perform multiple imputation, we first define a single imputation procedure, through the estimation of model (4.2). For a set of parameters $(\alpha, \beta, \epsilon, \boldsymbol{\Theta})$, consider the Poisson negative log-likelihood defined by:

$$\mathcal{L}(\boldsymbol{Y}; \alpha, \beta, \epsilon, \boldsymbol{\Theta}) = \sum_{(i,j)} \Omega_{i,j}[-\boldsymbol{Y}_{i,j}(\alpha_i + \beta_j + \sum_{k=1}^{K} \epsilon_k \boldsymbol{U}_k^{i,j} + \boldsymbol{\Theta}_{i,j}) + \exp(\alpha_i + \beta_j + \sum_{k=1}^{K} \epsilon_k \boldsymbol{U}_k^{i,j} + \boldsymbol{\Theta}_{i,j})]. \quad (4.4)$$

We estimate $(\alpha^0, \beta^0, \epsilon^0, \boldsymbol{\Theta}^0)$ by minimizing the data-fitting term (4.4) penalized by a hybrid penalty:

$$(\hat{\alpha}, \hat{\beta}, \hat{\epsilon}, \hat{\boldsymbol{\Theta}}) \in \operatorname{argmin} \mathcal{L}(\boldsymbol{Y}; \alpha, \beta, \epsilon, \boldsymbol{\Theta}) + \lambda_1 \|\boldsymbol{\Theta}\|_* + \lambda_2 \left(\|\alpha\|_1 + \|\beta\|_1 + \|\epsilon\|_1\right), \quad (4.5)$$

with λ_1 and λ_2 two positive regularization parameters. The first regularization term ($\lambda_1 \|\boldsymbol{\Theta}\|_*$) is a nuclear norm penalty for the matrix of row-column interactions, which induces low-rank solutions: this may be interpreted as assuming a few latent factors summarize the interactions. The second penalty ($\lambda_2(\|\alpha\|_1 + \|\beta\|_1 + \|\epsilon\|_1)$) is a LASSO-type regularization term (Tibshirani, 1996), which induces sparse solutions for the vectors of main effects (the vectors $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\epsilon}$ contain many zeros). The intuition behind (4.5) is to fit the data as much as possible, by minimizing the negative log-likelihood, while enforcing models of low-complexity, through the additional penalties which constrain the parameter space and induce automatic model selection. The trade-off between fitting the data and producing low-complexity solutions is controlled by the regularization parameters λ_1 and λ_2. As λ_1 increases, the rank of the solution $\hat{\boldsymbol{\Theta}}$ decreases (the number of latent factors decreases). As λ_2 increases, the number of nonzero values in $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\epsilon}$ decreases (the number of active rows, columns and covariates decreases). Statistical guarantees for such estimators will be provided in Chapter 6.

Problem (4.5) is solved using a mixed coordinate gradient descent procedure (MCGD), where $\phi := (\alpha, \beta, \epsilon)$ on the one hand, and $\boldsymbol{\Theta}$ on the other hand, are updated alternatively. The vector ϕ is updated along a proximal gradient direction. Denote by $\nabla \mathcal{L}_\phi(\phi, \boldsymbol{\Theta})$ the gradient of $\mathcal{L}$ with respect to ϕ and evaluated at $(\phi, \boldsymbol{\Theta})$. At iteration t, we update ϕ as follows:

$$\begin{aligned} \phi^{(t)} &= \operatorname{prox}_{\gamma\lambda_2\|\cdot\|_1}\left(\phi^{(t-1)} - \gamma \nabla \mathcal{L}_\phi(\phi, \boldsymbol{\Theta})\right) \\ &= \mathrm{T}_{\gamma\lambda_2}\left(\phi^{(t-1)} - \gamma \nabla \mathcal{L}_\phi(\phi, \boldsymbol{\Theta})\right), \end{aligned} \quad (4.6)$$

where γ is a step sized computed with a line search, and for $\lambda > 0$, $\mathrm{T}_\lambda(x) = \operatorname{sign}(x) \odot (x - \lambda \mathbb{1})_+$ is the component-wise soft-thresholding operator at level λ. The matrix $\boldsymbol{\Theta}$, on the other hand, is updated along a conditional gradient direction. Denote $F(\phi, \boldsymbol{\Theta}) = \mathcal{L}(\boldsymbol{Y}; \phi, \boldsymbol{\Theta}) + \lambda_1 \|\boldsymbol{\Theta}\|_* + \lambda_2 \|\phi\|_1$. We define at iteration t the quantity $R^{(t)} = \lambda_1^{-1} F(\phi^{(t-1)}, \boldsymbol{\Theta}^{(t-1)})$. We also denote $\nabla \mathcal{L}_{\boldsymbol{\Theta}}(\phi^{(t-1)}, \boldsymbol{\Theta}^{(t-1)})$ the gradient of $\mathcal{L}$ with respect to $\boldsymbol{\Theta}$ evaluated at $(\phi^{(t-1)}, \boldsymbol{\Theta}^{(t-1)})$. Furthermore, we denote by

$\sigma_1(\nabla \mathcal{L}_{\mathbf{\Theta}}(\phi^{(t-1)}, \mathbf{\Theta}^{(t-1)}))$ its largest singular value, and u_1 and v_1 its first left and right singular vectors. The conditional gradient update consists in the following operation:

$$\mathbf{\Theta}^{(t)} = \begin{cases} 0 & \text{if } \lambda_1 \geq \sigma_1(\nabla \mathcal{L}_{\mathbf{\Theta}}(\phi^{(t-1)}, \mathbf{\Theta}^{(t-1)})), \\ \mathbf{\Theta}^{(t-1)} - \delta R^{(t)} u_1 v_1^\top & \text{if } \lambda_1 < \sigma_1(\nabla \mathcal{L}_{\mathbf{\Theta}}(\phi^{(t-1)}, \mathbf{\Theta}^{(t-1)})). \end{cases} \tag{4.7}$$

In the update (4.7), δ is a step size which we determine with a line search. A sketch of the MCGD algorithm is provided in Algorithm 2, and its convergence properties are studied in Robin et al. (2018).

Algorithm 2 MCGD algorithm for (4.5).

1: **Initialize:** — $\mathbf{\Theta}^{(0)}, \phi^{(0)}, R^{(0)}$. E.g., $(\mathbf{\Theta}^{(0)}, \phi^{(0)}, R^{(0)}) = (\mathbf{0}, \mathbf{0}, 0)$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: *// Update for ϕ //*
Compute the proximal update using (4.6) to obtain $\phi^{(t)}$.
4: Perform line search to compute the step size γ.
5: *// Update for $\mathbf{\Theta}$ //*
Compute as $R^{(t)} := \lambda_1^{-1} F(\phi^{(t)}, \mathbf{\Theta}^{(t-1)})$.
6: Compute the update direction, $\hat{\mathbf{\Theta}}^{(t)}$, using (4.7).
7: Perform line search to compute the step size δ.
8: **end for**
9: **Return:** $\mathbf{\Theta}^{(T)}, \phi^{(T)}$.

Based on estimation problem (4.5), we define a single stochastic imputation procedure as follows. For $(i, j) \in [\![m_1]\!] \times [\![m_2]\!]$, denote $\hat{\boldsymbol{X}}_{i,j} = \hat{\alpha}_i + \hat{\beta}_j + \sum_{k=1}^{q} \hat{\epsilon}_k \boldsymbol{U}_k^{i,j} + \hat{\mathbf{\Theta}}_{i,j}$. We define the imputed data set $\hat{\boldsymbol{Y}}$ by:

$$\begin{array}{ll} \hat{\boldsymbol{Y}}_{i,j} = \boldsymbol{Y}_{i,j} & \text{if } \Omega_{i,j} = 1, \\ \hat{\boldsymbol{Y}}_{i,j} \sim \mathcal{P}(\exp(\hat{\boldsymbol{X}}_{i,j})) & \text{if } \Omega_{i,j} = 0. \end{array} \tag{4.8}$$

The imputation model (4.8) is stochastic since we draw samples from Poisson distributions. In classical single imputation procedures when one seeks to predict the missing values as well as possible without measuring the uncertainty, one would replace the sampling in (4.8) by a prediction step with $\hat{\boldsymbol{Y}}_{i,j} = \exp(\hat{\boldsymbol{X}}_{i,j})$ if $\Omega_{i,j} = 0$. To perform multiple imputations, we apply the single imputation (4.8) to M incomplete data sets bootstrapped from the original data $\boldsymbol{Y}$. Furthermore, we apply procedure (4.8) several times on each incomplete data set, to model the uncertainty of the imputed values.

4.2.3 Multiple imputation

We define a resampling procedure to model the uncertainty of the single imputation model (4.8), which corresponds to uncertainty about the parameter matrix $\hat{\boldsymbol{X}}$. To do so, we interpret the count matrix as a contingency table, and perform nonparametric bootstrap by resampling the counts with a multinomial model. We "de-aggregate" the

count table $\boldsymbol{Y}$:

$$\boldsymbol{Y} = \begin{array}{c|cccc} & \mathbf{1990} & \mathbf{1991} & \mathbf{1992} & \mathbf{1993} \\ \hline \textbf{Site 1} & 9 & 7 & 1 & 3 \\ \textbf{Site 2} & 14 & 6 & NA & 5 \\ \textbf{Site 3} & 5 & NA & 9 & NA \\ \textbf{Site 4} & 11 & 6 & 1 & 7 \\ \textbf{Site 5} & NA & 7 & 8 & 4 \end{array} \iff \boldsymbol{Z} = \begin{array}{c|c} \textbf{Site ID} & \textbf{Year} \\ \hline \text{Site 1} & 1990 \\ \text{Site 1} & 1990 \\ \vdots & \vdots \\ \text{Site 1} & 1990 \\ \text{Site 2} & 1990 \\ \vdots & \vdots \\ \text{Site 2} & 1990 \\ \vdots & \vdots \end{array}$$

In the de-aggregated table $\boldsymbol{Z}$, for all $(i, j) \in [\![m_1]\!] \times [\![m_2]\!]$, the row (Site i, Year j) is repeated $\boldsymbol{Y}_{i,j}$ times if $\boldsymbol{Y}_{i,j}$ is observed. If $\boldsymbol{Y}_{i,j}$ is not observed, the row (Site i, Year j) does not appear at all. We assume that the rows of the de-aggregated table $\boldsymbol{Z}$ are i.i.d. This amounts to assuming that, for each bird species, each individual bird is observed in site i during the year j independently of the other individuals, and with the same probability $\pi_{i,j}$ (however the $\pi_{i,j}$ vary across sites and years). This is a simplistic model, as birds are known to have gregarious behaviors, which challenges the independence assumption: we leave improvement in this direction to future work. Let n be the number of rows in $\boldsymbol{Z}$, i.e. the total number of birds counted in $\boldsymbol{Y}$, and let M be a predefined integer number. We perform nonparametric bootstrap by sampling n rows of $\boldsymbol{Z}$ with equal probability and with replacement. Furthermore, we repeat this procedure M times, thus obtaining M new tables $\tilde{\boldsymbol{Z}}^1, \ldots, \tilde{\boldsymbol{Z}}^M$. Then, each $\tilde{\boldsymbol{Z}}^m$, $m \in [\![M]\!]$ is "re-aggregated" to form a new count table $\tilde{\boldsymbol{Y}}^m$. Finally we obtain M count tables $\tilde{\boldsymbol{Y}}^1, \ldots, \tilde{\boldsymbol{Y}}^M$, with the same missing data pattern as the original table $\boldsymbol{Y}$.

The nonparametric bootstrap procedure described here in fact amounts to sampling new counts from a multinomial distribution with frequencies equal for each entry to $Y_{i,j}/N$, $N = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \boldsymbol{Y}_{i,j}$, and with total number of trials N. This is not the only way to generate bootstrap samples: other options include, for example, performing the same multinomial sampling in every row (or column) of the count table.

For each of the M incomplete data sets, we estimate (using our algorithm) a set of parameters

$$(\tilde{\alpha}^m, \tilde{\beta}^m, \tilde{\epsilon}^m, \tilde{\boldsymbol{\Theta}}^m),$$

and obtain a parameter matrix $\hat{\boldsymbol{X}}^m$. These parameter matrices are used to produce M imputation models, with the single imputation procedure described in (4.8). These multiple models reflect the uncertainty about the imputation procedure (learned from incomplete data). For each of these models, we produce a completed data set $\tilde{\boldsymbol{Y}}^m$, $m \in [\![M]\!]$. Then, uncertainty in the missing values is estimated. Since we produced completed data sets, we may now incorporate "new" missing values, to generate new missing data patterns. To do so, we add missing completely at random (MCAR) missing values to each $\tilde{\boldsymbol{Y}}^m$, with the same proportion of missing values as in the original data set. We then re-estimate an imputation model $(\hat{\alpha}^m, \hat{\beta}^m, \hat{\epsilon}^m, \hat{\boldsymbol{\Theta}}^m)$. Finally, we model the variability of the imputation with a parametric bootstrap. For each model $(\hat{\alpha}^m, \hat{\beta}^m, \hat{\epsilon}^m, \hat{\boldsymbol{\Theta}}^m)$, we generate D imputed data sets, using the same stochastic imputation procedure (4.8). Finally, we obtain MD imputed data sets $(\hat{\boldsymbol{Y}}_1^1, \ldots, \hat{\boldsymbol{Y}}_D^1, \hat{\boldsymbol{Y}}_1^2, \ldots, \hat{\boldsymbol{Y}}_D^2 \ldots, \hat{\boldsymbol{Y}}_1^M, \ldots, \hat{\boldsymbol{Y}}_D^M)$. The complete multiple imputation procedure is summarized in Figure 4.2. Note that our single imputation procedure is biased: the ℓ_1 norm and nuclear norm regularization terms

are precisely meant to trade variance for bias, in order to reduce the estimation error. As a result, the intervals of variability that we compute are not confidence intervals, but reflect the variability of our estimates in relation to the variability of the observations.

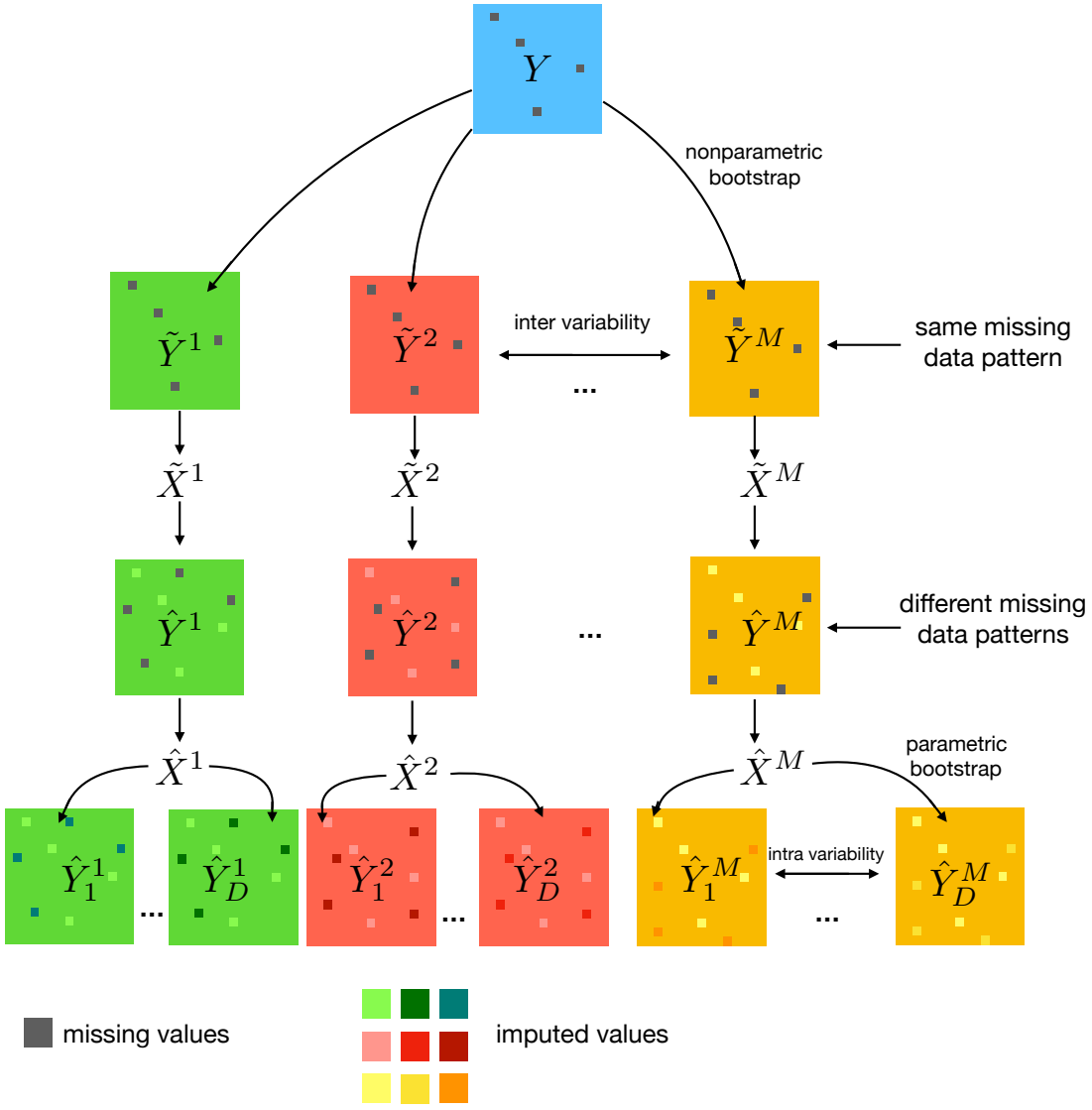


Figure 4.2: Multiple imputation procedure: nonparametric bootstrap (M samples); estimation (M estimates); parametric bootstrap (MD imputed data sets). Gray cells correspond to missing values: the same missing data pattern is shared across all incomplete data sets in the first level ($\tilde{\boldsymbol{Y}}^1, \ldots, \tilde{\boldsymbol{Y}}^M$). In the second level the missing data patterns are different ($\hat{\boldsymbol{Y}}^1, \ldots, \hat{\boldsymbol{Y}}^M$). The colored cells on the bottom line (which differ from the background color) correspond to imputed missing values, with different imputed values across the multiple imputed data sets.

4.3 Empirical performance

In this section, we conduct an empirical study to assess the performance of our imputation method, and compare it to state-of-the-art imputation techniques in three settings. In Section 4.3.1, we generate synthetic data under the assumed Poisson log-linear model

and show that, under this model, our method outperforms competitors, particularly when the proportion of missing values is large. In Section 4.3.2, we compare imputation performances under model misspecification, with experiments on non-Poisson count data. Finally in Section 4.3.3, we demonstrate that the method performs favourably on a subsample of the waterbird data set. Throughout the experiment section, we refer to our method as LORI (LOw-Rank Interactions).

4.3.1 Imputation of synthetic data

In a first experiment, we generate synthetic data under the LORI model (4.2). We sample a covariate matrix $\boldsymbol{U}$ as follows. We sample a row covariate matrix $\boldsymbol{R} \in \mathbb{R}^{300\times 2}$, a column covariate matrix $\boldsymbol{C} \in \mathbb{R}^{30\times 2}$, and a row-column covariate matrix $\boldsymbol{E} \in \mathbb{R}^{9,000\times 2}$. Then, we combine them in a large covariate matrix $\boldsymbol{U} \in \mathbb{R}^{9,000\times 6}$, where the rows of $\boldsymbol{R}$ and $\boldsymbol{C}$ are replicated. We set $\epsilon = (1.2, 0, -1.2, 0, 1)$. The interaction matrix $\boldsymbol{\Theta}$ is set to have a rank of 2, and its Euclidean norm $\|\boldsymbol{\Theta}\|_F$ fixed to $0.1\|\boldsymbol{U}\epsilon\|_2$, meaning that the scale of the interactions is small compared to the main effects. Finally, we generate a count table $\boldsymbol{Y} \in \mathbb{N}^{300\times 30}$ according to the Poisson model (4.2). To assess the performance of LORI in terms of imputation, we artificially remove entries from $\boldsymbol{Y}$, using two different missing data mechanisms. In the first mechanism, referred to as "random" in the figures, we remove entries uniformly at random. In the second mechanism, referred to as "pattern", we first select (at random), a fraction of rows and columns, and then remove some of their entries, so that the missing values are concentrated in these rows and columns. This is a more plausible mechanism as, in practice, the most important sites are almost always visited, while the less important sites are often neglected. In addition, during the years when less sampling effort was made (for financial reasons for instance), only a few sites are sampled. Finally, we predict the missing entries with LORI and several competitors:

- Imputation with Correspondence Analysis (CA) implemented in the package `missMDA` (Josse and Husson, 2016)
- Imputation with Trends and Indices in Monitoring data (TRIM) implemented in the package `rtrim` (Pannekoek and van Strien, 2001). We use the "model 3" in their implementation, which corresponds to a Poisson log-linear model with row and column effects (and no covariate).
- Imputation with TRIM using "model 3" and additional categorical covariates (obtained by cutting the quantitative covariates in $\boldsymbol{U}$). We only use this option for 10% of missing values. For larger percentages, the method failed most of the time, because not enough observations were available for every level combination of the categorical covariates.
- Imputation with a Generalized Linear Mixed Model (GLMM) with random row and column effects, and using the covariates $\boldsymbol{U}$ as predictors. We use the package `glmmTMB` (Brooks et al., 2017).

The results of the experiment are displayed in Figure 4.3, where we represent boxplots of the relative RMSE, i.e. the imputation error $\sqrt{\|\hat{\boldsymbol{Y}} - \boldsymbol{Y}\|_F^2}$ divided by the number of missing values in $\boldsymbol{Y}$. In other words, this corresponds to the average error per missing entry. As expected since we simulate under our model, LORI achieves smaller

imputation errors. The more remarkable result is that its imputation performance is quite stable across different proportions of missing values. Furthermore, the improvement with respect to other methods increases with the proportion of missing values. This is an encouraging result which seems to indicate that we can improve the imputation results by taking advantage of side information in settings where a large proportion of the data is missing, as in the waterbirds data set. Second, we evaluate the coverage

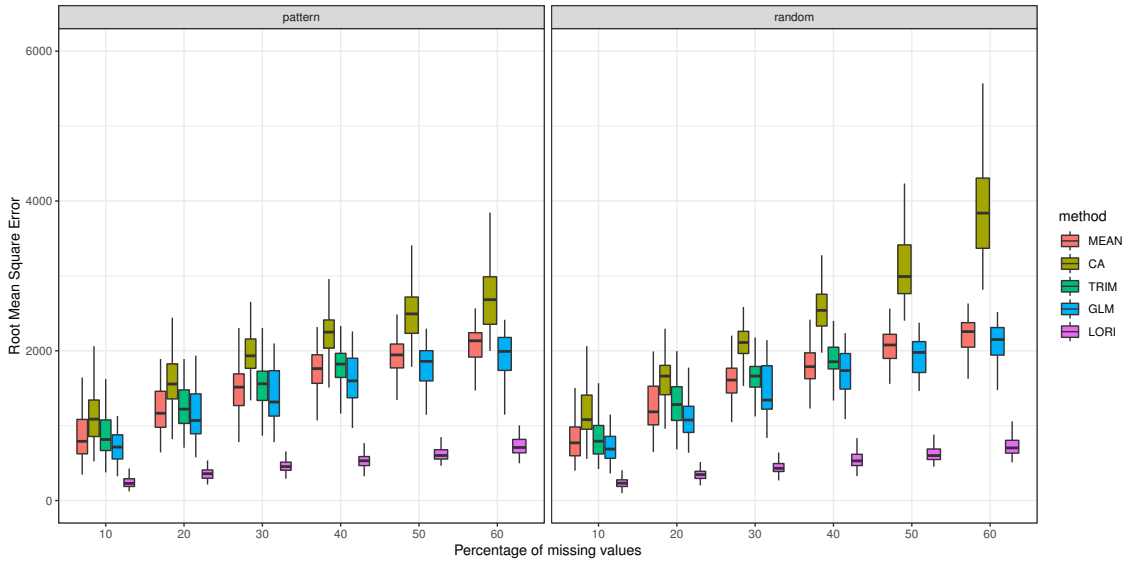


Figure 4.3: Average imputation relative RMSE (100 replications) for synthetic Poisson data, and increasing percentages of missing values (10%, 20%, 30%, 40%, 50%, 60%). Compared methods: imputation by the column mean (MEAN), Correspondence Analysis (CA), Trends and Indices in Monitoring data (TRIM), Generalized linear mixed model (GLMM), Low-rank Interactions (LORI).

of the multiple imputation method. With the same simulation setting, we estimate (from incomplete data) the column-wise sums as well as intervals of variability using the multiple imputation method described in Section 4.2.3. Repeating the experiment 100 times, we evaluate how often the true column-wise sum falls in our interval of variability. A plot of the result for one of the 100 trials is displayed in Figure 4.4, where we represent the estimated column sums, the computed intervals of variability, and the true column means. The horizontal axis corresponds to the columns of the count tables, and the point by point empirical coverage is displayed above each column-wise sum. Overall, the coverage is below 95%, and is close to 95% for only a small fraction of the columns (coverage above 80% for 7 columns out of 30). In some cases, the coverage is very poor, even though the prediction error is simultaneously very small (see columns 5 or 13 for example). This illustrates the fact that, as our imputation method is biased, the estimated intervals of variability are not valid confidence intervals, but may be interpreted as reflecting the variability of our predictions (i.e. they estimate the variance of our estimates, but our estimates are not consistent). For one of the years (column 16), corresponding to a large total count compared to the average, the total sum is way above the interval; again, this observation may result from the shrinkage induced by our regularized imputation procedure.

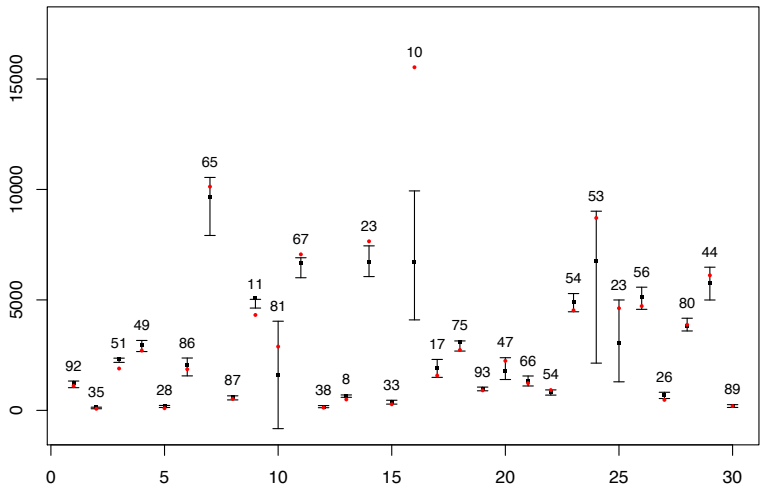


Figure 4.4: Estimated yearly abundances (black squares), intervals of variability (black segments), and true yearly abundances (red) points, for an example simulated under the LORI model with 30% of missing values. The displayed numbers correspond to the point by point empirical coverage for each interval of variability.

4.3.2 Robustness to model misspecification

In many cases in abundance data analysis, Poisson models are expected to be simplistic, as abundance data are often overdispersed and zero-inflated. We conduct a second similar experiment, this time generating synthetic data from a zero-inflated negative binomial distribution, where each entry is sampled from a negative binomial distribution with probability 0.9, and set to zero with probability 0.1. The compared methods are the same as in the previous section, except for the generalized linear mixed model. Here, we use a GLMM with a zero-inflated negative binomial distribution. In other words, here, GLMM is the only methods which imputes using the correct model.

Since all the methods (except the GLMM) are based on a Poisson model, they have worse performance in this misspecified experiment. However, we observe that LORI seems more robust to model misspecification than CA and TRIM. This difference in behavior may be explained by the fact that LORI estimates interactions, which are an alternative way of modeling overdispersion, since larger (resp. smaller) interactions yield larger (res. smaller) counts for the same set of covariate values.

4.3.3 Imputation of northern shoveler abundance data

To evaluate our method in the most realistic setting, we finally challenge it on the imputation of a subsample of the waterbirds data set. We focus on the northern shoveler (*Spatula clypeata*), one of the most abundant species-specific subsample in our data set, and which displayed the less missing values. We select 209 sites where the shoveler was counted in more than 13 years among the 28 years in total. In the end, we obtain a count table of size 209×28, and containing around 30% of missing values. In addition, we also have access to side information about the sites (altitude, longitude, water surface, etc.), the years (temperature anomalies and rainfall, etc.) and the site-year pairs (yearly economical index by country, etc.). In this experiment, we remove an increasing amount of entries in the count table, and seek to impute them back, using the available side information. We compare our method to imputation using CA and TRIM. The results

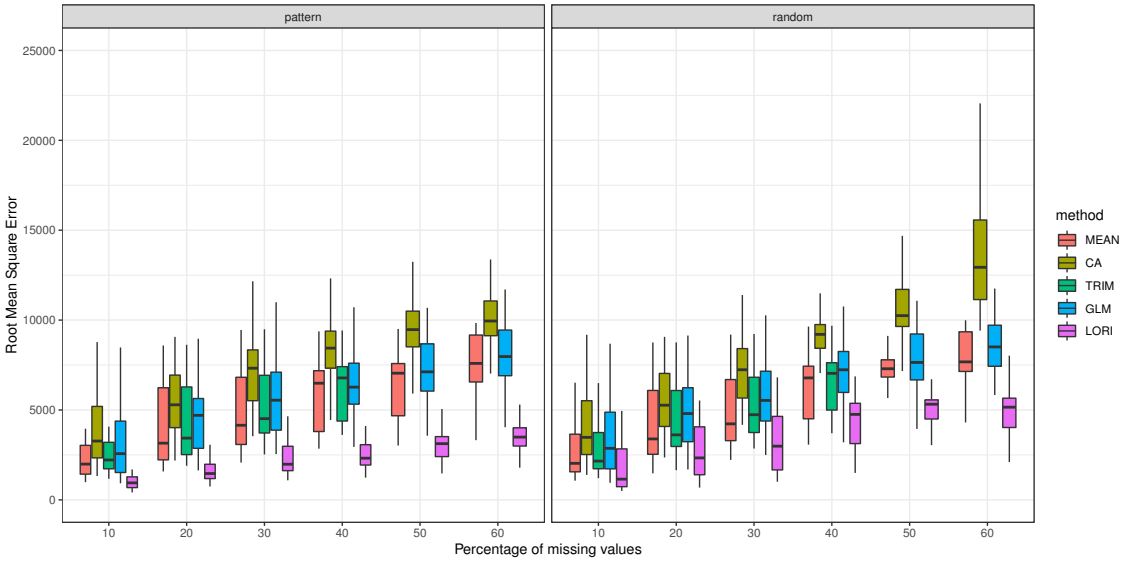


Figure 4.5: Average imputation relative RMSE (100 replications) for synthetic negative binomial data, and increasing percentages of missing values (10%, 20%, 30%, 40%, 50%, 60%). Compared methods: imputation by the column mean (MEAN), Correspondence Analysis (CA), Trends and Indices in Monitoring data (TRIM), Generalized linear mixed model (GLMM), Low-rank Interactions (LORI).

are displayed in Figure 4.6, where the percentage of missing values indicated was added to the already 30% of missing values. For example, in the first experiment with 15% of additional missing values, the count table has in total 40% of missing values, while in the last experiment where 30% is indicated, the table has 50% of missing values.

We observe that, compared to the experiments on synthetic data with a Poisson model, all the methods are closer to one another. However, LORI has smaller prediction errors, and the improvement between LORI and two state-of-the-art methods (CA and TRIM) is of the same order of magnitude as between these two methods and imputation by the column means. Furthermore, LORI is more robust to large proportions of missing values, with imputation errors stable in average and variability for increasing proportions, contrary to CA and TRIM. Finally, in the last setting with 30% of (additional) missing entries, all the methods have a large variability. Note that, in addition, TRIM does not impute all the missing entries: some entries corresponding to sites with large proportions of missing entries are automatically removed. Therefore, average RMSE and its variability are biased low for TRIM and thus not fully comparable to other methods. Because of the covariates included in the LORI model, we are able to impute all entries. The results of Figure 4.6, which are more similar to the results of the experiment on zero-inflated and overdispersed data than to those on Poisson data, probably indicate that there is room for improvement through more complex models which account for zero excess and overdispersion (known to occur in bird abundance data). We discuss this point in more details in Section 4.5, and leave such extensions to future work.

4.4 Estimation of waterbirds population trends

We finish by applying the LORI multiple imputation procedure to analyze the abundance of three waterbirds species, and produce estimates and intervals of variability of

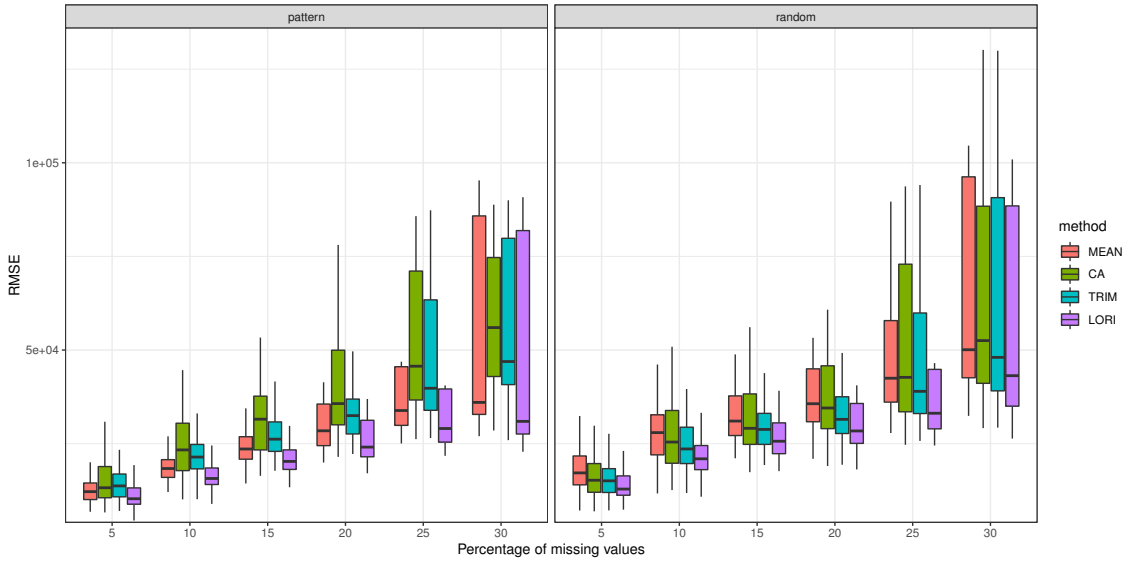


Figure 4.6: Average imputation RMSE (100 replications) for the northern shoveler data, and increasing percentages of missing values (5%, 10%, 15%, 20%, 25%, 30%): this proportion is added to the 30% of missing values originally present in the data set. Compared methods: imputation by the column mean (MEAN), Correspondence Analysis (CA), Trends and Indices in Monitoring data (TRIM), Low-rank Interactions (LORI).

Figure 4.7: northern shoveler (left), and common pochard (middle) and eurasian coot (right).

the total yearly abundances. Abundance data for these species were collected in Algeria, Egypt, Libya, Morocco and Tunisia between 1990 and 2017, by ornithologists and volunteers from different organizations and institutes including the Mediterranean Waterbirds Network (MNW), Groupe de Recherche pour la Protection des Oiseaux au Maroc/BirdLife Morocco, Direction Générale des Forêts (Algeria), Association "les Amis des Oiseaux"/BirdLife (Tunisia), Libyan Society for Birds, Egyptian Environment Affairs Agency, Office National de la Chasse et de la Faune Sauvage (ONCFS, France), and the Tour du Valat Institute (France). In total, 785 sites were visited. For every site and every year, we have access to a set of covariates, of which an excerpt is given in Table 4.1. In this table, we center and scale the quantitative covariates; indeed, this is a necessary step as we use the same regularization parameter for all the covariates. The side information includes categorical variables such as the country, that we code as dummy indicator variables (with 4 dummy variables to code the 5 countries). The other variables are quantitative, and we center and scale them before imputation. From one species to the next, the set of sites is different, because not all species are present in every site. However, the set of years, and the side information as well, remain constant.

In the next sections, we analyze the abundance data sets of the northern shoveler, common pochard, and Eurasian coot.

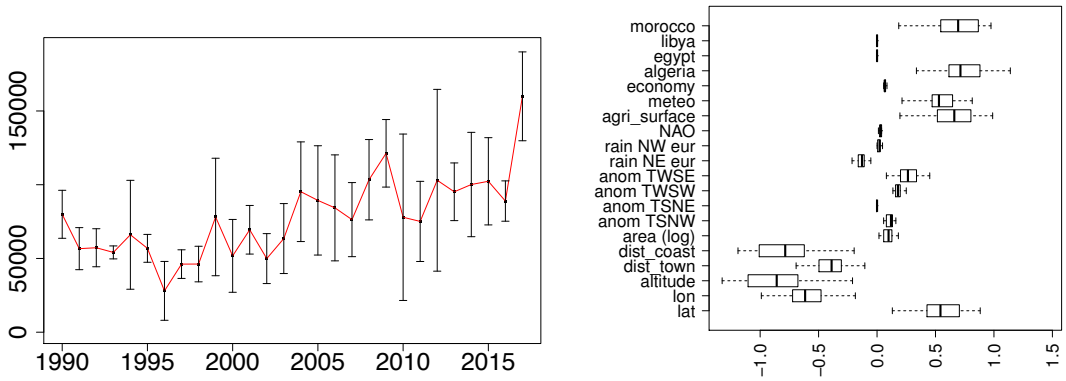
Site	1	2	3	4	5	6
Year	1990	1990	1990	1990	1990	1990
algeria	1	1	1	1	1	1
egypt	0	0	0	0	0	0
libya	0	0	0	0	0	0
morocco	0	0	0	0	0	0
latitude	0.88	-0.69	0.56	0.10	0.36	0.74
longitude	0.15	0.50	0.37	0.66	0.01	-0.06
altitude	-0.66	-0.60	1.14	1.33	0.67	-0.48
area (log)	-1.27	-0.58	0.99	0.55	1.17	0.54
dam	0	0	1	1	1	1
NAO	-0.20	-0.20	-0.20	-0.20	-0.20	-0.20
rainfall	-1.91	-0.20	-1.55	0.90	-0.19	-1.20
economy	-0.71	-0.71	-0.71	-0.71	-0.71	-0.71
agriculture surface	0.98	0.98	0.98	0.98	0.98	0.98

Table 4.1: Excerpt of the side information about the sites and years of the waterbird abundance data (after centering and scaling the quantitative variables).

4.4.1 Northern shoveler

The northern shoveler (*Spatula clypeata*), is a common duck species, breeding in the northern areas of Eurasia, and across North America; part of the European population of the northern shoveler winters in Africa. In this section, we analyze the whole northern shoveler abundance data (contrary to the previous section where we selected a subsample with less missing values). In total, we have access to the abundance of the shoveler across 513 sites, between 1990 and 2017, and 60% of the entries are missing. We apply our multiple imputation procedure to compute yearly total abundances, and display the estimated yearly totals, as well as the corresponding intervals of variability on Figure 4.8a. Based on our results, the total population of the northern shoveler in North Africa is increasing.

We can also look at the estimated covariate coefficients in the Poisson log-linear model across the multiple imputations, as displayed them in a boxplot in Figure 4.8b. Indeed, although we advertised LORI as a method to impute count data using side information, it also has the advantage of estimating covariate coefficients, which is useful for interpretation purposes independently of the imputation. We observe a clear country effect: Algeria and Morocco have positive effects, everything else being equal, compared to Egypt, Libya and Tunisia (which is the reference category). Overall, covariates describing the sites (latitude, longitude, altitude, distance to towns and coast, and area) have larger effects than the covariates describing the years, such as the rain index in North-East Europe (rain NE eur) or the winter temperature anomaly in South-East Europe (anom TWSE). This is expected as there is a large variability in the counts from one sites to the other, but most sites are relatively stable across years, in comparison to the variability across sites. Some of the year covariates have mostly zero effect across the imputation models, such as the North Atlantic Oscillations (NAO), a meteorological index, and the spring temperature anomaly in North-East Europe (anom TSNE). Among the covariates which describe both the sites and the years, the economical index has a small but positive effect across all imputation models, indicating that favorable economical conditions have a positive impact on the bird counts.



(a) Estimation of yearly abundances of the northern shoveler (red line) and intervals of variability (black segments).

(b) Boxplot of estimated effects of covariates on the northern shoveler counts ($\hat{\epsilon}$) across multiple imputations.

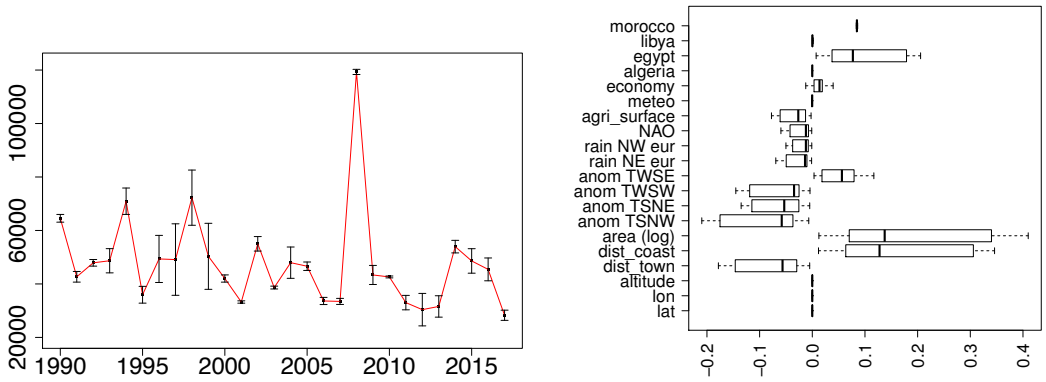
Figure 4.8: Results of yearly totals of multiple imputation for the northern shoveler data

Note that we may not interpret the estimated main effects in a straightforward manner. Indeed, first, the covariates are correlated. For instance, the covariates related to meteorological anomalies. In addition, some of the covariates (related to rain for example), are constant across sites, meaning that they are correlated to the year effects. Secondly, main effects and interactions may superimpose, as we do not enforce orthogonality between the two. However, for imputation purposes, the relevant parameter is the sum of all main effects and interactions, and thus this identifiability issue is not really a problem. In practice, we also observed that the estimated interaction matrix was approximately orthogonal to the main effects (up to residual means much smaller than the main effects), even without forcing identifiability.

4.4.2 Common pochard

The common pochard (*Aythya ferina*), is a diving duck breeding in wetlands across Europe and Asia, who migrates to south Europe and Africa to spend the winter there. In our data set, the common pochard was observed in 338 sites, between 1990 and 2017, and 55% of the entries are missing. Estimated yearly totals, as well as the corresponding intervals of variability on Figure 4.9a, where the North African common pochard population seems to decrease slowly, with an exception in 2008 where an excess of pochards were observed. This observation may correspond to a particular year where more pochards migrated from Europe, for some unknown reason.

As in the previous section, we can also look at the estimated covariate coefficients in the Poisson log-linear model across the multiple imputations. These results are displayed in a boxplot in Figure 4.9b. Here again, we observe a country effect: Egypt and Morocco have large positive effects on the pochard counts. The area also have a positive effect on the counts, which is expected. Most indicators of temperature anomalies in Europe (anom TWSW, anom TWSW, anom TSNW) are associated to smaller counts. Indeed, if the temperature is higher than expected in Europe, the ducks may be less encouraged to migrate to Africa.



(a) Estimation of yearly abundances of the common pochard (red line) and intervals of variability (black segments).

(b) Boxplot of estimated effects of covariates on the common pochard counts ($\hat{\epsilon}$) across multiple imputations.

Figure 4.9: Results of multiple imputation for the common pochard data

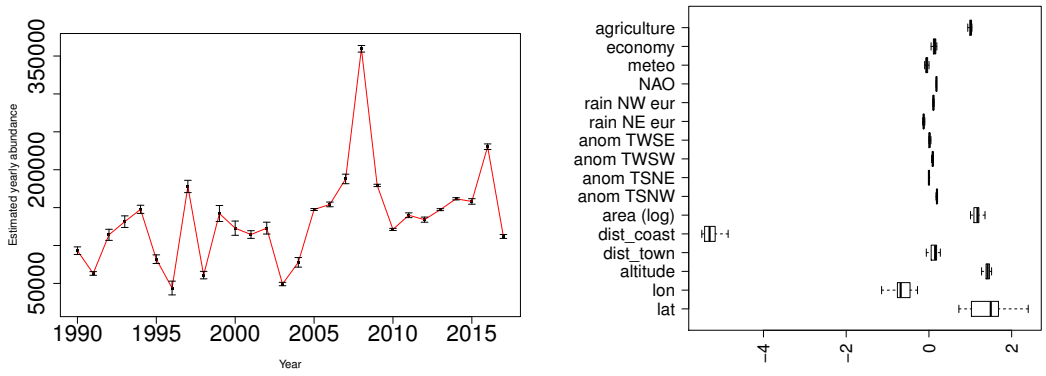
4.4.3 Eurasian coot

The Eurasian coot (*Fulica atra*), breeds in most parts of Europe and Asia, and is found in freshwater lakes and ponds. In winter, it migrates to North Africa and South-East Asia. In our data set, the coot was observed in 498 sites, between 1990 and 2017, and 60% of the entries are missing. Estimated yearly totals, as well as the corresponding intervals of variability on Figure 4.10a, where the North African coot population seems to increase slowly. The same peak is observed in 2008, as for the common pochard. This backs up the hypothesis that a particular condition in 2008 pushed more waterbirds to migrate to North Africa. The estimated covariate coefficients in the Poisson log-linear model across the multiple imputations are displayed in a boxplot in Figure 4.10b. The distance to coast has the largest estimated effect on the coot counts, with a negative effect indicating that sites located near the sea tend to have larger bird counts. On the other hand, the distance to town has a positive effect, meaning that sites located close to towns tend to have smaller bird counts. As for the two previous species, the area of the site also has a positive effect on the counts. Finally the agricultural surface is associated to larger counts, which may indicate that wetlands created by or close to agricultural areas became more favourable because eutrophication increased foraging opportunities on aquatic vegetation or water levels possibly became higher and more regular for irrigation purposes.

4.5 Conclusion

In this chapter, we introduced a new multiple imputation procedure for count data with supplementary covariates. The method also has the advantage of estimating the covariate coefficients in a Poisson log-linear model, and of selecting important covariates. We evaluated the method on synthetic data generated with different models for count data, and on a waterbird abundance data set. We also illustrated the method with the trend analysis of abundance data for three species of waterbirds, and released an open source library where the method is available on the CRAN (R package `lori`).

This work paves the way to several directions of future research. To begin with,



(a) Estimation of yearly abundances of the Eurasian coot (red line) and intervals of variability (black segments).

(b) Boxplot of estimated effects of covariates on the Eurasian coot counts ($\hat{\epsilon}$) across multiple imputations.

Figure 4.10: Results of multiple imputation for the Eurasian coot data

our experiments revealed that improvement may be obtained by extending the model to richer frameworks accounting for zero-inflation and over-dispersion. A first step could be to incorporate a scale parameter in the model, to estimate the variance and the mean of each entry with a quasi-Poisson model. Another important extension would be the derivation of valid confidence intervals. A possible direction would be to extend post-selection inference–which allows to compute valid intervals after ℓ_1 penalized variable selection in generalized linear models (Taylor and Tibshirani, 2018)–to our setting where an additional low-rank term is involved.

Chapter 5

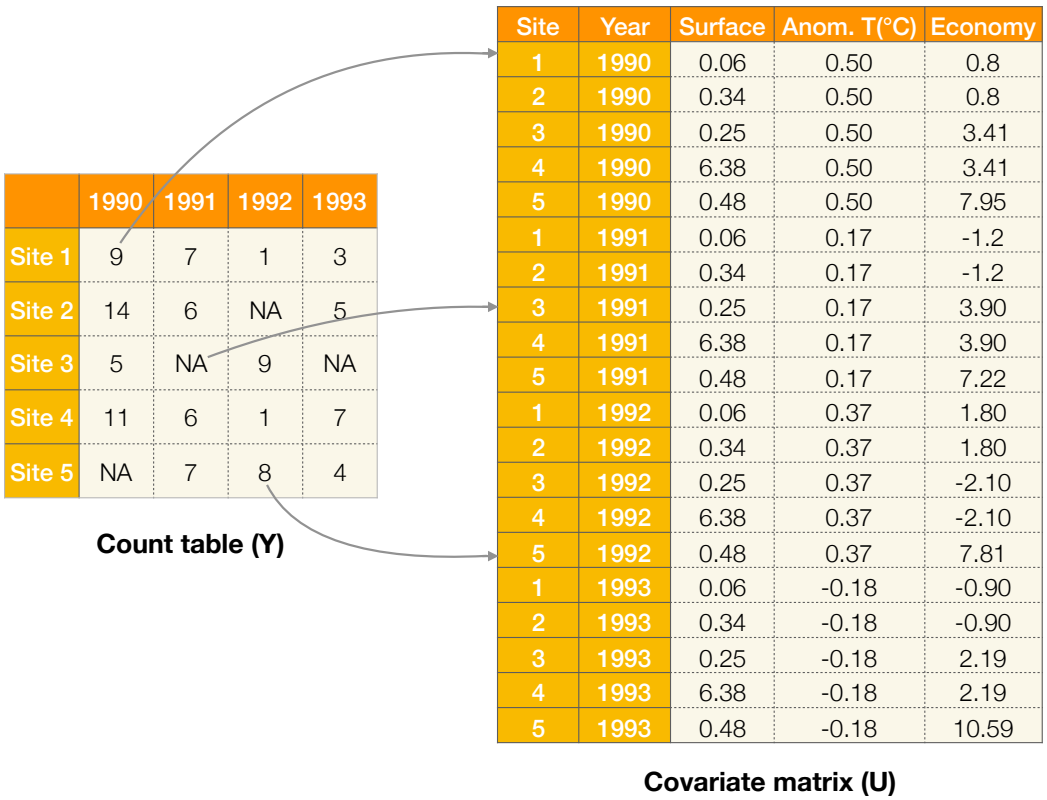
Tutorial: R package `lori`

Contents

The R package `lori` (https://CRAN.R-project.org/package=lori) contains methods for the analysis, single imputation, and multiple imputation of incomplete count data with side information. In this chapter, we display the main functionalities of the `lori` package, namely a single imputation procedure–which also returns estimates of main effects and interactions–a cross-validation function to select the regularization parameters, and a function to perform multiple imputation. We also provide reusable code using a simulated toy example and a public use data set from the `ade4` package. Finally we provide technical details about the implementation in the last section, and describe the two implemented optimization algorithms. All the code reported here is available as R files at `https://github.com/genevievelrobin/Lori-tutorial`. The implemented functions correspond to the methods described in Chapters 3 and 4, but this tutorial is also meant to be read and used independently. Thus, the models and procedures are recalled, and some of the figures from the previous chapters are replicated here: they will be indicated so that the reader may skip them.

5.1 Simulated example

We start with a synthetic example with a data set generated under the lori model. The goal is to describe and provide intuitions about the model, and to display the important functions of the package. Thus, we will start by generating some synthetic data, and add missing values. Then, we will estimate the parameters of the model using the `lori` function, and impute them. Before going on with the tutorial, here are the necessary



	1990	1991	1992	1993
Site 1	9	7	1	3
Site 2	14	6	NA	5
Site 3	5	NA	9	NA
Site 4	11	6	1	7
Site 5	NA	7	8	4

Count table (Y)

Site	Year	Surface	Anom. T(°C)	Economy
1	1990	0.06	0.50	0.8
2	1990	0.34	0.50	0.8
3	1990	0.25	0.50	3.41
4	1990	6.38	0.50	3.41
5	1990	0.48	0.50	7.95
1	1991	0.06	0.17	-1.2
2	1991	0.34	0.17	-1.2
3	1991	0.25	0.17	3.90
4	1991	6.38	0.17	3.90
5	1991	0.48	0.17	7.22
1	1992	0.06	0.37	1.80
2	1992	0.34	0.37	1.80
3	1992	0.25	0.37	-2.10
4	1992	6.38	0.37	-2.10
5	1992	0.48	0.37	7.81
1	1993	0.06	-0.18	-0.90
2	1993	0.34	-0.18	-0.90
3	1993	0.25	-0.18	2.19
4	1993	6.38	-0.18	2.19
5	1993	0.48	-0.18	10.59

Covariate matrix (U)

Figure 5.1: Incomplete count table and covariate matrix.

commands to install and load the package `lori`:

```
install.packages("lori")
library(lori)
```

To obtain the exact results that are presented here, you may set the following seed:

```
set.seed(123)
```

The lori model (already described in Chapter 4) Consider an abundance table $\boldsymbol{Y} \in \mathbb{N}^{m_1 \times m_2}$, with missing values. Assume that covariates about the row, columns, and row-column pairs of $\boldsymbol{Y}$ are available in a supplementary data matrix $\boldsymbol{U} \in \mathbb{R}^{m_1 \times m_2}$, as displayed in Figure 5.1. The columns of $\boldsymbol{U}$ correspond to variables describing either the rows, the columns, or the row-column pairs of $\boldsymbol{Y}$. For instance, in Figure 4.1, the first column of $\boldsymbol{U}$ indicates the surface of the sites (in km^2), a variable which depends only on the site index. Thus, the first column takes the same value in the rows that correspond to the same site (e.g. rows 1, 6, 11, 16). On the other hand, the second column of $\boldsymbol{U}$ indicates a global temperature anomaly, which only depends on the years. Thus, the second column takes the same value in the rows that correspond to the same year (e.g. rows 1-5, 6-10, etc.). Finally, the third column indicates a regional economical index, which depends on the location of the site *and* on the year. Note that the covariate vector associated to the (i, j)-th entry $\boldsymbol{Y}_{i,j}$ is the $(j-1)m_1 + i$-th row of $\boldsymbol{U}$, $\boldsymbol{U}_{(j-1)m_1+i,\cdot} \in \mathbb{R}^q$. For simplicity, we denote this vector

$$\boldsymbol{U}^{i,j} := \boldsymbol{U}_{(j-1)m_1+i,\cdot} \tag{5.1}$$

We also denote by $\boldsymbol{U}_k^{i,j}$ the k-th entry of $\boldsymbol{U}^{i,j}$. The model implemented in the `lori` package is the following log-linear model:

$$\boldsymbol{Y}_{i,j} \sim \mathcal{P}(\exp(\boldsymbol{X}_{i,j}^0)), \quad \boldsymbol{X}_{i,j}^0 = \alpha_i^0 + \beta_j^0 + \sum_{k=1}^{K_1+K_2} \boldsymbol{U}_k^{i,j} \epsilon_k^0 + \boldsymbol{\Theta}_{i,j}^0, \tag{5.2}$$

where $\mathcal{P}(\lambda)$ denotes the Poisson distribution of intensity λ. In (5.2), α_i^0 is a row effect, β_j^0 a column effect, ϵ_k^0 is the effect of the k-th covariate, and $\boldsymbol{\Theta}_{i,j}^0$ an interaction term between row i and column j. The model is saturated, and we make two main assumptions to constrain the parameter space. First, we assume that the vectors α^0, β^0 and ϵ^0 are sparse (contain zero values), meaning that not all sites, years or covariates have an effect on the counts. Second, we assume the interaction matrix $\boldsymbol{\Theta}^0$ has low-rank, meaning that the row-column interactions may be summarized by multiplicative interactions between a few latent row and column factors. To impose such structure to the parameters, we use an estimation procedure which includes sparsity and low-rank inducing regularization terms, as described in Section 4.2.2.

Remarks Note that, in the package, options are implemented to remove row and column effects: this will be detailed later on. The package is also designed to accomodate the case where several covariate tables are available separately. For instance, often, three supplementary tables are available: $\boldsymbol{R} \in \mathbb{R}^{m_1 \times K_1}$, containing covariates about the rows of $\boldsymbol{Y}$ (geographical information for instance), $\boldsymbol{C} \in \mathbb{R}^{m_2 \times K_2}$, containing covariates about the columns of $\boldsymbol{Y}$ (yearly meteorological indices), and $\boldsymbol{E} \in \mathbb{R}^{m_1 m_2 \times K_3}$ containing covariates about the row-column pairs (yearly meteorological information at the sites' scale, yearly economical indices of the sites' country, etc.). For $i \in [\![m_1]\!]$, the vector $\boldsymbol{R}_{i,\cdot} = (\boldsymbol{R}_{i,1}, \ldots, \boldsymbol{R}_{i,K_1})$ contains all the information about the row i. Similarly, for $j \in [\![m_2]\!]$, the vector $\boldsymbol{C}_{j,\cdot} = (\boldsymbol{C}_{i,1}, \ldots, \boldsymbol{C}_{j,K_1})$ contains all the information about the column j. Finally, the vector $\boldsymbol{E}_{(j-1)m_1+i,\cdot} = (\boldsymbol{E}_{(j-1)m_1+i,1}, \ldots, \boldsymbol{E}_{(j-1)m_1+i,K_3})$ contains all the information about the row-column pair (i, j). The package lori contains a function to construct the table $\boldsymbol{U}$ containing *all* these covariates, from $\boldsymbol{R}$, $\boldsymbol{C}$ and $\boldsymbol{E}$ (or any singleton or pair of these three matrices).

5.1.1 Generating simulated data

We now generate covariate tables and incomplete count data, using model (5.2). Consider the following simulated example.

```
## covariates
m1 <- 30 # number of rows
m2 <- 10 # number of columns
K1 <- 2 # number of row covariates
K2 <- 2 # number of column covariates
K3 <- 3 # number of (rowxcolumn) covariates
q <- K1+K2+K3
R <- matrix(rnorm(m1*K1), nrow=m1) # matrix of row covariates
C <- matrix(rnorm(m2*K2), nrow=m2) # matrix of column
   covariates
E <- matrix(rnorm(m1*m2*K3), nrow=m1*m2) # matrix of  (
   rowxcolumn) covariates
```

To compute the matrix $\boldsymbol{U}$, a simple command is available in the `lori` package:

```
U <- covmat(m1, m2, R, C, E)
```

Note that the function also supports creating $\boldsymbol{U}$ from any singleton or pair of the matrices $(\boldsymbol{R}, \boldsymbol{C}, \boldsymbol{E})$, provided that the arguments are in the correct order. In other words, all the commands below may also be used:

```
U <- covmat(m1, m2, R)
U <- covmat(m1, m2, C)
U <- covmat(m1, m2, R, C)
U <- covmat(m1, m2, R=R, E=E)
U <- covmat(m1, m2, C=C, E=E)
```

An important practical point is that the rows of $\boldsymbol{R}$, $\boldsymbol{C}$ and $\boldsymbol{E}$ should be sorted in the correct order. If $\boldsymbol{R}$ and $\boldsymbol{C}$ are sorted as displayed in Tables 5.1 and 5.2, then, $\boldsymbol{E}$ should be sorted as displayed in Table 5.3, with the row indices varying *inside* the colum indices.

Row ID	Covariate 1	Covariate 2
Row 1	-0.6	0.4
Row 2	-0.2	-0.3
Row 3	1.6	0.9

Table 5.1: First three rows of table $\boldsymbol{R}$

Column ID	Covariate 1	Covariate 2
Column 1	0.4	-0.5
Column 2	-0.5	-2.3
Column 3	-0.3	1.0

Table 5.2: First three rows of table $\boldsymbol{C}$

Row ID	Column ID	Covariate 1	Covariate 2	Covariate 3
Row 1	Column 1	0.0	0.0	0.4
Row 2	Column 1	0.4	0.2	1.0
Row 3	Column 1	-0.4	0.2	-0.7
⋮	⋮	⋮	⋮	
Row 1	Column 2	-0.6	0.5	1.1
Row 2	Column 2	0.6	0.3	-0.2
Row 3	Column 2	-1.6	0.7	-0.1
⋮	⋮	⋮	⋮	

Table 5.3: Covariate matrix $\boldsymbol{E}$

Now that we have our covariate matrices in the correct ordering, let us construct artificial row, column, and covariate effects, α^0, β^0 and ϵ^0, and artificial interactions $\boldsymbol{\Theta}^0$.

```
## parameters
alpha0 <- rep(0, m1);alpha0[1:6] <- 1
beta0 <- rep(0, m2);beta0[1:4] <- 1
epsilon0 <- rep(0, q);epsilon0[5:6] <- 0.2
r <- 2 #rank of interaction matrix theta0
theta0 <- 0.1*matrix(rnorm(m1*r), nrow=m1)%*%diag(r:1)%*%
   matrix(rnorm(m2*r), nrow=r)
theta0 <- sweep(theta0, 2, colMeans(theta0))
theta0 <- sweep(theta0, 1, rowMeans(theta0))
```

Finally, we can construct the parameter matrix $\boldsymbol{X}^0$ and sample our incomplete count data $\boldsymbol{Y}$, after centering and scaling the covariate matrix.

```
## construct x0
U <- scale(U)#center and normalize the covariates
x0 <- matrix(rep(alpha0,m2),nrow=m1)#row effects
x0 <- x0 + matrix(rep(beta0,each=m1),nrow=m1)#add col effects
x0 <- x0 + matrix(U%*%epsilon0,nrow=m1) #add cov effects
x0 <- x0 + theta0 #add interactions
```

```
## sample count data y
y0 <- matrix(rpois(m1*m2, lambda = c(exp(x0))), nrow = m1)

## add missing values
p <- 0.2
y <- y0
y[sample(1:(m1*m2), round(p*m1*m2))] <- NA
```

5.1.2 Estimation and single imputation

Model and estimation The purpose of the `lori` package is dual: to estimate the parameters α^0, β^0, ϵ^0, $\boldsymbol{\Theta}^0$, and to impute the count table $\boldsymbol{Y}$. The corresponding input are the incomplete count data $\boldsymbol{Y}$ and (optionally) the covariate matrix $\boldsymbol{U}$. The estimation procedure is:

$$(\hat{\alpha}, \hat{\beta}, \hat{\epsilon}, \hat{\boldsymbol{\Theta}}) \in \operatorname{argmin} \mathcal{L}(\boldsymbol{Y}; \alpha, \beta, \epsilon, \boldsymbol{\Theta}) + \lambda_1 \|\boldsymbol{\Theta}\|_* + \lambda_2 \left(\|\alpha\|_1 + \|\beta\|_1 + \|\epsilon\|_1\right), \quad (5.3)$$

where $\mathcal{L}(\boldsymbol{Y}; \alpha, \beta, \epsilon, \boldsymbol{\Theta})$ is the Poisson negative log-likelihood:

$$\sum_{(i,j)} \Omega_{i,j}[-\boldsymbol{Y}_{i,j}(\alpha_i + \beta_j + \sum_{k=1}^{K} \epsilon_k \boldsymbol{U}(i,j)_k + \boldsymbol{\Theta}_{i,j}) + \exp(\alpha_i + \beta_j + \sum_{k=1}^{K} \epsilon_k \boldsymbol{U}(i,j)_k + \boldsymbol{\Theta}_{i,j})],$$

and $\Omega_{i,j} = 1$ if $\boldsymbol{Y}_{i,j}$ is observed, and 0 otherwise. The intuition behind (5.3) is that the negative log-likelihood is minimized (the parameters should fit the data as much as possible) with additional regularization terms which constrain the parameters and perform model selection automatically. The first penalty term, $\|\boldsymbol{\Theta}\|_*$ corresponds to the nuclear norm of $\boldsymbol{\Theta}$ and induces low-rank solutions: this is interpreted as assuming that a few latent factors summarize the interactions. The second term, $\|\alpha\|_1 + \|\beta\|_1 + \|\epsilon\|_1$, corresponds to the sum of the ℓ_1 norms of the vectors α, β and ϵ:

$$\|\alpha\|_1 + \|\beta\|_1 + \|\epsilon\|_1 = \sum_{i=1}^{m_1} |\alpha_i| + \sum_{j=1}^{m_2} |\beta_j| + \sum_{k=1}^{K} |\epsilon_k|.$$

This penalty induces sparse solutions (vectors α, β and ϵ containing many zeros), meaning that not all rows, columns and covariates have an effect on the counts. The trade-off between the data-fitting term and the penalties is controlled by the regularization parameters λ_1 and λ_2. As λ_1 increases, the rank of the solution $\hat{\boldsymbol{\Theta}}$ decreases (the number of latent factors decreases). As λ_2 increases, the number of nonzero values in $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\epsilon}$ decreases (the number of active rows, columns and covariates decreases).

In R, the estimation is done as follows, for some predefined regularization parameters λ_1 and λ_2:

```
## lori estimation
lambda1 <- 0.1
lambda2 <- 0.1
res.lori <- lori(y, U, lambda1, lambda2)
```

Cross-validation and estimation However, the quality of the estimation is highly dependent on the choice of λ_1 and λ_2, and thus we provide a cross-validation function to select them automatically:

```
## cross-validation
res.cv <- cv.lori(y, U, trace.it = T) #this takes a few
   minutes, the trace.it argument states that information
   about the progress should be printed
## estimation
res.lori <- lori(y, U, res.cv$lambda1, res.cv$lambda2)
```

```
res.lori$alpha
res.lori$beta
res.lori$espilon
res.lori$theta
```

In the present case, we obtain point estimates for the main effects (with a one-digit precision), displayed in Table 5.4 for $\hat{\alpha}$, Table 5.5 for $\hat{\beta}$ and Table 5.6 for $\hat{\epsilon}$.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1.0	1.0	0.1	0.6	0.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 5.4: First 20 coefficients in $\hat{\alpha}$ (res.lori$alpha[1:20]). The cells colored in green correspond to the nonzero coefficients in α^0 ($\alpha_1^0 = \ldots = \alpha_6^0 = 1$).

1	2	3	4	5	6	7	8	9	10
1.2	0.7	0.9	0.6	0.0	0.0	0.0	0.0	0.0	0.2

Table 5.5: $\hat{\beta}$ (res.lori$beta). The cells colored in green correspond to the nonzero coefficients in β^0 ($\beta_1^0 = \ldots = \beta_4^0 = 1$).

1	2	3	4	5	6	7
0.1	0.1	-0.1	0.0	0.3	0.2	0.0

Table 5.6: $\hat{\epsilon}$ (res.lori$epsilon). The cells colored in green correspond to the nonzero coefficients in ϵ^0 ($\epsilon_5^0 = \epsilon_6^0 = 0.2$).

Single imputation The function `lori` also returns imputed values for the missing entries in $\boldsymbol{Y}$, based on these point estimates. For each missing entry, the predicted value is given by

$$\hat{\boldsymbol{Y}}_{i,j} = \exp\left(\hat{\alpha}_i + \hat{\beta}_j + \sum_{k=1}^{K_1+K_2} \boldsymbol{U}_{(j-1)m_1+i,k}\hat{\epsilon}_k + \hat{\boldsymbol{\Theta}}_{i,j}\right). \tag{5.4}$$

The imputed data set $\hat{\boldsymbol{Y}}$, such that $\hat{\boldsymbol{Y}}_{i,j} = \boldsymbol{Y}_{i,j}$ if $\boldsymbol{Y}_{i,j}$ is observed, and $\hat{\boldsymbol{Y}}_{i,j}$ given by (5.4) otherwise, is accessed as follows:

```
res.lori$imputed
```

	1	2	3	4	5	6	7	8	9	10
1	9	7	1	3	12	8	3	5	1	6
2	14	6	6.6	5	4	5	2.2	1	3	6
3	5	3.2	9	3.7	3	5	3	1	2	1
4	11	6	1	7	3.7	1.8	5	4	4	2
5	8.8	7	8	4	2	3	7	10	2	2.9
6	17	4.2	8	4.1	1	0	3	4	5	4

Table 5.7: First six rows of the imputed data set $\hat{\boldsymbol{Y}}$ (res.lori$imputed[1:5,]). The red cells correspond to imputed values, originally missing in $\boldsymbol{Y}$.

Removing row and column effects Note that, by default, the row and column effects α^0 and β^0 are estimated. To estimate model (5.2) without row and column effects, one can force them to zero by using the arguments `reff` (boolean indicating whether row effects should be fit) and `ceff` (boolean indicating whether column effects should be fit) in `lori`, which are set to TRUE by default:

```
res.noreff <- lori(y, U, lambda1, lambda2, reff=F)
res.noceff <- lori(y, U, lambda1, lambda2, ceff=F)
res.noreffceff <- lori(y, U, lambda1, lambda2, reff=F, ceff=F)
```

5.1.3 Multiple imputation

The function `mi.lori` performs multiple imputation, based on the single imputation procedure described above. To produce M imputed data sets, a two-step bootstrap procedure is applied.

Resampling procedure (already in Chapter 4) We define a resampling procedure to model the uncertainty of the single imputation model (4.8), which corresponds to uncertainty about the parameters $\hat{\alpha}, \hat{\beta}, \hat{\epsilon}$ and $\hat{\boldsymbol{\Theta}}$. To do so, we interpret the count matrix as a contingency table, and perform nonparametric bootstrap by resampling the counts with a multinomial model. We "de-aggregate" the count table $\boldsymbol{Y}$:

$$\boldsymbol{Y} = \begin{array}{c|cccc} & \textbf{Col 1} & \textbf{Col 2} & \textbf{Col 3} & \textbf{Col 4} \\ \hline \textbf{Row 1} & 9 & 7 & 1 & 3 \\ \textbf{Row 2} & 14 & 6 & NA & 5 \\ \textbf{Row 3} & 5 & NA & 9 & NA \\ \textbf{Row 4} & 11 & 6 & 1 & 7 \\ \textbf{Row 5} & NA & 7 & 8 & 4 \end{array} \iff \boldsymbol{Z} = \begin{array}{c|c} \textbf{Row ID} & \textbf{Col ID} \\ \hline \text{Row } 1 & \text{Col } 1 \\ \text{Row } 2 & \text{Col } 1 \\ \vdots & \vdots \\ \text{Row } 5 & \text{Col } 1 \\ \text{Row } 1 & \text{Col } 2 \\ \vdots & \vdots \\ \text{Row } 5 & \text{Col } 1 \\ \vdots & \vdots \end{array}$$

In the de-aggregated table $\boldsymbol{Z}$, for all $(i, j) \in [\![m_1]\!] \times [\![m_2]\!]$, the row (Row i, Col j) is repeated $\boldsymbol{Y}_{i,j}$ times if $\boldsymbol{Y}_{i,j}$ is observed. If $\boldsymbol{Y}_{i,j}$ is not observed, the row (Row i, Col j) does not appear at all. We assume that the rows of the de-aggregated table $\boldsymbol{Z}$ are i.i.d. In spatio-temporal species monitoring, where rows correspond to sites, columns to years, and entry to the number of individuals counted in each site at each time point,

this amounts to assuming that, each individual is observed in site i during the year j independently of the other individuals, and with the same probability $\pi_{i,j}$. Let n be the number of rows in $\boldsymbol{Z}$, i.e. the total number of individuals counted in $\boldsymbol{Y}$, and let M be a predefined integer number. We perform nonparametric bootstrap by sampling n rows of $\boldsymbol{Z}$ with equal probability and with replacement. Furthermore, we repeat this procedure M times, thus obtaining M new tables $\tilde{\boldsymbol{Z}}^1, \ldots, \tilde{\boldsymbol{Z}}^M$. Then, each $\tilde{\boldsymbol{Z}}^m$, $m \in [\![M]\!]$ is "re-aggregated" to form a new count table $\tilde{\boldsymbol{Y}}^m$. Finally we obtain M count tables $\tilde{\boldsymbol{Y}}^1, \ldots, \tilde{\boldsymbol{Y}}^M$, with the same missing data pattern as the original table $\boldsymbol{Y}$, and where the observed counts are sampled from a multinomial distribution with n trials, and with vector of probabilities given by the frequencies of each entry.

For each of the M incomplete data sets, we estimate a set of parameters

$$(\hat{\alpha}^m, \hat{\beta}^m, \hat{\epsilon}^m, \hat{\boldsymbol{\Theta}}^m).$$

These parameters are used to produce M imputation models. These multiple models reflect the uncertainty about the imputation procedure (learned from incomplete data). Then, uncertainty in the missing values is estimated, for each imputation model, with a parametric bootstrap. For each model, we produce D imputed data sets, using the same stochastic imputation procedure. Finally, we obtain MD imputed data sets $(\hat{\boldsymbol{Y}}_1^1, \ldots, \hat{\boldsymbol{Y}}_D^1, \hat{\boldsymbol{Y}}_1^2, \ldots, \hat{\boldsymbol{Y}}_D^2 \ldots, \hat{\boldsymbol{Y}}_1^M, \ldots, \hat{\boldsymbol{Y}}_D^M)$. The complete multiple imputation procedure is summarized in Figure 5.2. The entire procedure goes as follows in R. The `mi.lori` function performs multiple imputations, and the `pool.lori` function aggregates the results using Rubin's rule (Rubin, 1987).

```
res.mi <- mi.lori(y, U, res.cv$lambda1, res.cv$lambda2)
res.pool <- pool.lori(res.mi)
```

The function `pool.lori` returns estimates for the mean and variance of the imputed values, and for all the parameters involved in the model ($\hat{\alpha}$, $\hat{\beta}$, $\hat{\epsilon}$, $\hat{\theta}$). Table 5.8 shows the result obtained for the imputed values. Compared to the single imputation result displayed in Table 5.7, we obtain intervals of variability for the predicted values.

	1	2	3	4	5	6	7	8	9	10
1	9	7	1	3	12	8	3	5	1	6
2	14	6	5.5(3.1)	5	4	5	1.7(1.6)	1	3	6
3	5	3.1(2.7)	9	3.1(2.1)	3	5	3	1	2	1
4	11	6	1	7	3.2(2.0)	1.4(1.7)	5	4	4	2
5	8.3(3.8)	7	8	4	2	3	7	10	2	2.2(2.1)
6	17	3.9(2.8)	8	3.4(2.4)	1	0	3	4	5	4

Table 5.8: Results for the first six rows of the imputed data set $\hat{\boldsymbol{Y}}$ after multiple imputation and pooling. The red cells correspond to imputed values, originally missing in $\boldsymbol{Y}$. The standard deviation of the multiple imputation (computed via Rubin's formula (Rubin, 1987)) is given between parenthesis.

One may also visualize the variability of the parameter estimates with boxplots. Indeed, for every single data set generated by the bootstrap procedure, denoted $\boldsymbol{Y}^1, \ldots, \boldsymbol{Y}^M$, one obtains point estimates $\hat{\alpha}^k$, $\hat{\beta}^k$, $\hat{\epsilon}^k$ and $\hat{\boldsymbol{\Theta}}^k$, $k \in [\![M]\!]$. For example for the column effects:

```
boxplot(res.mi$mi.beta, pch="", names=paste("col", 1:10))
```

The resulting boxplot is displayed in Figure 5.3.

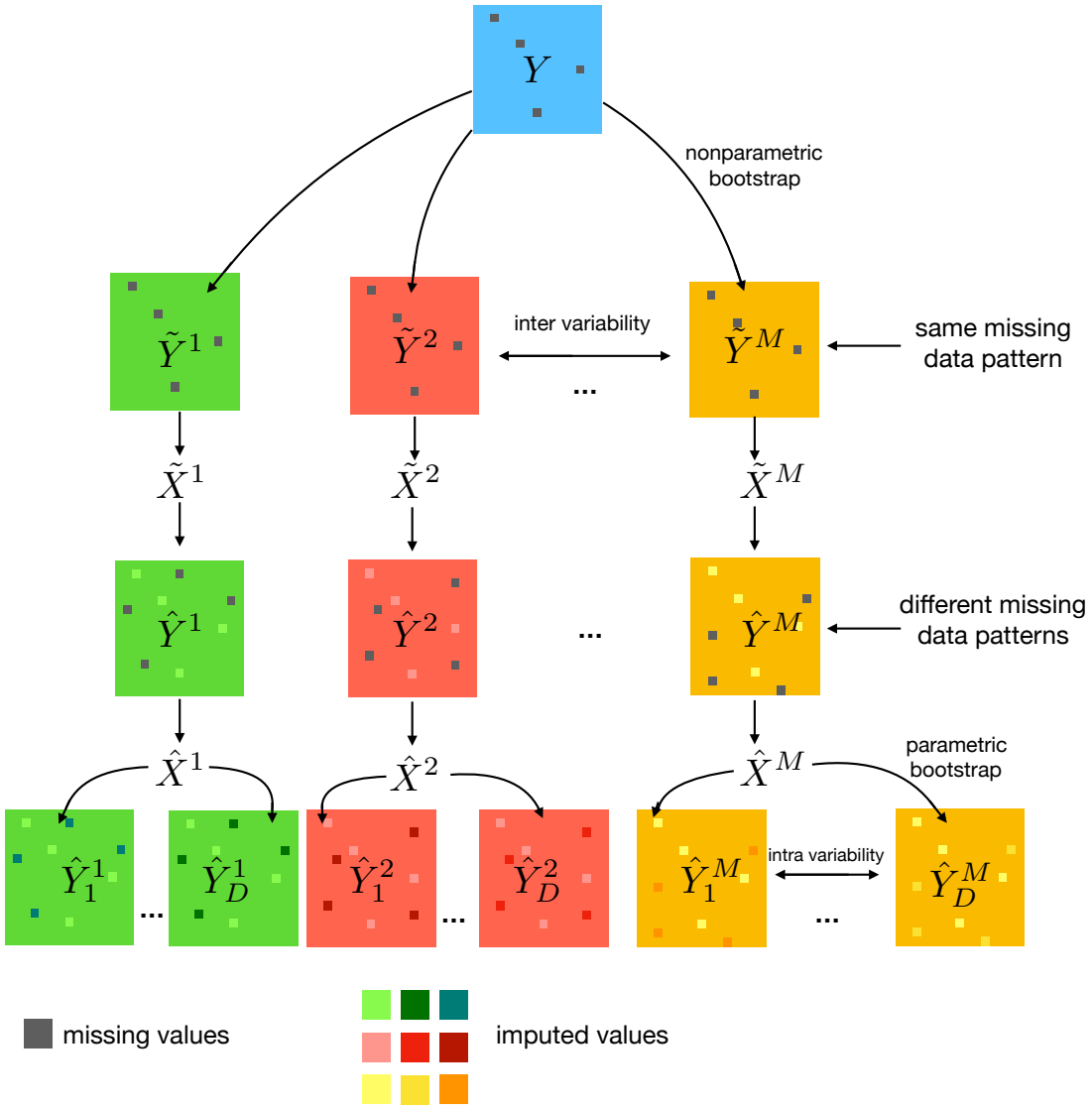


Figure 5.2: Multiple imputation procedure: nonparametric bootstrap (M samples); estimation (M estimates); parametric bootstrap (MD imputed data sets). Gray cells correspond to missing values: the same missing data pattern is shared across all incomplete data sets in the first level ($\tilde{\boldsymbol{Y}}^1, \ldots, \tilde{\boldsymbol{Y}}^M$). In the second level the missing data patterns are different ($\hat{\boldsymbol{Y}}^1, \ldots, \hat{\boldsymbol{Y}}^M$). The colored cells on the bottom line (which differ from the background color) correspond to imputed missing values, with different imputed values across the multiple imputed data sets.

5.2 Analysis of the Aravo data set

5.2.1 Loading the data

The Aravo data set (Choler, 2005) consists of three main data tables. First, a count table collecting the abundance of 82 species of alpine plants in 75 sites in France (the rows correspond to the environments, and the column to species). We will denote this abundance table, displayed in Table 5.9, $\boldsymbol{Y} \in \mathbb{R}^{m_1 \times m_2}$. Second, a matrix containing 6 geographical and meteorological characteristics of the sites. Third, a matrix containing 8 species traits (height, spread, etc.). We denote $\boldsymbol{R}$ the matrix of row covariates, and

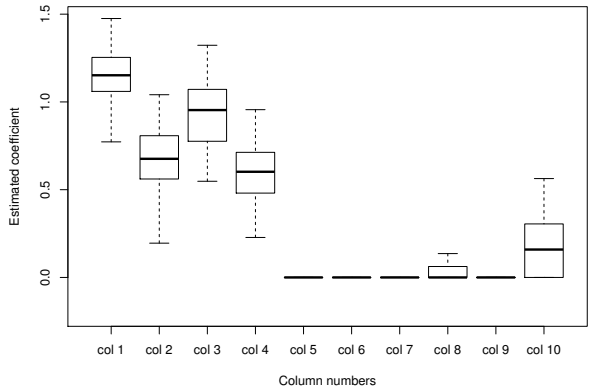


Figure 5.3: Boxplot of the estimators $\hat{\beta}^1, \ldots, \hat{\beta}^M$ produced by the multiple imputation procedure.

$\boldsymbol{C}$ the matrix of column covariates, respectively displayed in Tables 5.10 anc 5.11.

```
library(ade4)
data("aravo")
```

	Agro.rupe	Alop.alpi	Anth.nipp	Heli.sede	Aven.vers	Care.rosa
AR07	0	0	0	0	0	1
AR71	0	0	0	0	0	2
AR26	3	0	1	0	1	2
AR54	0	0	0	2	0	2
AR60	0	0	0	0	0	0

Table 5.9: First 5 rows (environments) and 6 columns (species) of the Aravo count table (aravo$spe[1:5, 1:6]).

```
aravo$env[1:5,]
aravo$traits[1:6,]
```

	Aspect	Slope	Form	PhysD	ZoogD	Snow
AR07	7	2	1	50	no	140
AR71	1	35	3	40	no	140
AR26	5	0	3	20	no	140
AR54	9	30	3	80	no	140
AR60	9	5	1	80	no	140

Table 5.10: First 5 rows (environments) of the Aravo row covariates (aravo$env[1:5,]).

First we put the data in the right shape for the `lori` function.

```
Y <- aravo$spe # count table
R <- aravo$env # row covariates
R <- R[, c(1,2,4,6)] # keep quantitative variables (for
   simplicity)
C <- aravo$traits # column covariates
d <- dim(Y)
n <- d[1]
p <- d[2]
```

	Height	Spread	Angle	Area	Thick	SLA	N_mass	Seed
Agro.rupe	6	10	80	60.0	0.12	8.10	218.70	0.08
Alop.alpi	5	20	20	190.9	0.20	15.10	203.85	0.21
Anth.nipp	15	5	50	280.0	0.08	18.00	219.60	0.54
Heli.sede	0	30	80	600.0	0.20	10.60	233.20	1.72
Aven.vers	12	30	60	420.0	0.14	12.50	156.25	1.17
Care.rosa	30	20	80	180.0	0.40	6.50	208.65	1.68

Table 5.11: First 6 rows (species) of the Aravo column covariates (aravo$traits[1:6,]).

```
U <- covmat(n,p,R,C) # construct covariate matrix for lori
   input
U <- scale(U) # scale the covariates
```

5.2.2 Estimation of the lori model

Then, we tune the regularization parameters, and apply the lori function.

```
# Tune regularization parameter
res_cv <- cv.lori(Y, U, reff=F, ceff=F, trace.it=T, len=10)
res_lori <- lori(Y, U, lambda1 = res_cv$lambda1, lambda2=res_
   cv$lambda2, reff=F, ceff=F)
```

Aspect	Slope	PhysD	Snow	Height	Spread	Angle	Area	Thick	SLA	Nmass	Seed
0.00	0.02	-0.00	-0.01	0.00	-0.04	-0.02	-0.03	-0.02	-0.00	0.01	-0.01

Table 5.12: Estimated covariate effects in the Aravo data set using `lori`. The regularization parameters are selected using cross-validation.

The obtain covariates coefficients are reported in Table 5.12. The estimated rank of the interaction matrix $\hat{\mathbf{\Theta}}$ is 3. The rows and columns of the interaction matrix can be visualized on a biplot (de Rooij and Heiser, 2005, Section 2.5), where rows and columns are represented simultaneously in a normalized Euclidean space. In such plots, the dimensions of the Euclidean space are given by the principal directions of $\hat{\mathbf{\Theta}}$, scaled by the square root of the singular values of $\hat{\mathbf{\Theta}}$. Such displays can be interpreted in terms of distance between points: a species and an environment that are close interact highly, and two species or two environments that are close have similar profiles. Justifications for such a distance interpretation can be found in (de Rooij and Heiser, 2005, Section 2.5) or (Fithian and Josse, 2017, Section 2). To visualize the two-dimensional display, one may use the following command, whose output plot in given in Figure 5.4, where the "axes" argument indicates which dimensions to use.

```
plot(res_lori, axes = c(1,2)
```

5.2.3 Bootstrap and intervals of variability

Finally, even though the table $\boldsymbol{Y}$ in complete, the multiple imputation function may be used to obtain intervals of variability for the coefficients reported in Table 5.12. The following command performs multiple imputation (which may be seen as bootstrap in this case), with 20 replications. Then, one can visualize the variability of the main effects coefficients with a boxplot (given in Figure 5.5).

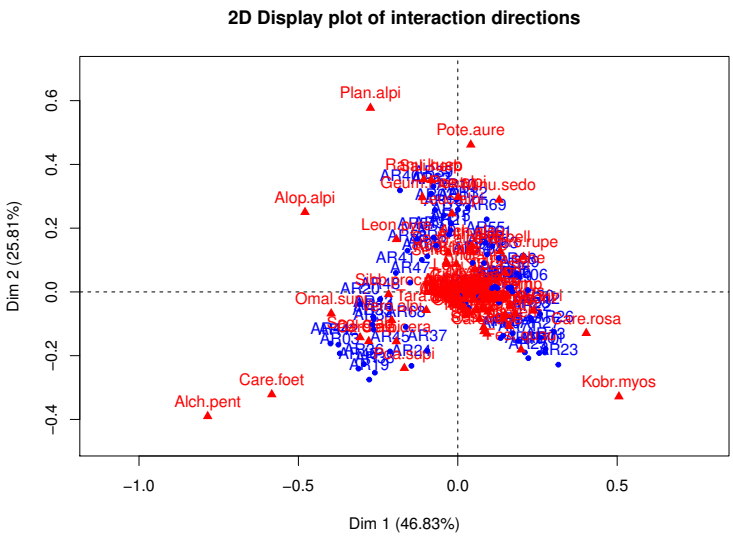


Figure 5.4: Two-dimensional display of the first dimensions of interaction.

```
res_mi <- mi.lori(Y, U, lambda1 = res_cv$lambda1, lambda2=res_
   cv$lambda2, reff=F, ceff=F, M=20)
boxplot(res_mi$mi.epsilon)
```

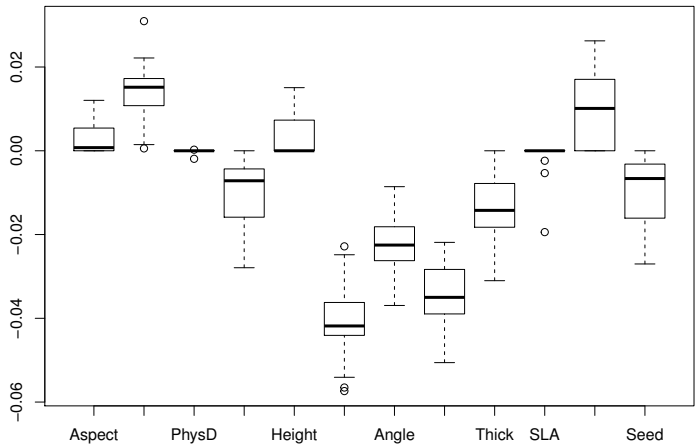


Figure 5.5: Boxplot of the main effects coefficients of the Aravo data set estimated with lori, across 20 bootstrap replications.

5.3 Implementation

To solve the minimization problem (5.3), we implemented two different algorithms in the `lori` package. The first one is based on alternating minimization (AM), and involves solving a LASSO problem and computing a *full-rank* SVD at each iteration. The second one is a mixed coordinate gradient descent (MCGD) algorithm, and involves solving a LASSO problem and computing a *rank-1* SVD at each iteration. In small dimensions, we

recommend using the AM option, which involves costly iterations (full-rank SVDs), but converges in fewer iterations. On the other hand, in large dimensions, we recommend using the MCGD option, which converges at a (slower) sublinear rate (Robin et al., 2018), but involves less costly iterations (rank-1 SVDs). By default, alternating minimization is used in the package `lori`. This option may be changed using the "algo" argument:

```
res <- lori(Y, U, lambda1 = 0.1, lambda2=0.1, algo = "alt") #
   alternating minimization (default)
res <- lori(Y, U, lambda1 = 0.1, lambda2=0.1, algo = "mcgd") #
    mixed coordinate gradient descent
```

5.3.1 Alternating minimization (AM)

Recall the lori estimation problem:

$$(\hat{\alpha}, \hat{\beta}, \hat{\epsilon}, \hat{\boldsymbol{\Theta}}) \in \operatorname{argmin} \mathcal{L}(\boldsymbol{Y}; \alpha, \beta, \epsilon, \boldsymbol{\Theta}) + \lambda_1 \|\boldsymbol{\Theta}\|_* + \lambda_2 \left(\|\alpha\|_1 + \|\beta\|_1 + \|\epsilon\|_1\right),$$

and denote $\phi = (\alpha, \beta, \epsilon)$ the vector of $\mathbb{R}^{m_1+m_2+q}$ containing all the main effects. Alternating minimization (AM) (Csiszár and Tusnády, 1984) consists in updating ϕ and $\boldsymbol{\Theta}$ alternatively, each time along a descent direction. At every iteration we update ϕ and $\boldsymbol{\Theta}$ both along proximal gradient directions. The update of ϕ involves entry-wise soft-thresholding, and the update of $\boldsymbol{\Theta}$ involves soft-thresholding of singular values. For each of the updates, we tune the step size using backtracking line search. The procedure is sketched in Algorithm 3 where, for $\lambda > 0$, $\mathcal{T}_\lambda$ denotes the operator of entry-wise soft-thresholding at level λ, and $\mathcal{D}_\lambda$ denotes the soft-thresholding of singular values operator at level λ.

Algorithm 3 Alternating minimization for problem (5.3)

Initialize $\phi^{[0]}$, $\boldsymbol{\Theta}^{[0]}$
for $t = 1, \dots, T$ **do**
  $\gamma = 1$
  $\phi^{(t)} = \operatorname{prox}_{\gamma\lambda_2\|\cdot\|_1}\left(\phi^{(t-1)} - \gamma\nabla\mathcal{L}_\phi(\boldsymbol{Y}; \phi, \boldsymbol{\Theta})\right)$
  $\qquad = \mathcal{T}_{\gamma\lambda_2}\left(\phi^{(t-1)} - \gamma\nabla\mathcal{L}_\phi(\boldsymbol{Y}; \phi, \boldsymbol{\Theta})\right).$
  while $\mathcal{L}(\boldsymbol{Y}; \phi^{(t)}, \boldsymbol{\Theta}^{(t-1)}) + \lambda_2\|\phi^{(t)}\|_1) > \mathcal{L}(\boldsymbol{Y}; \phi^{(t-1)}, \boldsymbol{\Theta}^{(t-1)}) + \lambda_2\|\phi^{(t-1)}\|_1)$ **do**
    $\gamma = \gamma/2$
    $\phi^{(t)} = \operatorname{prox}_{\gamma\lambda_2\|\cdot\|_1}\left(\phi^{(t-1)} - \gamma\nabla\mathcal{L}_\phi(\phi, \boldsymbol{\Theta})\right)$
    $\qquad = \mathcal{T}_{\gamma\lambda_2}\left(\phi^{(t-1)} - \gamma\nabla\mathcal{L}_\phi(\boldsymbol{Y}; \phi, \boldsymbol{\Theta})\right).$
  end while
  $\gamma = 1$
  $\boldsymbol{\Theta}^{(t)} = \mathcal{D}_{\gamma\lambda_1}(\boldsymbol{\Theta}^{(t-1)} - \gamma\nabla_{\boldsymbol{\Theta}}\mathcal{L}(\boldsymbol{Y}; \phi^{(t)}, \boldsymbol{\Theta}^{(t-1)}))$
  while $\mathcal{L}(\boldsymbol{Y}; \phi^{(t)}, \boldsymbol{\Theta}^{(t)}) + \lambda_1\|\boldsymbol{\Theta}^{(t)}\|_*) > \mathcal{L}(\boldsymbol{Y}; \phi^{(t)}, \boldsymbol{\Theta}^{(t-1)}) + \lambda_1\|\boldsymbol{\Theta}^{(t-1)}\|_*)$ **do**
    $\gamma = \gamma/2$
    $\boldsymbol{\Theta}^{(t)} = \mathcal{D}_{\gamma\lambda_1}(\boldsymbol{\Theta}^{(t-1)} - \gamma\nabla_{\boldsymbol{\Theta}}\mathcal{L}(\boldsymbol{Y}; \phi^{(t)}, \boldsymbol{\Theta}^{(t-1)})).$
  end while
end for
Return $\phi^{(T)}$, $\boldsymbol{\Theta}^{(T)}$

5.3.2 Mixed coordinate gradient descent (MCGD)

When the dimensions m_1 and m_2 are large, SVDs are extremely computationally heavy. Thus, we also implemented an alternative to AM, adapted to large data frames. In this

alternative algorithm, problem (5.3) is solved using a mixed coordinate gradient descent procedure (MCGD), where ϕ on the one hand, and $\mathbf{\Theta}$ on the other hand, are updated alternatively. The vector ϕ is, as before, updated along a proximal gradient direction with a step size γ computed using a line search. The matrix $\mathbf{\Theta}$, on the other hand, is updated along a conditional gradient direction: this requires computing only the first singular value and vectors of $\mathbf{\Theta}$, instead of the full SVD. Denote $F(\phi, \mathbf{\Theta}) = \mathcal{L}(\boldsymbol{Y}; \phi, \mathbf{\Theta}) + \lambda_1 \|\mathbf{\Theta}\|_* + \lambda_2 \|\phi\|_1$. We define at iteration t the quantity $R^{(t)} = \lambda_1^{-1} F(\phi^{(t-1)}, \mathbf{\Theta}^{(t-1)})$. We also denote $\nabla \mathcal{L}_{\mathbf{\Theta}}(\boldsymbol{Y}; \phi^{(t-1)}, \mathbf{\Theta}^{(t-1)})$ the gradient of $\mathcal{L}$ with respect to $\mathbf{\Theta}$ evaluated at $(\phi^{(t-1)}, \mathbf{\Theta}^{(t-1)})$. Furthermore, we denote by $\sigma_1(\nabla \mathcal{L}_{\mathbf{\Theta}}(\boldsymbol{Y}; \phi^{(t-1)}, \mathbf{\Theta}^{(t-1)}))$ its largest singular value, and u_1 and v_1 its first left and right singular vectors. The conditional gradient update consists in the following operation:

$$\mathbf{\Theta}^{(t)} = \begin{cases} 0 & \text{if } \lambda_1 \geq \sigma_1(\nabla \mathcal{L}_{\mathbf{\Theta}}(\boldsymbol{Y}; \phi^{(t-1)}, \mathbf{\Theta}^{(t-1)})), \\ \mathbf{\Theta}^{(t-1)} - \delta R^{(t)} u_1 v_1^\top & \text{if } \lambda_1 < \sigma_1(\nabla \mathcal{L}_{\mathbf{\Theta}}(\boldsymbol{Y}; \phi^{(t-1)}, \mathbf{\Theta}^{(t-1)})). \end{cases} \tag{5.5}$$

In the update (5.5), δ is a step size which we determine with a line search. A sketch of the MCGD algorithm is provided in Algorithm 4, and its convergence properties are studied in Robin et al. (2018).

Algorithm 4 MCGD algorithm for (5.3).

Initialize: — $\mathbf{\Theta}^{(0)}, \phi^{(0)}, R^{(0)}$. E.g., $(\mathbf{\Theta}^{(0)}, \phi^{(0)}, R^{(0)}) = (\mathbf{0}, \mathbf{0}, 0)$.
for $t = 1, 2, \ldots, T$ **do**
  $\gamma = 1$
  $\phi^{(t)} = \operatorname{prox}_{\gamma\lambda_2\|\cdot\|_1}\left(\phi^{(t-1)} - \gamma \nabla \mathcal{L}_\phi(\boldsymbol{Y}; \phi, \mathbf{\Theta})\right)$
  $\qquad = \mathcal{T}_{\gamma\lambda_2}\left(\phi^{(t-1)} - \gamma \nabla \mathcal{L}_\phi(\boldsymbol{Y}; \phi, \mathbf{\Theta})\right).$
  while $\mathcal{L}(\boldsymbol{Y}; \phi^{(t)}, \mathbf{\Theta}^{(t-1)}) + \lambda_2 \|\phi^{(t)}\|_1) > \mathcal{L}(\boldsymbol{Y}; \phi^{(t-1)}, \mathbf{\Theta}^{(t-1)}) + \lambda_2 \|\phi^{(t-1)}\|_1)$ **do**
    $\gamma = \gamma/2$
    $\phi^{(t)} = \operatorname{prox}_{\gamma\lambda_2\|\cdot\|_1}\left(\phi^{(t-1)} - \gamma \nabla \mathcal{L}_\phi(\phi, \mathbf{\Theta})\right)$
    $\qquad = \mathcal{T}_{\gamma\lambda_2}\left(\phi^{(t-1)} - \gamma \nabla \mathcal{L}_\phi(\boldsymbol{Y}; \phi, \mathbf{\Theta})\right).$
  end while
  $\delta = 1$
  $R^{(t)} = \lambda_1^{-1} F(\phi^{(t-1)}, \mathbf{\Theta}^{(t-1)})$
  $\mathbf{\Theta}^{(t)} = \begin{cases} 0 & \text{if } \lambda_1 \geq \sigma_1(\nabla \mathcal{L}_{\mathbf{\Theta}}(\boldsymbol{Y}; \phi^{(t-1)}, \mathbf{\Theta}^{(t-1)})), \\ \mathbf{\Theta}^{(t-1)} - \delta R^{(t)} u_1 v_1^\top & \text{if } \lambda_1 < \sigma_1(\nabla \mathcal{L}_{\mathbf{\Theta}}(\boldsymbol{Y}; \phi^{(t-1)}, \mathbf{\Theta}^{(t-1)})). \end{cases}$
  while $\mathcal{L}(\boldsymbol{Y}; \phi^{(t)}, \mathbf{\Theta}^{(t)}) + \lambda_1 \|\mathbf{\Theta}^{(t)}\|_*) > \mathcal{L}(\boldsymbol{Y}; \phi^{(t)}, \mathbf{\Theta}^{(t-1)}) + \lambda_1 \|\mathbf{\Theta}^{(t-1)}\|_*)$ **do**
    $\delta = \delta/2$
    $R^{(t)} = \lambda_1^{-1} F(\phi^{(t-1)}, \mathbf{\Theta}^{(t-1)})$
    $\mathbf{\Theta}^{(t)} = \begin{cases} 0 & \text{if } \lambda_1 \geq \sigma_1(\nabla \mathcal{L}_{\mathbf{\Theta}}(\boldsymbol{Y}; \phi^{(t-1)}, \mathbf{\Theta}^{(t-1)})), \\ \mathbf{\Theta}^{(t-1)} - \delta R^{(t)} u_1 v_1^\top & \text{if } \lambda_1 < \sigma_1(\nabla \mathcal{L}_{\mathbf{\Theta}}(\boldsymbol{Y}; \phi^{(t-1)}, \mathbf{\Theta}^{(t-1)})). \end{cases}$
  end while
end for
Return: $\mathbf{\Theta}^{(T)}, \phi^{(T)}$.

Chapter 6

Main effects and interactions in mixed and incomplete data frames

Contents

6.1 Introduction

Mixed data frames (MDF) (see Pagès (2015); Udell et al. (2016)) are tables collecting categorical, numerical and count data. In most applications, each row is an example or a subject and each column is a feature or an attribute. A distinctive characteristic of MDF is that column entries may be of different types, and most often many entries are missing. MDF appear in numerous applications including patient records in health care (survival values at different time points, quantitative and categorical clinical features like blood pressure, gender, disease stage, see e.g. Murdoch and Detsky (2013)), survey data (Heeringa et al., 2010, Chapters 5 and 6), abundance tables in ecology (Legendre

et al., 1997), and recommendation systems (Agarwal et al., 2011).

In all these applications, data analysis is often made in the light of additional information, such as sites and species traits in ecology, or users and items characteristics in recommendation systems. This caused the introduction of the two central concepts of interest in this chapter: *main effects* and *interactions*. This terminology is classically used to distinguish between effects of covariates on the observations which are independent of the other covariates (main effects), and effects of covariates on the observations which depend on the value of one or more other covariates (interactions). For example, in health care, a treatment might extend survival for all patients–this is a main effect–or extend survival for young patients but shorten it for older patients–this is an interaction.

Many statistical models have been developed to estimate such types of data. Abundance tables counting species across environments are for instance classically analyzed using the log-linear model (Agresti, 2013, Chapter 4). This model decomposes the logarithms of the expected abundances into the sum of species (rows) and environment (columns) effects, plus a low-rank interaction term. Other examples include multilevel models (Gelman and Hill, 2007) to analyze hierarchically structured data where examples (patients, students, etc.) are nested within groups (hospitals, schools, etc.).

At the same time, low-rank models, which embed rows and columns into low-dimensional spaces, have been widely used for exploratory analysis of MDF (Kiers, 1991; Pagès, 2015; Udell et al., 2016). Despite the abundance of results in low-rank matrix estimation (see Kumar and Schneider (2017) for a literature survey), to the best of our knowledge most of the existing methods for MDF analysis do not provide a statistically sound way to account for *main effects* in the data. In most applications, estimation of main effects in MDF has been done heuristically as a preprocessing step (Hastie et al., 2015; Udell et al., 2016; Landgraf and Lee, 2015). Fithian and Mazumder (2018) incorporate row and column covariates in their model, but mainly focus on optimization procedures and did not provide statistical guarantees concerning the main effects. Mao et al. (2017) propose a procedure to estimate jointly main effects and a low-rank matrix, which may be interpreted as an interaction matrix, but the procedure is based on a least squares loss, and is therefore not suitable to mixed data types.

On the other hand, several approaches are available in the matrix completion literature to model non-Gaussian, and particularly discrete data, but they do not consider mixed observations or main effects. Davenport et al. (2012) introduced one-bit matrix completion, where the observations are binary such as yes/no answers, and provide nearly optimal upper and lower bounds on the mean square error of estimation. One-bit matrix completion was also studied in Cai and Zhou (2013). In Klopp et al. (2015), the authors introduce multinomial matrix completion, where the observations are allowed to take more than two values, such as ratings in recommendation systems, and propose a minimax optimal estimator. Unbounded non-Gaussian observations have also been studied before. For instance, Cao and Xie (2016) extended the approach of Davenport et al. (2012) to Poisson matrix completion, and Lafond (2015) proves optimal convergence rates for exponential family matrix completion.

In this chapter, we propose a new framework for incomplete and mixed data which allows to account for main effects and interactions. We start in Section 6.2 with a con-

crete example from medical data analysis, before introducing a general model for MDF with sparse main effects and low-rank interactions. Then, we propose in Section 6.3 an estimation procedure based on the minimization of a doubly penalized negative quasi log-likelihood. We also propose a block coordinate gradient descent algorithm to compute our estimator, and provide a convergence result. In Section 6.4.1 we discuss the statistical guarantees of our procedure, with two simultaneous upper bounds on the estimation errors of the sparse and low-rank components. To assess the tightness of our convergence rates, we also derive lower bounds in Section 6.4.2, and show that in a number of situations, our upper bounds are near optimal. In addition, we specialize our results to three examples of interest in applications in Section 6.4.3. Numerical results are presented in Section 6.5 to support our theoretical claims. We also show that our method performs comparably to state-of the art mixed data imputation methods in terms of prediction of the missing values. The proofs are postponed to the supplementary material; the method is available in the R (R Core Team, 2017) package `mimi` on the Comprehensive R Archive Network.

Notation We denote the Frobenius norm on $\mathbb{R}^{m_1 \times m_2}$ by $\|\cdot\|_F$, the operator norm by $\|\cdot\|$, the nuclear norm by $\|\cdot\|_*$ and the sup norm $\|\cdot\|_\infty$. $\|\cdot\|_2$ is the usual Euclidean norm, $\|\cdot\|_0$ the number of non zero coefficients, and $\|\cdot\|_\infty$ the infinity norm. For $n \in \mathbb{N}$, denote $[\![n]\!] = \{1, \dots, n\}$. We denote the support of $\alpha \in \mathbb{R}^N$ by $\mathrm{supp}(\alpha) = \{k \in [\![N]\!], \alpha_k \neq 0\}$. For $I \subseteq [\![m_1]\!]$, we denote $\mathbb{1}_I$, defined by $\mathbb{1}_I(i) = 1$ if $i \in I$ and 0 otherwise, the indicator of set I_h.

6.2 General model and examples

6.2.1 Traumabase data set

Before introducing our general model, we start by giving a concrete example. The Traumabase registry (`http://www.traumabase.eu/en_US`) gathers information about severe trauma patients distributed across 15 trauma centers. As shown in Table 6.1, this results in a highly heterogeneous and incomplete data collection.

Center	Lung X-ray	Pelvic X-ray	Accident	Time in critical care (h)
Bicêtre	NA	NA	Falling from a height	NA
HEGP	NA	NA	Falling from a height	2
Pitié Salpétrière	NA	NA	Car-pedestrian accident	NA
Lille	Normal	NA	Falling from a height	2
Beaujon	NA	NA	Falling (from own height)	NA
Lille	NA	NA	Falling (from own height)	NA

Table 6.1: Excerpt of the Traumabase data set.

Here, the Center variable categorizes the patients in groups, depending on the hospital where they were treated. In an exploratory data analysis perspective, a question of interest may be: is the trauma center related to the values of other variables? For example the type of accident, survival, etc. Furthermore, as we do not expect the groups (trauma centers) to be sufficient to explain the observations, can we also model residuals, or *interactions*?

Denote $\boldsymbol{Y} = (\boldsymbol{Y}_{i,j})$ the data frame containing the patients in rows and the attributes in columns. If the j-th column is continuous (systolic blood pressure for instance), one might model the group effects and interactions as follows:

$$\mathbb{E}[\boldsymbol{Y}_{i,j}] = \alpha^0_{c(i)j} + \boldsymbol{\Theta}^0_{i,j},$$

where $c(i)$ indicates the group to which individual i belongs, and $\alpha^0_{c(i)j}$ and $\boldsymbol{\Theta}^0_{i,j}$ are fixed group effects and interactions respectively. This corresponds to the so-called *multilevel regression* framework (Gelman and Hill, 2007). If the j-th column is binary (result of pelvic X-ray for instance, which is either normal or abnormal), one might model

$$\mathbb{P}(\boldsymbol{Y}_{i,j} = 1) = \frac{\mathrm{e}^{\boldsymbol{X}^0_{i,j}}}{1 + \mathrm{e}^{\boldsymbol{X}^0_{i,j}}}, \quad \boldsymbol{X}^0_{i,j} = \alpha^0_{c(i)j} + \boldsymbol{\Theta}^0_{i,j},$$

corresponding to a logistic regression framework.

The goal is then to estimate the vector of group effects α^0 and the matrix of interactions $\boldsymbol{\Theta}^0$ simultaneously, from the mixed and incomplete data frame $\boldsymbol{Y}$. We propose a method assuming the vector of main effects α^0 is sparse and the matrix of interactions $\boldsymbol{\Theta}^0$ has low-rank. The sparsity assumption means that groups affect a small number of variables. On the other hand, the low-rank assumption means the population can be represented by a few archetypical individuals and summary features (Udell et al., 2016, Section 5.4), which interact in a multiplicative manner. In fact, if $\boldsymbol{\Theta}^0$ is of rank r, then it can be decomposed as the sum of r rank-1 matrices as follows:

$$\boldsymbol{\Theta}^0 = \sum_{k=1}^{r} u_k v_k^\top,$$

where u_k (resp. v_k) is a vector of $\mathbb{R}^{m_1}$ (resp. $\mathbb{R}^{m_2}$). Thus, using the above example, we obtain

$$\mathbb{E}[\boldsymbol{Y}_{i,j}] = \alpha^0_{c(i)j} + \sum_{k=1}^{r} u_{ik} v_{jk},$$

where the last term $\sum_{k=1}^{r} u_{ik} v_{jk}$ can be interpreted as the sum of multiplicative interaction terms between latent individual types and features.

6.2.2 General model

We now introduce a new framework generalizing the above example to other types of data and main effects. Consider an MDF $\boldsymbol{Y} = (\boldsymbol{Y}_{i,j})$ of size $m_1 \times m_2$. The entries in each column $j \in [\![m_2]\!]$ belong to an observation space, denoted $\mathbb{Y}_j$. For example, for numerical data, the observation space is $\mathbb{Y}_j = \mathbb{R}$, and for count data, $\mathbb{Y}_j = \mathbb{N}$ is the set of natural integers. For binary data, the observation space is $\mathbb{Y}_j = \{0, 1\}$. In the entire paper, we assume that the random variables $(\boldsymbol{Y}_{i,j})$ are independent and that for each $(i, j) \in [\![m_1]\!] \times [\![m_2]\!]$, $\boldsymbol{Y}_{i,j} \in \mathbb{Y}_j$ and $\mathbb{E}\left[|\boldsymbol{Y}_{i,j}|\right] < \infty$. Furthermore, we will assume that $\boldsymbol{Y}_{i,j}$ is sub-exponential with scale γ and variance σ^2: for all $(i, j) \in [\![m_1]\!] \times [\![m_2]\!]$ and $|z| < \gamma$, $\mathbb{E}\left[\mathrm{e}^{z(\boldsymbol{Y}_{i,j} - \mathbb{E}[\boldsymbol{Y}_{i,j}])}\right] \leq \mathrm{e}^{\sigma^2 z^2/2}$.

In our estimation procedure, we will use a data-fitting term based on heterogeneous exponential family quasi-likelihoods. Let $(\mathbb{Y}, \mathcal{Y}, \mu)$ be a measurable space, $h : \mathbb{Y} \to \mathbb{R}_+$,

and $g : \mathbb{R} \to \mathbb{R}$ be functions. Denote by $\mathrm{Exp}^{(h,g)} = \{f_x^{(h,g)} : x \in \mathbb{R}\}$ the canonical exponential family. Here, h is the base function, g is the link function, and $f_x^{(h,g)}$ is the density with respect to the base measure μ given by

$$f_x^{(h,g)}(y) = h(y) \exp\left(yx - g(x)\right), \tag{6.1}$$

for $y \in \mathbb{Y}$. For simplicity, we assume $\int h(y) \exp(yx)\mu(\mathrm{d}y) < \infty$ for all $x \in \mathbb{R}$.

The exponential family is a flexible framework for different data types. For example, for numerical data, we set $g(x) = x^2\sigma^2/2$ and $h(y) = (2\pi\sigma^2)^{-1/2} \exp(-y^2/\sigma^2)$. In this case, $\mathrm{Exp}^{(h,g)}$ is the family of Gaussian distributions with mean $\sigma^2 x$ and variance σ^2. For count data, we set $g(x) = \exp(ax)$ and $h(y) = 1/y!$, where $a \in \mathbb{R}$. In this case, $\mathrm{Exp}^{(h,g)}$ is the family of Poisson distributions with intensity $\exp(ax)$. For binary data, $g(x) = \log(1 + \exp(x))$ and $h(y) = 1$. Here, $\mathrm{Exp}^{(h,g)}$ is the family of Bernoulli distributions with success probability $1/(1 + \exp(-x))$.

In our estimation procedure, we choose a collection $\{(g_j, h_j),\ j \in [\![m_2]\!]\}$ of link functions and base functions corresponding to the observation spaces $\{(\mathbb{Y}_j, \mathcal{Y}_j, \mu_j)\,,\ j \in [\![m_2]\!]\}$. For each $(i, j) \in [\![m_1]\!] \times [\![m_2]\!]$, we denote by $\boldsymbol{X}^0_{i,j}$ the value of the parameter minimizing the divergence between the distribution of $\boldsymbol{Y}_{i,j}$ and the exponential family $\mathrm{Exp}^{(h_j,g_j)}$, $j \in [\![m_2]\!]$:

$$\boldsymbol{X}^0_{i,j} = \mathrm{argmin}_{x\in\mathbb{R}} \left\{-\mathbb{E}\left[\boldsymbol{Y}_{i,j}\right] x + g_j(x)\right\}. \tag{6.2}$$

To model main effects and interactions we assume the matrix of parameters $\boldsymbol{X}^0 = (\boldsymbol{X}^0_{i,j}) \in \mathbb{R}^{m_1\times m_2}$ can be decomposed as the sum of sparse main effects and low-rank interactions:

$$\boldsymbol{X}^0 = \sum_{k=1}^{N} \alpha_k^0 \boldsymbol{U}^k + \boldsymbol{\Theta}^0. \tag{6.3}$$

Here, $\mathcal{U} = (\boldsymbol{U}^1, \ldots, \boldsymbol{U}^N)$ is a fixed dictionary of $m_1 \times m_2$ matrices, α^0 is a sparse vector with unknown support $\mathcal{I} = \{k \in [\![N]\!]; \alpha_k^0 \neq 0\}$ and $\boldsymbol{\Theta}^0$ is an $m_1 \times m_2$ matrix with low-rank. The decomposition introduced in (6.3) is a general model combining regression on a dictionary and low-rank design.

Such decompositions have been studied before in the literature. In particular, a large body of work has tackled the problem of reconstructing a sparse and a low-rank term exactly from observation of their sum. Chandrasekaran et al. (2011) derived identifiability conditions under which exact reconstruction is possible when the sparse component is *entry-wise sparse*; the same model was also studied in Hsu et al. (2011). Candès et al. (2011) proved a similar results for entry-wise sparsity, when the location of the non-zero entries are chosen uniformly at random. Xu et al. (2010) extended the model to study *column-wise sparsity*. Mardani et al. (2013) studied an even more general case with general sparsity pattern, and determined conditions under which exact recovery is possible.

In this chapter, we consider the problem of estimating a (general) sparse component and a low-rank term from noisy and incomplete observation of their sum, when the noise is heterogeneous and in the exponential family. Because of this noisy setting, we do not seek to recover the two components exactly. Thus, we do not require strong

identifiability conditions as those derived in (Chandrasekaran et al., 2011; Hsu et al., 2011; Candès et al., 2011; Xu et al., 2010; Mardani et al., 2013). However, since decomposition (6.3) may not be unique, we restrict our model to the following class of possible decompositions, to which our estimator will be the closest. From all possible decompositions (α, L), consider $(\alpha', \boldsymbol{\Theta}')$ such that

$$(\alpha', \boldsymbol{\Theta}') \in \operatorname{argmin}_{\boldsymbol{X}^0 = \sum \alpha_k \boldsymbol{U}^k + \boldsymbol{\Theta}} \{\|\alpha\|_0 + \operatorname{rank} \boldsymbol{\Theta}\}. \tag{6.4}$$

Let $s' = \|\alpha'\|_0 + \operatorname{rank}(\boldsymbol{\Theta}')$. Finally let

$$(\alpha^0, \boldsymbol{\Theta}^0) \in \operatorname*{argmin}_{\substack{\boldsymbol{X}^0 = \sum \alpha_k \boldsymbol{U}^k + \boldsymbol{\Theta} \\ \|\alpha\|_0 + \operatorname{rank}(\boldsymbol{\Theta}) = s'}} \|\alpha\|_0 . \tag{6.5}$$

The decomposition satisfying (6.4) and (6.5) may also not be unique. Assume that there exists a pair $(\alpha, \boldsymbol{\Theta}) \neq (\alpha^0, \boldsymbol{\Theta}^0)$ satisfying (6.4) and (6.5). Then,

$$\|\boldsymbol{\Theta}^\star - \boldsymbol{\Theta}\|_F = \|\sum_k \alpha_k^\star \boldsymbol{U}^k - \sum_k \alpha_k \boldsymbol{U}^k\|_F \quad \leq 2a\|\alpha^0\|_0 \max_k \|\boldsymbol{U}_k\|_2 = R,$$

with a an upper bound on $\|\alpha^0\|_\infty$. This implies that for all such possible decompositions $(\alpha, \boldsymbol{\Theta})$ we have that $\boldsymbol{\Theta}$ and $\sum_k \alpha_k \boldsymbol{U}^k$ are in the small balls of radius R and centered at $\boldsymbol{\Theta}^0$ and $\sum_k \alpha_k^0 \boldsymbol{U}^k$ respectively. Our statistical guarantees in Section 6.4 show that our estimators of $\boldsymbol{\Theta}^0$ and $\sum_k \alpha_k^0 \boldsymbol{U}^k$ are in balls of radius $> R$, and also centered at $\boldsymbol{\Theta}^0$ and $\sum_k \alpha_k^0 \boldsymbol{U}^k$. Moreover, we also show that this error bound is minimax optimal in a number of situations. To summarize, the decomposition may not be unique, but all the possible decompositions are in a neighborhood of radius smaller than the optimal convergence rate.

6.2.3 Examples

We now provide three examples of dictionaries which can be used to deal with classical main effects.

Example 1. Group effects We assume the m_1 individuals are divided into H groups. For $h \in [\![H]\!]$ denote by $I_h \subset [\![m_1]\!]$ the h-th group containing n_h individuals. The size of the dictionary is $N = Hm_2$ and its elements are, for all $(h, q) \in [\![H]\!] \times [\![m_2]\!]$, $\boldsymbol{U}_{h,q} = (\mathbb{1}_{I_h}(i)\mathbb{1}_{\{q\}}(j))_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]}$. This example corresponds to the model discussed in Section 6.2.1; we develop it further in Section 6.5 with simulations and a survey data analysis.

Example 2. Row and column effects (see *e.g.* (Agresti, 2013, Chapter 4)) Another classical model is the log-linear model for count data analysis. Here, $\boldsymbol{Y}$ is a matrix of counts. Assuming a Poisson model, the parameter matrix $\boldsymbol{X}^0$, which satisfies $\mathbb{E}[\boldsymbol{Y}_{i,j}] = \exp(\boldsymbol{X}^0_{i,j})$ for all $(i, j) \in [\![m_1]\!] \times [\![m_2]\!]$, is assumed to be decomposed as follows:

$$\boldsymbol{X}^0_{i,j} = (\alpha_r^0)_i + (\alpha_c^0)_j + \boldsymbol{\Theta}^0_{i,j}, \tag{6.6}$$

where $\alpha_r^0 \in \mathbb{R}^{m_1}$, $\alpha_c^0 \in \mathbb{R}^{m_2}$ and $\boldsymbol{\Theta}^0 \in \mathbb{R}^{m_1 \times m_2}$ is low-rank. This model is often used to analyze abundance tables of species across environments (see, *e.g.*, ter Braak et al. (2017)). In this case the low-rank structure of $\boldsymbol{\Theta}^0$ reflects the presence of groups of similar species and environments. Model (6.6) can be re-written in our framework as

$$\boldsymbol{X}^0 = \sum_{k=1}^{N} \alpha_k^0 \boldsymbol{U}_k + \boldsymbol{\Theta}^0,$$

with $\alpha^0 = (\alpha_r^0, \alpha_c^0)$, $N = m_1 + m_2$ and where for $i \in [\![m_1]\!]$ and $j \in [\![m_2]\!]$ we have $\boldsymbol{U}_i = (\mathbb{1}_{\{i\}}(k))_{(k,l)\in[\![m_1]\!]\times[\![m_2]\!]}$ and $\boldsymbol{U}_{m_1+j} = (\mathbb{1}_{\{j\}}(l))_{(k,l)\in[\![m_1]\!]\times[\![m_2]\!]}$.

Example 3. Corruptions Our framework also embeds the well-known robust matrix completion problem (Hsu et al., 2011; Candès et al., 2011; Klopp et al., 2017) which is of interest, for instance, in recommendation systems. In this application, malicious users coexist with normal users, and introduce spurious perturbations. Thus, in robust matrix completion, we observe noisy and incomplete realizations of a low-rank matrix $\boldsymbol{\Theta}^0$ of fixed rank and containing zeros at the locations of malicious users, perturbed by corruptions. The sparse component corresponding to corruptions is denoted $\sum_{(i,j)\in\mathcal{I}} \alpha_k^0 \boldsymbol{U}_{i,j}$, where $\{\boldsymbol{U}_{ij}, (i,j) \in [\![m_1]\!] \times [\![m_2]\!]\}$, are the matrices of the canonical basis of $\mathbb{R}^{m_1\times m_2}$ $\boldsymbol{U}_{ij} = (\mathbb{1}_{\{i\}}(k)\mathbb{1}_{\{j\}}(l))_{(k,l)\in[\![m_1]\!]\times[\![m_1]\!]}$ and $\mathcal{I}$ is the set of indices of corrupted entries. Thus, the non-zero components of α^0 correspond to the locations where the malicious users introduced the corruptions.

For Example 3, the particular case of quadratic link functions $g_j(x) = x^2/2$ was studied in Klopp et al. (2017). We generalize these results in two directions: we consider mixed data types and general main effects.

6.2.4 Missing values

Finally, we consider a setting with missing observations. Let $\Omega = (\Omega_{i,j})$ be an observation mask with $\Omega_{i,j} = 1$ if $\boldsymbol{Y}_{i,j}$ is observed and $\Omega_{i,j} = 0$ otherwise. We assume that Ω and $\boldsymbol{Y}$ are independent, i.e. a Missing Completely At Random (MCAR) scenario (Little and Rubin, 2002): $(\Omega_{i,j})$ are independent Bernoulli random variables with probabilities π_{ij}, $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$. Furthermore for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, we assume there exists $p > 0$ allowed to vary with m_1 and m_2, such that

$$\pi_{ij} \geq p. \tag{6.7}$$

For $j \in [\![m_2]\!]$, denote by $\pi_{\cdot j} = \sum_{i=1}^{m_1} \pi_{ij}$, $j \in [\![m_2]\!]$ the probability of observing an element in the j-th column. Similarly, for $i \in [\![m_1]\!]$, denote by $\pi_{i\cdot} = \sum_{j=1}^{m_2} \pi_{ij}$ the probability of observing an element in the i-th row. We define the following upper bound:

$$\max_{i,j}(\pi_{i\cdot}, \pi_{\cdot j}) \leq \beta \tag{6.8}$$

6.3 Estimation procedure

Consider the data-fitting term defined by the heterogeneous exponential family negative quasi log-likelihood

$$\mathcal{L}(\boldsymbol{X}; \boldsymbol{Y}, \Omega) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \Omega_{i,j} \left\{-\boldsymbol{Y}_{i,j}\boldsymbol{X}_{i,j} + g_j(\boldsymbol{X}_{i,j})\right\}, \tag{6.9}$$

and define the function

$$f(\alpha, \boldsymbol{\Theta}) = \mathcal{L}(\mathsf{f}_U(\alpha) + \boldsymbol{\Theta}; \boldsymbol{Y}, \Omega), \tag{6.10}$$

where for $\alpha \in \mathbb{R}^N$, $\mathsf{f}_U(\alpha) = \sum_{k=1}^N \alpha_k \boldsymbol{U}_k$. We assume $\|\alpha^0\|_\infty \leq a$ and $\|\boldsymbol{\Theta}^0\|_\infty \leq a$ where $a > 0$ is a known upper bound. We use the nuclear norm $\|\cdot\|_*$ (the sum of

singular values) and ℓ_1 norm $\|\cdot\|_1$ penalties as convex relaxations of the rank and sparsity constraints respectively:

$$(\hat{\alpha}, \hat{\mathbf{\Theta}}) \in \operatorname{argmin}_{(\alpha,\mathbf{\Theta})} F(\alpha, \mathbf{\Theta}) \tag{6.11}$$

$$\text{s. t. } \|\alpha\|_\infty \leq a, \|\mathbf{\Theta}\|_\infty \leq a, \tag{6.12}$$

$$F(\alpha, \mathbf{\Theta}) = f(\alpha, \mathbf{\Theta}) + \lambda_1 \|\mathbf{\Theta}\|_* + \lambda_2 \|\alpha\|_1, \tag{6.13}$$

with $\lambda_1 > 0$ and $\lambda_2 > 0$. In the sequel, for all $(\hat{\alpha}, \hat{\mathbf{\Theta}})$ in the set of solutions, we denote by $\hat{\boldsymbol{X}} = \mathsf{f}_U(\hat{\alpha}) + \hat{\mathbf{\Theta}}$.

6.3.1 Block coordinate gradient descent (BCGD)

To solve (6.11) we develop a block coordinate gradient descent algorithm where the two components α and $\mathbf{\Theta}$ are updated alternatively in an iterative procedure. At every iteration, we compute a (strictly convex) quadratic approximation of the data fitting term and apply block coordinate gradient descent to generate a search direction. The BCGD algorithm we describe is a special instance of the coordinate gradient descent method for non-smooth separable minimization developed in Tseng and Yun (2009). Note first that the upper bound a on $\|\alpha\|_\infty$ and $\|\mathbf{\Theta}\|_\infty$ is mainly required to derive the statistical guarantees; thus we did not implement it for simplicity. In practice, we solve the following relaxed problem:

$$(\hat{\alpha}, \hat{\mathbf{\Theta}}) \in \operatorname{argmin}_{(\alpha,\mathbf{\Theta})} F(\alpha, \mathbf{\Theta}). \tag{6.14}$$

Quadratic approximation. For any $(\alpha, \mathbf{\Theta}) \in \mathbb{R}^N \times \mathbb{R}^{m_1 \times m_2}$ and for any direction $(d_\alpha, d_{\mathbf{\Theta}}) \in \mathbb{R}^N \times \mathbb{R}^{m_1 \times m_2}$, consider the following local approximation of the data fitting term

$$f(\alpha + d_\alpha, \mathbf{\Theta} + d_{\mathbf{\Theta}}) = f(\alpha, \mathbf{\Theta}) + \mathcal{A}(\mathsf{f}_U(\alpha) + \mathbf{\Theta}, d_\alpha, d_{\mathbf{\Theta}}) + o(\|d_\alpha\|_2^2 + \|d_{\mathbf{\Theta}}\|_F^2)\,, \tag{6.15}$$

where we have set

$$\begin{aligned}\mathcal{A}(X, d_\alpha, d_{\mathbf{\Theta}}) = &-2 \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} w_{ij}[\boldsymbol{X}_{i,j}] Z_{ij}[\boldsymbol{X}_{i,j}] (\mathsf{f}_U(d_\alpha)_{i,j} + d_{\mathbf{\Theta} i,j}) \\ &+ \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} w_{ij}[\boldsymbol{X}_{i,j}] (\mathsf{f}_U(d_\alpha)_{i,j} + d_{\mathbf{\Theta} i,j})^2 + \nu \|d_\alpha\|_2^2 + \nu \|d_{\mathbf{\Theta}}\|_F^2. \end{aligned} \tag{6.16}$$

In (6.16), $\nu > 0$ is a positive constant and for $x \in \mathbb{R}$ and $(i, j) \in [\![m_1]\!] \times [\![m_2]\!]$,

$$w_{ij}[x] = \Omega_{i,j} g_j''(x)/2\,, \quad Z_{ij}[x] = (\boldsymbol{Y}_{i,j} - g_j'(x))/g_j''(x)\,. \tag{6.17}$$

Note that the approximation (6.16) is simply a Taylor expansion of $\mathcal{L}$ around $\boldsymbol{X}$, with an additional quadratic term $\nu \|d_\alpha\|_2^2 + \nu \|d_{\mathbf{\Theta}}\|_F^2$ ensuring its strong convexity. Denote by $(\alpha^{[t]}, \mathbf{\Theta}^{[t]})$ the fit of the parameter at iteration t and set $\boldsymbol{X}^{[t]} = \mathsf{f}_U(\alpha^{[t]}) + \mathbf{\Theta}^{[t]}$. We update α and L alternatively as follows.

α-Update. We first solve

$$d_\alpha^{[t]} \in \underset{d\in\mathbb{R}^N}{\arg\min} \left\{ \mathcal{A}(\boldsymbol{X}^{[t]}, d, 0) + \lambda_2 \|\alpha^{[t]} + d\|_1 \right\} . \tag{6.18}$$

Problem (6.18) may be rewritten as a weighted Lasso problem:

$$\underset{\alpha\in\mathbb{R}^d}{\arg\min} \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_{ij}[\boldsymbol{X}_{i,j}^{[t]}](Z_{ij}^{[t]} - [\mathsf{f}_U(\alpha)]_{i,j})^2 + \nu\|\alpha^{[t]} - \alpha\|_2^2 + \lambda_2\|\alpha\|_1 ,$$

where for $i, j \in [\![m_1]\!] \times [\![m_2]\!]$ we have set $Z_{ij}^{[t]} := Z_{ij}[\boldsymbol{X}_{i,j}^{[t]}] + \mathsf{f}_U(\alpha^{[t]})$. Efficient numerical solutions to this problem are available (see, e.g., Friedman et al. (2010)). To update $\alpha^{[t]}$, we select a step size with an Armijo line search. The procedure goes as follows. We choose $\tau_{\mathsf{init}} > 0$ and we let $\tau_\alpha^{[t]}$ be the largest element of $\{\tau_{\mathsf{init}}\beta^j\}_{j=0}^\infty$ satisfying

$$f(\alpha^{[t]} + \tau_\alpha^{[t]} d^{[t]}, \boldsymbol{\Theta}^{[t]}) + \lambda_2\|\alpha^{[t]} + \tau_\alpha^{[t]} d^{[t]}\|_1 \leq f(\alpha^{[t]}, \boldsymbol{\Theta}^{[t]}) + \lambda_2\|\alpha^{[t]}\|_1 + \tau_\alpha^{[t]}\zeta\Gamma_\alpha^{[t]},$$

where $0 < \beta < 1$, $0 < \zeta < 1$, $0 \leq \theta < 1$, and

$$\Gamma_\alpha^{[t]} := -2\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_{ij}[\boldsymbol{X}_{i,j}^{[t]}] Z_{ij}[\boldsymbol{X}_{i,j}^{[t]}][\mathsf{f}_U(d^{[t]})]_{i,j} + \theta \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_{ij}[\boldsymbol{X}_{i,j}^{[t]}] \left[\mathsf{f}_U(d^{[t]})\right]_{i,j}^2 + \nu\|d^{[t]}\|_2^2 + \lambda_2 \left\{ \|\alpha^{[t]} + d^{[t]}\|_1 - \|\alpha^{[t]}\|_1 \right\} .$$

We set $\alpha^{[t+1]} = \alpha^{[t]} + \gamma^{[t]} d_\alpha^{[t]}$ and $\boldsymbol{X}^{[t+1/2]} = \mathsf{f}_U(\alpha^{[t+1]}) + \boldsymbol{\Theta}^{[t]}$.

L-Update. We first solve

$$d_{\boldsymbol{\Theta}}^{[t]} := \underset{d\in\mathbb{R}^{m_1\times m_2}}{\arg\min} \left\{ \mathcal{A}(\boldsymbol{X}^{[t+1/2]}, 0, d) + \lambda_1\|\boldsymbol{\Theta}^{[t]} + d\|_* \right\} , \tag{6.19}$$

which is equivalent to

$$\underset{\boldsymbol{\Theta}\in\mathbb{R}^{m_1\times m_2}}{\arg\min} \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} (\nu + w_{ij}[\boldsymbol{X}_{i,j}^{[t+1/2]}])(Z_{ij}^{[t+1/2]} - \boldsymbol{\Theta}_{i,j})^2 + \lambda_1\|\boldsymbol{\Theta}\|_*, \tag{6.20}$$

where for $i, j \in [\![m_1]\!] \times [\![m_2]\!]$ we have set

$$Z_{ij}^{[t+1/2]} = \frac{w_{ij}[\boldsymbol{X}_{i,j}^{[t+1/2]}](Z_{ij}[\boldsymbol{X}^{[t+1/2]}] + \boldsymbol{\Theta}_{i,j}^{[t]}) + \nu\boldsymbol{\Theta}_{i,j}^{[t]}}{\nu + w_{ij}[\boldsymbol{X}_{i,j}^{[t+1/2]}]}.$$

The minimisation problem (6.20) may be seen as a weighted version of `softImpute` (Hastie et al., 2015). Srebro and Jaakkola (2003) proposed to solve (6.20) using an EM algorithm where the weights in $[0, 1]$ are viewed as frequencies of observations in a missing value framework (see also Mazumder et al. (2010)). We use this procedure, which involves soft-thresholding of the singular values of $\boldsymbol{\Theta}$, by adapting the `softImpute` package (Hastie et al., 2015). To update $\boldsymbol{\Theta}^{[t]}$, we choose the step size using again the Armijo line search. We set $\tau_{\mathsf{init}} > 0$ and let $\tau_{\boldsymbol{\Theta}}^{[t]}$ be the largest element of $\{\tau_{\mathsf{init}}\beta^j\}_{j=0}^\infty$ satisfying

$$f(\alpha^{[t+1]}, \boldsymbol{\Theta}^{[t]} + \tau_L^{[t]} d_{\boldsymbol{\Theta}}^{[t]}) + \lambda_1\|\boldsymbol{\Theta}^{[t]} + \gamma^{[t]} d_{\boldsymbol{\Theta}}^{[t]}\|_* \leq f(\alpha^{[t+1]}, \boldsymbol{\Theta}^{[t]}) + \lambda_1\|\boldsymbol{\Theta}^{[t]}\|_* + \tau_{\boldsymbol{\Theta}}^{[t]}\zeta\Gamma_{\boldsymbol{\Theta}}^{[t]},$$

$$\Gamma_{\boldsymbol{\Theta}}^{[t]} := -2\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_{ij}[\boldsymbol{X}_{i,j}^{[t+1/2]}] Z_{ij}[\boldsymbol{X}_{i,j}^{[t+1/2]}] d_{\boldsymbol{\Theta}_{i,j}}^{[t]} + \theta \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} w_{ij}[\boldsymbol{X}_{i,j}^{[t+1/2]}] {d_{\boldsymbol{\Theta}_{i,j}}^{[t]}}^2 + \lambda_1 \left\{ \|\boldsymbol{\Theta}^{(t)} + d_{\boldsymbol{\Theta}}^{[t]}\|_* - \|d_{\boldsymbol{\Theta}}^{[t]}\|_* \right\} .$$

We finally set $\boldsymbol{\Theta}^{[t+1]} = \boldsymbol{\Theta}^{[t]} + \tau_{\boldsymbol{\Theta}}^{[t]} d_{\boldsymbol{\Theta}}^{[t]}$.

6.3.2 Convergence of the BCGD algorithm

The algorithm described in Section 6.3.1 is a particular case of the coordinate gradient descent method for nonsmooth minimisation introduced in Tseng and Yun (2009). In the aforementioned paper, the authors studied (in the general case) the convergence of the iterate sequence to a stationary point of the objective function. Here, we apply their general result (Tseng and Yun, 2009, Theorem 1) to our problem to obtain global convergence guarantees. Consider the following assumption on the dictionary $\mathcal{U}$.

H4. *For all $k \in [\![N]\!]$ and $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, $\boldsymbol{U}_{i,j}^k \in [-1,1]$ and there exists $æ > 0$ such that for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, $\sum_{k=1}^N |\boldsymbol{U}_{i,j}^k| \leq æ$.*

Assumption **H**4 is satisfied in the three models introduced in Examples 1, 2 and 3: for group effects and corruptions with $æ = 1$ and for row and column effects with $æ = 2$. In particular, it guarantees that $\boldsymbol{X}^0 = \mathsf{f}_U(\alpha^0) + \boldsymbol{\Theta}^0$ satisfies $\|\boldsymbol{X}^0\|_\infty \leq (1+æ)a$. Plugging this in the definition of $\boldsymbol{X}^0$ in (6.2), this assumption also implies that $\mathbb{E}[\boldsymbol{Y}_{i,j}] \in g_j'([-(1+æ)a, (1+æ)a])$ for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$. Note that **H**4 can be relaxed by $\|\boldsymbol{U}_k\|_\infty \leq \rho$, with ρ an arbitrary constant. Consider also the following assumption on the link functions.

H5. *For all $j \in [\![m_2]\!]$ the functions g_j twice continuously differentiable. Moreover, there exist $0 < \sigma_-, \sigma_+ < +\infty$ such that for all $|x| \leq (1+æ)a$ and $j \in [\![m_2]\!]$, $\sigma_-^2 \leq g_j''(x) \leq \sigma_+^2$.*

Assumptions **H**4–5 imply that the data-fitting term has Lipschitz gradient. Furthermore, the quadratic approximation defined in (6.16) is strictly convex at every iteration. We obtain the following convergence result.

Theorem 8. *Assume **H**4–5 and let $\{(\alpha^{[k]}, \boldsymbol{\Theta}^{[k]})\}$ be the iterate sequence generated by the BCGD algorithm. Then the following results hold.*

(a) $\{(\alpha^{[k]}, \boldsymbol{\Theta}^{[k]})\}$ has at least one accumulation point. Furthermore, all the accumulation points of $\{(\alpha^{[k]}, \boldsymbol{\Theta}^{[k]})\}$ are global optima of F.

(b) $\{F(\alpha^{[k]}, \boldsymbol{\Theta}^{[k]})\} \to F(\hat{\alpha}, \hat{\boldsymbol{\Theta}})$.

Proof. See Section 6.7.2. □

6.4 Statistical guarantees

We now state our main statistical results. Denote by $\langle \cdot, \cdot \rangle$ the usual trace scalar product in $\mathbb{R}^{m_1 \times m_2}$. For $a \geq 0$ and a sparsity pattern $\mathcal{I} \subset [\![N]\!]$, define the following sets

$$\begin{aligned}
\mathcal{E}_1(a, \mathcal{I}) &= \left\{\alpha \in \mathbb{R}^N, \|\alpha\|_\infty \leq a, \operatorname{supp}(\alpha) \subset \mathcal{I}\right\}, \\
\mathcal{E}_2(a, \mathcal{I}) &= \left\{\boldsymbol{\Theta} \in \mathbb{R}^{m_1 \times m_2}, \|\boldsymbol{\Theta}\|_\infty \leq a, \max_{k \in \mathcal{I}} |\langle \boldsymbol{\Theta}, \boldsymbol{U}_k \rangle| = 0, \right\}, \\
\mathcal{X}(a, \mathcal{I}) &= \left\{\boldsymbol{X} = \mathsf{f}_U(\alpha) + \boldsymbol{\Theta}; (\alpha, \boldsymbol{\Theta}) \in \mathcal{E}_1(a, \mathcal{I}) \times \mathcal{E}_2(a, \mathcal{I})\right\}.
\end{aligned} \tag{6.21}$$

H6. *There exist $a > 0$ and $\mathcal{I} \subset [\![N]\!]$ such that $(\alpha^0, \boldsymbol{\Theta}^0) \in \mathcal{E}_1(a, \mathcal{I}) \times \mathcal{E}_2(a, \mathcal{I})$.*

Assumption **H**6 can be relaxed to allow upper bounds to depend on the entries of α^0 and $\boldsymbol{\Theta}^0$, but we stick to **H**6 for simplicity.

6.4.1 Upper bounds

We now derive upper bounds for the Frobenius and ℓ_2 norms of the estimation errors $\boldsymbol{\Theta}^0 - \hat{\boldsymbol{\Theta}}$ and $\alpha^0 - \hat{\alpha}$ respectively. In Theorem 9 we give a general result under conditions on the regularization parameters λ_1 and λ_2, which depend on the random matrix $\nabla\mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)$. Then, Lemma 7 and 8 allow us to compute values of λ_1 and λ_2 that satisfy the assumptions of Theorem 9 with high probability. Finally we combine these results in Theorem 10.

We denote $\vee$ and $\wedge$ the $\max$ and $\min$ operators respectively, $M = m_1 \vee m_2$, $m = m_1 \wedge m_2$ and $d = m_1 + m_2$. We also define $r = \operatorname{rank}(\boldsymbol{\Theta}^0)$, $s = \|\alpha^0\|_0$ and $u = \max_k \|\boldsymbol{U}_k\|_1$. Let $(E_{ij})_{(i,j)\in[\![m_1]\!]\times[\![m_2]\!]}$ be the canonical basis of $\mathbb{R}^{m_1\times m_2}$ and $\{\epsilon_{ij}\}$ an i.i.d. Rademacher sequence independent of $\boldsymbol{Y}$ and Ω. Define

$$\Sigma_R = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \Omega_{i,j}\epsilon_{ij}E_{ij} \quad \text{and} \quad \nabla\mathcal{L}(\boldsymbol{X}; \boldsymbol{Y}, \Omega) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \Omega_{i,j}\left\{-\boldsymbol{Y}_{i,j} + g_j'\left(\boldsymbol{X}_{i,j}\right)\right\}E_{ij}. \tag{6.22}$$

Σ_R is a random matrix associated with the missingness pattern and $\nabla\mathcal{L}(\boldsymbol{X}; \boldsymbol{Y}, \Omega)$ is the gradient of $\mathcal{L}$ with respect to $\boldsymbol{X}$. Define also

$$\begin{aligned}
\theta_1 &= \frac{\lambda_2}{\sigma_-^2} + a^2 u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + \frac{p}{\|\alpha^0\|_1}\left(\frac{a}{p}\right)^2 \log(d),\\
\theta_2 &= \lambda_1^2 + (1+æ)a\mathbb{E}\left[\|\Sigma_R\|^2\right],\\
\theta_3 &= \frac{\lambda_2}{\lambda_1} + 2\theta_1.
\end{aligned}$$

Theorem 9. *Assume **H**4-6 and let*

$$\lambda_1 \geq 2\|\nabla\mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\| \quad \textit{and} \quad \lambda_2 \geq 2u\left(\|\nabla\mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\|_\infty + 2\sigma_+^2(1+æ)a\right).$$

Then, with probability at least $1 - 8d^{-1}$,

$$\|\mathsf{f}_U(\alpha^0) - \mathsf{f}_U(\hat{\alpha})\|_F^2 \leq \frac{as}{p}C_1\theta_1 \quad \textit{and} \quad \|\boldsymbol{\Theta}^0 - \hat{\boldsymbol{\Theta}}\|_F^2 \leq \frac{r}{p^2}C_2\theta_2 + \frac{as}{p}C_3\theta_3, \tag{6.23}$$

where C_1, C_2 *and* C_3 *are numerical constants independent of* m_1, m_2 *and* p.

Proof. See Section 6.7.2. □

We now give deterministic upper bounds on $\mathbb{E}\left[\|\Sigma_R\|\right]$ and $\mathbb{E}\left[\|\Sigma_R\|_\infty\right]$ in Lemma 7, and probabilistic upper bounds on $\|\nabla\mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\|$ and $\|\nabla\mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\|_\infty$ in Lemma 8. We will use them to select values of λ_1 and λ_2 which satisfy the assumptions of Theorem 9 and compute the corresponding upper bounds.

Lemma 7. *There exists an absolute constant* C^* *such that the two following inequalities hold*

$$\mathbb{E}\left[\|\Sigma_R\|_\infty\right] \leq 1 \quad \textit{and} \quad \mathbb{E}\left[\|\Sigma_R\|\right] \leq C^*\left\{\sqrt{\beta} + \sqrt{\log m}\right\}.$$

Proof. See Section 6.7.2 □

Lemma 8. *Assume **H**4-6. Then, there exists an absolute constant c^* such that the following two inequalities hold with probability at least $1 - d^{-1}$:*

$$\begin{aligned} \|\nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\|_\infty &\leq 6 \max\left\{\sigma_+ \sqrt{\log d}, \frac{\log d}{\gamma}\right\}, \\ \|\nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\| &\leq c^* \max\left\{\sigma_+ \sqrt{\beta \log d}, \frac{\log d}{\gamma} \log\left(\frac{1}{\sigma_-}\sqrt{\frac{m_1 m_2}{\beta}}\right)\right\}, \end{aligned} \tag{6.24}$$

*where $d = m_1 + m_2$, σ_+ and γ are defined in **H** 5, and β in* (6.8)*.*

Proof. See Section 6.7.2. □

We now combine Theorem 9, Lemma 7 and 8 with a union bound argument to derive upper bounds on $\|\mathsf{f}_U(\alpha^0) - \mathsf{f}_U(\hat{\alpha})\|_F^2$ and $\|\boldsymbol{\Theta}^0 - \hat{\boldsymbol{\Theta}}\|_F^2$. We assume that $M = (m_1 \vee m_2)$ is large enough, that is

$$M \geq \max\left\{\frac{4\sigma_+^2}{\gamma^6} \log^2\left(\frac{\sqrt{m}}{p\gamma\sigma_-}\right), 2\exp\left(\sigma_+^2/\gamma^2 \vee \sigma_+^2\gamma(1 + æa)\right)\right\}.$$

Define

$$\begin{aligned} \phi_1 &= a^2 + \frac{\log(d)}{u\sigma_-^2\gamma} + \frac{a^2\log(d)}{pu\|\alpha^0\|_1}, \\ \phi_2 &= \frac{\sigma_+^2}{\sigma_-^4}\log(d) + (1 + æ)a\left(1 \vee (\log m/\beta)\right), \\ \phi_3 &= \frac{12p\sqrt{\log(d)}}{\gamma(1 + æ)a\sigma_+\sqrt{\beta}} + \frac{1}{\sigma_-^2}\left(\frac{\log d}{\gamma}\right) + \frac{p}{u}\frac{\log(d)}{u\sigma_-^2\gamma} + \frac{a^2\log(d)}{pu\|\alpha^0\|_1}, \end{aligned}$$

and recall that $s = \|\alpha^0\|_0$, $r = \operatorname{rank}(\boldsymbol{\Theta}^0)$, $\beta \geq \max_{i,j}\left(\sum_{l=1}^{m_2}\pi_{il}, \sum_{k=1}^{m_1}\pi_{kj}\right)$ and that the entries $\boldsymbol{Y}_{i,j}$ are sub-exponential with scale parameter γ.

Theorem 10. *Assume **H**4-6 and let*

$$\lambda_1 = 2c^*\sigma_+\sqrt{\beta\log d}, \quad \lambda_2 \geq \frac{24u\log(d)}{\gamma},$$

where c_* is the absolute constant defined in Lemma 8. Then, with probability at least $1 - 10d^{-1}$,

$$\|\mathsf{f}_U(\alpha^0) - \mathsf{f}_U(\hat{\alpha})\|_F^2 \leq C\frac{sau}{p}\phi_1, \text{ and } \|\boldsymbol{\Theta}^0 - \hat{\boldsymbol{\Theta}}\|_F^2 \leq C\left(\frac{r\beta}{p^2}\phi_2 + \frac{sau}{p}\phi_3\right), \tag{6.25}$$

with C an absolute constant.

Denoting by $\lesssim$ the inequality up to constant and logarithmic factors we get:

$$\|\mathsf{f}_U(\alpha^0) - \mathsf{f}_U(\hat{\alpha})\|_F^2 \lesssim \frac{su}{p}, \text{ and } \|\boldsymbol{\Theta}^0 - \hat{\boldsymbol{\Theta}}\|_F^2 \lesssim \frac{r\beta}{p^2} + \frac{su}{p},$$

In the case of almost uniform sampling, *i.e.*, for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$ and two positive constants c_1 and c_2, $c_1 p \leq \pi_{ij} \leq c_2 p$ we obtain that $\beta \leq c_2 M p$ and the following simplified bound:

$$\|\boldsymbol{\Theta}^0 - \hat{\boldsymbol{\Theta}}\|_F^2 \lesssim \frac{rM}{p} + \frac{su}{p}. \tag{6.26}$$

The rate given in (6.26) is the sum of the usual convergence rate of low-rank matrix completion rM/p and of the usual sparse vector convergence rate s (Bühlmann and van de Geer, 2011; Tsybakov, 2008) multiplied by u/p. This additional factor accounts for missing observations (p^{-1}) and interplay between main effects and interactions (u). Furthermore, the estimation risk of $\mathsf{f}_U(\alpha^0)$ is also the usual sparse vector convergence rate, with an additional up^{-1} factor accounting for interactions and missing values.

Note that whenever the dictionary $\mathcal{U}$ is linearly independent, Theorem 10 also provides an upper bound on the estimation error $\alpha^0 - \hat{\alpha}$. Let $G \in \mathbb{R}^{N\times N}$ be the Gram matrix of the dictionary $\mathcal{U}$ defined by $G_{kl} = \langle U_k, U_l\rangle$ for all $(k,l) \in [\![N]\!] \times [\![N]\!]$.

H7. *For $\kappa > 0$ and all $\alpha \in \mathbb{R}^N$, $\alpha^\top G\alpha \geq \kappa^2\|\alpha\|_2^2$.*

Recall that in the group effects model, we denote by I_h the set of rows which belong to group h. **H**7 is satisfied for the group effects model with $\kappa^2 = \min_h |I_h|$, the row and column effects model with $\kappa^2 = \min(m_1, m_2)$ and the corruptions model with $\kappa^2 = 1$. If **H**7 is satisfied then, Theorem 10 implies that (up to constant and logarithmic factors):

$$\|\alpha^0 - \hat{\alpha}\|_F^2 \lesssim \frac{su}{p\kappa^2}.$$

6.4.2 Lower bounds

To characterize the tightness of the convergence rates given in Theorem 10, we now provide lower bounds on the estimation errors. We need three additional assumptions.

H8. *The sampling of entries is uniform, i.e. for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, $\pi_{ij} = p$.*

H9. *There exists $\mathcal{I} \subset [\![N]\!]$, $a > 0$ and $\boldsymbol{X} \in \mathcal{X}_{\mathcal{I},a}$ such that for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, $\boldsymbol{Y}_{i,j} \sim \mathrm{Exp}^{(h_j,g_j)}(\boldsymbol{X}_{i,j})$.*

Denote $\tau = \max_k \sum_{l\neq k} |\langle \boldsymbol{U}_k, \boldsymbol{U}_l\rangle|$. Without loss of generality we assume $m_1 = m_1 \vee m_2 = M$. For all $\boldsymbol{X} \in \mathbb{R}^{m_1\times m_2}$ we denote $\mathbb{P}_{\boldsymbol{X}}$ the product distribution of $(\boldsymbol{Y}, \Omega)$ satisfying **H**8 and 9. Consider two integers $s \leq (m_1 \wedge m_2)/2$ and $r \leq (m_1 \wedge m_2)/2$. We define the following set

$$\mathcal{F}(r,s) = \bigcup_{|\mathcal{I}|\leq s} \left\{(\alpha, \boldsymbol{\Theta}) \in \mathcal{E}_1(a,\mathcal{I}) \times \mathcal{E}_2(a,\mathcal{I}); \mathrm{rank}(\boldsymbol{\Theta}) \leq r\right\}. \tag{6.27}$$

Theorem 11. *Assume **H**4-8 and $p \geq \frac{r}{m_1\wedge m_2}$. Then, there exists a constant $\delta > 0$ such that*

$$\inf_{\hat{\boldsymbol{\Theta}},\hat{\alpha}} \sup_{(\boldsymbol{\Theta}^0,\alpha^0)\in\mathcal{F}(r,s)} \mathbb{P}_{\boldsymbol{X}^0}\left(\|\boldsymbol{\Theta}^0 - \hat{\boldsymbol{\Theta}}\|_F^2 + \|\mathsf{f}_U(\alpha^0) - \mathsf{f}_U(\hat{\alpha})\|_F^2 > \psi_1\frac{rM}{p} + \psi_2\frac{s\kappa^2}{p}\right) \geq \delta, \tag{6.28}$$

$$\begin{aligned}\psi_1 &= C\min\left(\sigma_+^{-2}, \min(a,\sigma_+)^2\right),\\ \psi_2 &= C\left(\frac{1}{\sigma_+^2\left(\max_k\|\boldsymbol{U}^k\|_F^2 + 2\tau\right)} \wedge (a\wedge\sigma_+)^2\right).\end{aligned} \tag{6.29}$$

Proof. See Section 6.7.2. □

Model	Group effects	Row & col effects	Corruptions
u	$\max_h \lvert I_h \rvert$	M	1
$\lVert \boldsymbol{\Theta}^0 - \hat{\boldsymbol{\Theta}} \rVert_F^2 + \lVert \mathsf{f}_U(\alpha^0) - \mathsf{f}_U(\hat{\alpha}) \rVert_F^2$	$\frac{rM}{p} + \frac{s \max \lvert I_h \rvert}{p}$	$\frac{rM}{p} + \frac{sM}{p}$	$\frac{rM}{p} + \frac{s}{p}$

Table 6.2: Order of magnitude of the upper bound for Examples 1, 2 and 3 (up to logarithmic factors).

Model	Group effects	Row & col effects	Corruptions
u	$\max_h \lvert I_h \rvert$	M	1
$\max_k \lVert \boldsymbol{U}_k \rVert_F^2$	$\max_h \lvert I_h \rvert$	M	1
κ^2	$\min_h \lvert I_h \rvert$	m	1
$\lVert \boldsymbol{\Theta}^0 - \hat{\boldsymbol{\Theta}} \rVert_F^2 + \lVert \mathsf{f}_U(\alpha^0) - \mathsf{f}_U(\hat{\alpha}) \rVert_F^2$	$\frac{rM}{p} + \frac{s \min \lvert I_h \rvert}{p \max \lvert I_h \rvert}$	$\frac{rM}{p} + \frac{sm}{pM}$	$\frac{rM}{p} + \frac{s}{p}$

Table 6.3: Order of magnitude of the lower bound for Examples 1, 2 and 3.

6.4.3 Examples

We now specialize our theoretical results to Examples 1, 2 and 3 presented in Section 6.2.2. We compute the values of u, τ and $\max_k \lVert \boldsymbol{U}^k \rVert_F^2$ for the group effects, row and column effects and corruption models, and obtain the rates of Theorem 10 and Theorem 11 for these particular cases. Recall that in the group effects model, we denote by I_h the set of rows which belong to group h. The orders of magnitude are summarized in Table 6.2 for the upper bound and in Table 6.3 for the lower bound. Comparing Table 6.2 and Table 6.3 we see that the convergence rates obtained in Theorem 10 are minimax optimal across the three examples whenever $s < r$. Furthermore, in the corruptions model our rates are optimal (up to constant and logarithmic factors) for any values of r, s and M, and equal to the minimax rates derived in Klopp et al. (2017). In the case of group effects, the rates are optimal when $r > s \max_h \lvert I_h \rvert / M$ or when $\max_h \lvert I_h \rvert$ is of the order of a constant. When $s > rM / \max_h \lvert I_h \rvert$, we have an additional factor of the order $(\max_h \lvert I_h \rvert)^2 / \min_h \lvert I_h \rvert$ in the upper bound. Note that the bounds have the same dependence in the sparsity pattern s. In the row and column model, when $r < s$, we have an additional factor of the order s/r in the upper bound.

6.5 Numerical results

6.5.1 Estimation of main effects and interactions

We start by evaluating our method (referred to as "mimi": Main effects and Interactions in Mixed and Incomplete data) in terms of estimation of main effects and interactions. In this experiment, we focus on the group effects model presented in Section 6.2.1, with $H = 5$ groups of equal size. We select at random s non-zero coefficients in α^0, and construct a matrix $\boldsymbol{\Theta}^0$ of rank k. Then, $\boldsymbol{X}^0 = \sum_{h=1}^{H} \sum_{j=1}^{m_2} \alpha_{hj}^0 \boldsymbol{U}^{h,j} + \boldsymbol{\Theta}^0$, with $\boldsymbol{U}_{h,j}$, $1 \leq h \leq H$ and $1 \leq j \leq m_2$ defined in Example 1. Finally, every entry of the matrix is observed with probability p. We then evaluate the estimation errors $\lVert \alpha^0 - \hat{\alpha} \rVert_2^2$ and $\lVert \boldsymbol{\Theta}^0 - \hat{\boldsymbol{\Theta}} \rVert_F^2$ of estimator (6.11).

In a first experiment, we consider only numeric variables, and compare mimi to the following two-step method. In this alternative method, the main effects α^0 are estimated by the means of the variables taken by group; this corresponds to the preprocessing step

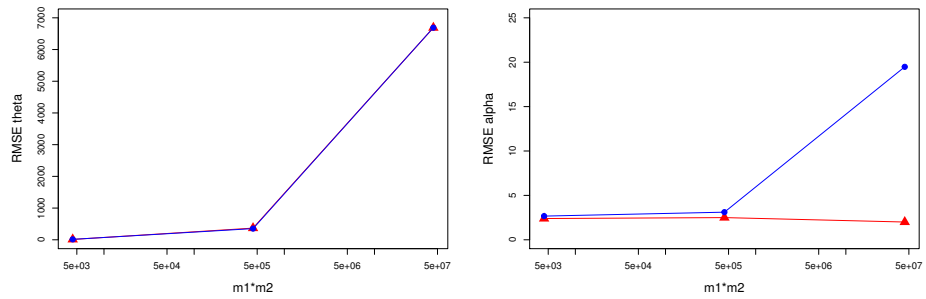


Figure 6.1: Estimation error of mimi (red triangles) and of groups means + SVD (blue points) for increasing problem sizes ($m_1 m_2$, in log scale).

performed in Udell et al. (2016) and Landgraf and Lee (2015) for instance. Then, $\mathbf{\Theta}^0$ is estimated using `softImpute` (Hastie et al., 2015); we refer to this method as "group mean + SVD". The regularization parameters of both methods are selected with cross-validation. The results are displayed in Figure 6.1 where we plot the estimation errors $\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}^0\|_F^2$ and $\|\hat{\alpha} - \alpha^0\|_2^2$ for increasing problem sizes. We observe that the excess risk $\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}^0\|_F^2$ the two methods are similar. Furthermore, the estimation error of mimi increases linearly with the largest dimension $m_1 \vee m_2$ as predicted by Theorem 10, when the rank r and probability p are fixed. In terms of estimation of α^0, mimi is superior. The estimation error of mimi is constant as the problem size increases but the sparsity level of α^0 is kept constant, as predicted by Theorem 10. On the contrary, we observe that estimating α^0 in a preprocessing step yields large errors in high dimensions.

6.5.2 Imputation of mixed data

To evaluate mimi in a mixed data setting, we compare it in terms of imputation of missing values to five state-of-the-art methods:

- softImpute (Hastie et al., 2015), a method based on soft-thresholding of singular values to impute numeric data implemented in the R package `softImpute`.
- Generalized Low-Rank Model (GLRM, Udell et al. (2016)), a matrix factorization framework for mixed data implemented in R in the `h2o` package.
- Factorial Analysis of Mixed Data (FAMD, Pagès (2015)), a principal component method for mixed data implemented in the R package `missMDA` (Josse and Husson, 2016).
- Multilevel Factorial Analysis of Mixed Data (MLFAMD, Husson et al. (2018)), an extension of FAMD to impute *multilevel* data, i.e. when individual are nested within groups. The method is also implemented in `missMDA`.
- Multivariate Imputation by Chained Equations (mice, van Buuren and Groothuis-Oudshoorn (2011)), an implementation of multiple imputation using Fully Conditional Specification. In the package `mice`, different models can be set for each column to account for mixed data.

To do so, we fix a dictionary $\mathcal{U}$ of indicator matrices corresponding to group effects (see Example 1), and generate a parameter matrix satisfying the decomposition (6.3). Then,

columns are sampled from different data types, namely Gaussian and Bernoulli. For varying proportions of missing entries and values of the ratio $\rho = \|\mathsf{f}_U(\alpha^0)\|_F / \|\boldsymbol{\Theta}^0\|_F$, we evaluate the six methods in terms of imputation error of the two different data types. The parameters of all the methods (number of components for GLRM and FAMD and regularization parameters for softImpute and mimi) are selected using cross-validation. In addition, we use an optional ridge regularization in the h2o implementation of the GLRM method, which penalizes the ℓ_2 norm of the left and right principal components ($\boldsymbol{U}$ and $\boldsymbol{V}$), and improved the imputation in practice. Note that we also add a comparison to imputation by the column means, in order to have a baseline reference. The details are available in the associated code provided as supplementary material.

% missing	20		
ρ	**0.2**	**1**	**5**
mean	24.5(0.7)	23.3(0.7)	22.9(0.4)
mimi	18.6(0.4)	**18.3**(0.3)	**17.7**(0.3)
GLRM	21.5(0.7)	22.0(0.8)	19.9(0.5)
softImpute	**18.5**(0.3)	18.5(0.2)	17.9(0.3)
FAMD	**18.5**(0.4)	18.9(0.4)	18.1(0.4)
MLFAMD	**18.5**(0.4)	19.2(0.4)	18.3(0.4)
mice	22.3(0.8)	22.6(0.6)	22.1(0.6)

Table 6.4: Imputation error (MSE) of mimi, GLRM, softImpute and FAMD for 20% of missing entries and different values of the ratio $\|\mathsf{f}_U(\alpha^0)\|_F / \|\boldsymbol{\Theta}^0\|_F$ $(0.2,\ 1,\ 5)$. The values are averaged across 100 replications and the standard deviation is given between parenthesis. In this simulation $m_1 = 150$, $m_2 = 30$, $s = 3$ and $r = 2$.

% missing	40		
ρ	**0.2**	**1**	**5**
mean	24.4(1.15)	33.2(1.1)	31.0(1.0)
mimi	18.8(0.3)	27.0(0.5)	**24.8**(0.6)
GLRM	21.5(0.7)	31.7(1.2)	31.0(0.9)
softImpute	18.6(0.3)	**26.8**(0.6)	24.9(0.5)
FAMD	18.7(0.3)	28.3(0.6)	25.6(0.7)
MLFAMD	**18.5**(0.5)	27.7(0.6)	26.3(0.5)
mice	22.7(0.6)	32.9(0.6)	30.1(0.9)

Table 6.5: Imputation error (MSE) of mimi, GLRM, softImpute and FAMD for 40% of missing entriesand different values of the ratio $\|\mathsf{f}_U(\alpha^0)\|_F / \|\boldsymbol{\Theta}^0\|_F$ $(0.2,\ 1,\ 5)$. The values are averaged across 100 replications and the standard deviation is given between parenthesis. In this simulation $m_1 = 150$, $m_2 = 30$, $s = 3$ and $r = 2$.

The results, presented in Tables 6.4-6.6, reveal that mimi, softImpute, FAMD and MLFAMD yield imputation errors of comparable order. In this simulation setting, our method mimi improves on these existing methods when the ratio $\rho = \|\mathsf{f}_U(\alpha^0)\|_F / \|\boldsymbol{\Theta}^0\|_F$ is large, i.e. when the scale of the main effects is large compared to the interactions. The size of this improvement also increases with the amount of missing values. The imputation error by data type (quantitative and qualitative) are given in Section 6.7.1, along with average experimental computational times of all the compared methods.

% missing	60		
ρ	**0.2**	**1**	**5**
mean	42.1(1.2)	40.7(1.2)	39.9(0.6)
mimi	36.0(1.0)	**33.7**(0.8)	**30.6**(0.4)
GLRM	44.5(10.8)	49.4(16.2)	50.7(3.2)
softImpute	34.9(1.0)	34.9(0.8)	32.2(0.5)
FAMD	36.0(1.5)	40.6(0.8)	32.7(0.5)
MLFAMD	**34.9**(1.3)	40.7(1.0)	33.5(0.6)
mice	48.1(2.4)	48.1(0.9)	44.7(1.4)

Table 6.6: Imputation error (MSE) of mimi, GLRM, softImpute and FAMD for 60% of missing entries and different values of the ratio $\|\mathsf{f}_U(\alpha^0)\|_F/\|\mathbf{\Theta}^0\|_F$ $(0.2,\ 1,\ 5)$. The values are averaged across 100 replications and the standard deviation is given between parenthesis. In this simulation $m_1 = 150$, $m_2 = 30$, $s = 3$ and $r = 2$.

6.5.3 Analysis of the Traumabase data set

We next apply our method on the Traumabase data presented in Section 6.2.1. We focus on a subsample of the original registry, with 2,120 patients treated for head trauma in six French hospitals from the Paris area (Beaujon, Bicêtre, Hôpital Européen George Pompidou (HEGP), Henri Mondor, Percy, and Pitié Salpêtrière). We also restrict ourselves to 9 variables (4 quantitative and 5 qualitative variables):

- Age of the patient: numeric, missing for 235 patients
- Sex of the patient: binary (male=1, female=0), missing for 235 patients
- Weight of the patient: numeric, missing for 513 patients
- Height of the patient: numeric, missing for 588 patients
- Body mass index (BMI): numeric, missing for 599 patients
- On-call: binary variable indicating whether the patient was treated during the day (1) or during on-call periods (0), missing for 203 patients
- KTV/TDM: binary variable indicating whether a catheter was placed before doing a CT scan (yes=1, no=0), missing for 711 patients
- PIC: binary variable indicating whether a small sensing device was placed to measure the intracranial pressure (yes=1, no=0), missing for 191 patients
- Death: binary variable indicating whether the patient died in intensive care (death=1, no death = 0), missing for 278 patients

We model the quantitative attributes using Gaussian distributions, and the binary attributes with Bernoulli distributions. Using the same notations as in Section 6.2.1, $c(i)$ denotes the group (the hospital center) to which patient i belongs. Thus, if the j-th column is continuous (Weight), our model implies:

$$\mathbb{E}[\boldsymbol{Y}_{i,j}] = \alpha^0_{c(i)j} + \mathbf{\Theta}^0_{i,j}.$$

If the j-th column is binary (PIC for instance), we model

$$\mathbb{P}(\boldsymbol{Y}_{i,j}=1)=\frac{\mathrm{e}^{\boldsymbol{X}^0_{i,j}}}{1+\mathrm{e}^{\boldsymbol{X}^0_{i,j}}},\quad \boldsymbol{X}^0_{i,j}=\alpha^0_{c(i)j}+\boldsymbol{\Theta}^0_{i,j}.$$

In Table 6.7, we display the value of the parameter $\alpha_{c(i)j}$ for all possible groups $c(i)$ and variables j. The value of $\alpha_{c(i)j}$ is related to the expected value $\mathbb{E}[\boldsymbol{Y}_{i,j}]$: everything else being fixed, $\mathbb{E}[\boldsymbol{Y}_{i,j}]$ is an increasing function of $\alpha_{c(i)j}$. Thus, in terms of interpretation, the "group effect" $\alpha_{c(i)j}$ indicates (everything else being equal) whether being treated in hospital $c(i)$ yields larger or smaller values for $\mathbb{E}[\boldsymbol{Y}_{i,j}]$ compared to other hospital centers.

	Age	Sex	Weight	Height	BMI	On-call	KTV/TDM	PIC	Death
Beaujon	0.51	0.16	0.12	0.20	0.18	-0.10	0	-0.02	-0.11
Bicêtre	-0.56	0.14	0.16	-0.06	-0.25	-0.14	0.08	-0.05	-0.12
HEGP	0	0.01	-0.09	-0.09	-0.04	0	0	-0.02	-0.04
Henri Mondor	0.2	0.04	0	0.02	0	0	0	0	0
Percy	0	0	0.03	0	0	0	0	0	0
Pitié Salpêtrière	0.02	0.11	-0.19	-0.30	-0.02	-0.07	0	-0.03	-0.09

Table 6.7: Main effect of hospital centers on other variables in the Traumabase (estimated with MIMI).

In Table 6.7, we observe that the main effects α_{cj} take nonzero values for some of the hospitals and some of the variables. For instance, for the Age variable, the effect of the Beaujon hospital is positive, and the effect of the Bicêtre hospital is negative. This indicates that, compared to the overall average, patients from Beaujon are older, while patients from Bicêtre are younger. Looking at the Sex variable, observe that some of the hospitals have positive effects (Beaujon and Bicêtre), which means that they treat more males (which are indicated by 1 in the Sex variable). The hospitals Pitié-Salpêtrière, Beaujon and Bicêtre also tend to treat more patients during on-call times (indicated by Day = 0) than the average of hospitals. These three hospitals also seem to resort less to the PIC procedure, and to have less deaths.

One may also use the MIMI results as dimensionality reduction tools to visualize the individuals. Figure 6.2 is a two-dimensional display, where each individual is represented in a Euclidean plane defined by the first principal components of the interaction matrix $\hat{\boldsymbol{\Theta}}$. The purple squares correspond to the center of gravity of each hospital center. We observe that these centers of gravity are very close to the origin, indicating that the hospital effects have been captures in the main effects (α), and that the interactions are now centered in each group. Then, on Figure 6.3, we represent the correlation between the original variables and the first principal components of the interaction matrix $\hat{\boldsymbol{\Theta}}$. The first direction is highly correlated with the Death, On-call and PIC variable. On the other hand, the second direction seems to correspond to a summary variable for physical characteristics of patients: it is highly correlated with the Weight, Height and BMI.

6.6 Conclusions and perspectives

This article introduces a general framework to analyze high-dimensonal, mixed and incomplete data frames with main effects and interactions. Upper bounds on the estimation error of main effects and interactions are derived; these bounds match with the

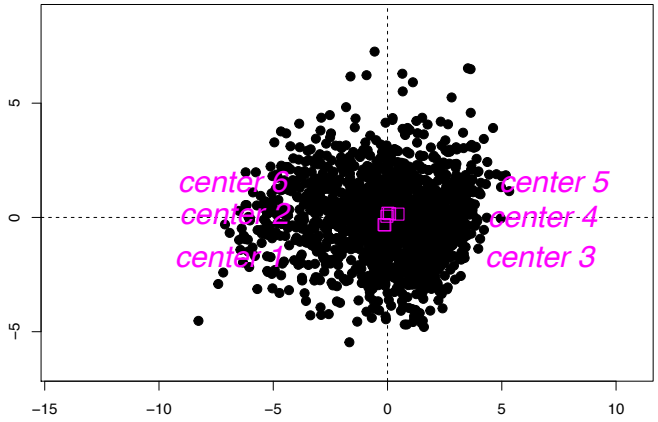


Figure 6.2: Two-dimensional display of the individuals (patients) in a Euclidean plane defined by the principal directions of the interaction matrix.

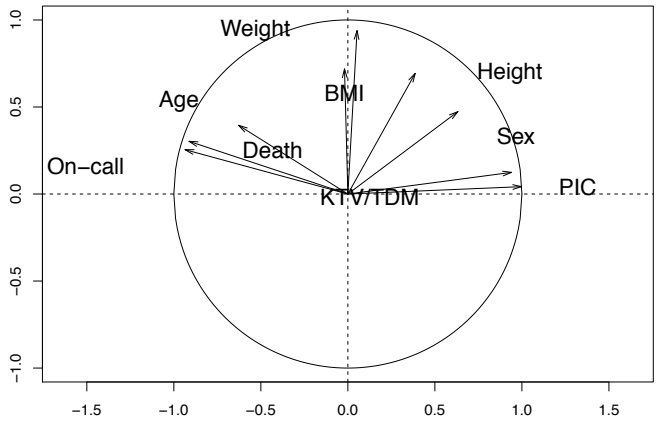


Figure 6.3: Correlation circle between the original variables and the principal directions of the interaction matrix.

lower-bounds under weak additional assumptions Our theoretical results are supported by a numerical experiments on synthetic and survey data, showing that the introduced method performs best when the proportion of missing values is large and the main effects and interactions are of comparable size.

Our work opens several directions of future research. A natural extension would be to consider the inference problem, i.e. to derive confidence intervals for the main effects coefficients. Another useful direction would be to consider exponential family distributions with multi-dimensional parameters, for example multinomial distributions, to incorporate categorical variables with more than two categories. One could also learn the scale parameter (which we currently assume fixed) adaptively.

6.7 Supplementary material

6.7.1 Additional experiments

In this section we provide more details on the simulations of Section 6.5.2. Tables 6.8, 6.9 and 6.10 present the imputation errors of the compared methods for quantitative

variables only, and the tables 6.11, 6.12 and 6.13 for binary variables. For the quantitative variables, mimi and MLFAMD, which both model main group effects, perform best. As already noticed in Section 6.5.2, mimi has smaller imputation errors than other methods when the size of the main effects compared to the interactions, and the proportion of missing entries, are both large. For the binary variables, suprisingly, softImpute outperforms consistently the other methods, although it is not designed for mixed data. Finally, Table 6.14 shows the average computational times of the different compared methods. We observe that the computational times of mimi, GLRM, FAMD and MLFAMD are of comparable order. The aforementioned methods are an order of magnitude slower than softImpute and mice.

% missing	**20**		
ρ	**0.2**	**1**	**5**
mean	20.7(1.3)	19.8(0.7)	19.6(0.6)
mimi	13.0(0.4)	**12.3**(0.4)	**11.4**(0.3)
GLRM	16.1(1.0)	16.9(0.7)	13.8(0.4)
softImpute	14.0(0.5)	14.0(0.4)	13.3(0.4)
FAMD	12.7(0.5)	12.9(0.6)	12.1(0.3)
MLFAMD	**12.6**(0.6)	13.7(0.6)	12.2(0.4)
mice	17.3(0.8)	17.2(1.0)	16.9(0.6)

Table 6.8: Quantitative variables: Imputation error (MSE) of mimi, GLRM, softImpute and FAMD for 20% of missing entries and different values of the ratio $\|\mathsf{f}_U(\alpha^0)\|_F/\|\mathbf{\Theta}^0\|_F$ (0.2, 1, 5). The values are averaged across 100 replications and the standard deviation is given between parenthesis.

% missing	**40**		
ρ	**0.2**	**1**	**5**
mean	28.0(2.6)	28.2(1.3)	26.9(1.1)
mimi	19.8(1.1)	**19.0**(0.7)	**16.1**(0.5)
GLRM	24.0(5.3)	24.5(1.5)	23.4(1.1)
softImpute	20.3(1.2)	20.9(0.7)	18.5(0.8)
FAMD	**19.2**(1.3)	20.2(0.6)	17.3(0.6)
MLFAMD	18.8(1.0)	19.7(0.6)	17.6(0.7)
mice	25.1(1.2)	26.0(0.7)	23.1(1.0)

Table 6.9: Quantitative variables: Imputation error (MSE) of mimi, GLRM, softImpute and FAMD for 40% of missing entries and different values of the ratio $\|\mathsf{f}_U(\alpha^0)\|_F/\|\mathbf{\Theta}^0\|_F$ (0.2, 1, 5). The values are averaged across 100 replications and the standard deviation is given between parenthesis.

6.7.2 Proofs

Proof of Theorem 8

To prove global convergence of the BCGD algorithm, we use a result from (Tseng and Yun, 2009, Theorem 1) summarized below in Theorem 12, combined with the compacity of the level sets of the objective F, proved using Lemma 9 and Lemma 10.

% missing	60		
ρ	**0.2**	**1**	**5**
mean	35.5(1.6)	34.2(1.3)	34.1(0.5)
mimi	27.1(1.0)	**24.3**(1.1)	**20.2**(0.4)
GLRM	36.5(12.3)	41.9(18.0)	44.1(3.7)
softImpute	27.3(1.2)	27.4(1.0)	24.4(0.5)
FAMD	26.9(1.8)	31.2(1.0)	22.7(0.4)
MLFAMD	**25.4**(1.5)	26.2(1.2)	23.5(0.6)
mice	40.7(2.8)	40.1(0.9)	36.8(1.8)

Table 6.10: Quantitative variables: Imputation error (MSE) of mimi, GLRM, softImpute and FAMD for 60% of missing entries and different values of the ratio $\|\mathsf{f}_U(\alpha^0)\|_F/\|\mathbf{\Theta}^0\|_F$ $(0.2,\ 1,\ 5)$. The values are averaged across 100 replications and the standard deviation is given between parenthesis.

% missing	20		
ρ	**0.2**	**1**	**5**
mean	13.0(0.3)	12.4(0.3)	**11.8**(0.4)
mimi	13.5(0.3)	13.5(0.3)	13.5(0.3)
GLRM	14.2(0.4)	14.1(0.6)	14.2(0.5)
softImpute	**12.2**(0.1)	**12.0**(0.3)	12.0(0.6)
FAMD	13.6(0.4)	13.8(0.4)	13.5(0.3)
MLFAMD	13.6(0.5)	13.5(0.4)	13.6(0.4)
mice	14.6(0.3)	14.5(0.4)	14.4(0.4)

Table 6.11: Binary variables: Imputation error (MSE) of mimi, GLRM, softImpute and FAMD for 20% of missing entries and different values of the ratio $\|\mathsf{f}_U(\alpha^0)\|_F/\|\mathbf{\Theta}^0\|_F$ $(0.2,\ 1,\ 5)$. The values are averaged across 100 replications and the standard deviation is given between parenthesis.

% missing	40		
ρ	**0.2**	**1**	**5**
mean	18.33(0.4)	17.4(0.3)	16.9(0.3)
mimi	18.9(0.5)	19.1(0.3)	18.9(0.6)
GLRM	20.0(0.4)	20.2(0.4)	20.4(0.3)
softImpute	**17.0**(0.3)	**16.7**(0.2)	**16.6**(0.4)
FAMD	19.2(0.5)	19.8(0.3)	18.8(0.6)
MLFAMD	19.4(0.5)	19.5(0.4)	19.6(0.5)
mice	20.5(0.4)	20.3(0.2)	20.5(0.4)

Table 6.12: Binary variables: Imputation error (MSE) of mimi, GLRM, softImpute and FAMD for 40% of missing entries and different values of the ratio $\|\mathsf{f}_U(\alpha^0)\|_F/\|\mathbf{\Theta}^0\|_F$ $(0.2,\ 1,\ 5)$. The values are averaged across 100 replications and the standard deviation is given between parenthesis.

Theorem 12. *Let* $\{(\alpha^{[k]}, \mathbf{\Theta}^{[k]})\}$ *be the current iterates,* $\{(d_\alpha^{[k]}, d_{\mathbf{\Theta}}^{[k]})\}$ *the descent directions and* $\{(\Gamma_\alpha^{[k]}, \Gamma_{\mathbf{\Theta}}^{[k]})\}$ *the functionals generated by the BCGD algorithm. Then the following results hold.*

% missing	60		
ρ	**0.2**	**1**	**5**
mean	22.6(0.5)	22.0(0.6)	**20.8**(0.6)
mimi	23.7(0.6)	23.4(0.5)	23.1(0.4)
GLRM	24.9(0.5)	25.1(0.6)	24.9(0.3)
softImpute	**21.6**(0.4)	**21.6**(0.3)	21.0(0.5)
FAMD	24.0(0.5)	25.0(0.4)	23.6(0.4)
MLFAMD	24.0(0.5)	24.1(0.4)	23.9(0.4)
mice	25.7(0.4)	25.7(0.6)	25.3(0.2)

Table 6.13: Binary variables: Imputation error (MSE) of mimi, GLRM, softImpute and FAMD for 60% of missing entries and different values of the ratio $\|\mathsf{f}_U(\alpha^0)\|_F/\|\boldsymbol{\Theta}^0\|_F$ (0.2, 1, 5). The values are averaged across 100 replications and the standard deviation is given between parenthesis.

method	mean	mimi	GLRM	softImpute	FAMD	MLFAMD	mice
time (s)	1.7e-4	6.6	5.5	0.1	2.6	3.5	0.2

Table 6.14: Computation time of the seven compared methods (averaged across 100 simulations).

(a) $\{F(\alpha^{[k]}, \boldsymbol{\Theta}^{[k]})\}$ *is nonincreasing and for all* k, $(\Gamma_\alpha^{[k]}, \Gamma_{\boldsymbol{\Theta}}^{[k]})$ *satisfies*

$$-\Gamma_\alpha^{[k]} \geq (1-\theta)\nu\|d_\alpha^{[k]}\|_2^2 \text{ and } -\Gamma_{\boldsymbol{\Theta}}^{[k]} \geq (1-\theta)\nu\|d_{\boldsymbol{\Theta}}^{[k]}\|_F^2.$$

(b) Every cluster point of $\{(\alpha^{[k]}, \boldsymbol{\Theta}^{[k]})\}$ *is a stationary point of* F.

Assumptions **H**4 and 5, combined with the separability of the ℓ_1 and nuclear norm penalties, guarantee that the conditions of (Tseng and Yun, 2009, Theorem 1) are satisfied. We now show that the data-fitting term $\mathcal{L}(\mathsf{f}_U(\alpha)+\boldsymbol{\Theta}; \boldsymbol{Y}, \Omega)$ is lower-bounded.

Lemma 9. *There exists a constant* $c > -\infty$ *such that, for all* $\boldsymbol{X} \in \mathbb{R}^{m_1 \times m_2}$, $\mathcal{L}(\boldsymbol{X}; \boldsymbol{Y}, \Omega) \geq c$.

Proof. Recall that $\mathcal{L}(\boldsymbol{X}; \boldsymbol{Y}, \Omega) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \Omega\{-\boldsymbol{Y}_{i,j}\boldsymbol{X}_{i,j} + g_j(\boldsymbol{X}_{i,j})\}$. Thus, we only need to prove that for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, the function $x \mapsto -\boldsymbol{Y}_{i,j}x + g_j(x)$ is lower bounded by a constant $c_{ij} > -\infty$. Assume that this is not the case; by the convexity of $x \mapsto -\boldsymbol{Y}_{i,j}x + g_j(x)$ we have that either $-\boldsymbol{Y}_{i,j}x + g_j(x) \underset{x\to+\infty}{\to} -\infty$ or $-\boldsymbol{Y}_{i,j}x + g_j(x) \underset{x\to-\infty}{\to} -\infty$. Assume without loss of generality that $-\boldsymbol{Y}_{i,j}x + g_j(x) \underset{x\to+\infty}{\to} -\infty$. Then, there exists $x_0 \in \mathbb{R}$ such that for all $x \geq x_0$, $-\boldsymbol{Y}_{i,j}x + g_j(x) < \log \int_{\substack{y\in\mathcal{Y}_j \\ y\geq \boldsymbol{Y}_{i,j}}} h_j(y)\mu_j(d_y)$. Thus, for all $x \geq \max(x_0, 0)$, we have that

$$\begin{aligned}\int_{y\in\mathcal{Y}_j} h_j(y)\mathrm{e}^{yx-g_j(x)}\mu_j(d_y) &= \int_{\substack{y\in\mathcal{Y}_j \\ y<\boldsymbol{Y}_{i,j}}} h_j(y)\mathrm{e}^{yx-g_j(x)}\mu_j(d_y) + \int_{\substack{y\in\mathcal{Y}_j \\ y\geq\boldsymbol{Y}_{i,j}}} h_j(y)\mathrm{e}^{yx-g_j(x)}\mu_j(d_y) \\ &> \int_{\substack{y\in\mathcal{Y}_j \\ y<\boldsymbol{Y}_{i,j}}} h_j(y)\mathrm{e}^{yx-g_j(x)}\mu_j(d_y) + 1 > 1,\end{aligned}$$

contradicting normality of the density $h_j(y)\mathrm{e}^{yx-g_j(x)}$. Thus, there exists $c_{ij} > -\infty$, such that for all $x \in \mathbb{R}$, $-\boldsymbol{Y}_{i,j}x + g_j(x) \geq c_{ij}$. Finally we obtain that $\mathcal{L}(\boldsymbol{X}; \boldsymbol{Y}, \Omega) \geq c = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} c_{ij}$. □

Finally, we use Lemma 9 to show the compactness of the level sets of the objective function F, defined for $C \in \mathbb{R}$ by

$$L_C = \{(\alpha, \mathbf{\Theta}) \in \mathbb{R}^N \times \mathbb{R}^{m_1 \times m_2}; F(\alpha, \mathbf{\Theta}) \leq C\}.$$

Lemma 10. *The level sets of the objective function F are compact.*

Proof. For all $(\alpha, \mathbf{\Theta}) \in \mathbb{R}^N \times \mathbb{R}^{m_1 \times m_2}$, $F(\alpha, \mathbf{\Theta}) \geq c + \lambda_1 \|\mathbf{\Theta}\|_* + \lambda_2 \|\alpha\|_1$, where c is the constant defined in Lemma 9. Thus, for all $C \in \mathbb{R}$, the level set L_C is included in the compact set

$$\left\{(\alpha, \mathbf{\Theta}) \in \mathbb{R}^N \times \mathbb{R}^{m_1 \times m_2}; \|\mathbf{\Theta}\|_* \leq \frac{C-c}{2\lambda_1} \text{ and } \|\alpha\|_1 \leq \frac{C-c}{2\lambda_2}\right\}.$$

Furthermore, by the continuity of F, the level set L_C is also a closed set. Thus we obtain that for all $C \in \mathbb{R}$, the level set L_C is compact. □

We can now combine Theorem 12, Lemma 9 and Lemma 10 to prove Theorem 8. Let $(\alpha^{[0]}, \mathbf{\Theta}^{[0]})$ be an initialization point. Theorem 12 (a) implies that the sequence $(\alpha^{[k]}, \mathbf{\Theta}^{[k]})$ generated by the BCGD algorithm lies in the level set of F

$$L_{F(\alpha^{[0]}, \mathbf{\Theta}^{[0]})} = \left\{(\alpha, \mathbf{\Theta}) \in \mathbb{R}^N \times \mathbb{R}^{m_1 \times m_2}; F(\alpha, \mathbf{\Theta}) \leq F(\alpha^{[0]}, \mathbf{\Theta}^{[0]})\right\}.$$

Furthermore, $L_{F(\alpha^{[0]}, \mathbf{\Theta}^{[0]})}$ is compact by Lemma 10, showing that the sequence $(\alpha^{[k]}, \mathbf{\Theta}^{[k]})$ has at least one accumulation point. Combined with Theorem 12 (b) and the convexity of F, this shows Theorem 8 (a).

Theorem 12 (a) and Lemma 9 combined imply that the sequence $\{F(\alpha^{[k]}, \mathbf{\Theta}^{[k]})\}$ converges to a limit F^*. Furthermore, Theorem 8 (a) and the continuity of F imply that there exists a sub-sequence $\{F(\alpha^{[k]}, \mathbf{\Theta}^{[k]})\}_{k \in \mathcal{K}}$ such that $\{F(\alpha^{[k]}, \mathbf{\Theta}^{[k]})\}_{k \in \mathcal{K}} \to F(\hat{\alpha}, \hat{\mathbf{\Theta}})$. Thus, $F^* = F(\hat{\alpha}, \hat{\mathbf{\Theta}})$, which proves Theorem 8 (b).

Proof of Theorem 9

Let $\Pi = (\pi_{ij})_{(i,j) \in [\![m_1]\!] \times [\![m_2]\!]}$ be the distribution of the mask Ω. For $B \in \mathbb{R}^{m_1 \times m_2}$ we denote B_Ω the projection of B on the set of observed entries. We define $\|B\|_\Omega^2 = \|B_\Omega\|_F^2$, and $\|B\|_\Pi^2 = \mathbb{E}\left[\|B\|_\Omega^2\right]$, where the expectation is taken with respect to Π. The proof of Theorem 9 will follow the subsequent two steps. We first derive an upper bound on the Frobenius error restricted to the observed entries $\|\Delta \boldsymbol{X}\|_\Omega^2$, then show that the expected Frobenius error $\|\Delta \boldsymbol{X}\|_\Pi^2$ is upper bounded by $\|\Delta \boldsymbol{X}\|_\Omega^2$ with high probability, and up to a residual term defined later on.

Let us derive the upper bound on $\|\Delta \boldsymbol{X}\|_\Omega^2$. By definition of $\hat{\mathbf{\Theta}}$ and $\hat{\alpha}$: $\mathcal{L}(\hat{\boldsymbol{X}}; \boldsymbol{Y}, \Omega) - \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega) \leq \lambda_1 \left(\|\mathbf{\Theta}^0\|_* - \|\hat{\mathbf{\Theta}}\|_*\right) + \lambda_2 \left(\|\alpha^0\|_1 - \|\hat{\alpha}\|_1\right)$. Recall that, for $\alpha \in \mathbb{R}^N$, we use the notation $\mathsf{f}_U(\alpha) = \sum_{k=1}^N \alpha_k \boldsymbol{U}^k$. Adding $\langle \nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega), \Delta \boldsymbol{X} \rangle$ on both sides of the last inequality, we get

$$\begin{aligned}\mathcal{L}(\hat{\boldsymbol{X}}; \boldsymbol{Y}, \Omega) - \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega) + \langle \nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega), \Delta \boldsymbol{X} \rangle \leq \\ \lambda_1 \left(\|\mathbf{\Theta}^0\|_* - \|\hat{\mathbf{\Theta}}\|_*\right) - \langle \nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega), \Delta \mathbf{\Theta} \rangle \\ + \lambda_2 \left(\|\alpha^0\|_1 - \|\hat{\alpha}\|_1\right) - \langle \nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega), \mathsf{f}_U(\Delta \alpha) \rangle. \quad (6.30)\end{aligned}$$

Assumption H5 implies that for any pair of matrices $\boldsymbol{X}^1$ and $\boldsymbol{X}^2$ in $\mathbb{R}^{m_1 \times m_2}$ satisfying $\|\boldsymbol{X}^1\|_\infty \vee \|\boldsymbol{X}^2\|_\infty \leq (1 + æ)a$, the two following inequalities hold for all Ω:

$$\mathcal{L}(\boldsymbol{X}; \boldsymbol{Y}, \Omega) - \mathcal{L}(\tilde{X}; \boldsymbol{Y}, \Omega) - \langle \nabla \mathcal{L}(\tilde{\boldsymbol{X}}; \boldsymbol{Y}, \Omega), \boldsymbol{X} - \tilde{\boldsymbol{X}} \rangle \geq \frac{\sigma_-^2}{2} \|\boldsymbol{X} - \tilde{\boldsymbol{X}}\|_\Omega^2, \quad (6.31)$$

$$\|\nabla \mathcal{L}(\boldsymbol{X}; \boldsymbol{Y}, \Omega) - \nabla \mathcal{L}(\tilde{\boldsymbol{X}}; \boldsymbol{Y}, \Omega)\|_F \leq \sigma_+^2 \|\boldsymbol{X} - \tilde{\boldsymbol{X}}\|_\Omega. \quad (6.32)$$

Plugging (6.31) into (6.30) allows to construct a lower bound on the left hand side term and obtain $\sigma_-^2 \|\Delta \boldsymbol{X}\|_\Omega^2 / 2 \leq \mathsf{A}_1 + \mathsf{A}_2$,

$$\begin{aligned} \mathsf{A}_1 &= \lambda_1 \left(\|\boldsymbol{\Theta}^0\|_* - \|\hat{\boldsymbol{\Theta}}\|_* \right) + \left| \langle \nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega), \Delta \boldsymbol{\Theta} \rangle \right|, \\ \mathsf{A}_2 &= \lambda_2 \left(\|\alpha^0\|_1 - \|\hat{\alpha}\|_1 \right) + \left| \langle \nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega), \mathsf{f}_U(\Delta \alpha) \rangle \right|. \end{aligned} \quad (6.33)$$

Let us upper bound A_1. The duality of the norms $\|\cdot\|_*$ and $\|\cdot\|$ implies that

$$\left| \langle \nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega), \Delta \boldsymbol{\Theta} \rangle \right| \leq \|\nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\| \|\Delta \boldsymbol{\Theta}\|_*.$$

Denote by S_1 and S_2 the linear subspaces spanned respectively by the left and right singular vectors of $\boldsymbol{\Theta}^0$, and $P_{S_1^\perp}$ and $P_{S_2^\perp}$ the orthogonal projectors on the orthogonal of S_1 and S_2, $P_{\boldsymbol{\Theta}^{0\perp}} : X \mapsto P_{S_1^\perp} \boldsymbol{X} P_{S_2^\perp}$ and $P_{\boldsymbol{\Theta}^0} : \boldsymbol{X} \mapsto \boldsymbol{X} - P_{S_1^\perp} \boldsymbol{X} P_{S_2^\perp}$. The triangular inequality yields

$$\begin{aligned} \|\hat{\boldsymbol{\Theta}}\|_* = \|\boldsymbol{\Theta}^0 - P_{\boldsymbol{\Theta}^{0\perp}}(\Delta \boldsymbol{\Theta}) - P_{\boldsymbol{\Theta}^0}(\Delta \boldsymbol{\Theta})\|_* \geq \\ \|\boldsymbol{\Theta}^0 + P_{\boldsymbol{\Theta}^{0\perp}}(\Delta \boldsymbol{\Theta})\|_* - \|P_{\boldsymbol{\Theta}^0}(\Delta \boldsymbol{\Theta})\|_*. \end{aligned} \quad (6.34)$$

Moreover, by definition of $P_{\boldsymbol{\Theta}^{0\perp}}$, the left and right singular vectors of $P_{\boldsymbol{\Theta}^{0\perp}}(\Delta \boldsymbol{\Theta})$ are respectively orthogonal to the left and right singular spaces of $\boldsymbol{\Theta}^0$, implying $\|\boldsymbol{\Theta}^0 + P_{\boldsymbol{\Theta}^{0\perp}}(\Delta \boldsymbol{\Theta})\|_* = \|\boldsymbol{\Theta}^0\|_* + \|P_{\boldsymbol{\Theta}^{0\perp}}(\Delta \boldsymbol{\Theta})\|_*$. Plugging this identity into (6.34) we obtain

$$\|\boldsymbol{\Theta}^0\|_* - \|\hat{\boldsymbol{\Theta}}\|_* \leq \|P_{\boldsymbol{\Theta}^0}(\Delta \boldsymbol{\Theta})\|_* - \|P_{\boldsymbol{\Theta}^{0\perp}}(\Delta \boldsymbol{\Theta})\|_*, \quad (6.35)$$

and $\mathsf{A}_1 \leq \lambda_1 \left(\|P_{\boldsymbol{\Theta}^0}(\Delta \boldsymbol{\Theta})\|_* - \|P_{\boldsymbol{\Theta}^{0\perp}}(\Delta \boldsymbol{\Theta})\|_* \right) + \|\nabla \mathcal{L}(\boldsymbol{X}^0; Y, \Omega)\| \|\Delta \boldsymbol{\Theta}\|_*$.

Using $\|\Delta \boldsymbol{\Theta}\|_* \leq \|P_{\boldsymbol{\Theta}^0}(\Delta \boldsymbol{\Theta})\|_* + \|P_{\boldsymbol{\Theta}^{0\perp}}(\Delta \boldsymbol{\Theta})\|_*$ and the assumption

$$\lambda_1 \geq 2 \|\nabla \mathcal{L}(\boldsymbol{X}^0; Y, \Omega)\|$$

we get $\mathsf{A}_1 \leq 3\lambda_1 \|P_{\boldsymbol{\Theta}^0}(\Delta \boldsymbol{\Theta})\|_* / 2$. In addition,

$$\|P_{\boldsymbol{\Theta}^0}(\Delta \boldsymbol{\Theta})\|_* \leq \sqrt{\operatorname{rank}\left(P_{\boldsymbol{\Theta}^0}(\Delta \boldsymbol{\Theta})\right)} \|P_{\boldsymbol{\Theta}^0}(\Delta \boldsymbol{\Theta})\|_F$$

, and $\operatorname{rank}\left(P_{\boldsymbol{\Theta}^0}(\Delta \boldsymbol{\Theta})\right) \leq 2 \operatorname{rank}\left(\boldsymbol{\Theta}^0\right)$ (see, e.g., (Klopp, 2014, Theorem 3)). Together with $\|P_{\boldsymbol{\Theta}^0}(\Delta \boldsymbol{\Theta})\|_F \leq \|\Delta \boldsymbol{\Theta}\|_F$, this finally implies the following upper bound:

$$\mathsf{A}_1 \leq \frac{3\lambda_1}{2} \sqrt{2r} \|\Delta \boldsymbol{\Theta}\|_F. \quad (6.36)$$

We now derive an upper bound for A_2. The duality between $\|\cdot\|_1$ and $\|\cdot\|_\infty$ ensures

$$\left| \langle \nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega), \mathsf{f}_U(\Delta \alpha) \rangle \right| \leq \|\Delta \alpha\|_1 \|\nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\|_\infty u. \quad (6.37)$$

The assumption $\lambda_2 \geq 2\|\nabla\mathcal{L}(\boldsymbol{X}^0;\boldsymbol{Y},\Omega)\|_\infty u$ in conjunction with (6.37) and the triangular inequality $\|\Delta\alpha\|_1 \leq \|\alpha^0\|_1 + \|\hat{\alpha}\|_1$ yield

$$\mathsf{A}_2 \leq \frac{3\lambda_2}{2}\|\alpha^0\|_1. \tag{6.38}$$

Combining inequalities (6.33), (6.36) and (6.38) we obtain

$$\|\Delta\boldsymbol{X}\|_\Omega^2 \leq \frac{3\lambda_1}{\sigma_-^2}\sqrt{2r}\|\Delta\boldsymbol{\Theta}\|_F + \frac{3\lambda_2}{\sigma_-^2}\|\alpha^0\|_1. \tag{6.39}$$

We now show that when the errors $\Delta\boldsymbol{\Theta}$ and $\Delta\alpha$ belong to a subspace $\mathcal{C}$ and for a residual D - both defined later on - the following holds with high probability:

$$\|\Delta\boldsymbol{X}\|_\Omega^2 \geq \|\Delta\boldsymbol{X}\|_\Pi^2 - \mathsf{D}. \tag{6.40}$$

We start by defining our constrained set and prove that it contains the errors $\Delta\boldsymbol{\Theta}$ and $\Delta\alpha$ with high probability (Lemma 11-12); then we show that restricted strong convexity holds on this subspace (Lemma 13). For non-negative constants d_1, d_Π, $\rho < m$ and ε that will be specified later on, define the two following sets where $\Delta\alpha$ and $\Delta\boldsymbol{\Theta}$ should lie:

$$\mathcal{A}(d_1, d_\Pi) = \left\{\alpha \in \mathbb{R}^N : \|\alpha\|_1 \leq d_1,\ \|\mathsf{f}_U(\alpha)\|_\Pi^2 \leq d_\Pi\right\}. \tag{6.41}$$

$$\begin{aligned}\mathcal{L}(\rho,\varepsilon) \quad = \Bigg\{ &\boldsymbol{\Theta} \in \mathbb{R}^{m_1\times m_2}, \alpha \in \mathbb{R}^N : \|L + \mathsf{f}_U(\alpha)\|_\Pi^2 \geq \frac{72\log(d)}{p\log(6/5)}, \\ &\|\boldsymbol{\Theta} + \mathsf{f}_U(\alpha)\|_\infty \leq 1, \|\boldsymbol{\Theta}\|_* \leq \sqrt{\rho}\|L\|_F + \varepsilon\Bigg\}\end{aligned} \tag{6.42}$$

If $\|\Delta\boldsymbol{X}\|_\Pi^2$ is too small, the right hand side of (6.40) is negative. The first inequality in the definition of $\mathcal{L}(\rho,\varepsilon)$ prevents from this. Condition $\|\boldsymbol{\Theta}\|_* \leq \sqrt{\rho}\|\boldsymbol{\Theta}\|_F + \varepsilon$ is a relaxed form of the condition $\|\boldsymbol{\Theta}\|_* \leq \sqrt{\rho}\|\boldsymbol{\Theta}\|_F$ satisfied for matrices of rank ρ. Finally, we define the constrained set of interest:

$$\mathcal{C}(d_1, d_\Pi, \rho, \varepsilon) = \mathcal{L}(\rho,\varepsilon) \cap \left\{\mathbb{R}^{m_1\times m_2} \times \mathcal{A}(d_1, d_\Pi)\right\}.$$

Recall $u = \max_k \|\boldsymbol{U}_k\|_1$ and let

$$d_1 = 4\|\alpha^0\|_1, \text{ and } d_\Pi = \frac{3\lambda_2}{\sigma_-^2}\|\alpha^0\|_1 + 64a^2u\mathbb{E}\left[\|\Sigma_R\|_\infty\right]\|\alpha^0\|_1 + 3072a^2p^{-1} + \frac{72a^2\log(d)}{\log(6/5)}.$$

Lemma 11. *Let $\lambda_2 \geq 2u\left(\|\nabla\mathcal{L}(\boldsymbol{X}^0;\boldsymbol{Y},\Omega)\|_\infty + 2\sigma_+^2(1+u)a\right)$ and assume **H**4-5 hold. Then, with probability at least $1 - 8d^{-1}$, $\Delta\alpha \in \mathcal{A}(d_1, d_\Pi)$.*

Proof. See Section 6.7.2. □

Lemma 11 implies the upper bound on $\|\Delta\alpha\|_2^2$ of Theorem 9. Thus, we only need to prove the upper bound on $\|\Delta\boldsymbol{\Theta}\|_F^2$. Let $\rho = 32r$ and $\varepsilon = 3\lambda_2/\lambda_1\|\alpha^0\|_1$.

Lemma 12. *Assume **H**5 and let*

$$\lambda_1 \geq 2\|\nabla\mathcal{L}(\boldsymbol{X}^0;\boldsymbol{Y},\Omega)\|, \quad \lambda_2 \geq 2u\left(\|\nabla\mathcal{L}(\boldsymbol{X}^0;\boldsymbol{Y},\Omega)\|_\infty + 2\sigma_+^2(1+u)a\right).$$

Then $\|\Delta\boldsymbol{\Theta}\|_* \leq \sqrt{\rho}\|\Delta\boldsymbol{\Theta}\|_F + \varepsilon$.

Proof. See Section 6.7.2 □

As a consequence, under the conditions on the regularization parameters λ_1 and λ_2 given in Lemma 12 and whenever $\|\Delta\mathbf{\Theta} + \mathsf{f}_U(\Delta\alpha)\|_\Pi^2 \geq 72\log(d)/(p\log(6/5))$, the error terms $(\Delta\mathbf{\Theta}, \Delta\alpha)$ belong to the constrained set $\mathcal{C}(d_1, d_\Pi, \rho, \varepsilon)$ with high probability.

Case 1: Suppose $\|\Delta\mathbf{\Theta} + \mathsf{f}_U(\Delta\alpha)\|_\Pi^2 < 72\log(d)/(p\log(6/5))$. Then, Lemma 11 combined with the fact that $\|\boldsymbol{M}\|_F^2 \leq p^{-1}\|\boldsymbol{M}\|_\Pi^2$ for all $\boldsymbol{M}$, and the identity $(a+b)^2 \geq a^2/4 - 4b^2$ ensures that $\|\Delta\mathbf{\Theta}\|_F^2 \leq 4\|\Delta\mathbf{\Theta} + \mathsf{f}_U(\Delta\alpha)\|_F^2 + 16\|\mathsf{f}_U(\Delta\alpha)\|_F^2$. Therefore we obtain (ii) of Theorem 9:

$$\|\Delta\mathbf{\Theta}\|_F^2 \leq \frac{288a^2\log(d)}{\log(6/5)} + 16\frac{\|\alpha^0\|_1}{p}\theta_1.$$

Case 2: Suppose $\|\Delta\mathbf{\Theta} + \mathsf{f}_U(\Delta\alpha)\|_\Pi^2 \geq 72\log(d)/(p\log(6/5))$. Then, Lemma 11 and 12 yield that with probability at least $1 - 8d^{-1}$,

$$\left(\frac{\Delta\mathbf{\Theta}}{2(1+\text{æ})a}, \frac{\Delta\alpha}{2(1+\text{æ})a}\right) \in \mathcal{C}(d_1', d_\Pi', \rho', \varepsilon'), \text{where}$$

$$d_1' = \frac{d_1}{2(1+\text{æ})a}, \quad d_\Pi' = \frac{d_\Pi}{4(1+\text{æ})^2a^2}, \quad \rho' = \rho, \quad \varepsilon' = \frac{\varepsilon}{2(1+\text{æ})a},$$

and where d_1, d_Π, ρ and ε are the same as in Lemma 11 and 12. We use the following result, proven in Section 6.7.2. Recall that we assume for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, $\mathbb{P}(\Omega_{i,j} = 1) \geq p$ and define:

$$\tilde{\mathcal{A}}(d_1) = \left\{\alpha \in \mathbb{R}^N : \quad \|\alpha\|_\infty \leq 1; \quad \|\alpha\|_1 \leq d_1; \quad \|\mathsf{f}_U(\alpha)\|_\Pi^2 \geq \frac{18\log(d)}{p\log(6/5)}\right\},$$

$$\begin{aligned} \mathsf{D}_\alpha &= 8\text{æ}d_1u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + 768p^{-1}, \\ \mathsf{D}_{\boldsymbol{X}} &= \frac{112\rho}{p}\mathbb{E}\left[\|\Sigma_R\|\right]^2 + 8\text{æ}\varepsilon\mathbb{E}\left[\|\Sigma_R\|\right] + 8\text{æ}d_1u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + d_\Pi + 768p^{-1}. \end{aligned} \tag{6.43}$$

Lemma 13. *(i) For any $\alpha \in \tilde{\mathcal{A}}(d_1)$, with probability at least $1 - 8d^{-1}$,*

$$\|\mathsf{f}_U(\alpha)\|_\Omega^2 \geq \frac{1}{2}\|\mathsf{f}_U(\alpha)\|_\Pi^2 - \mathsf{D}_\alpha.$$

(ii) For any pair $(\mathbf{\Theta}, \alpha) \in \mathcal{C}(d_1, d_\Pi, \rho, \varepsilon)$, with probability at least $1 - 8d^{-1}$

$$\|\mathbf{\Theta} + \mathsf{f}_U(\alpha)\|_\Omega^2 \geq \frac{1}{2}\|\mathbf{\Theta} + \mathsf{f}_U(\alpha)\|_\Pi^2 - \mathsf{D}_{\boldsymbol{X}}. \tag{6.44}$$

Proof. See Section 6.7.2. □

Lemma 13 (ii) applied to $\left(\frac{\Delta\mathbf{\Theta}}{2(1+\text{æ})a}, \frac{\Delta\alpha}{2(1+\text{æ})a}\right)$ implies that with probability at least $1 - 8d^{-1}$, $\|\Delta\boldsymbol{X}\|_\Pi^2 \leq 2\|\Delta\boldsymbol{X}\|_\Omega^2 + 4(1+\text{æ})a\mathsf{D}_{\boldsymbol{X}}$. Combined with (6.39), $\|\Delta\boldsymbol{X}\|_F^2 \leq p^{-1}\|\Delta\boldsymbol{X}\|_\Pi^2$, $\|\Delta\boldsymbol{X}\|_F^2 \geq \|\Delta\mathbf{\Theta}\|_F^2/2 - \|\mathsf{f}_U(\Delta\alpha)\|_F^2$ and $6\sqrt{2r}\lambda_1/(p\sigma_-^2)\|\Delta\mathbf{\Theta}\|_F \leq \|\Delta\mathbf{\Theta}\|_F^2/4 + 288r\lambda_1^2/(p^2\sigma_-^4)$, we obtain the result of Theorem 9 (ii):

$$\|\Delta\mathbf{\Theta}\|_F^2 \leq \frac{1152r\lambda_1^2}{p^2\sigma_-^4} + \frac{24\lambda_2\|\alpha^0\|_1}{p\sigma_-^2} + 4(1+\text{æ})a\mathsf{D}_{\boldsymbol{X}} + 4\frac{\|\alpha^0\|}{p}\Theta_1.$$

Proof of Theorem 11

We will establish separately two lower bounds of order rM/p and s/p respectively. Define

$$\tilde{\mathcal{L}} = \left\{ \tilde{\mathbf{\Theta}} \in \mathbb{R}^{m_1 \times r} : \tilde{\mathbf{\Theta}}_{i,j} \in \left\{ 0, \eta \min(a, \sigma_+) \left(\frac{r}{pm} \right)^{1/2} \right\}, \forall (i,j) \in [\![m_1]\!] \times [\![r]\!] \right\},$$

where $0 \leq \eta \leq 1$ will be chosen later. Define also the associated set of block matrices

$$\mathcal{L} = \left\{ \mathbf{\Theta} = (\tilde{\mathbf{\Theta}} | \ldots | \tilde{\mathbf{\Theta}} | O) \in \mathbb{R}^{m_1 \times m_2} : \tilde{\mathbf{\Theta}} \in \tilde{\mathcal{L}} \right\},$$

where O denotes the $m_1 \times (m_2 - r \lfloor m_2/r \rfloor)$ null matrix and, for some $x \in \mathbb{R}$, $\lfloor x \rfloor$ is the integer part of x. We also define the following set of vectors

$$\mathcal{A} = \left\{ \alpha = (\tilde{O} | \tilde{\alpha}) \in \mathbb{R}^N, \tilde{\alpha}_k \in \{0, \tilde{\eta} \min(a, \sigma_+)\} \ \forall 1 \leq k \leq s \right\},$$

with $\tilde{O} \in \mathbb{R}^{m_2 - s}$ denoting the null vector. Finally, we set

$$\mathcal{X} = \left\{ \boldsymbol{X} = \mathbf{\Theta} + \mathsf{f}_U(\alpha) \in \mathbb{R}^{m_1 \times m_2}, \alpha \in \mathcal{A}, \mathbf{\Theta} \in \mathcal{L} \right\}.$$

For any $\boldsymbol{X} \in \mathcal{X}$ there exists a matrix $\mathbf{\Theta} \in \mathcal{L}$ of rank at most r and a vector α with at most s non-zero components satisfying $\boldsymbol{X} = \mathbf{\Theta} + \mathsf{f}_U(\alpha)$. Furthermore, for any $\tilde{\boldsymbol{X}} \in \mathcal{X}$ there exists a matrix $\tilde{\mathbf{\Theta}} \in \mathcal{L}$ of rank at most r and a vector $\tilde{\alpha}$ with at most s non-zero components satisfying $\boldsymbol{X} - \tilde{\boldsymbol{X}} = \tilde{\mathbf{\Theta}} + \mathsf{f}_U(\tilde{\alpha})$. Finally, for all $\boldsymbol{X} \in \mathcal{X}$ and $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, $0 \leq \boldsymbol{X}_{i,j} \leq (1 + \text{æ})a$. Thus, $\mathcal{X} \subset \mathcal{F}(r,s)$, where $\mathcal{F}(r,s)$ is defined in (6.27).

Lower bound of order rM/p. Consider the set

$$\mathcal{X}_L = \{ \boldsymbol{X} = \mathbf{\Theta} + \mathsf{f}_U(\alpha) \in \mathcal{X}; \alpha = 0 \}.$$

Lemma 2.9 in Tsybakov (2008) (Varshamov Gilbert bound) implies that there exists a subset $\mathcal{X}_L^0 \subset \mathcal{X}_L$ satisfying $\mathrm{Card}(\mathcal{X}_L^0) \geq 2^{rM/8} + 1$, such that the zero $m_1 \times m_2$ matrix $\mathbf{0} \in \mathcal{X}_L^0$, and that for any two $\boldsymbol{X}$ and $\boldsymbol{X}'$ in $\mathcal{X}_L^0$, $\boldsymbol{X} \neq \boldsymbol{X}'$ we have

$$\|\boldsymbol{X} - \boldsymbol{X}'\|_F^2 \geq \frac{Mr}{8} \left(\eta^2 \min(a, \sigma_+)^2 \frac{r}{pm} \left\lfloor \frac{m_2}{r} \right\rfloor \right) \geq \frac{\eta^2}{16} \min(a^2, \sigma_+^2) \frac{rM}{p}. \tag{6.45}$$

For $\boldsymbol{X} \in \mathcal{X}_L^0$ we compute the Kullback-Leibler divergence $\mathrm{KL}(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_X)$ between $\mathbb{P}_{\mathbf{0}}$ and $\mathbb{P}_X$. Using Assumption **H**5 we obtain

$$\mathrm{KL}(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_X) = \sum_{i,j} \pi_{ij} \left(g_j(\boldsymbol{X}_{i,j}) - g_j(0) - g_j'(0) \boldsymbol{X}_{i,j} \right) \leq \frac{\sigma_+^2 \eta^2 \min(a, \sigma_+)^2 Mr}{2}. \tag{6.46}$$

Inequality (6.46) implies that

$$\frac{1}{\mathrm{Card}(\mathcal{X}_L^0) - 1} \sum_{\boldsymbol{X} \in \mathcal{X}_L^0} \mathrm{KL}(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_X) \leq \frac{1}{16} \log(\mathrm{Card}(\mathcal{X}_L^0) - 1) \tag{6.47}$$

is satisfied for $\tilde{\eta} = \min\left\{1, (8\sigma_+ \min(a, \sigma_+))^{-1}\right\}$. Then, conditions (6.45) and (6.46) guarantee that we can apply Theorem 2.5 from Tsybakov (2008). We obtain that for some constant $\delta > 0$ and with $\Psi_1 = C \min\left(\sigma_+^{-2}, \min(a, \sigma_+^2)\right)$:

$$\inf_{\hat{\mathbf{\Theta}}, \hat{\alpha}} \sup_{(\mathbf{\Theta}^0, \alpha^0) \in \mathcal{E}} \mathbb{P}_{\boldsymbol{X}^0} \left(\|\Delta \mathbf{\Theta}\|_F^2 + \|\Delta \alpha\|_2^2 > \frac{\Psi_1 rM}{p} \right) \geq \delta, \tag{6.48}$$

Lower bound of order s/p. Using again the Varshamov-Gilbert bound (Tsybakov (2008), Lemma 2.9) we obtain that there exists a subset $\mathcal{A}^0 \in \mathcal{A}$ satisfying $\mathrm{Card}(\mathcal{A}^0) \geq 2^{s/8}+1$ and containing the null vector $\mathbf{0} \in \mathbb{R}^N$ and such that, for any α and α' of $\mathcal{A}^0$, $\alpha \neq \alpha'$,

$$\|\alpha - \alpha'\|_2^2 \geq \frac{s}{8}\tilde{\eta}^2 \min(a, \sigma_+)^2. \tag{6.49}$$

Define $\mathcal{X}_\alpha \subset \mathcal{X}$ the set of matrices $X = \mathsf{f}_U(\alpha)$ such that $\alpha \in \mathcal{A}^0$ and $L = 0$. For any $X \in \mathcal{X}_\alpha$ we compute the Kullback-Leibler divergence $\mathrm{KL}(\mathbb{P}_\mathbf{0}, \mathbb{P}_X)$ between $\mathbb{P}_\mathbf{0}$ and $\mathbb{P}_X$

$$\mathrm{KL}(\mathbb{P}_\mathbf{0}, \mathbb{P}_X) = \sum_{i,j} \pi_{ij}(g_j(\boldsymbol{X}_{i,j}) - g_j(0) - g_j'(0)\boldsymbol{X}_{i,j} \leq \sigma_+^2 \|\mathsf{f}_U(\alpha)\|_\Pi^2 \leq \sigma_+^2 p \|\mathsf{f}_U(\alpha)\|_F^2. \tag{6.50}$$

Using Assumption **H**5

$$\begin{aligned} \mathrm{KL}(\mathbb{P}_0, \mathbb{P}_X) &\leq \sigma_+^2 p \left(\max_k \|\boldsymbol{U}^k\|_F^2 + 2\tau\right) \|\alpha\|_2^2 \\ &\leq s\sigma_+^2 p \left(\max_k \|\boldsymbol{U}^k\|_F^2 + 2\tau\right) \tilde{\eta}^2 \min(a, \sigma_+)^2. \end{aligned} \tag{6.51}$$

From (6.51) we deduce that

$$\frac{1}{\mathrm{Card}(\mathcal{A}^0) - 1} \sum_{\mathcal{A}^0} \mathrm{KL}(\mathbb{P}_0, \mathbb{P}_X) \leq sp \left(\max_k \|\boldsymbol{U}^k\|_F^2 + 2\tau\right) \sigma_+^2 \tilde{\eta}^2 \min(a, \sigma_+)^2. \tag{6.52}$$

Choosing $\tilde{\eta} = \min\left\{1, \left(\sqrt{p}\sigma_+ \max_k(\|\boldsymbol{U}^k\|_F + 2\tau)\min(a, \sigma_+)\right)^{-1}\right\}$, we now use Tsybakov (2008), Theorem 2.5 which implies for some constant $\delta > 0$

$$\inf_{\hat{\boldsymbol{\Theta}}, \hat{\alpha}} \sup_{(\boldsymbol{\Theta}^0, \alpha^0) \in \mathcal{E}} \mathbb{P}_{\boldsymbol{X}^0} \left\{ \|\Delta\boldsymbol{\Theta}\|_F^2 + \|\sum_{k=1}^N (\alpha_k^0 - \hat{\alpha}_k)\boldsymbol{U}^k\|_F^2 > \Psi_2 \frac{s\kappa^2}{p} \right\} \geq \delta, \tag{6.53}$$

$$\Psi_2 = C\left(\frac{1}{\sigma_+^2\left(\max_k \|\boldsymbol{U}^k\|_F^2 + 2\tau\right)} \wedge (a \wedge \sigma_+)^2\right),$$

where we have used that $\|\sum_{k=1}^N (\alpha_k^0 - \hat{\alpha}_k)\boldsymbol{U}^k\|_F^2 \geq \kappa^2 \|\hat{\alpha} - \alpha^0\|_2^2$. We finally obtain the result by combining (6.48) and (6.53).

Proof of Lemma 11

We start by proving $\|\Delta\alpha\|_1 \leq 4\|\alpha^0\|_1$. By the optimality conditions over a convex set (Aubin and Ekeland, 1984, Chapter 4, Section 2, Proposition 4), there exist two subgradients $\hat{f}_{\boldsymbol{\Theta}}$ in the subdifferential of $\|\cdot\|_*$ taken at $\hat{\boldsymbol{\Theta}}$ and $\hat{f}_\alpha$ in the subdifferential of $\|\cdot\|_1$ taken at $\hat{\alpha}$, such that for all feasible pairs $(\boldsymbol{\Theta}, \alpha)$ we have

$$\langle \nabla\mathcal{L}(\hat{\boldsymbol{X}}; \boldsymbol{Y}, \Omega), \boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}} + \sum_{k=1}^N (\alpha_k - \hat{\alpha}_k)\boldsymbol{U}^k \rangle + \lambda_1 \langle \hat{f}_{\boldsymbol{\Theta}}, \boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}} \rangle + \lambda_2 \langle \hat{f}_\alpha, \alpha - \hat{\alpha} \rangle \geq 0. \tag{6.54}$$

Applying inequality (6.54) to the pair $(\hat{\boldsymbol{\Theta}}, \alpha^0)$ we obtain $\langle \nabla\mathcal{L}(\hat{\boldsymbol{X}}; \boldsymbol{Y}, \Omega), \sum_{k=1}^N \Delta\alpha_k \boldsymbol{U}^k\rangle + \lambda_2\langle \hat{f}_\alpha, \Delta\alpha\rangle \geq 0$. Denote $\tilde{\boldsymbol{X}} = \hat{\boldsymbol{\Theta}} + \sum_{k=1}^N \alpha_k^0 \boldsymbol{U}^k$. The last inequality is equivalent to

$$\underbrace{\langle \nabla\mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega), \sum_{k=1}^N \Delta\alpha_k \boldsymbol{U}^k\rangle}_{\mathsf{B}_1} + \underbrace{\langle \nabla\mathcal{L}(\tilde{\boldsymbol{X}}; \boldsymbol{Y}, \Omega) - \nabla\mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega), \sum_{k=1}^N \Delta\alpha_k \boldsymbol{U}^k\rangle}_{\mathsf{B}_2}$$
$$+ \underbrace{\langle \nabla\mathcal{L}(\hat{\boldsymbol{X}}; \boldsymbol{Y}, \Omega) - \nabla\mathcal{L}(\tilde{\boldsymbol{X}}; \boldsymbol{Y}, \Omega), \sum_{k=1}^N \Delta\alpha_k \boldsymbol{U}^k\rangle}_{\mathsf{B}_3} + \lambda_2\langle \hat{f}_\alpha, \Delta\alpha\rangle \geq 0.$$

We now derive upper bounds on the three terms B_1, B_2 and B_3 separately. Recall that we denote $u = \max_k \|\boldsymbol{U}^k\|_1$ and use (6.37) to bound B_1:

$$\mathsf{B}_1 \leq \|\Delta\alpha\|_1 \|\nabla\mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\|_\infty u. \tag{6.55}$$

The duality between $\|\cdot\|_\infty$ and $\|\cdot\|_1$ gives $\mathsf{B}_2 \leq \|\Delta\alpha\|_1 \|\nabla\mathcal{L}(\tilde{\boldsymbol{X}}; \boldsymbol{Y}, \Omega) - \nabla\mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\|_\infty u$. Moreover, $\nabla\mathcal{L}(\tilde{\boldsymbol{X}}; \boldsymbol{Y}, \Omega) - \nabla\mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)$ is a matrix with entries $g_j'(\tilde{\boldsymbol{X}}_{i,j}) - g_j'(\boldsymbol{X}^0_{i,j})$, therefore assumption **H**5 ensures $\|\nabla\mathcal{L}(\tilde{\boldsymbol{X}}; \boldsymbol{Y}, \Omega) - \nabla\mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\|_\infty \leq 2\sigma_+^2(1 + æ)a$, and finally we obtain

$$\mathsf{B}_2 \leq \|\Delta\alpha\|_1 2\sigma_+^2(1 + æ)au. \tag{6.56}$$

We finally bound B_3 as follows. We have that $\mathsf{B}_3 = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \Omega_{i,j}(g_j'(\hat{\boldsymbol{X}}_{i,j}) - g_j'(\tilde{\boldsymbol{X}}_{i,j}))(\tilde{\boldsymbol{X}}_{i,j} - \hat{\boldsymbol{X}}_{i,j})$. Now, for all $j \in [\![m_2]\!]$, g_j' is increasing therefore $(g_j'(\hat{\boldsymbol{X}}_{i,j}) - g_j'(\tilde{\boldsymbol{X}}_{i,j}))(\tilde{\boldsymbol{X}}_{i,j} - \hat{\boldsymbol{X}}_{i,j}) \leq 0$, which implies $\mathsf{B}_3 \leq 0$. Combined with (6.55) and (6.56) this yields

$$\lambda_2\langle \hat{f}_\alpha, \hat{\alpha} - \alpha^0\rangle \leq \|\Delta\alpha\|_1 u \left(\|\nabla\mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\|_\infty + 2\sigma_+^2(1 + æ)a\right).$$

Besides, the convexity of $\|\cdot\|_1$ gives $\langle \hat{f}_\alpha, \hat{\alpha} - \alpha^0\rangle \geq \|\hat{\alpha}\|_1 - \|\alpha^0\|_1$, therefore

$$\left\{\lambda_2 - u\left(\|\nabla\mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\|_\infty + 2\sigma_+^2(1 + æ)a\right)\right\}\|\hat{\alpha}\|_1 \leq$$
$$\left\{\lambda_2 + u\left(\|\nabla\mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\|_\infty + 2\sigma_+^2(1 + æ)a\right)\right\}\|\alpha^0\|_1,$$

and the condition $\lambda_2 \geq 2\left\{u\left(\|\nabla\mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\|_\infty + 2\sigma_+^2(1 + æ)a\right)\right\}$ gives $\|\hat{\alpha}\|_1 \leq 3\|\alpha^0\|_1$ and finally

$$\|\Delta\alpha\|_1 \leq 4\|\alpha^0\|_1. \tag{6.57}$$

Case 1: $\|\mathsf{f}_U(\Delta\alpha)\|_\Pi^2 < 72a^2\log(d)/(p\log(6/5))$. Then the result holds trivially.

Case 2: $\|\mathsf{f}_U(\Delta\alpha)\|_\Pi^2 \geq 72a^2\log(d)/(p\log(6/5))$. For $d_1 > 0$ recall the definition of the set

$$\tilde{\mathcal{A}}(d_1) = \left\{\alpha \in \mathbb{R}^N : \quad \|\alpha\|_\infty \leq 1; \quad \|\alpha\|_1 \leq d_1; \quad \|\mathsf{f}_U(\alpha)\|_\Pi^2 \geq \frac{18\log(d)}{p\log(6/5)}\right\}.$$

Inequality (6.57) and $\|\Delta\alpha\|_\infty \leq 2a$ imply that $\Delta\alpha/(2a) \in \tilde{\mathcal{A}}(2\|\alpha^0\|_1/a)$. Therefore we can apply Lemma 13(i) and obtain that with probability at least $1 - 8d^{-1}$,

$$\|\mathsf{f}_U(\Delta\alpha)\|_\Pi^2 \leq 2\|\mathsf{f}_U(\Delta\alpha)\|_\Omega^2 + 64æa\|\alpha^0\|_1 u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + 3072a^2p^{-1}. \tag{6.58}$$

We now must upper bound the quantity $\|\mathsf{f}_U(\Delta\alpha)\|_\Omega^2$. Recall that $\tilde{\boldsymbol{X}} = \sum_{k=1}^N \alpha_k^0 \boldsymbol{U}^k + \hat{\boldsymbol{X}}$. By definition, $\mathcal{L}(\hat{\boldsymbol{X}}; \boldsymbol{Y}, \Omega) + \lambda_1 \|\hat{\boldsymbol{\Theta}}\|_* + \lambda_2 \|\hat{\alpha}\|_1 \leq \mathcal{L}(\tilde{\boldsymbol{X}}; \boldsymbol{Y}, \Omega) + \lambda_1 \|\hat{\boldsymbol{\Theta}}\|_* + \lambda_2 \|\alpha^0\|_1$, i.e.

$$\mathcal{L}(\hat{X}; \boldsymbol{Y}, \Omega) - \mathcal{L}(\tilde{\boldsymbol{X}}; \boldsymbol{Y}, \Omega) \leq \lambda_2 \left(\|\alpha^0\|_1 - \|\hat{\alpha}\|_1 \right).$$

Substracting $\langle \nabla \mathcal{L}(\tilde{\boldsymbol{X}}; \boldsymbol{Y}, \Omega), \hat{\boldsymbol{X}} - \tilde{\boldsymbol{X}} \rangle$ on both sides and using the restricted strong convexity ((6.31)), we obtain

$$\begin{aligned} \frac{\sigma_-^2}{2} \|\mathsf{f}_U(\Delta\alpha)\|_\Omega^2 &\leq \lambda_2 \left(\|\alpha^0\|_1 - \|\hat{\alpha}\|_1 \right) + \langle \nabla \mathcal{L}(\tilde{\boldsymbol{X}}; \boldsymbol{Y}, \Omega), \mathsf{f}_U(\Delta\alpha) \rangle \\ &\leq \lambda_2 \left(\|\alpha^0\|_1 - \|\hat{\alpha}\|_1 \right) + \underbrace{\left| \langle \nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega), \mathsf{f}_U(\Delta\alpha) \rangle \right|}_{\mathsf{C}_1}. \\ &\quad + \underbrace{\left| \langle \nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega) - \nabla \mathcal{L}(\tilde{\boldsymbol{X}}; Y), \mathsf{f}_U(\Delta\alpha) \rangle \right|}_{\mathsf{C}_2} \end{aligned} \tag{6.59}$$

The duality of $\|\cdot\|_1$ and $\|\cdot\|_\infty$ yields $\mathsf{C}_1 \leq \|\nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\|_\infty u \|\Delta\alpha\|_1$, and

$$\mathsf{C}_2 \leq \|\nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega) - \nabla \mathcal{L}(\tilde{\boldsymbol{X}}; \boldsymbol{Y}, \Omega)\|_\infty u \|\Delta\alpha\|_1.$$

Furthermore, $\|\nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega) - \nabla \mathcal{L}(\tilde{\boldsymbol{X}}; \boldsymbol{Y}, \Omega)\|_\infty \leq 2\sigma_+^2 a$, since for all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$ $|\tilde{\boldsymbol{X}}_{i,j} - \boldsymbol{X}_{i,j}^0| \leq 2a$ and $g_j''(\tilde{\boldsymbol{X}}_{i,j}) \leq \sigma_+^2$. The last three inequalities plugged in (6.59) give

$$\frac{\sigma_-^2}{2} \|\mathsf{f}_U(\Delta\alpha)\|_\Omega^2 \leq \lambda_2 \left(\|\alpha^0\|_1 - \|\hat{\alpha}\|_1 \right) + u \|\Delta\alpha\|_1 \left\{ \|\nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\|_\infty + 2\sigma_+^2 a \right\}.$$

The triangular inequality gives

$$\begin{aligned} \frac{\sigma_-^2}{2} \|\mathsf{f}_U(\Delta\alpha)\|_\Omega^2 &\leq \left\{ u \left(\|\nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\|_\infty + 2\sigma_+^2 a \right) + \lambda_2 \right\} \|\alpha^0\|_1 \\ &\quad + \left\{ u \left(\|\nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\|_\infty + 2\sigma_+^2 a \right) - \lambda_2 \right\} \|\hat{\alpha}\|_1. \end{aligned}$$

Then, the assumption $\lambda_2 \geq 2u \left(\|\nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\|_\infty + 2\sigma_+^2 (1 + æ) a \right)$ gives

$$\|\mathsf{f}_U(\Delta\alpha)\|_\Omega^2 \leq \frac{3\lambda_2}{\sigma_-^2} \|\alpha^0\|_1.$$

Plugged into (6.58), this last inequality implies that with probability at least $1 - 8d^{-1}$

$$\|\mathsf{f}_U(\Delta\alpha)\|_\Pi^2 \leq \frac{3\lambda_2}{\sigma_-^2} \|\alpha^0\|_1 + 64æa \|\alpha^0\|_1 u \mathbb{E}\left[\|\Sigma_R\|_\infty \right] + 3072 a^2 p^{-1}. \tag{6.60}$$

Combining (6.57) and (6.60) gives the result.

Proof of Lemma 12

Using (6.54) for $\boldsymbol{\Theta} = \boldsymbol{\Theta}^0$ and $\alpha = \alpha^0$ we obtain

$$\langle \nabla \mathcal{L}(\hat{\boldsymbol{X}}; \boldsymbol{Y}, \Omega), \Delta\boldsymbol{\Theta} + \sum_{k=1}^N (\Delta\alpha_k) \boldsymbol{U}^k \rangle + \lambda_1 \langle \hat{f}_{\boldsymbol{\Theta}}, \Delta\boldsymbol{\Theta} \rangle + \lambda_2 \langle \hat{f}_\alpha, \Delta\alpha \rangle \geq 0.$$

Then, the convexity of $\|\cdot\|_*$ and $\|\cdot\|_1$ imply that $\|\boldsymbol{\Theta}^0\|_* \geq \|\hat{\boldsymbol{\Theta}}\|_* + \langle \partial\|\hat{\boldsymbol{\Theta}}\|_*, \Delta L\rangle$ and $\|\alpha^0\|_1 \geq \|\hat{\alpha}\|_* + \langle \partial\|\hat{\alpha}\|_1, \Delta\alpha\rangle$. The last three inequalities yield

$$\begin{aligned}\lambda_1\left(\|\hat{\boldsymbol{\Theta}}\|_* - \|\boldsymbol{\Theta}^0\|_*\right) + \lambda_2\left(\|\hat{\alpha}\|_1 - \|\alpha^0\|_1\right) &\leq \langle \nabla\mathcal{L}(\hat{\boldsymbol{X}}; \boldsymbol{Y}, \Omega), \Delta\boldsymbol{\Theta}\rangle \\ &\quad + \langle \nabla\mathcal{L}(\hat{\boldsymbol{X}}; \boldsymbol{Y}, \Omega), \sum_{k=1}^{N}(\Delta\alpha_k)\boldsymbol{U}^k\rangle \\ &\leq \|\nabla\mathcal{L}(\hat{\boldsymbol{X}}; \boldsymbol{Y}, \Omega)\|\|\Delta\boldsymbol{\Theta}\|_* + u\|\nabla\mathcal{L}(\hat{\boldsymbol{X}}; \boldsymbol{Y}, \Omega)\|_\infty\|\Delta\alpha\|_1.\end{aligned}$$

Using (6.35) and the conditions

$$\lambda_1 \geq 2\|\nabla\mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\|, \quad \lambda_2 \geq 2u\left\{\|\nabla\mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)\|_\infty + 2\sigma_+^2(1+æ)a\right\},$$

we get

$$\begin{aligned}\lambda_1\left(\|P^\perp_{\boldsymbol{\Theta}^0}(\Delta\boldsymbol{\Theta})\|_* - \|P_{\boldsymbol{\Theta}^0}(\Delta\boldsymbol{\Theta})\|_*\right) + \lambda_2\left(\|\hat{\alpha}\|_1 - \|\alpha^0\|_1\right) \leq \\ \frac{\lambda_1}{2}\left(\|P^\perp_{\boldsymbol{\Theta}^0}(\Delta\boldsymbol{\Theta})\|_* + \|P_{\boldsymbol{\Theta}^0}(\Delta\boldsymbol{\Theta})\|_*\right) + \frac{\lambda_2}{2}\|\Delta\alpha\|_1,\end{aligned}$$

which implies $\|P^\perp_{\boldsymbol{\Theta}^0}(\Delta\boldsymbol{\Theta})\|_* \leq 3\|P_{\boldsymbol{\Theta}^0}(\Delta\boldsymbol{\Theta})\|_* + 3\lambda_2/\lambda_1\|\alpha^0\|_1$. Now, using

$$\|\Delta\boldsymbol{\Theta}\|_* \leq \|P^\perp_{\boldsymbol{\Theta}^0}(\Delta\boldsymbol{\Theta})\|_* + \|P_{\boldsymbol{\Theta}^0}(\Delta\boldsymbol{\Theta})\|_*, \quad \|P_{\boldsymbol{\Theta}^0}(\Delta\boldsymbol{\Theta})\|_F \leq \|\Delta\boldsymbol{\Theta}\|_F$$

and $\mathrm{rank}(P_{\boldsymbol{\Theta}^0}(\Delta\boldsymbol{\Theta})) \leq 2r$, we get $\|\Delta\boldsymbol{\Theta}\|_* \leq \sqrt{32r}\|\Delta L\|_F + 3\lambda_2/\lambda_1\|\alpha^0\|_1$. This completes the proof of Lemma 12.

Proof of Lemma 13

Proof of (i): Recall $\mathsf{D}_\alpha = 8æd_1u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + 768p^{-1}$ and

$$\tilde{\mathcal{A}}(d_1) = \left\{\alpha \in \mathbb{R}^N : \quad \|\alpha\|_\infty \leq 1; \quad \|\alpha\|_1 \leq d_1; \quad \|\mathsf{f}_U(\alpha)\|^2_\Pi \geq \frac{18\log(d)}{p\log(6/5)}\right\}.$$

We will show that the probability of the following event is small:

$$\mathcal{B} = \left\{\exists\alpha \in \tilde{\mathcal{A}}(d_1) \text{ such that } \left|\|\mathsf{f}_U(\alpha)\|^2_\Omega - \|\mathsf{f}_U(\alpha)\|^2_\Pi\right| > \frac{1}{2}\|\mathsf{f}_U(\alpha)\|^2_\Pi + \mathsf{D}_\alpha\right\}.$$

Indeed, $\mathcal{B}$ contains the complement of the event we are interested in. We use a peeling argument to upper bound the probability of event $\mathcal{B}$. Let $\nu = 18\log(d)/(p\log(6/5))$ and $\eta = 6/5$. For $l \in \mathbb{N}$ set

$$\mathcal{S}_l = \left\{\alpha \in \tilde{\mathcal{A}}(d_1) : \quad \eta^{l-1}\nu \leq \|\mathsf{f}_U(\alpha)\|^2_\Pi \leq \eta^l\nu\right\}.$$

Under the event $\mathcal{B}$, there exists $l \geq 1$ and $\alpha \in \tilde{\mathcal{A}}(d_1) \cap S_l$ such that

$$\left|\|\mathsf{f}_U(\alpha)\|^2_\Omega - \|\mathsf{f}_U(\alpha)\|^2_\Pi\right| \quad > \frac{1}{2}\|\mathsf{f}_U(\alpha)\|^2_\Pi + \mathsf{D}_\alpha > \frac{1}{2}\eta^{l-1}\nu + \mathsf{D}_\alpha = \frac{5}{12}\eta^l\nu + \mathsf{D}_\alpha. \tag{6.61}$$

For $T > \nu$, consider the set of vectors

$$\tilde{\mathcal{A}}(d_1, T) = \left\{\alpha \in \tilde{\mathcal{A}}(d_1) : \|\mathsf{f}_U(\alpha)\|^2_\Pi \leq T\right\}$$

and the event

$$\mathcal{B}_l = \left\{ \exists \alpha \in \tilde{\mathcal{A}}(d_1, \eta^l \nu) : \left| \|\mathsf{f}_U(\alpha)\|_\Omega^2 - \|\mathsf{f}_U(\alpha)\|_\Pi^2 \right| > \frac{5}{12} \eta^l \nu + \mathsf{D}_\alpha \right\}.$$

If $\mathcal{B}$ holds, then (6.61) implies that $\mathcal{B}_l$ holds for some $l \leq 1$. Therefore ,$\mathcal{B} \subset \cup_{l=1}^{+\infty} \mathcal{B}_l$, and it is enough to estimate the probability of the events $\mathcal{B}_l$ and then apply the union bound. Such an estimation is given in the following lemma, adapted from Lemma 10 in Klopp (2015).

Lemma 14. *Define* $Z_T = \sup_{\alpha \in \tilde{\mathcal{A}}(d_1, T)} \left| \|\mathsf{f}_U(\alpha)\|_\Omega^2 - \|\mathsf{f}_U(\alpha)\|_\Pi^2 \right|$. *Then,*

$$\mathbb{P}\left(Z_T \geq \mathsf{D}_\alpha + \frac{5}{12} T \right) \leq 4\mathrm{e}^{-pT/18}.$$

Proof. By definition,

$$Z_T = \sup_{\alpha \in \tilde{\mathcal{A}}(d_1, T)} \left| \sum_{(i,j)} \Omega_{i,j} \mathsf{f}_U(\alpha)_{i,j}^2 - \mathbb{E}\left[\sum_{(i,j)} \Omega_{i,j} \mathsf{f}_U(\alpha)_{i,j}^2 \right] \right|.$$

We use the following Talagrand's concentration inequality, proven in Talagrand (1996) and **?**.

Lemma 15. *Assume* $f : [-1,1]^n \mapsto \mathbb{R}$ *is a convex Lipschitz function with Lipschitz constant* L. *Let* $\Xi_1, \ldots, \Xi_n$ *be independent random variables taking values in* $[-1,1]$. *Let* $Z := f(\Xi_1, \ldots, \Xi_n)$. *Then, for any* $t \geq 0$, $\mathbb{P}\left(|Z - \mathbb{E}[Z]| \geq 16L + t \right) \leq 4\mathrm{e}^{-t^2/2L^2}$.

We apply this result to the function

$$f(x_{11}, \ldots, x_{m_1 m_2}) = \sup_{\alpha \in \tilde{\mathcal{A}}(d_1, T)} \left| \sum_{(i,j)} (x_{ij} - \pi_{ij}) \mathsf{f}_U(\alpha)_{i,j}^2 \right|,$$

which is Lipschitz with Lipschitz constant $\sqrt{p^{-1}T}$. Indeed, for any $(x_{11}, \ldots, x_{m_1 m_2}) \in$

$\mathbb{R}^{m_1 \times m_2}$ and $(z_{11}, \ldots, z_{m_1 m_2}) \in \mathbb{R}^{m_1 \times m_2}$:

$$
\begin{aligned}
&|f(x_{11}, \ldots, x_{m_1 m_2}) - f(z_{11}, \ldots, z_{m_1 m_2})| \\
&\quad = \left| \sup_{\alpha \in \tilde{\mathcal{A}}(d_1, T)} \left| \sum_{(i,j)} (x_{ij} - \pi_{ij}) \mathsf{f}_U(\alpha)_{i,j}^2 \right| - \sup_{\alpha \in \tilde{\mathcal{A}}(d_1, T)} \left| \sum_{(i,j)} (z_{ij} - \pi_{ij}) \mathsf{f}_U(\alpha)_{i,j}^2 \right| \right| \\
&\quad \leq \sup_{\alpha \in \tilde{\mathcal{A}}(d_1, T)} \left| \left| \sum_{(i,j)} (x_{ij} - \pi_{ij}) \mathsf{f}_U(\alpha)_{i,j}^2 \right| - \left| \sum_{(i,j)} (z_{ij} - \pi_{ij}) \mathsf{f}_U(\alpha)_{i,j}^2 \right| \right| \\
&\quad \leq \sup_{\alpha \in \tilde{\mathcal{A}}(d_1, T)} \left| \sum_{(i,j)} (x_{ij} - \pi_{ij}) \mathsf{f}_U(\alpha)_{i,j}^2 - \sum_{(i,j)} (z_{ij} - \pi_{ij}) \mathsf{f}_U(\alpha)_{i,j}^2 \right| \\
&\quad \leq \sup_{\alpha \in \tilde{\mathcal{A}}(d_1, T)} \left| \sum_{(i,j)} (x_{ij} - z_{ij}) \mathsf{f}_U(\alpha)_{i,j}^2 \right| \\
&\quad \leq \sup_{\alpha \in \tilde{\mathcal{A}}(d_1, T)} \sqrt{\sum_{(i,j)} \pi_{ij}^{-1} (x_{ij} - z_{ij})^2} \sqrt{\sum_{(i,j)} \pi_{ij} \mathsf{f}_U(\alpha)_{i,j}^4} \\
&\quad \leq \sup_{\alpha \in \tilde{\mathcal{A}}(d_1, T)} \sqrt{p^{-1}} \sqrt{\sum_{(i,j)} (x_{ij} - z_{ij})^2} \sqrt{\sum_{(i,j)} \pi_{ij} \mathsf{f}_U(\alpha)_{i,j}^2} \\
&\quad \leq \sqrt{p^{-1} T} \sqrt{\sum_{(i,j)} (x_{ij} - z_{ij})^2},
\end{aligned}
$$

where we used $||a| - |b|| \leq |a - b|, \|\mathsf{f}_U(\alpha)\|_\infty \leq 1$ and $\|A\|_\Pi^2 \leq T$. Thus, Lemma 15 and the identity $\sqrt{p^{-1} T} \leq \frac{96 p^{-1}}{2} + \frac{T}{2 \times 96}$ imply

$$
\mathbb{P}\left(|Z - \mathbb{E}[Z]| \geq 768 p^{-1} + \frac{1}{12} T + t \right) \leq 4 e^{-t^2 p / 2T}.
$$

Taking $t = T/3$ we get

$$
\mathbb{P}\left(|Z - \mathbb{E}[Z]| \geq 768 p^{-1} + \frac{5}{12} T \right) \leq 4 e^{-pT/18}. \tag{6.62}
$$

Now we must bound the expectation $\mathbb{E}[Z_T]$. To do so, we use a symmetrization argument (Ledoux, 2001) which gives

$$
\begin{aligned}
\mathbb{E}[Z_T] &= \mathbb{E}\left[\sup_{\alpha \in \tilde{\mathcal{A}}(d_1, T)} \left| \sum_{(i,j)} \Omega_{ij} \mathsf{f}_U(\alpha)_{i,j}^2 - \mathbb{E}\left[\sum_{(i,j)} \Omega_{ij} \mathsf{f}_U(\alpha)_{i,j}^2 \right] \right| \right] \\
&\leq 2\mathbb{E}\left[\sup_{\alpha \in \tilde{\mathcal{A}}(d_1, T)} \left| \sum_{(i,j)} \epsilon_{ij} \Omega_{ij} \mathsf{f}_U(\alpha)_{i,j}^2 \right| \right],
\end{aligned}
$$

where $\{\epsilon_{ij}\}$ is an i.i.d. Rademacher sequence independent of $\{\Omega_{ij}\}$. We apply an extension Talagrand's contraction inequality to Lipschitz functions (see Koltchinskii (2011a),

Theorem 2.2) and obtain

$$\mathbb{E}\left[Z_T\right] = \mathbb{E}\left[\sup_{\boldsymbol{A}\in\mathcal{T}}\left|\sum_{i,j}\epsilon_{ij}\Omega_{ij}\boldsymbol{A}_{i,j}^2\right|\right] \leq 4\text{æ}\mathbb{E}\left[\sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)}\left|\sum_{(i,j)}\epsilon_{ij}\Omega_{ij}\boldsymbol{A}_{i,j}\right|\right]$$
$$= 4\text{æ}\mathbb{E}\left[\sup_{\alpha\in\tilde{\mathcal{A}}(d_1,T)}\left|\langle\Sigma_R, \mathsf{f}_U(\alpha)\rangle\right|\right],$$

where $\Sigma_R = \sum_{(i,j)}\epsilon_{ij}\Omega_{ij}E_{ij}$. Moreover, for $\alpha\in\tilde{\mathcal{A}}(d_1,T)$ we have

$$\left|\langle\Sigma_R, \mathsf{f}_U(\alpha)\rangle\right| \quad = \left|\langle\Sigma_R, \sum_{k=1}^{N}\alpha_k\boldsymbol{U}^k\rangle\right| \leq \|\alpha\|_1 u\|\Sigma_R\|_\infty.$$

Finally, we get $\mathbb{E}\left[Z_T\right] \leq 4\text{æ}d_1u\mathbb{E}\left[\|\Sigma_R\|_\infty\right]$. Combining this with the concentration inequality (6.62) we complete the proof of Lemma 14:

$$\mathbb{P}\left(Z_T \geq 8\text{æ}d_1u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + 768p^{-1} + \frac{5}{12}T\right) \leq 4\mathrm{e}^{-pT/18}.$$

□

Lemma 14 gives that $\mathbb{P}\left(\mathcal{B}_l\right) \leq 4\exp(-p\eta^l\nu/18)$. Applying the union bound we obtain

$$\mathbb{P}\left(\mathcal{B}\right) \quad \leq \sum_{l=1}^{\infty}\mathbb{P}\left(\mathcal{B}_l\right) \leq 4\sum_{l=1}^{\infty}\exp(-p\eta^l\nu/18)$$
$$\leq 4\sum_{l=1}^{\infty}\exp(-p\log(\eta)l\nu/18),$$

where we used $e^x \geq x$. Finally, for $\nu = 18\log(d)/(p\log(6/5))$ we obtain

$$\mathbb{P}\left(\mathcal{B}\right) \leq \frac{4\exp(-p\nu\log(\eta)/18)}{1-\exp(-p\nu\log(\eta)/18)} \leq \frac{4\exp(-\log(d))}{1-\exp(-\log(d))} \leq \frac{8}{d},$$

since $d-1 \geq d/2$, which concludes the proof of (i).

Proof of (ii): The proof is similar to that of (i); we recycle some of the notations for simplicity. Recall $\mathsf{D}_{\boldsymbol{X}} = 112\rho p^{-1}\mathbb{E}\left[\|\Sigma_R\|\right]^2 + 8\text{æ}\varepsilon\mathbb{E}\left[\|\Sigma_R\|\right] + 8\text{æ}d_1u\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + d_\Pi + 768p^{-1}$, and let

$$\mathcal{B} = \Big\{\exists(\boldsymbol{\Theta},\alpha)\in\mathcal{C}(d_1,d_\Pi,\rho,\varepsilon);$$
$$\left|\|\boldsymbol{\Theta}+\mathsf{f}_U(\alpha)\|_\Omega^2 - \|\boldsymbol{\Theta}+\mathsf{f}_U(\alpha)\|_\Pi^2\right| > \frac{1}{2}\|\boldsymbol{\Theta}+\mathsf{f}_U(\alpha)\|_\Pi^2 + \mathsf{D}_{\boldsymbol{X}}\Big\},$$

$\nu = 72\log(d)/(p\log(6/5))$, $\eta = 6/5$ and for $l\in\mathbb{N}$

$$\mathcal{S}_l = \left\{(\boldsymbol{\Theta},\alpha)\in\mathcal{C}(d_1,d_\Pi,\rho,\varepsilon): \quad \eta^{l-1}\nu \leq \|\boldsymbol{\Theta}+\mathsf{f}_U(\alpha)\|_\Pi^2 \leq \eta^l\nu\right\}.$$

As before, if $\mathcal{B}$ holds, then there exist $l\geq 2$ and $(\boldsymbol{\Theta},\alpha)\in\mathcal{C}(d_1,d_\Pi,\rho,\varepsilon)\cap S_l$ such that

$$\left|\|\boldsymbol{\Theta}+\mathsf{f}_U(\alpha)\|_\Omega^2 - \|\boldsymbol{\Theta}+\mathsf{f}_U(\alpha)\|_\Pi^2\right| \quad > \frac{5}{12}\eta^l\nu + \mathsf{D}_{\boldsymbol{X}}. \tag{6.63}$$

For $T > \nu$, consider the set $\tilde{\mathcal{C}}(T) = \{(\mathbf{\Theta}, \alpha) \in \mathcal{C}(d_1, d_\Pi, \rho, \varepsilon) : \|\mathbf{\Theta} + \mathsf{f}_U(\alpha)\|_\Pi^2 \leq T\}$, and the event

$$\mathcal{B}_l = \left\{ \exists (\mathbf{\Theta}, \alpha) \in \tilde{\mathcal{C}}(\eta^l \nu) : \quad \left| \|\mathbf{\Theta} + \mathsf{f}_U(\alpha)\|_\Omega^2 - \|\mathbf{\Theta} + \mathsf{f}_U(\alpha)\|_\Pi^2 \right| > \frac{5}{12}\eta^l \nu + \mathsf{D}_{\boldsymbol{X}} \right\}.$$

Then, (6.63) implies that $\mathcal{B}_l$ holds and $\mathcal{B} \subset \cup_{l=1}^{+\infty} \mathcal{B}_l$. Thus, we estimate in Lemma 16 the probability of the events $\mathcal{B}_l$, and then apply the union bound.

Lemma 16. *Let* $W_T = \sup_{(\mathbf{\Theta},\alpha)\in\tilde{\mathcal{C}}(T)} \left| \|\mathbf{\Theta} + \mathsf{f}_U(\alpha)\|_\Omega^2 - \|\mathbf{\Theta} + \mathsf{f}_U(\alpha)\|_\Pi^2 \right|$.

$$\mathbb{P}\left(W_T \geq \mathsf{D}_{\boldsymbol{X}} + \frac{5}{12} T \right) \leq 4\mathrm{e}^{-pT/72}.$$

Proof. The proof is two-fold: first we show that W_T concentrates around its expectation, then bound its expectation. By definition,

$$W_T = \sup_{(\mathbf{\Theta},\alpha)\in\tilde{\mathcal{C}}(T)} \left| \sum_{(i,j)} \Omega_{ij} (\mathbf{\Theta}_{i,j} + \mathsf{f}_U(\alpha)_{i,j})^2 - \mathbb{E}\left[\sum_{(i,j)} \Omega_{ij} (\mathbf{\Theta}_{i,j} + \mathsf{f}_U(\alpha)_{i,j})^2 \right] \right|.$$

The concentration proof is exactly similar to the proof in Lemma 14, but we choose $t = T/6$, and we obtain

$$\mathbb{P}\left(|W_T - \mathbb{E}[W_T]| \geq 768 p^{-1} + \frac{3}{12} T \right) \leq 4e^{-pT/72}. \tag{6.64}$$

Let us now bound the expectation $\mathbb{E}[W_T]$. Again, we use a standard symmetrization argument (Ledoux, 2001) which gives

$$\mathbb{E}[W_T] \quad \leq 2\mathbb{E}\left[\sup_{(\mathbf{\Theta},\alpha)\in\tilde{\mathcal{C}}(T)} \left| \sum_{(i,j)} \epsilon_{ij} \Omega_{ij} (\mathbf{\Theta}_{i,j} + \mathsf{f}_U(\alpha)_{i,j})^2 \right| \right],$$

where $\{\epsilon_{ij}\}$ is an i.i.d. Rademacher sequence independent of Ω_{ij}. Then, the contraction inequality (see Koltchinskii (2011a), Theorem 2.2) yields

$$\mathbb{E}[W_T] \leq 4\text{æ}\mathbb{E}\left[\sup_{(\mathbf{\Theta},\alpha)\in\tilde{\mathcal{C}}(T)} |\langle \Sigma_R, \mathbf{\Theta} + \mathsf{f}_U(\alpha) \rangle| \right],$$

where $\Sigma_R = \sum_{(i,j)} \epsilon_{ij} \Omega_{ij} E_{ij}$. Moreover

$$\begin{aligned} |\langle \Sigma_R, \mathbf{\Theta} + \mathsf{f}_U(\alpha) \rangle| \quad &\leq |\langle \Sigma_R, \mathbf{\Theta} \rangle| + |\langle \Sigma_R, \mathsf{f}_U(\alpha) \rangle| \\ &\leq \|\mathbf{\Theta}\|_* \|\Sigma_R\| + \|\alpha\|_1 u \|\Sigma_R\|_\infty. \end{aligned}$$

For $(\mathbf{\Theta}, \alpha) \in \tilde{\mathcal{C}}(T)$ we have by assumption $\|\alpha\|_1 \leq d_1$, $\|\mathsf{f}_U(\alpha)\|_\Pi \leq \sqrt{d_\Pi}$ and $\|\mathbf{\Theta}\|_* \leq \sqrt{\rho}\|\mathbf{\Theta}\|_F + \varepsilon$. We obtain

$$\begin{aligned} \|\mathbf{\Theta}\|_* \quad &\leq \sqrt{\frac{\rho}{p}} \|\mathbf{\Theta}\|_\Pi + \varepsilon \leq \sqrt{\frac{\rho}{p}} \left(\|\mathbf{\Theta} + \mathsf{f}_U(\alpha)\|_\Pi + \|\mathsf{f}_U(\alpha)\|_\Pi \right) + \varepsilon \\ &\leq \sqrt{\frac{\rho}{p}} \left(\sqrt{T} + \sqrt{d_\Pi} \right) + \varepsilon. \end{aligned}$$

This gives

$$\begin{aligned}\mathbb{E}\left[W_T\right] &\leq 4æ\left\{\sqrt{\frac{\rho}{p}}\left(\sqrt{T}+\sqrt{d_\Pi}\right)+\varepsilon\right\}\|\Sigma_R\|+4æd_1u\|\Sigma_R\|_\infty \\ &\leq \frac{T}{12}+\frac{d_\Pi}{2}+56æ^2\frac{\rho}{p}\|\Sigma_R\|^2+4æ\varepsilon\|\Sigma_R\|+4æd_1u\|\Sigma_R\|_\infty.\end{aligned}$$

Combining this with the concentration inequality (6.64) we finally obtain:

$$\mathbb{P}\left(W_T \geq \mathsf{D}_{\boldsymbol{X}}+\frac{5}{12}T\right) \leq 4\mathrm{e}^{-pT/72}.$$

□

Lemma 16 gives that $\mathbb{P}\left(\mathcal{B}_l\right) \leq 4\exp(-p\eta^l\nu/72)$. Applying the union bound we obtain

$$\begin{aligned}\mathbb{P}\left(\mathcal{B}\right) &\leq \sum_{l=1}^{\infty}\mathbb{P}\left(\mathcal{B}_l\right) \leq 4\sum_{l=1}^{\infty}\exp(-p\eta^l\nu/72) \\ &\leq 4\sum_{l=1}^{\infty}\exp(-p\log(\eta)l\nu/72),\end{aligned}$$

where we used $e^x \geq x$. Finally, for $\nu = 72\log(d)/(p\log(6/5))$ we obtain

$$\mathbb{P}\left(\mathcal{B}\right) \leq \frac{4\exp(-p\nu\log(\eta)/72)}{1-\exp(-p\nu\log(\eta)/72)} \leq \frac{4\exp(-\log(d))}{1-\exp(-\log(d))} \leq 8d^{-1},$$

since $d-1 \geq d/2$, which concludes the proof of (ii).

Proof of Lemma 7

The first inequality is trivially true using that $\|\Sigma\|_\infty = \max_{i,j}|\Omega_{ij}\epsilon_{ij}| \leq 1$. We prove the second inequality using an extension to rectangular matrices via self-adjoint dilation of Corollary 3.3 in Bandeira and van Handel (2016).

Proposition 2. *Let $\boldsymbol{A}$ be an $m_1 \times m_2$ rectangular matrix with $\boldsymbol{A}_{i,j}$ independent centered bounded random variables. then, there exists a universal constant C^* such that*

$$\mathbb{E}\left[\|\boldsymbol{A}\|\right] \leq C^*\left\{\sigma_1 \vee \sigma_2 + \sigma_*\sqrt{\log(m_1 \wedge m_2)}\right\},$$

$$\sigma_1 = \max_i\sqrt{\sum_j \mathbb{E}\left[\boldsymbol{A}_{i,j}^2\right]}, \quad \sigma_2 = \max_j\sqrt{\sum_i \mathbb{E}\left[\boldsymbol{A}_{i,j}^2\right]}, \quad \sigma_* = \max_{i,j}|\boldsymbol{A}_{i,j}|.$$

Applying Proposition 2 to Σ_R with $\sigma_1 \vee \sigma_2 \leq \sqrt{\beta}$ and $\sigma_* \leq 1$ we obtain

$$\mathbb{E}\left[\|\Sigma_R\|\right] \leq C^*\left\{\sqrt{\beta}+\sqrt{\log(m_1 \wedge m_2)}\right\}.$$

Proof of Lemma 8

Denote $\Sigma = \nabla \mathcal{L}(\boldsymbol{X}^0; \boldsymbol{Y}, \Omega)$. Definition (6.2) implies that $\mathbb{E}[\boldsymbol{Y}_{i,j}] = g_j'(\boldsymbol{X}^0_{i,j})$, $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$. Combined with the sub-exponentiality of the entries $\boldsymbol{Y}_{i,j}$, we obtain that for all i, j, $\boldsymbol{Y}_{i,j} - g_j'(\boldsymbol{X}^0_{i,j})$ is sub-exponential with scale and variance parameters $1/\gamma$ and σ_+^2 respectively. Then, noticing that $|\Omega_{ij}| \leq 1$ implies that for all $t \geq 0$,

$$\mathbb{P}\left\{\left|\Omega_{ij}\left(\boldsymbol{Y}_{i,j} - g_j'(\boldsymbol{X}^0_{i,j})\right)\right| \geq t\right\} \leq \mathbb{P}\left\{\left|\boldsymbol{Y}_{i,j} - g_j'(\boldsymbol{X}^0_{i,j})\right| \geq t\right\},$$

we obtain that the random variables $\Sigma_{i,j} = \Omega_{ij}\left(\boldsymbol{Y}_{i,j} - g_j'(\boldsymbol{X}^0_{i,j})\right)$ are also sub-exponential. Thus, for all i, j and for all $t \geq 0$ we have that $|\Sigma_{i,j}| \leq t$ with probability at least $1 - \max\left\{2e^{-t^2/2\sigma_+^2}, 2e^{-\gamma t/2}\right\}$. A union bound argument then yields

$$\|\Sigma\|_\infty \leq t \quad \text{w. p. at least } 1 - \max\left\{2m_1m_2e^{-t^2/2\sigma_+^2}, 2m_1m_2e^{-\gamma t/2}\right\},$$

where γ and σ_+ are defined in **H**5. Using $\log(m_1m_2) \leq 2\log d$, where $d = m_1 + m_2$ and setting $t = 6\max\left\{\sigma_+\sqrt{\log d}, \gamma^{-1}\log d\right\}$, we obtain that with probability at least $1 - d^{-1}$,

$$\|\Sigma\|_\infty \leq 6\max\left\{\sigma_+\sqrt{\log d}, \gamma^{-1}\log d\right\},$$

which proves the first inequality. Now we prove the second inequality using the following result obtained by extension of Theorem 4 in Tropp (2012) to rectangular matrices.

Proposition 3. *Let $W_1, \ldots, W_n$ be independent random matrices with dimensions $m_1 \times m_2$ that satisfy $\mathbb{E}[W_i] = 0$. Suppose that*

$$\delta_* = \sup_{i \in [\![n]\!]} \inf_{\delta > 0} \left\{\mathbb{E}\left[\exp\left(\|W_i\|/\delta\right)\right] \leq \mathrm{e}\right\} < +\infty. \tag{6.65}$$

Then, there exists an absolute constant c^* such that, for all $t > 0$ and with probability at least $1 - \mathrm{e}^{-t}$ we have

$$\left\|\frac{1}{n}\sum_{i=1}^n W_i\right\| \leq c^* \max\left\{\sigma_W\sqrt{\frac{t + \log d}{n}}, \delta_*\left(\log\frac{\delta_*}{\sigma_W}\right)\frac{t + \log d}{n}\right\},$$

where

$$\sigma_W = \max\left\{\left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[W_iW_i^\top\right]\right\|^{1/2}, \left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[W_i^\top W_i\right]\right\|^{1/2}\right\}.$$

For all $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$ define $\boldsymbol{Z}_{ij} = -\Omega_{ij}\left(\boldsymbol{Y}_{i,j} - g_j'(\boldsymbol{X}^0_{i,j})\right)E_{ij}$. The sub-exponentiality of the variables $\Omega_{ij}\left(\boldsymbol{Y}_{i,j} - g_j'(\boldsymbol{X}^0_{i,j})\right)$ implies that for all $i, j \in [\![m_1]\!] \times [\![m_2]\!]$

$$\delta_{ij} = \inf_{\delta > 0} \quad \left\{\mathbb{E}\left[\exp\left(\left|\Omega_{ij}\left(\boldsymbol{Y}_{i,j} - g_j'(\boldsymbol{X}^0_{i,j})\right)\right|/\delta\right)\right] \leq \mathrm{e}\right\} \leq \frac{1}{\gamma}.$$

We can therefore apply Proposition 3 to the matrices $\boldsymbol{Z}_{ij}$ defined above, with the quantity

$$\sigma_Z = \max\left\{\left\|\frac{1}{m_1m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\mathbb{E}\left[\boldsymbol{Z}_{ij}\boldsymbol{Z}_{ij}^\top\right]\right\|^{1/2}, \left\|\frac{1}{m_1m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\mathbb{E}\left[\boldsymbol{Z}_{ij}^\top\boldsymbol{Z}_{ij}\right]\right\|^{1/2}\right\}.$$

We obtain that for all $t \geq 0$ and with probability at least $1 - \mathrm{e}^{-t}$,

$$\|\Sigma\| \leq c^* \max\left\{\sigma_Z \sqrt{m_1 m_2 (t + \log d)}, \left(\log \frac{1}{\gamma \sigma_Z}\right) \frac{t + \log d}{\gamma}\right\}.$$

We bound σ_Z from above and below as follows.

$$\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathbb{E}\left[\boldsymbol{Z}_{ij} \boldsymbol{Z}_{ij}^\top\right] = \sum_{i=1}^{m_1} \left\{\sum_{j=1}^{m_2} \mathbb{E}\left[\Omega_{ij}^2\right] \mathbb{E}\left[\left(\boldsymbol{Y}_{i,j} - g_j'(\boldsymbol{X}_{i,j}^0)\right)^2\right]\right\} E_{ii}(m_1),$$

where $E_{ii}(n)$, $i, n \geq 1$ denotes the $n \times n$ square matrix with 1 in the (i,i)-th entry and zero everywhere else. Therefore

$$\left\|\frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathbb{E}\left[\boldsymbol{Z}_{ij} \boldsymbol{Z}_{ij}^\top\right]\right\|^{1/2} = \sqrt{\frac{1}{m_1 m_2} \max_i \sum_{j=1}^{m_2} \mathbb{E}\left[\Omega_{ij}^2\right] \mathbb{E}\left[\left(\boldsymbol{Y}_{i,j} - g_j'(\boldsymbol{X}_{i,j}^0)\right)^2\right]}.$$

Then, assumption **H**5 gives

$$\left\|\frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathbb{E}\left[\boldsymbol{Z}_{ij} \boldsymbol{Z}_{ij}^\top\right]\right\|^{1/2} \leq \sigma_+ \sqrt{\frac{1}{m_1 m_2} \left(\max_i \sum_{j=1}^{m_2} \mathbb{E}\left[\Omega_{ij}^2\right]\right)},$$

and

$$\left\|\frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathbb{E}\left[\boldsymbol{Z}_{ij} \boldsymbol{Z}_{ij}^\top\right]\right\|^{1/2} \geq \sigma_- \sqrt{\frac{1}{m_1 m_2} \left(\max_i \sum_{j=1}^{m_2} \mathbb{E}\left[\Omega_{ij}^2\right]\right)}.$$

Similarly, we obtain

$$\left\|\frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathbb{E}\left[\boldsymbol{Z}_{ij}^\top \boldsymbol{Z}_{ij}\right]\right\|^{1/2} \leq \sigma_+ \sqrt{\frac{1}{m_1 m_2} \left(\max_j \sum_{i=1}^{m_1} \mathbb{E}\left[\Omega_{ij}^2\right]\right)},$$

and

$$\left\|\frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathbb{E}\left[\boldsymbol{Z}_{ij}^\top \boldsymbol{Z}_{ij}\right]\right\|^{1/2} \geq \sigma_- \sqrt{\frac{1}{m_1 m_2} \left(\max_j \sum_{i=1}^{m_1} \mathbb{E}\left[\Omega_{ij}^2\right]\right)}.$$

Combining the last four inequalities, we obtain

$$\sigma_- \sqrt{\frac{\beta}{m_1 m_2}} \leq \sigma_Z \leq \sigma_+ \sqrt{\frac{\beta}{m_1 m_2}},$$

and setting $t = \log d$, we further obtain for all $t \geq 0$ and with probability at least $1 - d^{-1}$:

$$\|\Sigma\| \leq c^* \max\left\{\sigma_+ \sqrt{2\beta \log d}, \frac{2 \log d}{\gamma} \log\left(\frac{1}{\sigma_-} \sqrt{\frac{m_1 m_2}{\beta}}\right)\right\},$$

which proves the result.

Chapter 7

R tutorial

Contents

This chapter consists in a tutorial for the R package `mimi`, which may be used independently of the rest of this dissertation. We describe the three models implemented in the package, namely generalized low-rank models (GLRM) for mixed and incomplete data in Section 7.1, as well as two extensions of GLRM to multilevel structures in Section 7.2, and to side information in Section 7.3. Then, we describe in Section 7.4 the implementation of the package, with two optimization algorithms, adapted to small and large scale problems respectively.

7.1 Generalized low-rank models (GLRM)

In its general form, the package `mimi` assumes the following model. Consider a mixed data frame $\boldsymbol{Y} \in \mathbb{R}^{m_1 \times m_2}$, whose columns may be any combination of *numeric*, *binary*, or *count* variables, and with missing values. We place ourselves in the framework of exponential family models, assume that the entries of $\boldsymbol{Y}$ are independent, and approximate the distribution of $\boldsymbol{Y}_{i,j}$, $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$ by:

- a Gaussian distribution if the column $\boldsymbol{Y}_{\cdot,j}$ contains numerical values,
- a Bernoulli distribution if the column $\boldsymbol{Y}_{\cdot,j}$ contains binary variables,
- a Poisson distribution if the column $\boldsymbol{Y}_{\cdot,j}$ contains counts.

In particular, for every entry, we assume a model of the form:

$$\boldsymbol{Y}_{i,j} \sim \mathrm{Exp}_{h_j,g_j}(\boldsymbol{X}^0_{i,j}), \boldsymbol{X}^0_{i,j} \in \mathbb{R}, \tag{7.1}$$

where different choices of h_j and g_j induces either Gaussian, Bernoulli, or Poisson distributions. The package `mimi` aims to estimate the parameter matrix $\boldsymbol{X}^0$ based on a low-rank assumption.

The first and simplest models available in the package `mimi` are Generalized low-rank models (GLRM). In these models, the data matrix $\boldsymbol{Y} \in \mathbb{R}^{m_1 \times m_2}$ is generated (or

approximately generated) by model (7.1). We estimate $\boldsymbol{X}^0$ with the following convex program:

$$\hat{\boldsymbol{X}} \in \operatorname{argmin} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \Omega_{i,j} \mathcal{L}_j(\boldsymbol{X}_{i,j}; \boldsymbol{Y}_{i,j}) + \lambda_1 \|\boldsymbol{X}\|_*. \tag{7.2}$$

In (7.2), $\Omega_{i,j}$ is a binary variable indicating the observed entries: $\Omega_{i,j} = 1$ if $\boldsymbol{Y}_{i,j}$ is observed and 0 otherwise. This amounts to assuming a Missing At Random mechanism with independent entries. On the other hand, $\mathcal{L}_j$ is a loss function adapted to the type of the column j:

- Least squares loss if $\boldsymbol{Y}_{\cdot,j}$ is numeric,
- Logistic loss if $\boldsymbol{Y}_{\cdot,j}$ is binary,
- Poisson loss if $\boldsymbol{Y}_{\cdot,j}$ contains count data.

In fact, problem (7.2) is an extension of `softImpute` (Hastie et al., 2015) for mixed data. The main function of the package estimates the underlying parameter matrix $\boldsymbol{X}^0$ from the noisy and incomplete mixed data $\boldsymbol{Y}$, by solving the optimization problem (7.2).

The package is installed and loaded with the following commands.

```
install.packages("mimi")
library(mimi)
```

Let us now generate a synthetic example. We start by producing a low-rank matrix $\boldsymbol{X}^0$ using the package `denoiseR` (Josse et al., 2017).

```
install.packages("denoiseR")
library(denoiseR)
m1 <- 50 # number of rows
m2 <- 30 # number of columns
r <- 3 # rank of x0
SNR <- 10 # signal to noise ratio (required for function LRsim
    but plays no role)
x0 <- LRsim(m1, m2, r, SNR)$mu
```

Then, we sample heterogeneous data using Gaussian, Bernoulli and Poisson distributions.

```
ngaus <- 10
nber <- 10
npois <- 10
y <- matrix(0, m1, m2)
# The 10 first columns are Gaussian
y[, 1:ngaus] <- matrix(rnorm(m1*ngaus, c(x0[, 1:ngaus])), nrow
   =m1)
# The columns 11-20 are Bernoulli
probs <-
  exp(x0[, (ngaus+1):(ngaus+nber)])/(1+exp(x0[, (ngaus+1):(
     ngaus+nber)]))
y[, (ngaus+1):(ngaus+nber)] <-
  matrix(rbinom(m1*ngaus, 1, prob=c(probs)), nrow=m1)
# The columns 21-30 are Poisson
lambdas <- exp(x0[, (ngaus+nber+1):(ngaus+nber+npois)])
```

```
y[, (ngaus+nber+1):(ngaus+nber+npois)] <-
  matrix(rpois(m1*ngaus, lambda=c(lambdas)), nrow=m1)
```

Finally, we add (MCAR) mising values in the data set $\boldsymbol{Y}$.

```
prob.mis <- 0.3 # proportion of missing values
y[sample(1:(m1*m2), round(prob.mis*m1*m2))] <- NA
```

Now we can estimate the matrix $\boldsymbol{X}^0$ using the `mimi` function. It takes four main arguments: the data matrix $\boldsymbol{Y}$, the type of model to fit (GLRM or extensions, as detailed later), the type of each column, and the value of the regularization parameter λ_1.

```
model <- "low-rank"
var.type <- c(rep("gaussian", 10), rep("binomial", 10), rep("
   poisson", 10))
lambda1 <- 1
res <- mimi(y, model=model, var.type=var.type, lambda1=lambda1
   )
```

Then, the function outputs two results: the estimated parameter matrix $\hat{\boldsymbol{X}}$, called `theta` in the `mimi` output, and an imputed data set: `y.imputed`.

```
head(res$y.imputed) # imputed data set
head(res$theta) # estimated parameter matrix
```

Of course, the output is highly dependent on the value of the regularization parameter λ_1. Thus, we select it by cross-validation, using the function `cv.mimi`. The function takes the same arguments as `mimi` (except λ_1 of course). Optionally, one may set the argument `trace.it` to TRUE, to have printed information about the progress of the computation; indeed, for large data sets, cross-validation can be heavy. The size of the grid of λ_1 values tried is controlled by the `len` parameter (by default 15).

```
rescv <- cv.mimi(y, model=model, var.type=var.type, trace.it=
   TRUE,len=100)
res <- mimi(y, model=model, var.type=var.type, lambda1=rescv$
   lambda)
head(res$y.imputed) # imputed data set
head(res$theta) # estimated parameter matrix
```

7.2 Multilevel GLRM

The `mimi` package also contains an extension of the previous GLRM model for multilevel data, i.e. when the rows of $\boldsymbol{Y}$ are nested within different groups. For instance, individuals coming from different schools or hospitals. Denote by N the total number of groups, and $k(i) \in [\![N]\!]$ the group to which individual i belongs. In this case, the multilevel `mimi` model is the following:

$$\boldsymbol{Y}_{i,j} \sim \text{Exp}_{h_j,g_j}(\boldsymbol{X}^0_{i,j}), \boldsymbol{X}^0_{i,j} = \alpha^0_{k(i),j} + \boldsymbol{\Theta}^0_{i,j}. \tag{7.3}$$

In (7.3), $\alpha^0_{k(i),j}$ is the effect of group $k(i)$ on variable j. For instance, if the j-th column corresponds to the age of individuals, and the individuals of the $k(i)$-th group are older on average than the rest of the population, $\alpha^0_{k(i),j}$ is positive. On the contrary, if the individuals of the $k(i)$-th group are younger on average than the rest of the population,

$\alpha^0_{k(i),j}$ is negative. If the $k(i)$-th group has no effect on the j-th variable, then $\alpha^0_{k(i),j} = 0$. Then, $\mathbf{\Theta}^0_{i,j}$ is an individual effect (which we assume fixed). The multilevel `mimi` option solves the following estimation problem.

$$(\hat{\alpha}, \hat{\mathbf{\Theta}}) \in \operatorname{argmin} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \Omega_{i,j} \mathcal{L}_j(\alpha_{k(i),j} + \mathbf{\Theta}_{i,j}; \boldsymbol{Y}_{i,j}) + \lambda_1 \|\mathbf{\Theta}\|_* + \lambda_2 \|\alpha\|_1. \tag{7.4}$$

In (7.4), the loss function $\mathcal{L}_j$ are defined as in the previous section. The nuclear norm penalty $\lambda_1\|\mathbf{\Theta}\|_*$ inducing low-rank solutions for the matrix of individual effects $\mathbf{\Theta}$, meaning that a few archetypical individuals and a few summary features summarize the individual effects. The ℓ_1 norm penalty $\lambda_2\|\alpha\|_1$ induces sparse solution for the matrix α of group effects, meaning that not all groups have an effect on all variables.

Let us now generate synthetic multilevel data to illustrate how to fit the multilevel mimi model.

```
ngroup <- m1/5 # number of individuals in each group
groups <- as.factor(rep(1:5, each = ngroup)) # factor
   indicating group memberships
N <- nlevels(groups) # number of groups (5)

# matrix of group effects
alpha0 <- matrix(rep(0, N * m2), nrow = N)
alpha0[sample(1:(N * m2), 15)] <- 2

# low-rank individual effects
theta0 <- LRsim(m1, m2, r, SNR)$mu

# parameter matrix (sum of group and individual effects)
x0 <- matrix(rep(as.matrix(alpha0), rep(ncenters, m2)), nrow =
    m1)+theta0
```

Then, we can generate mixed and incomplete data from the matrix $\boldsymbol{X}^0$ exactly as before.

```
ngaus <- 10
nber <- 10
npois <- 10
y <- matrix(0, m1, m2)
y[, 1:ngaus] <- matrix(rnorm(m1*ngaus, c(x0[, 1:ngaus])), nrow
   =m1)
probs <-
  exp(x0[, (ngaus+1):(ngaus+nber)])/(1+exp(x0[, (ngaus+1):(
     ngaus+nber)]))
y[, (ngaus+1):(ngaus+nber)] <-
  matrix(rbinom(m1*ngaus, 1, prob=c(probs)), nrow=m1)
lambdas <- exp(x0[, (ngaus+nber+1):(ngaus+nber+npois)])
y[, (ngaus+nber+1):(ngaus+nber+npois)] <-
  matrix(rpois(m1*ngaus, lambda=c(lambdas)), nrow=m1)
prob.mis <- 0.1
y[sample(1:(m1*m2), round(prob.mis*m1*m2))] <- NA
```

Finally, we estimate the multilevel `mimi` model using the same `mimi` with different arguments: `model` is set to "multilevel", and we must now also specify the value of λ_2, and the factor `groups` indicating the group memberships.

```
model <- "multilevel"
var.type <- c(rep("gaussian", 10), rep("binomial", 10), rep("
   poisson", 10))
lambda1 <- 10
lambda2 <- 5
res <- mimi(y, model=model, var.type=var.type,
            groups=groups, lambda1=10,lambda2=5)
```

There are four outputs to `mimi` in this case: the imputed data set `y.imputed`, the matrix of group effects `alpha`, the matrix of individual effects `theta`, and the parameter matrix (the sum of the group and individual effects) `param`. They are accessed as follows.

```
res$y.imputed # imputed data set
res$alpha # matrix of group effects
res$theta # matrix of individual effects
res$param # parameter matrix
```

Here again, the results are highly dependent of the parameters λ_1 and λ_2, and we may also select them with the `cv.mimi` function. Of course, because we must now go through a two-dimensional grid with different values of λ_1 and λ_2, this can be quite computationally heavy.

```
# this takes around 20 minutes on my computer
rescv <- cv.mimi(y, model=model, var.type=var.type, groups=
   groups, trace.it=T)
res <- mimi(y, model=model, var.type=var.type,
            groups=groups, lambda1=rescv$lambda1,lambda2=rescv
               $lambda2)
```

7.3 GLRM with side information

The last model implemented in the `mimi` package is a GLRM with side information. To give a motivating example, consider a recommendation system problem where users rate items with binary ratings (0 or 1). In this problem, the data is incomplete, and the goal is to recommend relevant items to users. The scope of the GLRM with side information is to incorporate knowledge about users and items attributes in the imputation model. The GLRM with side information in `mimi` may also be generalized to accommodate mixed data (binary, numeric, counts). The model is the following.

$$\boldsymbol{Y}_{i,j} \sim \mathrm{Exp}_{h_j,g_j}(\boldsymbol{X}^0_{i,j}), \boldsymbol{X}^0_{i,j} = \langle \alpha, \boldsymbol{U}_{i,j} \rangle + \boldsymbol{\Theta}^0_{i,j}, \tag{7.5}$$

where $\boldsymbol{U}_{i,j} \in \mathbb{R}^q$ is a vector of covariates which may contain information about the i-th row (e.g., user characteristics) as well as information about the j-th column (e.g., item characteristics). In (7.5), α^0 is a vector of covariate effects in a generalized linear regression framework. For instance, if the k-th covariate corresponds to the age of individuals, and young individuals tend to give higher scores to all items, then α^0_k is positive. On the contrary, if young individuals tend to give lower scores to all items, then α^0_k is negative. If the age has no effect on the ratings, then $\alpha^0_k = 0$. Then, $\boldsymbol{\Theta}^0_{i,j}$ is an individual effect, residual, or interaction. The side information `mimi` option solves the following estimation problem.

$$(\hat{\alpha}, \hat{\boldsymbol{\Theta}}) \in \mathrm{argmin} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \Omega_{i,j} \mathcal{L}_j(\langle \boldsymbol{U}_{i,j}, \alpha \rangle + \boldsymbol{\Theta}_{i,j}; \boldsymbol{Y}_{i,j}) + \lambda_1 \|\boldsymbol{\Theta}\|_* + \lambda_2 \|\alpha\|_1. \tag{7.6}$$

In (7.6), the loss function $\mathcal{L}_j$ are defined as in the previous sections. The nuclear norm penalty $\lambda_1\|\mathbf{\Theta}\|_*$ inducing low-rank solutions for the matrix of residuals $\mathbf{\Theta}$, meaning that a few archetypical individuals and a few summary features summarize them. The ℓ_1 norm penalty $\lambda_2\|\alpha\|_1$ induces sparse solution for the matrix α of group effects, meaning that not all covariates have an effect on the observations (on the ratings for instance).

Let us now generate a synthetic GLRM with side information.

```
# covariate matrix
N <- 4
U <- matrix(rnorm(m1*m2*N), nrow=m1*m2)

# vector of covariate effects
alpha0 <- rep(0,N)
alpha0[sample(1:N,2)] <- 2

# low-rank individual effects
theta0 <- LRsim(m1, m2, r, SNR)$mu

# parameter matrix (sum of group and individual effects)
x0 <- matrix(U%*%alpha0, nrow = m1)+theta0
```

Then, we can generate binary incomplete data from the matrix $\boldsymbol{X}^0$.

```
#binary and incomplete data
probs <- exp(x0)/(1+exp(x0))
y <- matrix(rbinom(m1*m2, 1, prob=c(probs)), nrow=m1)
prob.mis <- 0.1
y[sample(1:(m1*m2), round(prob.mis*m1*m2))] <- NA
```

Finally, we estimate the GLRM with side information with the same `mimi` with different arguments: `model` is set to "covariates", and we must specify the value of λ_2, and the matrix of covariates `x` containing the predictors.

```
model <- "covariates"
var.type <- c(rep("binomial", 30))
lambda1 <- 1
lambda2 <- 1
res <- mimi(y, model=model, var.type=var.type,
            x=U, lambda1=lambda1,lambda2=lambda2)
```

There are four outputs to `mimi` in this case: the imputed data set `y.imputed`, the vector of covariate effects `alpha`, the matrix of individual effects `theta`, and the parameter matrix (the sum of the group and individual effects) `param`. They are accessed as follows.

```
res$y.imputed # imputed data set
res$alpha # matrix of group effects
res$theta # matrix of individual effects
res$param # parameter matrix
```

Here again, the results are highly dependent of the parameters λ_1 and λ_2, and we may also select them with the `cv.mimi` function. Of course, because we must now go through a two-dimensional grid with different values of λ_1 and λ_2, this can be quite computationally heavy.

```
# this takes around 20 minutes on my computer
```

```
rescv <- cv.mimi(y, model=model, var.type=var.type, x=U, trace
   .it=T)
res <- mimi(y, model=model, var.type=var.type,
            groups=groups, lambda1=rescv$lambda1,lambda2=rescv
               $lambda2)
```

7.4 Implementations

To solve the minimization problems (7.2), (7.4) and (7.6), we implemented two different algorithms in the `mimi` package. The first one is based on block coordinate gradient descent, and involves solving a LASSO problem and computing a *full-rank* SVD at each iteration. The second one is a mixed coordinate gradient descent (MCGD) algorithm, and involves solving a LASSO problem and computing a *rank-1* SVD at each iteration. In small dimensions, we recommend using the BCGD option, which involves costly iterations (full-rank SVDs), but converges in fewer iterations. On the other hand, in large dimensions, we recommend using the MCGD option, which converges at a (slower) sublinear rate (Robin et al., 2018), but involves less costly iterations (rank-1 SVDs). When Poisson variables are included in the model, we recommend to used MCGD whatever the size of the data set, since quadratic approximations can be slow with Poisson loss functions. By default, MCGD is used in the package `mimi`. This option may be changed using the "algo" argument:

```
res <- mimi(y, model=model, var.type=var.type,
            groups=groups, lambda1=rescv$lambda1,
            lambda2=rescv$lambda2,
            algo = "mcgd") # mixed coordinate gradient descent
                (default)
res <- mimi(y, model=model, var.type=var.type,
            groups=groups, lambda1=rescv$lambda1,
            lambda2=rescv$lambda2,
            algo = "bcgd") # block coordinate gradient descent
```

Chapter 8

Imputation of mixed data with multilevel singular value decomposition

Contents

8.1 Introduction

Analyzing large data sets offers new opportunities to better understand underlying processes, and to increase statistical power. For this reason, practitioners are increasingly encouraged to share their data–after applying anonymization procedures if needed. Yet, data accumulation often implies relaxing acquisition procedures or compounding diverse sources. As a consequence, data sets often contain mixed data, i.e., both quantitative and qualitative, as well as many missing values. In addition, aggregated data often present a natural *multilevel* structure, where individuals, or samples, are nested within different groups, such as different countries or hospitals. For instance, we focus in this chapter on a running example in public health, with the imputation of a subsample of the Traumabase data set, a large severe trauma registry from Paris hospitals maintained by the Traumabase Group (`http://www.traumabase.eu/en_US`). In this data set, around 40% of the data is missing. Another distinctive characteristic of the Traumabase is that it contains mixed data, such as quantitative physiological measurements and qualitative socio-professional information. Lastly, the Traumabase data set results from the aggregation of multilple smaller data sets coming from several French hospitals. As a broad heterogeneity in the care process is known to occur across trauma centers, modeling this multilevel structure is crucial. Although we illustrate the method on a particular example, we emphasize that the method is quite general, as multi-source, heterogeneous and incomplete data are in fact found in many applications, such as census data, social surveys, and medical applications in general.

Imputation of multilevel data has therefore drawn some attention recently, but current solutions are not designed to handle mixed data, and suffer from important drawbacks such as their computational cost. In this chapter, we propose a single imputation method for multilevel data, which can be used to complete either quantitative, categorical, or mixed data. The method is based on multilevel singular value decomposition (SVD), which consists in decomposing the variability of the data into two components, the between and within groups variability, and performing an SVD on both parts. We show on a simulation study that, in comparison to competitors, the method has the advantage of scaling up to data sets of larger dimensions, and being computationally faster. Furthermore, it is, as far as we know, the first multilevel imputation method available for mixed data. We also apply the method to impute a subsample of the Traumabase medical registry. To overcome some of the obstacles associated to the aggregation of medical data, we turn to distributed computation. Indeed, such computations allow hospitals to benefit from data coming from other centers, which increases the chance of finding similar patients, without the need to explicitly share patients information. The method is implemented in the R package `missMDA`.

8.1.1 Problem formulation and related work

Consider a data set $\boldsymbol{Y} \in \mathbb{R}^{m_1 \times m_2}$, which is naturally the row concatenation of K smaller data sets $\boldsymbol{Y}_k \in \mathbb{R}^{n_k \times m_2}$, $k \in [\![K]\!]$. $\boldsymbol{Y}$ collects the measurements of m_2 variables across a population of m_1 individuals categorized in K groups, such that the k-th group contains

n_k individuals and $\sum_{k=1}^{K} n_k = m_1$:

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \hline \boldsymbol{Y}_2 \\ \hline \vdots \\ \hline \boldsymbol{Y}_K \end{pmatrix} \begin{matrix} \updownarrow n_1 \\ \updownarrow n_2 \\ \vdots \\ \updownarrow n_K \end{matrix} .$$

For a group $k \in [\![K]\!]$, an individual of the k-th group $i_k \in [\![n_k]\!]$ and a variable $j \in [\![m_2]\!]$, we denote by $y_{k,i_k,j}$ the value of variable j taken by individual i_k in group k. Such structure is often called *multilevel structure*, and occurs in many fields of applications. Famous examples include pupils nested within schools or patients within hospitals. Throughout this article, we focus on this latter example with a running application in public health. If some entries of $\boldsymbol{Y}$ are missing, we denote by $\boldsymbol{M}$ the indicator matrix of observations, with $\boldsymbol{M}_{k,i_k,j} = 1$ if $y_{k,i_k,j}$ is observed and $\boldsymbol{M}_{k,i_k,j} = 0$ otherwise. To handle missing values, corresponding to $\boldsymbol{M}_{k,i_k,j} = 0$, a popular approach (Little and Rubin, 2002) consists in imputing them, that is, in replacing the missing entries with plausible values to obtain a completed data set. To do so, several approaches have been developed, and a complete overview of state-of-the-art multilevel imputation methods is available in Audigier et al. (2018). Latest proposals have focused on handling both sporadically missing values, meaning that some variables are partly missing in some of the groups, and systematically missing values, that is, when some variables are completely unobserved in (at least) one of the groups. In most imputation methods, the hierarchical structure is modeled using a random effects regression model, as suggested for instance in Resche-Rigon and White (2016) and Quartagno and Carpenter (2016). However, current solutions suffer from important gaps that deserve further development. In particular, they are not designed to handle mixed data (quantitative and categorical), struggle with large dimensions and are extremely costly in terms of computations.

At the same time, imputation by iterative singular value decomposition (SVD) algorithms have proven excellent imputation capacities for quantitative (Hastie et al., 2015), qualitative (Audigier et al., 2018) and mixed data (Audigier et al., 2016). This can be explained in part because they assume an underlying low-rank structure for the data which is plausible for many large data sets, as discussed in Udell and Townsend (2018). These methods behave well compared to competitors in terms of prediction of the missing values, in particular when the number of observations is small with respect to the number of variables, and when the qualitative variables have many categories and some of them are rare. In addition, they are often competitive in terms of execution time. However, these methods are not dedicated to the multilevel data we address in this chapter. The work we present here can be cast as an extension of single imputation methods based on SVD to the multilevel framework.

8.1.2 Multilevel principal component analysis

For sake of clarity, we start by reviewing the multilevel extension of principal component analysis (PCA, Pearson (1901)) described in Timmerman (2006). Assume the data set $\boldsymbol{Y}$ contains only quantitative variables. The measured values can be decomposed, for a group $k \in [\![K]\!]$, an individual $i_k \in [\![n_k]\!]$ in the k-th group and a variable $j \in [\![m_2]\!]$, as

$$y_{k,i_k,j} = \underbrace{y_{.,.,j}}_{\text{offset}} + \underbrace{y_{k,.,j} - y_{.,.,j}}_{\text{between}} + \underbrace{y_{k,i_k,j} - y_{.,k,j}}_{\text{within}} .$$

Here,

$$y_{.,.,j} = \frac{1}{m_1} \sum_{k=1}^{K} \sum_{i_k=1}^{n_k} y_{k,i_k,j}$$

is the overall mean of variable j and

$$y_{k,.,j} = \frac{1}{n_k} \sum_{i_k=1}^{n_k} y_{k,i_k,j}$$

is the mean of variable j among individuals of group k. Then, $(y_{k,.,j} - y_{.,.,j})$ is the deviation of group k to the overall mean of variable j, and $(y_{k,i_k,j} - y_{k,.,j})$ is the deviation of individual i_k to the mean of variable j in group k. Written in matrix form, this gives

$$\boldsymbol{Y} = \mathbb{1}_{m_1} m^\top + \boldsymbol{Y}_b + \boldsymbol{Y}_w,$$

where $\mathbb{1}_{m_1}$ is the $m_1 \times 1$ vector of ones and m is the $m_2 \times 1$ vector containing the overall means of the m_2 variables, $\boldsymbol{Y}_b$ contains the variable means per group minus the overall means, and $\boldsymbol{Y}_w$ contains the residuals. Similarly to what is done in analysis of variance, we can split the sum of squares for each variable j as

$$\sum_{k=1}^{K} \sum_{i_k=1}^{n_k} y_{k,i_k,j}^2 = \sum_{k=1}^{K} n_k y_{.,.,j}^2 + \sum_{k=1}^{K} n_k (y_{k,.,j} - y_{.,.,j})^2 + \sum_{k=1}^{K} \sum_{i_k=1}^{n_k} (y_{k,i_k,j} - y_{k,.,j})^2.$$

In the classical framework where there is no multilevel structure, PCA yields the best fixed rank estimator of $\boldsymbol{Y}$ in terms of the least squares criterion. The multilevel extension naturally leads, for $(k, i_k, j) \in [\![K]\!] \times [\![n_k]\!] \times [\![m_2]\!]$, to modelling the offsets, the between and within terms separately by explaining as well as possible both the between and within sum of squares. Therefore, multilevel PCA (MLPCA) consists in assuming two low-rank models, for the between matrix $\boldsymbol{Y}_b = (y_{k,.,j} - y_{.,.,j})_{k,j}$, approximated by a matrix of rank Q_b, and for the within matrix $\boldsymbol{Y}_w = (y_{k,i_k,j} - y_{k,.,j})_{k,i_k,j}$, approximated by a matrix of rank Q_w. This yields the following decomposition:

$$\boldsymbol{Y} = \mathbb{1}_{m_1} m^\top + \boldsymbol{F}_b \boldsymbol{V}_b^\top + \boldsymbol{F}_w \boldsymbol{V}_w^\top + \boldsymbol{E}. \tag{8.1}$$

$\boldsymbol{F}_b$ is the matrix of size $m_1 \times Q_b$ containing the between component scores

$$\boldsymbol{F}_b = \begin{pmatrix} \boldsymbol{F}_{b,1} \\ \hline \boldsymbol{F}_{b,2} \\ \hline \vdots \\ \hline \boldsymbol{F}_{b,K} \end{pmatrix}, \tag{8.2}$$

where for all $k \in [\![K]\!]$, $\boldsymbol{F}_{b,k}$ is row-wise constant, with $f_{b,k}$ repeated on every row. Let $I_k \in \{0,1\}^{m_1}$ be the indicator vector of group k such that the i-th entry $I_{k,i} = 1$ if individual i belongs to group k and 0 otherwise. Representation (8.2) is equivalent to

$$\boldsymbol{F}_b = \sum_{k=1}^{K} I_k f_{b,k}^\top.$$

$\boldsymbol{V}_b$ is the $m_2 \times Q_b$ between loadings matrix, $\boldsymbol{F}_w$ $(m_1 \times Q_w)$ denotes the within component scores, and finally $\boldsymbol{V}_w$ $(m_2 \times Q_w)$ denotes the within loadings matrix, and $\boldsymbol{E}$

$(m_1 \times m_2)$ denotes the matrix of residuals. Note that in this model, the within loadings matrix $\boldsymbol{V}_w$ is constrained to be constant across groups. Model (8.1) is called multilevel simultaneous component analysis (MLSCA) in Timmerman (2006). We keep the name MLPCA for simplicity.

In terms of interpretation, the low rank structure on the between part implies that there are dimensions of variability to describe the hospitals: for instance the first dimension could oppose hospitals that resort to a large extent to pelvic and chest X-ray to hospitals where those examinations are not usually performed. The low rank structure on the within part implies that there are dimensions of variability to describe the patients: for instance the first dimension opposes patients with a head trauma (taking specific values for variables related to head trauma) to other patients. The constraint that the within loading matrix is the same across hospitals means that this dimension is the same from one hospital to the other but the strength of the dimension, i.e. the variability of patients on the dimension, can differ from one group to the other. This constraint also leads to fewer parameters to estimate.

The model is fitted by solving the least squares problem with respect to the parameters $(m, \boldsymbol{F}_b, \boldsymbol{V}_b, \boldsymbol{F}_w, \boldsymbol{V}_w)$:

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{Y} - \left(\mathbb{1}_{m_1} m^\top + \boldsymbol{F}_b \boldsymbol{V}_b^\top + \boldsymbol{F}_w \boldsymbol{V}_w^\top\right)\|_F^2 \\ \text{subject to} & \boldsymbol{F}_b = \sum_{k=1}^K I_k f_{b,k}^\top, \\ & \sum_{k=1}^K n_k f_{b,k} = 0_{Q_b}, \\ & \mathbb{1}_{m_1}^\top \boldsymbol{F}_w = 0_{Q_w}. \end{array} \tag{8.3}$$

where the last two constraints serve for identifiability. Since the three components $\mathbb{1}_{m_1} m^\top$, $\boldsymbol{F}_b \boldsymbol{V}_b^\top$ and $\boldsymbol{F}_w \boldsymbol{V}_w^\top$ are orthogonal, (8.3) is equivalent to solving the three following subproblems. Denote $\bar{\boldsymbol{Y}} \in \mathbb{R}^p$ the vector of column means of $\boldsymbol{Y}$.

$$\begin{array}{ll} \text{minimize}_m & \|\bar{\boldsymbol{Y}} - m\|_2^2, \\ \text{minimize}_{\boldsymbol{F}_b, \boldsymbol{V}_b} & \|\boldsymbol{Y}_b - \boldsymbol{F}_b \boldsymbol{V}_b^\top\|_F^2, \\ \text{minimize}_{\boldsymbol{F}_w, \boldsymbol{V}_w} & \|\boldsymbol{Y}_w - \boldsymbol{F}_w \boldsymbol{V}_w^\top\|_F^2. \end{array} \tag{8.4}$$

The constraints do not need to be specified: $\bar{\boldsymbol{Y}}$, $\boldsymbol{Y}_b$ and $\boldsymbol{Y}_w$ are orthogonal projections of $\boldsymbol{Y}$ on orthogonal subspaces, thus the solutions to Equation (8.4) belong to the same orthogonal spaces, and therefore satisfy the constraints of Equation (8.3). The solution is obtained in Timmerman (2006) by computing the variables mean to estimate m; then, truncated SVD of $\boldsymbol{Y}_b = \boldsymbol{U}_b \Lambda_b^{1/2} \boldsymbol{V}_b^\top$ at rank Q_b and of $\boldsymbol{Y}_w = \boldsymbol{U}_w \Lambda_w^{1/2} \boldsymbol{V}_w^\top$ at rank Q_w are performed to estimate the parameters $\boldsymbol{F}_b = \boldsymbol{U}_b \Lambda_b^{1/2}, \boldsymbol{V}_b, \boldsymbol{F}_w = \boldsymbol{U}_w \Lambda_w^{1/2}$ and $\boldsymbol{V}_w$. Such a solution is in agreement with the rationale of performing an SVD on the matrix of means per group to study the differences between groups and a SVD of the matrix centered by groups to study the differences between patients after discarding the hospital effects.

8.1.3 Contributions and outline of the chapter

In this chapter, we introduce two new multilevel component methods adapted to categorical and mixed data, respectively. Furthermore, we directly extend them to the missing data framework, and develop imputation procedures based on them. We conduct an empirical study to evaluate our methods in realistic settings, and apply them to the imputation of a subsample of a medical registry. The chapter is organized as follows.

In Section 8.2, we introduce Multilevel Multiple Correspondence Analysis (MLMCA) and Multilevel Factorial Analysis of Mixed Data (MLFAMD), two multilevel extensions of component methods designed to analyze qualitative and mixed data respectively. MLMCA and MLFAMD may also be seen as extensions of MLPCA to categorical and mixed data: to the best of our knowledge, we are the first to propose such methods. Our second main contribution is to propose, in Section 8.3, multilevel single imputation methods to impute categorical and mixed variables with multilevel structures, based on MLMCA and MLFAMD. In Section 8.4, we show on synthetic data that our methods have smaller prediction errors than competitors when the data are generated with a multilevel model. Finally, in Section 8.5, we illustrate the methods with the imputation of a large registry from Paris hospitals and discuss how to distribute the computation. The methods are implemented in the R (R Core Team, 2017) package `missMDA` (Josse and Husson, 2016).

8.2 Multilevel component methods

In the classical data analysis framework without multilevel structure, Multiple Correspondence Analysis (MCA) and Factorial Analysis of Mixed Data (FAMD), were developed as equivalents of PCA for categorical and mixed data, respectively. In short, MCA and FAMD proceed by applying geometric transformations to the data, and then performing PCA, with additional weights designed to balance the columns of different types. To analyze multilevel categorical and mixed data, it is thus natural to define extensions of MLPCA, based on the same geometric transformations. In this section, we introduce such extensions: MLMCA and MLFAMD.

8.2.1 Multilevel Multiple Correspondence Analysis (MLMCA)

We now propose a new extension of MLPCA to analyse categorical variables with multilevel structures; the method is based on MCA (Greenacre and Blasius, 2006; Husson et al., 2010). MCA is considered to be the counterpart of PCA for categorical data analysis, and has been successfully applied in many fields, such as survey data analysis, to visualize associations between categories. More precisely, categorical data are coded as a complete disjunctive table $\boldsymbol{Z}$ where all categories of all variables are represented as indicator vectors. In other words $z_{ic} = 1$ if individual i takes the category c and 0 otherwise. For example, if there are $m_2 = 2$ variables with 2 and 3 levels respectively, we have the following equivalent codings:

$$\boldsymbol{Y} = \begin{pmatrix} 1 & 1 \\ 2 & 3 \\ 1 & 2 \\ 2 & 3 \\ 2 & 2 \\ 2 & 2 \end{pmatrix} \quad \Longleftrightarrow \quad \boldsymbol{Z} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix}.$$

For $1 \leq j \leq m_2$ we denote by C_j the number of categories of variable j, and $C = \sum_{j=1}^{m_2} C_j$ the total number of categories. For $1 \leq c \leq C$, $\boldsymbol{Z}_{,c}$ is the c-th column of $\boldsymbol{Z}$ corresponding to the indicator of category c. We define $\pi_c = m_1^{-1} \mathbb{1}_{m_1}^\top \boldsymbol{Z}_{,c}$ the proportion of observations in category c, $\pi = (\pi_1, \ldots, \pi_C)^\top$ and $\boldsymbol{D}_\pi$ the $C \times C$ diagonal matrix

with π on its diagonal. Multiple correspondence analysis (MCA) is defined as the SVD of the matrix

$$\boldsymbol{A} = \frac{1}{m_1 m_2} \left(\boldsymbol{Z} - \mathbb{1}_{m_1} \pi^\top \right) \boldsymbol{D}_\pi{}^{-1/2}. \tag{8.5}$$

This specific transformation endows MCA with many properties: the distances between the rows and columns in the transformed matrix $\boldsymbol{A}$ coincide with the chi-squared distances, the first principal component (the scores) corresponds to the quantitative variable the most related to the categorical variables in the sense of the η^2 coefficient of analysis of variance (Husson et al., 2010, Section 3). This latter property justifies why MCA is considered as the equivalent of PCA for categorical data.

We introduce the following strategy for multilevel MCA (MLMCA). From the indicator matrix of dummy variables $\boldsymbol{Z}$, we start by defining a between part and a within part. MCA, in the sense of the SVD of a transformed matrix (8.5), will then be applied on each part. For $k \in [\![K]\!]$, define $\boldsymbol{Z}_k$ the sub-matrix of $\boldsymbol{Z}$ containing all categories and the rows corresponding to individuals of group k. The between part is defined block-wise as the mean of the indicator matrix per group k with the following $n_k \times m_2$ matrices, stacked below one another:

$$\boldsymbol{Z}_{b,k} = n_k^{-1} \mathbb{1}_{n_k} \mathbb{1}_{n_k}^\top \boldsymbol{Z}_k.$$

The entries of $\boldsymbol{Z}_{b,k}$ contain the proportion of observations taking each category in group k (n_{c_k}/n_k) (for instance the proportion of individuals carrying some disease in a particular hospital). Finally

$$\boldsymbol{Z}_b = \begin{pmatrix} \boldsymbol{Z}_{b,1} \\ \hline \boldsymbol{Z}_{b,2} \\ \hline \vdots \\ \hline \boldsymbol{Z}_{b,K} \end{pmatrix}.$$

MCA (8.5) is afterwards applied to the fuzzy indicator matrix $\boldsymbol{Z}_b$, i.e. SVD is applied to

$$(\boldsymbol{Z}_b - \mathbb{1}_{m_1} \pi^\top) \boldsymbol{D}_\pi^{-1/2}.$$

This results in obtaining between component scores $\boldsymbol{F}_b \in \mathbb{R}^{m_1 \times Q_b}$ and between loadings $\boldsymbol{V}_b \in \mathbb{R}^{m_1 \times Q_b}$. The estimated between matrix is then $\hat{\boldsymbol{Z}}_b = \boldsymbol{F}_b \boldsymbol{V}_b^\top \boldsymbol{D}_\pi^{1/2} + \mathbb{1}_{m_1} \pi^\top$. As for the within part, MCA is applied to the data where the between part has been swept out, i.e. SVD is applied to the following matrix:

$$(\boldsymbol{Z} - \boldsymbol{Z}_b) \, \boldsymbol{D}_\pi^{-1/2}. \tag{8.6}$$

Weighting by the inverse square root of the margins of the categories implies that more weight is given to categories which are rare over all groups (for instance a rare disease). We obtain within component scores $\boldsymbol{F}_w \in \mathbb{R}^{m_1 \times Q_w}$, within loadings $\boldsymbol{V}_w \in \mathbb{R}^{m_1 \times Q_w}$, and the estimated within matrix $\hat{\boldsymbol{Z}}_w = \boldsymbol{F}_w \boldsymbol{V}_w^\top \boldsymbol{D}_\pi^{1/2}$.

Finally, we estimate $\boldsymbol{Z}$ by $\hat{\boldsymbol{Z}} = \hat{\boldsymbol{Z}}_b + \hat{\boldsymbol{Z}}_w$. As with MCA (Josse et al., 2012), the reconstructed fuzzy indicator matrix $\hat{\boldsymbol{Z}} = \hat{\boldsymbol{Z}}_b + \hat{\boldsymbol{Z}}_w$ has the property that the sum of values for one individual and one variable is equal to one. Consequently, the estimated values can be considered as degrees of membership to the categories. This property will prove useful for the imputation.

Remark Another approach to define MLMCA would have been to directly apply MLPCA on the matrix $\boldsymbol{A}$ defined in (8.5). As it turns out, these two strategies are equivalent, which reinforces our definition of Multilevel MCA.

8.2.2 Multilevel Factorial Analysis of Mixed Data (MLFAMD)

Consider now a mixed data set $\boldsymbol{Y} = (\boldsymbol{Y}_q, \boldsymbol{Y}_c)$, where $\boldsymbol{Y}_q$ is a submatrix containing q quantitative variables, and $\boldsymbol{Y}_c$ a submatrix containing c categories:

$$\boldsymbol{Y} = \left(\begin{array}{ccc|cc|ccc} 0.3 & -3.4 & 0.1 & 0 & 1 & 0 & 1 & 0 \\ 1.4 & 0.4 & -2.8 & 1 & 0 & 0 & 0 & 1 \\ 9.2 & 1.8 & 7.1 & 0 & 1 & 1 & 0 & 1 \end{array} \right).$$
$$\underbrace{\qquad\qquad}_{\boldsymbol{Y}_q} \quad \underbrace{\qquad\qquad\qquad}_{\boldsymbol{Y}_c}$$

In the same flavour, we define a multilevel method for mixed data by extending a counterpart of PCA for mixed data, namely factorial analysis for mixed data (FAMD), presented in Pagès (2015). FAMD consists in transforming the categorical variables as in MCA (8.5) and concatenating them with the quantitative variables. Then, each quantitative variable is standardized (centered and divided by its standard deviation). Finally, SVD is applied to this weighted matrix. This specific weighting ensures that all quantitative and categorical variables play the same role in the analysis. More precisely, the principal components, denoted $\boldsymbol{F}_q$ for $q = 1, ..., Q$ maximize the link between the quantitative and categorical variables in the following sense:

$$\boldsymbol{F}_q = \underset{\boldsymbol{F}_q \in \mathbb{R}^{m_1}}{\arg\max} \sum_{j=1}^{q} r^2(\boldsymbol{F}_q, \boldsymbol{Y}_j) + \sum_{j_c=1}^{c} \eta^2(\boldsymbol{F}_q, \boldsymbol{Y}_{j_c}),$$

with the constraint that $\boldsymbol{F}_q$ is orthogonal to $\boldsymbol{F}_{q'}$ for all $q' \neq q$ and with $\boldsymbol{Y}_j$ being the variable j, r^2 the square of the correlation coefficient and η^2 the square of the correlation ratio. This formulation highlights that FAMD can be seen as the counterpart of PCA for mixed data. More details about the method are given in Pagès (2015).

The extension to a multilevel structure, named MLFAMD, is now straightforward following what is done for MCA and categorical data in the previous section. Denote C the number of categories, $\pi \in (0,1)^C$ the vector of categories proportions and $\boldsymbol{D}_\pi$ the $C \times C$ diagonal matrix with π on its diagonal. Denote $m \in \mathbb{R}^q$ the vector of means of the quantitative variables, and $\boldsymbol{\Sigma} \in \mathbb{R}^{q \times q}$ the diagonal matrix containing the standard deviations of $\boldsymbol{Y}_q$. MLFAMD consists in doing the following transformations.

$$\boldsymbol{W} \in \mathbb{R}^{m_1 \times (q+c)} \leftarrow \left((\boldsymbol{Y}_q - \mathbb{1}_{m_1} m^\top)\boldsymbol{\Sigma}^{-1}, \frac{1}{m_1 m_2} \left(\boldsymbol{Y}_c - \mathbb{1}_{m_1} \pi^\top \right) {\boldsymbol{D}_\pi}^{-1/2} \right). \tag{8.7}$$

Then, multilevel SVD is performed on the matrix $\boldsymbol{W}$. This boils down to computing the between and within part, and performing SVD on both separately:

$$\boldsymbol{W}_b = \sum_{k=1}^{K} \mathbb{1}_{n_k} \mathbb{1}_{n_k}^\top \boldsymbol{W}, \quad \boldsymbol{W}_w = \boldsymbol{W} - \boldsymbol{W}_b. \tag{8.8}$$

8.3 Multilevel imputation

We now focus on the case where some values in $\boldsymbol{Y}$ are missing, and describe how MLPCA, MLMCA and MLFAMD may be extended into multilevel imputation methods.

8.3.1 Imputation with MLPCA

Recall that $\boldsymbol{M}$ is the $m_1 \times m_2$ indicator matrix of observations with $\boldsymbol{M}_{k,i_k,j} = 1$ if $y_{k,i_k,j}$ is observed and $\boldsymbol{M}_{k,i_k,j} = 0$ otherwise. We denote by $\boldsymbol{M}_k$ the restriction of matrix $\boldsymbol{M}$ to the rows belonging to group $k \in [\![K]\!]$. Consider a Missing (Completely) At Random (M(C)AR) setting (Little and Rubin, 2002) where the process that generated the missing values can be ignored for likelihood based model. To impute the missing values using the multilevel model (8.1), we need to estimate its parameters from incomplete data. This can be done through low rank matrix estimation for incomplete data sets (Hastie et al., 2015) by weighting the least squares criterion (8.3) with $\{0, 1\}$ weights indicating the observed entries. The optimization problem is the following:

$$\begin{array}{ll} \text{minimize} & \| \boldsymbol{M} \odot \left(\boldsymbol{Y} - (\mathbb{1}_{m_1} m^\top + \boldsymbol{F}_b \boldsymbol{V}_b^\top + \boldsymbol{F}_w \boldsymbol{V}_w^\top) \right) \|_F^2 \\ \text{subject to} & \boldsymbol{F}_b = \sum_{k=1}^K I_k f_{b,k}^\top, \\ & \sum_{k=1}^K n_k f_{b,k} = 0_{Q_b}, \\ & \mathbb{1}_{m_1}^\top \boldsymbol{F}_w = 0_{Q_w}, \end{array} \tag{8.9}$$

where $\odot$ denotes the Hadamard entry-wise product. Note that, with this approach, it is possible to handle both sporadically and systematic missing values as it will be illustrated in the Section 8.4. In Josse et al. (2013), the authors solved program (8.9) using an iterative imputation algorithm. Note that the aim in Josse et al. (2013) was to perform MLPCA with missing values, that is, to estimate the parameters *in spite* of the missing values, and not to impute multilevel data. The distinction may appear tenuous as the algorithm involves an underlying imputation of the missing entries, but the quality of this imputation was never evaluated in itself. Let $\hat{m}^0$ be the mean vector of the non-missing entries. The algorithm works iteratively, as described in Algorithm 5. Such an algorithm starts by replacing the missing values by initial values (for example

Algorithm 5 Iterative MLPCA

1: **Initialize** missing values: $\hat{\boldsymbol{Y}} = \boldsymbol{Y} \odot \boldsymbol{M} + \mathbb{1}_{m_1} \hat{m}^{0\top} \odot (\mathbb{1}_{m_1} \mathbb{1}_{m_2}^\top - \boldsymbol{M})$.
2: Estimate $\boldsymbol{F}_b, \boldsymbol{V}_b, \boldsymbol{F}_w, \boldsymbol{V}_w$ with multilevel PCA (8.3);
3: Impute $\boldsymbol{Y} = \boldsymbol{Y} \odot \boldsymbol{M} + (\mathbb{1}_{m_1} m^\top + \boldsymbol{F}_b \boldsymbol{V}_b^\top + \boldsymbol{F}_w \boldsymbol{V}_w^\top) \odot (\mathbb{1}_{m_1} \mathbb{1}_{m_2}^\top - \boldsymbol{M})$;
4: Update means $m = m_1^{-1} \mathbb{1}_{m_1}^\top \boldsymbol{Y}$.
5: **Repeat** steps 1, 2, 3 until empirical stabilization of the prediction.

the mean of the non-missing entries), then the estimator (here MLPCA) is computed on the completed matrix and the predicted values of the missing entries are updated using the values given by the new estimation. The two steps of imputation and estimation are repeated until empirical stabilization of the prediction. The detailed algorithm for iterative MLPCA with missing values is given in Algorithm 6. In the end, it outputs both the between and within scores and loadings obtained from the incomplete data set, and a data set imputed using the MLPCA model (8.1). Thus, it is a single imputation method (Schafer, 1997; Little and Rubin, 2002) which takes into account the multilevel structure of the data. Note also that the algorithm corresponds to an expectation-maximization

Algorithm 6 Iterative MLPCA (detailed)

Input: $\boldsymbol{Y} = (\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}}) \in \mathbb{R}^{m_1 \times m_2}$, Q_b, Q_w
Inintialize: $\hat{m}^0$ the mean vector of the non-missing entries
for $(i, j) \in [\![m_1]\!] \times [\![m_2]\!]$ **do**
  if $\boldsymbol{M}_{ij} = 0$ **then**
    $\boldsymbol{Y}_{ij} \leftarrow \hat{m}^0_j$
  end if
end for
for t=1,... **do**
  Estimation of the between structure
  $\boldsymbol{Y}_b = \sum_{k=1}^{K} n_k^{-1} I_k \left(\mathbb{1}_{n_k}^\top \boldsymbol{Y}_k - \hat{m}^{0\top} \right)$
  $\boldsymbol{Y}_b = \boldsymbol{F} \boldsymbol{V}^\top$ (SVD)
  $\boldsymbol{F}_b \leftarrow \boldsymbol{F}[, 1 : Q_b]$; $\boldsymbol{V}_b \leftarrow \boldsymbol{V}[, 1 : Q_b]$
  $\hat{\boldsymbol{Y}}_b = \boldsymbol{F}_b \boldsymbol{V}_b'$
  Estimation of the within structure
  $\boldsymbol{Y}_w = \boldsymbol{Y} - \mathbb{1}_{m_1} \hat{m}^\top - \boldsymbol{Y}_b$
  $\boldsymbol{Y}_w = \boldsymbol{F} \boldsymbol{V}^\top$ (SVD)
  $\boldsymbol{F}_w \leftarrow \boldsymbol{F}[, 1 : Q_w]$; $\boldsymbol{V}_w \leftarrow \boldsymbol{V}[, 1 : Q_w]$
  $\hat{\boldsymbol{Y}}_w = \boldsymbol{F}_w \boldsymbol{V}_w'$
  Imputation of the missing values
  $\hat{\boldsymbol{Y}} = \mathbb{1}_{m_1} \hat{m}^\top \hat{\boldsymbol{Y}}_b + \hat{\boldsymbol{Y}}_w$
  $Y \leftarrow \boldsymbol{M} \odot \boldsymbol{Y} + (\mathbb{1}_{m_1} \mathbb{1}_{m_2}^\top - \boldsymbol{M}) \odot \hat{\boldsymbol{Y}}$
  $\hat{m} = m_1^{-1} \mathbb{1}_{m_1}^\top \boldsymbol{Y}$
end for

(EM) algorithm of the multilevel model (8.1) assuming Gaussian noise (see Appendix 8.7.1). Furthermore, we implemented an accelerated version of the algorithm where the between and the within parts are not updated simultaneously but one at a time. This corresponds to a generalized EM step, where the least-squares criterion is decreased at every iteration of the algorithm, but not entirely minimized. To prevent overfitting, the SVD step is replaced by a regularized SVD, i.e. where the singular values are shrunk, as described in Section 8.3.3.

Note that the criterion (8.9) does not have a unique solution in general, unless some assumptions are made on the missing data pattern $\boldsymbol{M}$ and when the weighted least squares are penalized by the nuclear norm for instance. Many theoretical results on the recovery of $\boldsymbol{Y}$ are then available (Candès and Tao, 2010; Candes and Plan, 2010; Klopp, 2014). Providing such guarantees is beyond the scope of this paper, but some intuitions about the conditions on $\boldsymbol{M}$ can be drawn from the aforementioned papers. The rationale is that the unknown low-rank matrix can be well approximated with high probability, if the probability of observing an entry is positive for every entry, and the number of observations is large enough. In particular, recovering $\boldsymbol{Y}$ when one row or one column is completely missing is hopeless. We would expect similar results and behaviors, if these previous works were extended to our framework.

8.3.2 Imputation with MLMCA and MLFAMD

Based on Algorithm 6 for imputation of multilevel quantitative data, we define two iterative imputation algorithms for multilevel MCA and multilevel FAMD. They are sketched together in Algorithm 7. Note that for categorical features, our algorithm does not

Algorithm 7 Iterative MLMCA and iterative MLFAMD

1: **Initialization**

   (a) Initialize missing values: mean imputation for quantitative data, proportion imputation for dummy variables.

   (b) Compute weights, standard deviations and column margins.

2: **Repeat until convergence:**

   (a) Estimate parameters (with MLFAMD or MLMCA)

   (b) Impute the missing entries with fitted values

   (c) Update means, standard deviations, column margins.[1]

output discrete categories but proportions, which can be interpreted as degrees of membership to each category. To impute, at the end of the algorithm, we assign the most plausible category.

8.3.3 Implementation

We now discuss some technical points related to the implementation of MLPCA, MLMCA and MLFAMD, available in the package `missMDA`. In particular, we describe how we may select the parameters, add a regularization term, and implement the algorithms in a distributed fashion.

Selecting the number of dimensions

The imputation methods described in Sections 8.3.1 and 8.3.2 require to select two parameters: the number of between and within components Q_b and Q_w. Furthermore, they must be selected from an incomplete data set. This is far from trivial, especially in the case of categorical variables. In fact, even in the complete case and without multilevel structure, not many options are available. Consequently, we advocate the use of cross-validation to select these components. More precisely, for quantitative data, leave-one-out cross-validation consists in removing each observed value $y_{k,i_k,j}$ of the data matrix Y one at a time. Then, for a fixed number of dimensions Q_b and Q_w, we predict its value using the multi-level method obtained from the data set that excludes this cell (using the iterative MLPCA on the incomplete data set). The predicted value is denoted $\left(\hat{y}_{k,i_k,j}^{-ki_kj}\right)^{(Q_b,Q_w)}$. Lastly, the prediction error is computed and the operation repeated for all observed cells in $\boldsymbol{Y}$ and for a number of dimensions varying for Q_b from 0 to $\min(K-2, m_2-1)$ and for Q_w from 0 to $\min(m_1-K, m_2-1)$. The numbers

Q_b and Q_w that minimize the mean square error of prediction (MSEP) are kept:

$$\text{MSEP}(Q_b, Q_w) = \frac{1}{m_1 m_2} \sum_{k=1}^{K} \sum_{i_k=1}^{n_k} \sum_{j=1}^{m_2} \left(y_{k,i_k,j} - \left(\hat{y}_{k,i_k,j}^{-ki_kj} \right)^{(Q_b,Q_w)} \right)^2.$$

This method is computationally costly, especially when the number of cells is large, since it requires to perform the iterative multilevel algorithm for each cell and for each number of between and within components. To reduce the computational cost, we implement a k-fold approach which consists in removing more than one value in the data set, for instance 5% of the cells and predict them simultaneously, combined with parallel computing and fast implementation of SVD. The same approach is used for categorical and mixed data using the coding with the indicator matrix of dummy variables and the iterative MLMCA and MLFAMD algorithms.

Regularization

Furthermore, to prevent overfitting, we actually perform a regularized SVD where singular values are shrunk. Many regularization are available for low-rank matrix estimation (Gavish and Donoho, 2014; Josse and Wager, 2016) and they have different regime of predilection. One of the most famous is soft-thresholding of singular values, which minimizes a least squares criterion penalized by the nuclear norm; it shows good estimation properties in low signal to noise ratio (SNR) regimes, but may struggle in other situations. The shrinkage rule we use applies a non-linear transformation of the singular values and shows good empirical performances in many regimes (Josse et al., 2017). Let λ_l, $1 \leq l \leq Q_b$, and ν_q, $1 \leq q \leq Q_w$, be the ordered singular values of $\boldsymbol{W}_b$ and $\boldsymbol{W}_w$ defined in (8.8). Let $\hat{\sigma}_b^2 = 1/(K - Q_b) \sum_{s=Q_b+1}^{K} \lambda_s$ and $\hat{\sigma}_w^2 = 1/(m_2 - Q_w) \sum_{s=Q_w+1}^{m_2} \nu_s$. We shrink the singular values as follows:

$$(\lambda_1, \ldots, \lambda_{Q_b}) \leftarrow \left(\lambda_1 - \frac{\hat{\sigma}_b^2}{\lambda_1}, \ldots, \lambda_{Q_b} - \frac{\hat{\sigma}_b^2}{\lambda_{Q_b}} \right),$$

$$(\nu_1, \ldots, \nu_{Q_w}) \leftarrow \left(\nu_1 - \frac{\hat{\sigma}_w^2}{\nu_1}, \ldots, \nu_{Q_w} - \frac{\hat{\sigma}_w^2}{\nu_{Q_w}} \right).$$

This regularization comes from Verbanck et al. (2013) and is a particular instance of the adaptive shrinkage estimator of Josse and Sardy (2015). It can be seen as a compromise between hard and soft thresholding. Indeed, when the noise variance is low (the SNR is high), it is equivalent to a hard thresholding, which behaves well when the SNR is high. When the noise variance is high (the SNR is low), it is equivalent to imputing using 0 dimensions, i.e. using the average of every variable. Between these two extremes, it shrinks the smallest singular values, which can be considered responsible for instability, more than the largest ones. In some particular cases, this shrinkage rule is equivalent to minimizing a penalized criterion, as shown in Appendix 8.7.4.

Distribution

Finally, the algorithms we present in this paper can be implemented in parallel across groups, provided that groups agree to share their mean values, standard deviations, sample sizes, and right singular vectors. Indeed, among other methods, SVD, which only involves inner products and sums, can be very straightforwardly implemented in a distributed manner. This is one main advantage of the methods we present. The procedure

to distribute the computation across sites is described in Section 8.7.3. Such a procedure is interesting in the framework of the medical application described in Section 8.5 as it allows each hospital to keep their data on site while benefiting from other hospitals data for the imputation.

8.4 Simulation study

8.4.1 Imputation of multilevel quantitative data

We conducted a comparative simulation study to contrast the performances of the multilevel imputation with PCA (MLPCA) to other single imputation methods, namely

1. mean imputation which consists in imputing by the mean of each variable, used as a benchmark method;
2. a separate PCA imputation where each group is imputed independently, using the R package missMDA (Josse and Husson, 2016);
3. a global imputation by PCA (which ignores the multilevel structure and the group variable) using the R package missMDA (Josse and Husson, 2016);
4. imputation with iterative conditional random effects regression models as implemented in the R package mice (van Buuren, 2012);
5. imputation by a joint model based on random effects models as implemented in the R package jomo (Quartagno and Carpenter, 2017);
6. imputation with iterative random forest (RF) as implemented in the R package missForest, (Stekhoven and Bühlmann, 2012). The group variable is included for the imputation.

Note that methods 4 and 5 are considered as the references to impute multilevel quantitative data (Audigier et al., 2018). However, these methods are defined as multiple imputation methods and used the imputed data as an intermediary to do statistical inference with missing values. Here, we compute the mean over 100 multiple imputed data to get one single imputed data set. The imputation based on random forests can handle mixed variables and is known to be a very powerful tool for imputation. It is not specifically designed to handle a multilevel structure, but is expected to perform well in such a hierarchical setting. Indeed, random forests can account for interactions between variables, and therefore in particular for interactions between the categorical variable indicating the group and the other variables. This is another way of handling the multilevel structure. In the same way, even though we focus here on quantitative variables, we also added imputation method for mixed data with FAMD (Audigier et al., 2016), where the group membership is used as a categorical variable.

We first simulate data according to the multilevel model (8.1) with Gaussian noise and set the true number of between and within components to 2. For global PCA and FAMD, we select the number of components resulting in the smallest errors, so in this example 4 dimensions. For MLPCA, we use the true number of dimensions, that is $Q_b = 2$ and $Q_w = 2$, and we use those estimated by cross-validation (Section 8.3.3). We use default parameters for the other methods. We start with $n_k = 20$ observations

per group k and we vary the number of groups K $(3, 5)$, the number of variables J (5, 10, 30), the intensity of the noise ($\sigma = 1, 2$) and the percentage of missing values (10%, 20%, 30%, 40%), which are missing completely at random (MCAR). We also consider the case of missing at random (MAR) values. The detail is available in the associated code provided as supplementary material. We then compute the mean squared error (MSE) of prediction, and repeat the process 100 times. Figure 8.1 is representative of

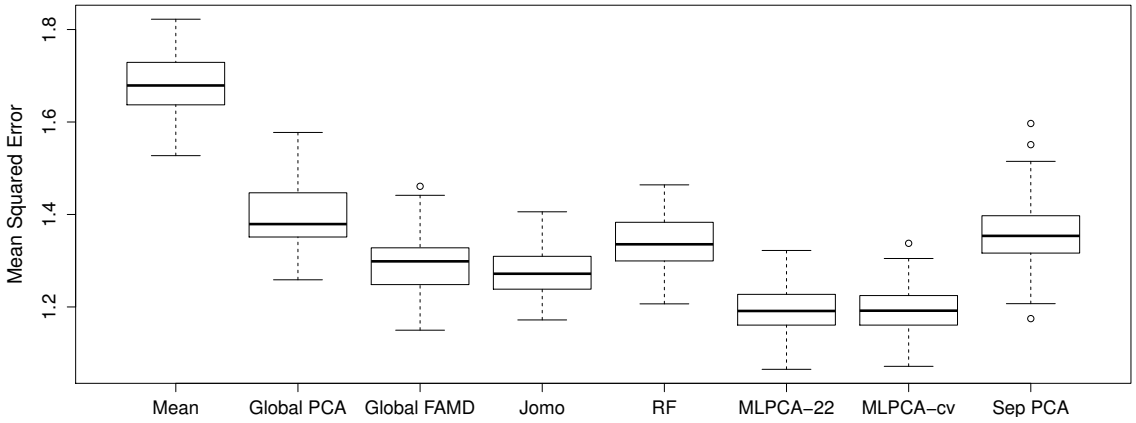


Figure 8.1: MSE of prediction for a data with $J = 10$ variables, $K = 5$ groups, $n_k = 20$ observations per group and 30% of missing values completely at random. MLPCA is performed with the true number of dimensions $Q_b = 2$ and $Q_w = 2$, and with the numbers of dimensions Q_b and Q_w estimated by cross-validation.

many results where multilevel imputation MLPCA improves both on global PCA imputation and separate PCA imputation but also on competitors. We have not included the results from the package mice as, using the default parameters, we encountered too many errors. It may be explained by the size of the data set, as the method does not behave well when there are not too many variables. More tuning is surely required to use the mice package seamlessly.

We summarize here our main findings with respect to all the simulations carried out. The results for MAR data are given in Appendix 8.7.5. All the results are in agreement to those obtained for MCAR values. Imputations with random forests and FAMD often perform similarly with a slight advantage for FAMD especially when the percentage of missing values is large. Imputation with jomo encounters many difficulties when the number of variables increases as well as when the noise increases. Finally imputation based on separate PCA collapses when the percentage of missing values increases and/or the number of observations per group decreases, which is not surprising as it operates on the smaller group data sets. The multilevel imputation is always the most accurate. This is expected (but still reassuring) as the data are simulated according to a multilevel model. We also simulated data without a multilevel structure, i.e. with one single group containing all individuals, and the performances of multilevel PCA are only slightly lower than those of global PCA.

All the methods have of course their strengths and weaknesses, and the properties of an imputation method depend on its inherent characteristics: an imputation method based on low rank assumption and linear relationships provides good prediction for data with strong linear relationships contrary to imputation using random forests which are designed for non-linear relationships. However, we observe that imputation with random

forests breaks down for small sample sizes in missing at random (MAR) cases, because extrapolation and prediction outside the range of the data seems difficult with random forests. Since the structure of the data is not known in advance, one could use cross-validation and select the method which best predicts the removed entries.

Figure 8.2 represents the differences, for each group, between imputing with a separate PCA and with MLPCA. The improvement of a multilevel imputation over a separate imputation may differ from one study to the other but still groups have interest in using a multilevel imputation. Indeed, the results presented in Figure 8.2 reveal that in terms of predicting the missing entries, multilevel PCA yields better results that separate PCA for every group, thus showing that as far as imputation is concerned, all groups benefit from participating in the study. This justifies the use of distributed multilevel methods in contexts where there are confidentiality issues at stake, by quantifying how much the different centers gain in terms of imputation accuracy, as further discussed in Section 8.5.

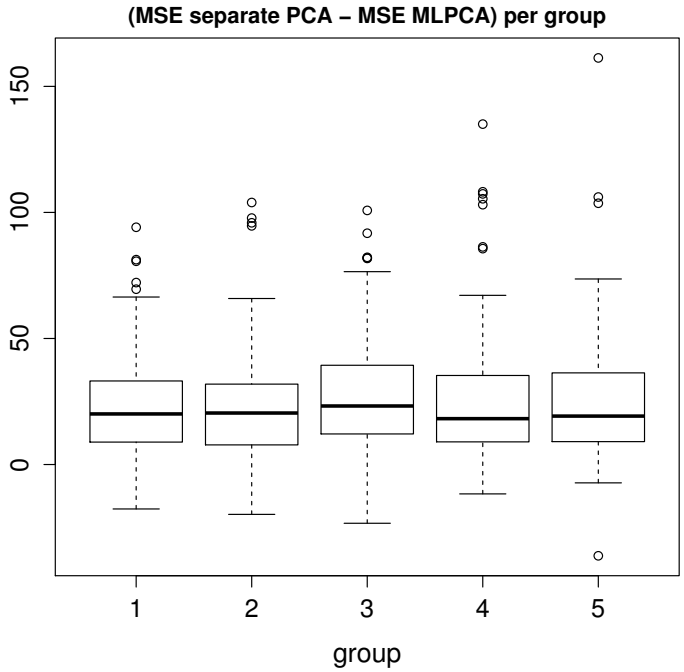


Figure 8.2: Difference between MSE obtained with separate PCA and with MLPCA for each group.

8.4.2 Imputation of multilevel mixed data

To simulate mixed data, we use the same design as for quantitative variables but cut some of the variables into categories. We vary the same parameters as for the quantitative variables but also the ratio of the number of quantitative over the number of categorical variables. We simulate either MCAR values or systematic missing values where all the values of a variable are missing for one group. Note that the methods implemented in the packages mice and jomo can handle mixed data when categorical variables are binary, but not when variables have more than two categories. This is why they are not included in the simulations. The global FAMD imputation is performed with 2, 3 and 4 dimensions (we display only the number of components which resulted in

	$J = 10$		$J = 30$		$J = 15$		$J = 35$	
n_k	50	200	50	200	50	200	50	200
Global PCA	0.009	0.021	0.013	0.037				
jomo	10.431	40.362	191.47	757.586				
Multilevel FAMD	0.017	0.030	0.025	0.057	0.037	0.067	0.045	0.108
Global FAMD	0.060	0.108	0.122	0.210	0.160	0.239	0.155	0.333
Random forest	2.073	15.143	8.88	59.953	1.953	14.488	10.326	64.466

Table 8.1: Time in seconds for a data set with 20% of missing values, $K = 5$ groups and $n_k = 50$ or $n_k = 200$ observations per groups, with 10 and 30 quantitative variables for the two left columns and with additional 5 categorical variables for the two right columns.

the lowest prediction error) whereas we add cross-validation for the multilevel method. Figure 8.3 shows again that imputing with the multilevel method gives better results than imputing with global FAMD or with random forests. This is especially true for the quantitative variables. Note that when the missing values are systematic, it is not possible to apply the separate imputation. We can evaluate the quality of the imputation of the multilevel method by comparing the distributions of the observed and imputed values (see Figure 8.9 in the appendix).

As far as the computational time is concerned, we compare in Table 8.1 the performances of the different approaches. Regarding this point, SVD based imputation methods have a clear advantage over jomo and random forests.

8.4.3 Robustness to model misspecification

In many cases we expect the data to be approximately rather than exactly of low-rank, that is, to have a few principal directions of large variability and many principal directions of very small variability. It is therefore crucial that the imputation method be robust to such model misspecification. To assess the performance of imputation with multilevel FAMD in approximately low-rank models, we use the same simulation scheme as before with 5 quantitative variables, 5 qualitative variables, and 5 groups of 200 individuals each. We simulate data from an underlying matrix with rank varying from 4 to 10. For each of these rank settings, we impute the data with MLFAMD, and compute the RMSE for quantitative variables, and the percentage of misclassification for qualitative variables.

The results are given in Figure 8.4, along with the same experiment with mean imputation, in order to have a baseline. As expected, the imputation error increases with the underlying rank. However, the degradation of the imputation is small compared to the difference between MLFAMD and the baseline for similar rank, which indicates the robustness of MLFAMD to such model misspecification.

8.5 Hospital data analysis

8.5.1 Trauma Register Traumabase

Expedient management of major trauma based on standardized and protocol-based care improves functional outcome and survival. Even in mature systems trauma care is still

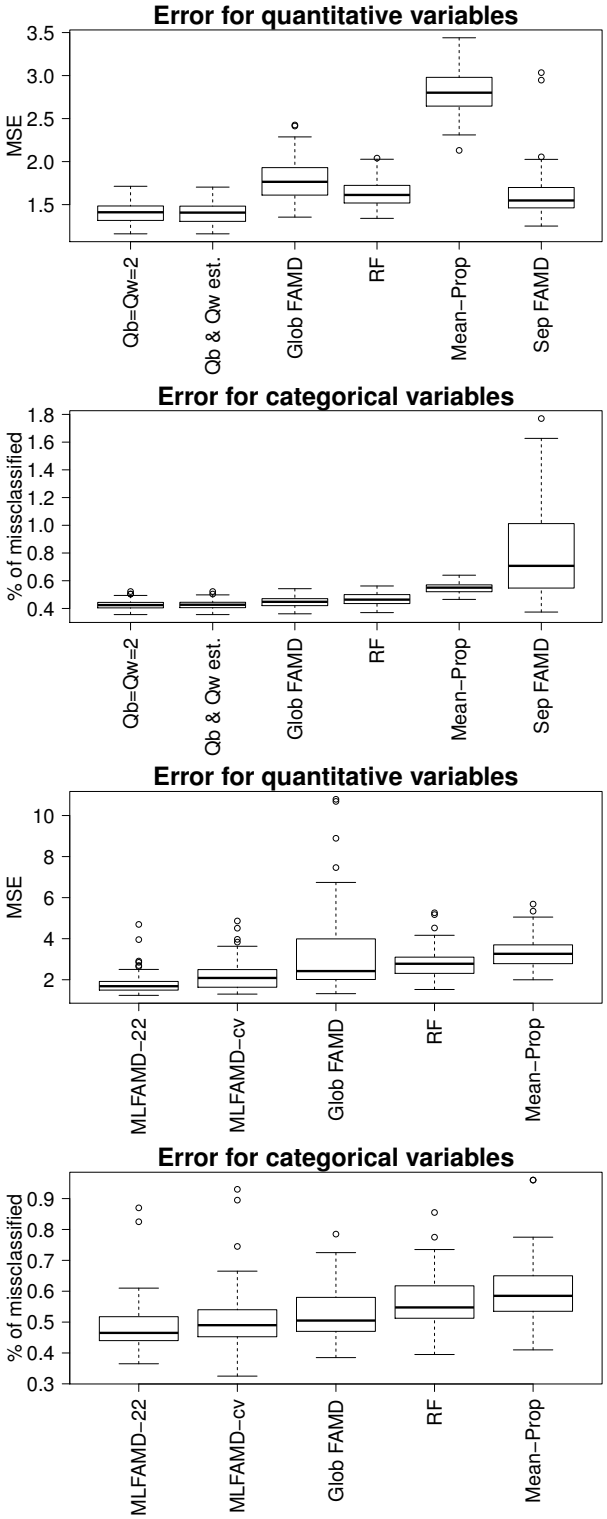


Figure 8.3: $J = 10$ variables, 5 quantitative and 5 categorical (4 categories each), $K = 5$ groups and $n_k = 30$ observations per group. Top: 20% of MCAR missing values, bottom: all values of one group are missing for two continuous variables and all the values of another group are missing for two categorical variables. Left: MSE for quantitative variables; right: percentage of misclassified categorical variables. Global FAMD and separate FAMD are represented with the number of dimensions that yield the smallest errors and MLFAMD is represented with values of Q_b and Q_w estimated by cross-validation. RF is imputation with random forest and Mean-Prop means the imputation is done by the mean for quantitative variables and the proportion for categorical ones.

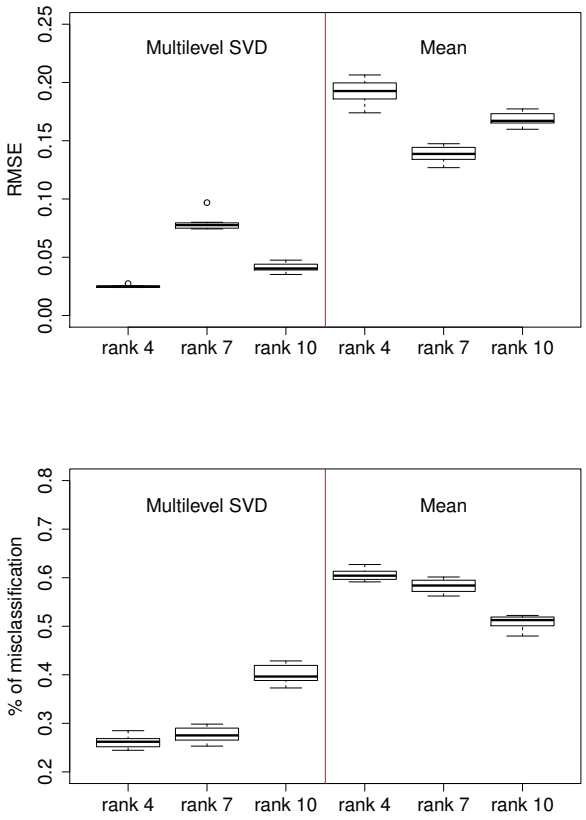


Figure 8.4: $J = 10$ variables, 5 quantitative and 5 categorical, 20% of missing values, $K = 5$ groups and $n_k = 200$ observations per group: on the left plot MSE for the quantitative variables; on the right plot percentage of misclassified for categorical variables. Left side: imputation with MLFAMD; right side: imputation by the column means.

hampered by delays, deviation from protocols and cognitive error. The present work is motivated by developing and providing data processing and analysis tools to inform clinical decision making. This study is conducted in cooperation with the Traumabase Group (`http://www.traumabase.eu/en_US`), a French multicenter trauma data base. The Traumabase consists currently of more than 15 French level-1 Trauma centers and holds detailed data of more than 20,000 trauma cases. These data are highly heterogeneous, multi-source, and contain many missing values. Furthermore, expert clinicians expect practice variation and lack of standardisation across different hospitals and regions to exert considerable influence on a number of variables.

For this work, an initial reduced data set containing eight features was analysed, that expert clinicians suggested to be prone to variation. The data set of interest consisted of 5 qualitative and 3 quantitative variables measured over 7,495 patients with around 11% of missing values and at least one missing entry for 49% of all the patients. Probably different mechanisms generate this level of missing values. For some variables, such as accident type and name of center, no data are missing, whereas for other variables such as chest and pelvis X-ray, a lot of entries are missing and may depend on local practice. In a first approximation, a Missing At Random (MAR) mechanism, where the probability of missingness is allowed to depend on the observed variables, seems satisfying.

Generally speaking, we focus on imputing medical data with iterative MLFAMD with two aims. First, the imputed data can be further analyzed with other statistical methods such as predictive models, to predict some outcome of interest. However, care must be taken when analysing an imputed data set, as discussed in Section 8.6. Secondly, the imputation of missing data from a hospital is improved when the hospital is integrated into the aggregated database. Therefore, this may encourage medical professionals to share their data and participate in the data aggregation projects. These projects are important because disposing of aggregated data is an opportunity to have more cases and to develop more robust and clinically pertinent modelling. Thus, more reliable and powerful imputation techniques may actually entice hospitals to share their data in order to facilitate evaluation and sequentially improve care for all patients.

However, there are technical and social barriers to the aggregation of medical data. The size of combined databases often makes computations and storage intractable, while institutions are usually reluctant to share their data due to privacy concerns and proprietary attitudes. Both obstacles can be overcome by turning to distributed computations, which consists in leaving the data on sites and distributing the calculations, so that hospitals only share some intermediate results instead of the raw data (Narasimhan et al., 2017). The distributed framework is presented in Section 8.7.3.

8.5.2 Simulated imputation of the Traumabase

To assess the quality of imputation and legitimate the use of iterative MLFAMD to impute the Traumabase, we first perform simulations by inserting an additional of 20% of MCAR values to the data set, predicting them with the different imputation methods described in Section 8.4, and computing the mean squared error of prediction for quantitative variables and the percentage of misclassification for categorical variables. We also conducted simulations with systematically missing data in both quantitative and categorical variables, which often happens when a hospital does not collect a measurement.

Figure 8.5 presents the results over 30 replications of the experiment. Number of components for all the methods is estimated by cross-validation. When the missing values

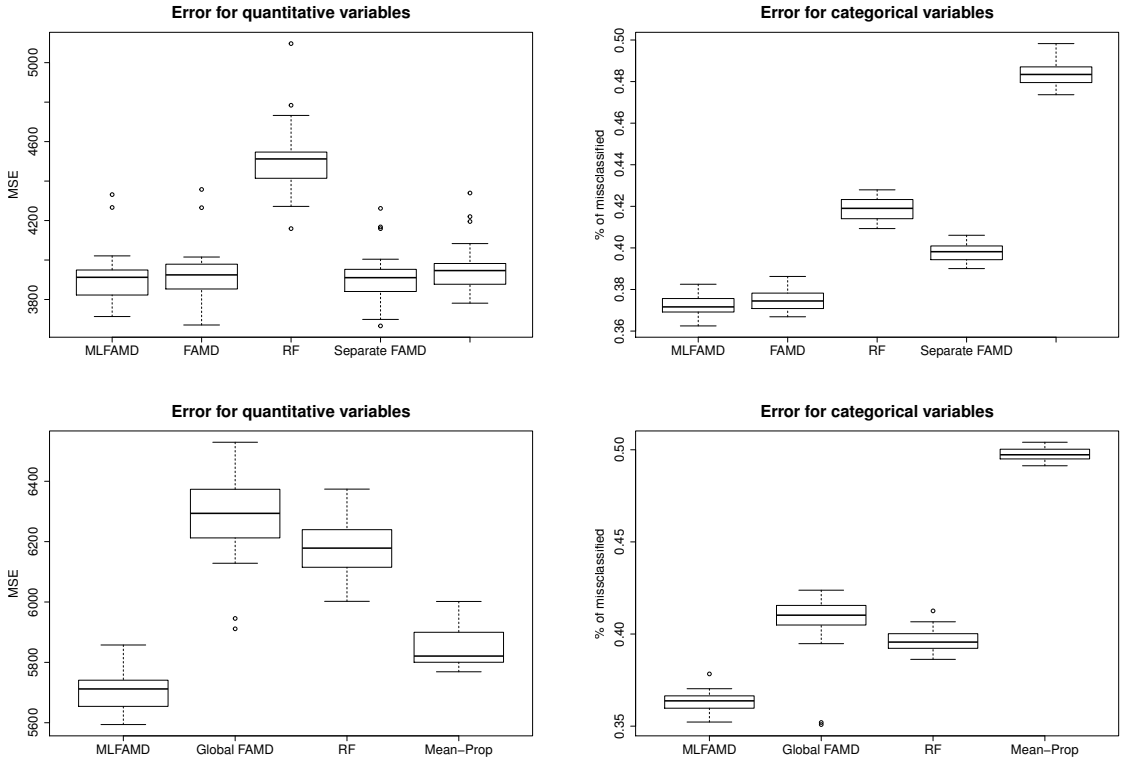


Figure 8.5: Traumabase: MSE of prediction and % of mis-classification. Top: 20% of MCAR values, bottom: systematic missing values.

are MCAR, in terms of prediction of quantitative variables, multilevel FAMD and global FAMD perform similarly and improve on the random forest imputation. We observe the same behavior for the categorical variables, with multilevel FAMD improving only slightly on global FAMD. Note that the data are quite difficult to impute and the relationship between variables weak. However, when we generate systematic missing values there is a large improvement when using the multilevel method.

8.6 Conclusion

We proposed a method dedicated to the imputation of multilevel mixed data based on an iterative SVD algorithm. To the best of our knowledge this is the first multilevel method available for mixed data and which can handle both sporadic and systematic missing values. We also believe the multilevel methods we have developped for mixed data can be useful for exploratory analysis and visualization. We are eager to investigate for future research a multiple imputation (Murray, 2018) procedure based on this multilevel component method, in order to further analyse the Traumabase data set with predictive models, for instance to study the occurence of diagnosis errors based on patients profiles. Indeed, the proposed method is a single imputation method. This is perfectly appropriate when the objective is to accurately predict the missing values, which was one of the objectives of the application on hospital data. However, great caution should be taken when analyzing the completed table. Indeed, like all simple imputation methods,

our method suffers from not taking into account the uncertainty associated with predicting missing values from observed values. Thus, if we apply a statistical method on a completed data table, the variability of estimators will be underestimated. To avoid this problem, one solution is to use multiple imputation where different values are predicted for each missing value, resulting in several imputed tables and the variability of imputations reflects the prediction variance. Multiple imputation then consists of applying an analysis to each of the completed tables combining the results. The proposed iterative imputation algorithm could be a first step in a multiple imputation method for multi-level mixed data. A first idea could be to combine a stratified bootstrap with our algorithms.

Finally, as discussed, the methods presented in this paper can be implemented in parallel across groups or sites. A following project we are currently involved in consists in exploiting this property to implement a real-time distributed and privacy preserving platform, dedicated to the imputation of health care data partitioned across several hospitals, without having to aggregate the data. One issue with the distribution technique described in Section 8.7.3 is that we use iterative procedures, therefore after N iterations each hospital has shared N summary statistics, which can lead to information leakage. A possible solution to this problem is to resort to homomorphic encryption (Gentry, 2009) which allows to perform computations on encrypted data.

8.7 Supplementary material

8.7.1 EM algorithm for multilevel Gaussian data

The iterative procedure described in Section 8.3.1 is equivalent to an EM algorithm based on a Gaussian model.
Let $\boldsymbol{Y} = \mathbb{1}_{m_1} m^\top + \boldsymbol{F}_b \boldsymbol{V}_b^\top + \boldsymbol{F}_w \boldsymbol{V}_w^\top + \boldsymbol{E}$, where $m \in \mathbb{R}^{m_2}$, $\boldsymbol{F}_b \in \mathbb{R}^{m_1 \times Q_b}$, $\boldsymbol{V}_b \in \mathbb{R}^{m_2 \times Q_b}$, $\boldsymbol{V}_b \in \mathbb{R}^{m_1 \times Q_w}$ and $\boldsymbol{F}_b \in \mathbb{R}^{m_2 \times Q_w}$ are parameters to be estimated, and $\boldsymbol{E}$ is a matrix with Gaussian entries with known variance $\boldsymbol{E}_{i,j} \sim \mathcal{N}(0, \sigma^2)$. Denote by $\boldsymbol{M} \in \{0,1\}^{m_1 \times m_2}$ the mask indicating (with a 1) the observed entries of $\boldsymbol{Y}$, $\boldsymbol{Y}_{\text{obs}}$ the observed entries and $\boldsymbol{Y}_{\text{mis}}$ the missing entries.

Denote $\theta = (m, \boldsymbol{F}_b, \boldsymbol{V}_b, \boldsymbol{F}_w, \boldsymbol{V}_w)$. Under the classical hypothesis that the missing values are missing at random, *i.e.*, $p(\boldsymbol{M}|\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}}; \phi) = p(\boldsymbol{M}|\boldsymbol{Y}_{\text{obs}}; \phi)$, the likelihood of the observed data $(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{M})$ is

$$\begin{aligned} p(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{M}; \theta, \phi) &= \int p(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}}, \boldsymbol{M}; \theta) d\boldsymbol{Y}_{\text{mis};\phi} \\ &= p(\boldsymbol{Y}_{\text{obs}}; \theta) p(\boldsymbol{M}|\boldsymbol{Y}_{\text{obs}}; \phi) \end{aligned}.$$

The EM algorithm (Dempster et al., 1977) maximizes the observed log-likelihood $\ell(\boldsymbol{Y}_{\text{obs}}; \theta)$ iteratively by alternatively computing the expectation of the complete likelihood $\ell(\boldsymbol{Y}; \theta)$ under the distribution of the missing values given the observed values and the current estimate θ^t (E step), and maximizing this expectation with respect to θ (M step). It goes as follows at iteration t:

E step: $\mathbb{E}_{\theta^t}\left[\ell(\boldsymbol{Y}; \theta)|\boldsymbol{Y}_{\text{obs}}\right] = \mathbb{E}_{\theta^t}\left[\left\|\boldsymbol{Y} - \mathbb{1}_{m_1} m^\top - \boldsymbol{F}_b \boldsymbol{V}_b^\top - \boldsymbol{F}_w \boldsymbol{V}_w^\top\right\|_F^2\right]$. To compute this expectation, one only needs to compute the two following sufficient statistics for all

$(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{Y}_{i,j}|\boldsymbol{Y}_{\text{obs}},\theta^t] &= \begin{cases} \boldsymbol{Y}_{i,j} & \text{if } \boldsymbol{M}_{i,j}=1 \\ ((\mathbb{1}_{m_1} m^\top + \boldsymbol{F}_b \boldsymbol{V}_b^\top + \boldsymbol{F}_w \boldsymbol{V}_w^\top)^t)_{i,j} & \text{if } \boldsymbol{M}_{i,j}=0 \end{cases}, \\
\mathbb{E}[\boldsymbol{Y}_{i,j}^2|\boldsymbol{Y}_{\text{obs}},\theta^t] &= \begin{cases} \boldsymbol{Y}_{i,j}^2 & \text{if } \boldsymbol{M}_{i,j}=1 \\ ((\mathbb{1}_{m_1} m^\top + \boldsymbol{F}_b \boldsymbol{V}_b^\top + \boldsymbol{F}_w \boldsymbol{V}_w^\top)^t)_{i,j}^2 + (\sigma^2)^t & \text{if } \boldsymbol{M}_{i,j}=0 \end{cases}.
\end{aligned} \tag{8.10}$$

M step: The M steps consists in maximizing the complete likelihood where the sufficient statistics are replaced by the conditional expectation above. However, estimation of the parameters in θ do not require knowledge of σ so that we do not need to compute $\mathbb{E}[\boldsymbol{Y}_{i,j}^2|\boldsymbol{Y}_{\text{obs}},\theta^t]$ in the E step. Denote for, $(i,j) \in [\![m_1]\!] \times [\![m_2]\!]$, $\hat{\boldsymbol{Y}}_{i,j}^t = \mathbb{E}_{\theta^t}[\boldsymbol{Y}_{i,j};\theta]$, and $\hat{\boldsymbol{Y}} = (\hat{\boldsymbol{Y}}_{i,j})_{i,j}$. The maximization of $\mathbb{E}_{\theta^t}\left[\ell(\boldsymbol{Y};\theta)|\boldsymbol{Y}_{\text{obs}}\right]$ with respect to θ is equivalent to solving:

$$\text{argmin}_{m,\boldsymbol{F}_b,\boldsymbol{V}_b,\boldsymbol{F}_w,\boldsymbol{F}_b} \frac{1}{2}\left\|\hat{\boldsymbol{Y}} - \mathbb{1}_{m_1} m^\top - \boldsymbol{F}_b \boldsymbol{V}_b^\top - \boldsymbol{F}_w \boldsymbol{V}_w^\top\right\|_2^2.$$

The EM algorithm consisting of the E and M step described above is exactly equivalent to Algorithm 5.

8.7.2 Distributed rank-Q PCA

We start by reminding the power method (Golub and Van Loan, 1996), which computes the first left and right singular vectors of a matrix $\boldsymbol{Y} \in \mathbb{R}^{m_1 \times m_2}$. Without loss of generality, we assume $m_1 \leq m_2$. Suppose $\boldsymbol{Y} = \boldsymbol{U}\Lambda^{1/2}\boldsymbol{V}^\top$, $\boldsymbol{U} = (u_1,\ldots,u_{m_1})$, $\boldsymbol{V} = (v_1,\ldots,v_{m_1})$ and $\Lambda = \text{diag}(\lambda_1^2,\ldots,\lambda_n^2)$ $|\lambda_1| \geq |\lambda_2| \ldots \geq |\lambda_{m_1}|$. The power method is iterative and produces sequences of vectors $z^{(t)}$ and $q^{(t)}$ converging to u_1 and v_1 respectively, with iterations detailed in Algorithm 8. Let $q^{(0)}$ be a starting point satisfying $\|q^{(0)}\|_2 = 1$. The sequences $q^{(t)}$ and $z^{(t)}$ converge to u_1 and v_1 respectively, when

Algorithm 8 Power method

for $t = 1, 2, \ldots$ **do**
  $z^{(t)} = \boldsymbol{Y}^\top q^{(t-1)}$
  $z^{(t)} = z^{(t)}/\|z^{(t)}\|_2$
  $q^{(t)} = \boldsymbol{Y} z^{(t)}$
  $\lambda^{(k)} = \|q^{(t)}\|_2$
  $q^{(t)} = q^{(t)}/\|q^{(t)}\|_2$
end for

$\langle q^{(0)}, u_1\rangle \neq 0$ and $|\lambda_1| > |\lambda_2|$; the rate of convergence is dictated by the ratio $|\lambda_2|/|\lambda_1|$. This directly extends to the computation of the rank-Q SVD. One can actually estimate u_1, v_1 and λ_1, then the second dimension by applying the same procedure to $\boldsymbol{Y} - u_1\lambda_1 v_1^\top$, and so on so forth. Moreover it is straightforward to distribute this procedure when the data are grouped in K different sites with

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \hline \boldsymbol{Y}_2 \\ \hline \vdots \\ \hline \boldsymbol{Y}_K \end{pmatrix}.$$

Indeed, all the computations in Algorithm 8 can be done in parallel with a master-slave architecture (Narasimhan et al., 2017), where a central server collects summary statistics computed locally on sites, as illustrated Figure 8.6. Here, the local right singular vectors v_j, $j \in [\![m_1]\!]$ are sent to the master. The corresponding algorithm is given in Algorithm 9, and leads exactly to applying the power method for rank-Q SVD to the entire data matrix $\boldsymbol{Y}$. The procedure is implemented in the distcomp R package (Narasimhan et al., 2017). This algorithm can in turn be extended to perform distributed PCA with

Algorithm 9 Distributed power method

Input: workers private data $\boldsymbol{Y}_k \in \mathbb{R}^{n_k \times m_2}$
Output: $\boldsymbol{F} \in \mathbb{R}^{m_1 \times Q}$, $\boldsymbol{V} \in \mathbb{R}^{m_2 \times Q}$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_Q$ $\boldsymbol{F} = 0$, $\lambda = 0$
for $k = 1, \ldots K$ **do**
  $\boldsymbol{F}_k = 0$
  transmit n_k to master
end for
for $i = 1, \ldots, Q$ **do**
  for $k = 1, \ldots K$ **do**
    $q_k = (1, 1, \ldots, 1)$
  end for
end for
$\|q\|_2 = \sqrt{\sum_{k=1}^{K} n_k}$
transmit $\|q\|_2$, V and λ to workers
while Not converged **do**
  for $k = 1, \ldots, K$ **do**
    $q_k = q^k / \|q\|_2$
    $r_k = (\boldsymbol{Y}_k - \boldsymbol{F}_k \boldsymbol{V}^\top)^\top q_k$
    transmit r_k to master
  end for
  $r = \sum_{k=1}^{K} r_k$
  $r = r / \|r\|_2$
  transmit r to workers
  for $k = 1, \ldots, K$ **do**
    $q_k = \boldsymbol{Y}_k r$
    transmit $\|q_k\|_2$ to master
  end for
  $\|q\|_2 = \sum_{k=1}^{K} \|q_k\|_2$
  transmit $\|q\|_2$ to workers
  $\lambda_i = \|q\|_2$
end while
$\boldsymbol{V} =$ combine by column $(\boldsymbol{V}, r)$
for $k = 1, \ldots, K$ **do**
  $\boldsymbol{F}_k =$ combine by column $(\boldsymbol{F}_k, q_k)$
end for

missing values, yielding the algorithm given in 8.7.3. Indeed, the iterative PCA algorithm iteratively performs SVD.

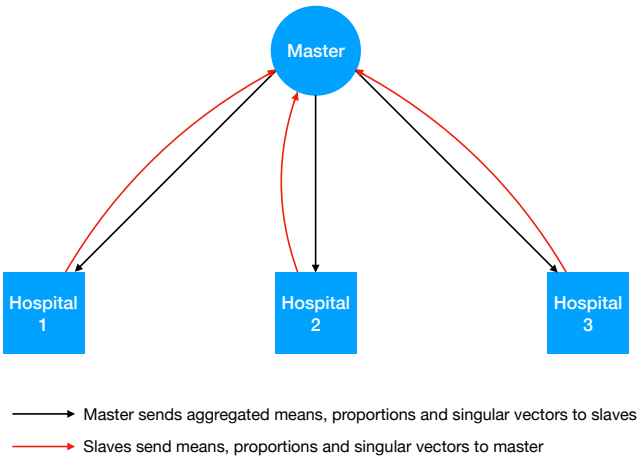


Figure 8.6: Master-slave distribution structure. The hospitals send their local means, proportions, sample size and right sigular vectors to the master. The master sends back the overall means, proportions, and right singular vectors to the hospitals.

8.7.3 Distributed algorithm for iterative multilevel PCA

In Section 8.7.2, we see how the power method (Golub and Van Loan, 1996), which computes the first left and right singular vectors of a matrix $\boldsymbol{Y} \in \mathbb{R}^{m_1 \times m_2}$, can be straightforwardly distributed over K different sites. This algorithm can then be used to perform a distributed rank-Q SVD, as shown in Algorithm 9. We take advantage of this property to develop a distributed version of the iterative PCA algorithm, presented in Algorithm 10. This algorithm imputes missing values with the iterative PCA algorithm in a distributed way. Indeed, iterative PCA imputation involves iterative SVD. Plugged in Algorithm 5, Algorithm 10 leads to a distributed version of the iterative multilevel PCA algorithm. In the same way, distributed iterative MLMCA and MLFAMD are implemented.

8.7.4 Minimization of a penalized criterion

The singular value shrinkage we employ is an extension of the shrinkage rule of Verbanck et al. (2013) to the multilevel settings. In Verbanck et al. (2013) the authors use a shrinkage rule for the first Q largest eigenvalues of the form:

$$\psi(\lambda_i) = \lambda_i \left(1 - \frac{\sigma^2}{\lambda_i^2}\right)$$

Their estimator is defined to minimize the asymptotic MSE of the resulting estimator in the Gaussian model $X = \mu + \varepsilon$ where μ is of rank Q. In their framework, the asymptotic corresponds to the variance of the Gaussian noise σ^2 going to 0. This rule is related to other works on singular value shrinkage such as (Gavish and Donoho, 2014) but in particular the one of Josse and Sardy (2015) which is:

$$\psi(\lambda_i) = \lambda_i \max\left(1 - \frac{\tau^\gamma}{\lambda_i^\gamma}, 0\right)$$

using $\tau = \hat{\sigma}$ and $\gamma = 2$. The rule (Josse and Sardy, 2015) is the solution of a penalized criterion where the least-squares is penalized by a weighted nuclear norm. Consequently,

Algorithm 10 Distributed iterative PCA

Input: $\boldsymbol{Y}_k \in \mathbb{R}^{n_k \times m_2}$, Q_b, Q_w
Output: $\hat{m}$, $\boldsymbol{F}_b$,$\boldsymbol{V}_b$,$\boldsymbol{F}_w$,$\boldsymbol{V}_w$
Initialization: impute missing values with initial values; $(m_1 \times m_2) = diag(\sqrt{n_k})$.

$R = 0$, $\lambda = 0$
for $k = 1, \dots K$ **do**
  $\boldsymbol{F}_k = 0$
  transmit n_k to master
end for
for $i = 1, \dots, Q$ **do**
  for $k = 1, \dots, K$ **do**
    $q_k = (1, 1, \dots, 1)$
  end for
  $\|q\|_2 = \sqrt{\sum_{k=1}^{K} n_k}$
  transmit $\|q\|_2$, $\boldsymbol{V}$ and λ to workers
  while Not converged **do**
    for $k = 1, \dots, K$ **do**
      $q_k = q_k / \|q\|_2$
      $r_k = (\boldsymbol{Y}_k - \boldsymbol{F}_k \boldsymbol{V}^\top)^\top q_k$
      transmit r_k to master
    end for
    $r = \sum_{k=1}^{K} r_k$
    $r = r / \|r\|_2$
    transmit r to workers
    for $k = 1, \dots, K$ **do**
      $q_k = \boldsymbol{Y}_k r$
      transmit $\|q_k\|_2$ to master
    end for
    $\|q\|_2 = \sum_{k=1}^{K} \|q_k\|_2$
    transmit $\|q\|_2$ to workers
    $\lambda_i = \|q\|_2$
  end while
  $\boldsymbol{V}$ = combine by column $(\boldsymbol{V}, \sqrt{\lambda_i} r)$
  for $k = 1, \dots, K$ **do**
    $\boldsymbol{F}_k$ = combine by column $(\boldsymbol{F}_k, q_k)$
  end for
end for

although the rationale of our shrinkage rule is not to minimize a penalized criterion, we observe that our shrinkage rule is in fact, at iteration $t+1$, the solution to a least squares problem penalized by a *weighted* nuclear norm

$$\begin{array}{ll} \text{minimize} & \left\{ \frac{1}{2} \left\| \hat{\boldsymbol{X}}^{(t)} - \mu \right\|^2 + \lambda^\gamma \|\mu\|_{*,w^{(t)}} \right\} \\ \text{such that} & \operatorname{rank}(\mu) \leq Q, \end{array} \tag{8.11}$$

where $\hat{\boldsymbol{X}}^{(t)}$ is the estimate from the previous iteration, $d_1, \dots, d_l$ are the singular values of μ, $\|\mu\|_{*,w^{(t)}} = \sum_{i=1}^{\min(m_1,m_2)} w_i^{(t)} d_i$, $\sigma_i(\hat{\boldsymbol{X}}^{(t)})$ is the i-th singular value of $\hat{\boldsymbol{X}}^{(t)}$, $\lambda = \hat{\sigma}$

and $w_i^{(t)} = 1/\sigma_i(\hat{\boldsymbol{X}}^{(t)})$ for all i. Extending this to the Multilevel PCA iterative algorithm, we solve such a problem twice at each iteration (one for the within component and one for the between component).

In the following toy example, the entries of $\boldsymbol{X}$ are Gaussian, the variance of the noise σ^2 is known and no entries are missing. With all these assumptions, the weights in (8.11) are constant, and we show empirically that the criterion (8.11) is indeed decreasing for the multilevel method. Figure 8.7 shows the criterion (8.11) is decreasing in this case.

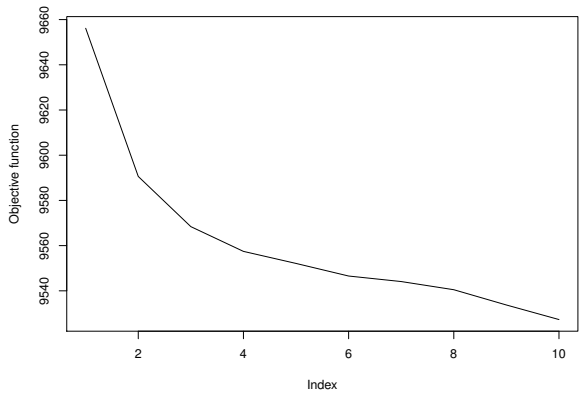


Figure 8.7: Objective function at every iteration until convergence of MLPCA.

8.7.5 MAR simulations

We performed simulations by putting Missing At Random (MAR) data as follows: in each group, we selected the most correlated pair of variables and put missing data for one of the 2 variables when the values of the other variable were greater than 1.1 times the mean. This leads to a relatively low percentage of missing values, approximately from 3% to 6% of missing values over the entire data set. Since this percentage is low, to show the method's effect, we centered by simulation the MSE (in an analysis of variance way, we removed the variability due to the simulations). Here we find results quite similar to those found when the missing data are MCAR, i.e. that the MLPCA algorithm is the most efficient. As expected, random forests are not very well suited for MAR values. Figure 8.8 is representative of many results where multilevel imputation MLPCA improves both on global PCA imputation and separate PCA imputation but also on competitors.

8.7.6 Representation of the imputed values

Whether before or after imputation, it is very important to perform descriptive statistics and graphical representations. The graphs in Figure 8.9 allow the distributions of observed and imputed values to be compared. The graph on the left shows for a variable the distribution of the predicted values in red and observed in black. Note that a difference between these distributions does not mean that the imputation model is unsuitable. Indeed, when the missing data mechanism is not MCAR, it could make sense to observe differences between the distribution of imputed values and the distribution of observed values. However, if differences occur, more investigations would be required to try to

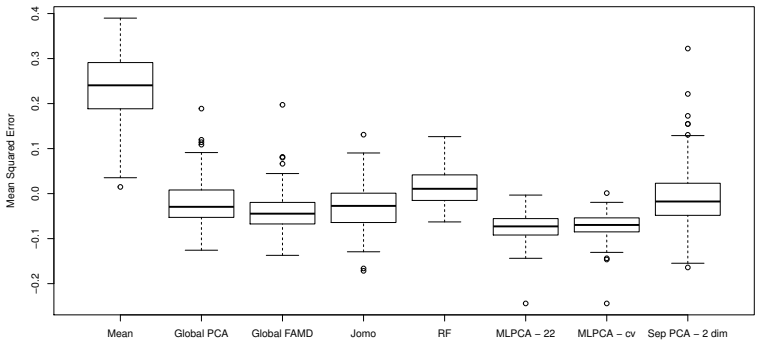


Figure 8.8: MSE centered by simulation for a data with $J = 10$ variables, $K = 5$ groups, $n_k = 20$ observations per group and missing values that are missing at random. MLPCA is performed with the true number of dimensions $Q_b = 2$ and $Q_w = 2$, and with the numbers of dimensions Q_b and Q_w estimated by cross-validation.

explain them. Here, the imputed values follow a distribution close to that of the observed values. The quality of the imputation can also be assessed by graphically representing two variables (the graph on the right). Here, it is clear that the correlation between the two variables is not destroyed by imputation.

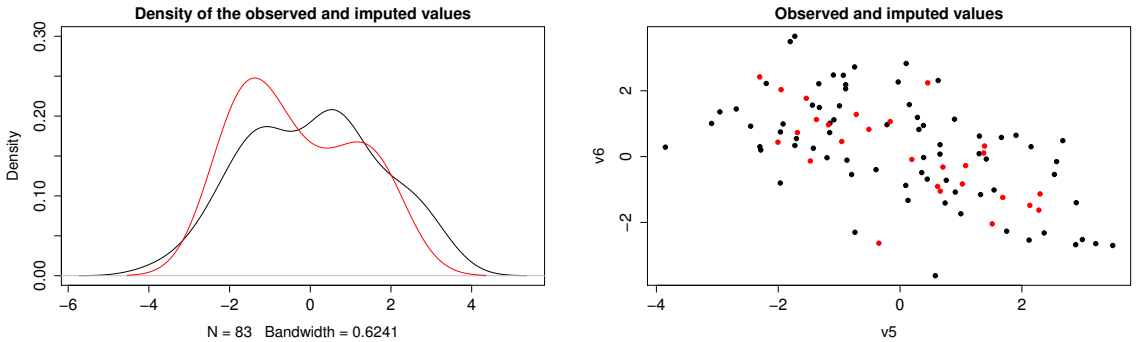


Figure 8.9: Density of the observed (in black) and imputed (in red) values. Scatterplot with observed values (in black) and imputed values (in red) for two variables.

Chapter 9

Conclusion

The objective of this thesis was to develop new data analysis tools adapted to modern data collection processes, which often compound information from diverse sources, and result in missingness and heterogeneity. In most fields of applications, because of such imperfections, data collections indeed fall out the classical frameworks for which substantial theory was already available. On the other hand, many methods used in practice to cope with multi-source, heterogeneous and incomplete data in fact do not benefit from any statistical guarantees. In this dissertation, we introduced a complete framework based on hybrid low-rank models and heterogeneous data fitting terms, to analyze and impute mixed data with missing values and side information. This new framework is rooted in theoretical aspects of the low-rank matrix completion literature, and blooms into the visual playground of principal components methods, from which it borrows interpretation tools.

We started with the special case of incomplete count data with side information, and developed in Chapter 3 a Poisson model which directly incorporates covariates in the inference procedure, by combining log-linear models and exponential family matrix completion. We demonstrated that this framework simultaneously inherited from the theoretical guarantees of convex low-rank methods, and the interpretation capabilities of model-based count data analysis.

Equipped with this new framework from count data analysis, we tackled in Chapter 4 the analysis of a challenging waterbird abundance data set, in order to estimate populations temporal trends and to detect important predictors of the bird counts. In the process, we extended the initial model, to incorporate a variable selection tool, and empirical assessment of uncertainty in our predictions. We implemented an open source R package for count data imputation and analysis, and provided a tutorial for potential users.

Building upon Chapter 3 and Chapter 4, we generalized in Chapter 6 the model from incomplete count data with side information, to heterogeneous incomplete data with hybrid structures, incorporating several models of interest in applications. We provided a theoretical study demonstrating that our procedure has near-optimal estimation errors and, in the process, generalizing theoretical results in noisy low-rank plus sparse matrix decomposition. We proposed an optimization procedure, which is implemented in a second R package adapted to mixed data with side information or multilevel structure.

Finally, in Chapter 8, we approached the imputation of multilevel mixed data with a different perspective, and introduced a counterpart of the methods of Chapter 6, based on component methods. This method stands out from the rest of this dissertation, as

it is not based on a probabilistic model, and thus does not benefit from the statistical background of Chapters 3 and 6. However, this lack of underlying model endows it with other advantages, and in particular, relieves it from the computational burden of heterogeneous likelihood based approaches. In practice, the method of Chapter 8 displayed good imputation properties, as well as a very competitive computational cost. In addition, it naturally leads to distributed implementations, which is attractive in applications where privacy is an important issue.

The contributions of this dissertation paved the way for future research in both theoretical and applied directions. In practice, there are three main limitations to the methods developed in Chapter 3, Chapter 4 and Chapter 6. First, these regularized approaches involve selecting two hyper-parameters with heavy cross-validation procedures. Even with efficient and scalable algorithms, this can be problematic for practitioners with limited computational resources. A useful extension would therefore be to develop an approximate cross-validation procedure. Second, although our methods automatically perform variable selection, the interpretation of the selected predictors is impaired by the lack of valid inference procedure which produces confidence intervals. This issue was tackled heuristically in Chapter 4, but still deserves further investigation. To do so, an option would be to consider the problem from a Bayesian perspective. Third, the mixed data methods introduced in Chapter 6 are based on heterogeneous data fitting terms, and suffer from scaling problems, with some variables taking more importance than others in the analysis. To overcome this, a solution would be to incorporate a scale parameter in our exponential family models.

On the theoretical side, there are also several directions of improvement. In particular, the general statistical guarantees derived in Chapter 6 depend on the geometry of a fixed dictionary of matrices embedding side information. More precisely, it indirectly depends on the sparsity of the dictionary, through the ℓ_1 norm of its elements. However, in several applications we have in mind, the dictionary is not sparse, but is rather generated from multivariate Gaussian distributions. An important extension of our our results would be to adapt them to this particular type of dictionaries. The second main limitation of our theoretical framework, is that we model heterogeneous data through univariate exponential family distributions. In other word, we are able to model data of different types such as Gaussian, binomial and Poisson data simultaneously, but we cannot model categorical data with more than two categories, or estimate simultaneously the mean and the variance of the variables. This is also an important point, as we observed in practice that fixing the scale of the variable could lead to poor estimation results in some cases.

Finally, there are very exciting perspectives to this dissertation in terms of applications. In particular, the developed methods received very positive responses from ecologists, and ongoing work includes the application of our methods to the analysis of several species abundance data sets.

Scientific production

Articles in peer-reviewed journals

Robin, G., C. Ambroise, and S. Robin (2018). Incomplete graphical model inference via latent tree aggregation. Statistical Modelling.

Husson, F. J. Josse, B. Narasimhan and G. Robin (2019). Imputation of mixed data with multilevel singular value decomposition. Journal of Computational and Graphical Statistics.

Robin, G., O. Klopp, J. Josse, É. Moulines and R. Tibshirani (2019). Main effects and interactions in mixed and incomplete data frames. Journal of American Statistical Association.

Robin, G., J. Josse, É. Moulines and S. Sardy (2019). Low-rank model with covariates for count data analysis. Journal of Multivariate Analysis.

Proceedings of international peer-reviewed conferences

Robin, G., H.-T. Wai, J. Josse, O. Klopp, and E. Moulines (2018). Low-rank interaction with sparse additive effects model for large data frames. Advances in Neural Information Processing Systems 31, pp. 5496–5506. Curran Associates, Inc.

Software

R package `lori` (2018).

R package `mimi` (2018).

Awards

Four-months visiting student researcher fellowship, France-Stanford Center for Interdisciplinary Studies (2017).

Student award, 31st Conference on Neural Information Processing Systems (2018).

Student award, Conference of the Eastern Mediterranean Region of the International Biometric Society (2018).

Bibliography

Abernethy, J., F. Bach, T. Evgeniou, and J.-P. Vert (2009, June). A new approach to collaborative filtering: Operator estimation with spectral regularization. *J. Mach. Learn. Res. 10*, 803–826.

Agarwal, A., S. Negahban, and M. J. Wainwright (2012, 04). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Ann. Statist. 40*(2), 1171–1197.

Agarwal, D. and B.-C. Chen (2009). Regression-based latent factor models. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, New York, NY, USA, pp. 19–28. ACM.

Agarwal, D., L. Zhang, and R. Mazumder (2011, September). Modeling item-item similarities for personalized recommendations on yahoo! front page. *Ann. Appl. Stat. 5*(3), 1839–1875.

Agresti, A. (2013). *Categorical Data Analysis, 3rd Edition*. Wiley.

Alon, N., T. Lee, A. Shraibman, and S. Vempala (2013). The approximate rank of a matrix and its algorithmic applications: Approximate rank. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, New York, NY, USA, pp. 675–684. ACM.

Amano, T., T. Székely, B. Sandel, S. Nagy, T. Mundkur, T. Langendoen, D. Blanco, C. U. Soykan, and W. J. Sutherland (2017, 12). Successful conservation of global waterbird populations depends on effective governance. *Nature 553*, 199.

Amat, J. A. and A. J. Green (2010). *Waterbirds as Bioindicators of Environmental Conditions*. Springer.

Angst, R., C. Zach, and M. Pollefeys (2011). The generalized trace-norm and its application to structure-from-motion problems. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, Washington, DC, USA, pp. 2502–2509. IEEE Computer Society.

Aubin, J.-P. and I. Ekeland (1984). *Applied nonlinear analysis*. Pure and applied mathematics. John Wiley, New-York. A Wiley-Interscience publication.

Audigier, V., F. Husson, and J. Josse (2016). A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification 10*(1), 5–26.

Audigier, V., I. White, S. Jolani, T. Debray, M. Quartagno, J. Carpenter, S. van Buuren, and M. Resche-Rigon (2018). Multiple imputation for multilevel data with continuous and binary variables. *Statistical Science*.

Bandeira, A. S. and R. van Handel (2016, July). Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Ann. Probab. 44*(4), 2479–2506.

Bell, R. M. and Y. Koren (2007, December). Lessons from the netflix prize challenge. *SIGKDD Explor. Newsl. 9*(2), 75–79.

Brooks, M. E., K. Kristensen, K. J. van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Maechler, and B. M. Bolker (2017). Modeling zero-inflated count data with glmmtmb. *bioRxiv*.

Brown, A. M., D. I. Warton, N. R. Andrew, M. Binns, G. Cassis, and H. Gibb (2014). The fourth-corner solution - using predictive models to understand how species traits interact with the environment. *Methods in Ecology and Evolution 5*(4), 344–352.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Cai, J.-F., E. J. Candès, and Z. Shen (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization 20*(4), 1956–1982.

Cai, T. and W.-X. Zhou (2013, December). A max-norm constrained minimization approach to 1-bit matrix completion. *J. Mach. Learn. Res. 14*(1), 3619–3647.

Candès, E. J., X. Li, Y. Ma, and J. Wright (2011, June). Robust principal component analysis? *J. ACM 58*(3), 11:1–11:37.

Candes, E. J. and Y. Plan (2010, June). Matrix completion with noise. *Proceedings of the IEEE 98*(6), 925–936.

Candès, E. J. and B. Recht (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics 9*(6), 717–772.

Candès, E. J. and T. Tao (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory 56*(5), 2053–2080.

Cannon, J. W. (2018). Hemorrhagic shock. *New England Journal of Medicine 378*(4), 370–379. PMID: 29365303.

Cao, Y. and Y. Xie (2016, March). Poisson matrix recovery and completion. *IEEE Transactions on Signal Processing 64*(6).

Chandrasekaran, V., P. A. Parrilo, and A. S. Willsky (2012, 08). Latent variable graphical model selection via convex optimization. *Ann. Statist. 40*(4), 1935–1967.

Chandrasekaran, V., S. Sanghavi, P. A. Parrilo, and A. S. Willsky (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization 21*(2), 572–596.

Chatterjee, S. (2015, February). Matrix estimation by universal singular value thresholding. *Ann. Statist. 43*(1), 177–214.

Chiquet, J., M. Mariadassou, and S. Robin (2018, 12). Variational inference for probabilistic poisson pca. *Ann. Appl. Stat. 12*(4), 2674–2698.

Choler, P. (2005, 1). Consistent shifts in alpine plant traits along a mesotopographical gradient. *Arctic, Antarctic, and Alpine Research 37*(4), 444–453.

Collins, M., S. Dasgupta, and R. E. Schapire (2001). A generalization of principal component analysis to the exponential family. In *Advances in Neural Information Processing Systems*. MIT Press.

Csiszár, I. and G. Tusnády (1984). Information Geometry and Alternating minimization procedures. *Statistics and Decisions Supplement Issue 1*.

Davenport, M. A., Y. Plan, E. van den Berg, and M. Wootters (2012, September). 1-Bit Matrix Completion. *ArXiv e-prints*.

David JS, Bouzat P, R. M. (2018). Evolution and organisation of trauma systems. *Anaesth Crit*.

de Falguerolles, A. (1998). Log-bilinear biplot in action. In J. Blasius and M. . Greenacre (Eds.), *Visualisation of categorical data*, pp. 527–533. Academic Press.

de Leeuw, J. (2006, January). Principal component analysis of binary data by iterated singular value decomposition. *Computational Statistics and Data Analysis 50*(1), 21–39.

de Rooij, M. and W. J. Heiser (2005, Mar). Graphical representations and odds ratios in a distance-association model for the analysis of cross-classified data. *Psychometrika 70*(1), 99–122.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society B 39*(1), 1–38.

Donoho, D. L. and I. M. Johnstone (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika 81*, 425–455.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics*, 1–26.

Escoufier, Y. (1982). The analysis of simple and multiple contingency tables. In *Proceedings of the international meeting of the analysis of multidimensional contingency tables*, Rome, Italy, pp. 53–77. R. Coppi.

Etayeb, K. S., A. Berbash, W. Bashimam, M. Bouzainen, A. Galidana, M. Saied, J. Yahia, and E. Bourass (2015). Results of the eighth winter waterbird census in libya in january 2012. *Biodiversity Journal 1*(6), 253–262.

Fithian, W. and J. Josse (2017). Multiple correspondence analysis & the multilogit bilinear model. *Journal of Multivariate Analysis*.

Fithian, W. and R. Mazumder (2018, 05). Flexible low-rank statistical modeling with missing data and side information. *Statist. Sci. 33*(2), 238–260.

Foygel, R. and N. Srebro (2011). Concentration-based guarantees for low-rank matrix reconstruction. In *COLT*.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1.

Gaiffas, S. and G. Lecué (2011, Oct). Sharp oracle inequalities for high-dimensional matrix prediction. *IEEE Transactions on Information Theory 57*(10), 6942–6957.

Galewski, T., B. Collen, L. McRae, J. Loh, P. Grillas, M. Gauthier-Clerc, and V. Devictor (2011). Long-term trends in the abundance of mediterranean wetland vertebrates: From global recovery to localized declines. *Biological Conservation 144*(5), 1392 – 1399. Ecoregional-scale monitoring within conservation areas, in a rapidly changing climate.

Gavish, M. and D. L. Donoho (2014). Optimal shrinkage of singular values. *arXiv:1405.7511v2*.

Gelman, A. and J. Hill (2007, June). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gentry, C. (2009). *A fully homomorphic encryption scheme*. Ph. D. thesis, Stanford University.

Giacobino, C., S. Sardy, J. Diaz Rodriguez, and N. Hengardner (2016). Quantile universal threshold for model selection. *arXiv:1511.05433v2*.

Golub, G. H. and C. F. Van Loan (1996). *Matrix Computations (3rd Ed.)*. Baltimore, MD, USA: Johns Hopkins University Press.

Goodman, L. A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Annals of Statistics 13*, 10–69.

Gopalan, P., F. J. R. Ruiz, R. Ranganath, and D. M. Blei (2014). Bayesian nonparametric poisson factorization for recommendation systems. In *In AISTATS*, pp. 275–283.

Gordon, G. J. (2002). Generalized^2 Linear^2 models. *Advances in Neural Information Processing Systems*, 577–584.

Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*. Acadamic Press.

Greenacre, M. J. and J. Blasius (2006). *Multiple correspondence analysis and related methods*. Boca Raton : Chapman & Hall/CRC. Formerly CIP.

Gross, D. (2011, March). Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory 57*(3), 1548–1566.

Hamada, S. R., T. Gauss, J. Pann, M. Dünser, M. Leone, and J. Duranteau (2015). European trauma guideline compliance assessment: the etrauss study. *Critical Care 19*(423).

Hamada, S. R., A. Rosa, T. Gauss, J.-P. Desclefs, M. Raux, A. Harrois, A. Follin, F. Cook, M. Boutonnet, T. Group, A. Attias, S. Ausset, G. Dhonneur, J. Duranteau, O. Langeron, C. Paugam-Burtz, R. Pirracchio, G. de St Maurice, B. Vigué, and A. Rouquette (2018, May). Development and validation of a pre-hospital "red flag" alert for activation of intra-hospital haemorrhage control response in blunt trauma. *Crit Care 22*(1), 113–113. 29728151[pmid].

Hastie, T., R. Mazumder, J. Lee, and R. Zadeh (2015, January). Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares. *The Journal of Machine Learning Research 16*, 3367–3402.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC.

Hay, S. I. e. a. (2017). Global, regional, and national disability-adjusted life-years (dalys) for 333 diseases and injuries and healthy life expectancy (hale) for 195 countries and territories, 1990-2016: a systematic analysis for the global burden of disease study 2016. *The Lancet 390*(10100), 1260–1344.

Heeringa, S., B. West, and P. Berlung (2010). *Applied Survey Data Analysis*. New York: Chapman & Hall/CRC.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology 24*(6), 417–441.

Hsu, D., S. M. Kakade, and T. Zhang (2011). Robust matrix decomposition with sparse corruptions. *EEE Transactions on Information Theory 57*(11), 7221–7234.

Husson, F., J. Josse, B. Narasimhan, and G. Robin (2018, April). Imputation of mixed data with multilevel singular value decomposition. *arXiv e-prints*, arXiv:1804.11087.

Husson, F., S. Lê, and J. Pagès (2010). *Exploratory Multivariate Analysis by Example Using R*. Chapman & Hall/CRC.

Josse, J., M. Chavent, B. Liquet, and F. Husson (2012). Handling missing values with regularized iterative multiple correspondence analysis. *Journal of Classification 29*(1), 91–116.

Josse, J. and F. Husson (2012). Handling missing values in exploratory multivariate data analysis methods. *Journal de la Societe Française de Statistique 153*(2), 79–99.

Josse, J. and F. Husson (2016). missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software 70*(1), 1–31.

Josse, J. and S. Sardy (2015). Adaptive shrinkage of singular values. *Statistics and Computing*, 1–10.

Josse, J., M. E. Timmerman, and H. A. Kiers (2013). Missing values in multi-level simultaneous component analysis. *Chemometrics and Intelligent Laboratory Systems 129*, 21 – 32. Multiway and Multiset Methods.

Josse, J. and S. Wager (2016). Bootstrap-based regularization for low-rank matrix estimation. *Journal of Machine Learning Research 17*(124), 1–29.

Josse, J., S. Wager, and S. Sardy (2017). denoiser: A package for low rank matrix estimation. *Journal of Statistical Software*.

Jouffroy, R., X. Bobbia, T. Gauss, P. Bouzat, and M. Pierre (2018). Process and organisation of in-hospital emergencies in france. *Anaesthesia Critical Care & Pain Medicine 37*(6), 629 – 631.

Kateri, M. (2014). *Contingency Table Analysis*. Springer New York.

Keshavan, R. H., A. Montanari, and S. Oh (2010, August). Matrix completion from noisy entries. *J. Mach. Learn. Res. 11*, 2057–2078.

Kiers, H. A. L. (1991, June). Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. *Psychometrika 56*(2), 197–212.

Kleinke, K. and J. Reinecke (2013). *countimp: Multiple Imputation of Incomplete Count Data*. R package version 1.0.

Klopp, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli 20*(1), 282–303.

Klopp, O. (2015). Matrix completion by singular value thresholding: sharp bounds. *Electronic journal of statistics 9*(2), 2348–2369.

Klopp, O., J. Lafond, É. Moulines, and J. Salmon (2015). Adaptive multinomial matrix completion. *Electronic Journal of Statistics 9*, 2950–2975.

Klopp, O., K. Lounici, and A. B. Tsybakov (2017, October). Robust matrix completion. *Probability Theory and Related Fields 169*(1), 523–564.

Koltchinskii, V. (2011a). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery*. Springer.

Koltchinskii, V. (2011b, 12). Von neumann entropy penalization and low-rank matrix estimation. *Ann. Statist. 39*(6), 2936–2973.

Koltchinskii, V. (2013). *A remark on low rank matrix recovery and noncommutative Bernstein type inequalities*, Volume Volume 9 of *Collections*, pp. 213–226. Beachwood, Ohio, USA: Institute of Mathematical Statistics.

Koltchinskii, V., K. Lounici, and A. B. Tsybakov (2011, 10). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist. 39*(5), 2302–2329.

Kumar, N. K. and J. Schneider (2017). Literature survey on low rank approximation of matrices. *Linear and Multilinear Algebra 65*(11), 2212–2244.

Lafond, J. (2015). Low rank matrix completion with exponential family noise. *Journal of Machine Learning Research: Workshop and Conference Proceedings 40*, 1–18.

Landgraf, A. J. and Y. Lee (2015, June). Generalized principal component analysis: Projection of saturated model parameters. Technical report, The Ohio State University, Department of Statistics.

Ledoux, M. (2001). *The concentration of measure phenomenon*, Volume 89 of *Mathematical Surveyx and Monographs*. American Mathematical Society, Providence.

Legendre, P., R. Galzin, and M. L. Harmelin-Vivien (1997). Relating behavior to habitat: solutions to the fourth-corner problem. *Ecology 78*(2), 547–562.

Leisch, F. and E. Dimitriadou (2010). mlbench: Machine learning benchmark problems. *R package version 2.1-1*.

Li, J. and D. Tao (2013). Simple exponential family PCA. *IEEE Transactions on Neural Networks and Learning Systems 24*(3), 485–497.

Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis with Missing Data*. New-York: John Wiley & Sons series in probability and statistics.

Liu, L., E. Dobriban, and A. Singer (2016). epca: High dimensional exponential family pca. *arXiv:1611.05550*.

Luisier, F., T. Blu, and M. Unser (2011). Image denoising in mixed poisson-gaussian noise. *IEEE Transactions on Image Processing 20*(3), 696–708.

Mao, X., S. X. Chen, and R. K. W. Wong (2017). Matrix completion with covariate information. *Journal of the American Statistical Association 0*(ja), 0–0.

Mardani, M., G. Mateos, and G. B. Giannakis (2013, Aug). Recovery of low-rank plus compressed sparse matrices with application to unveiling traffic anomalies. *IEEE Transactions on Information Theory 59*(8), 5186–5205.

Mazumder, R., T. Hastie, and R. Tibshirani (2010). Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research 11*, 2287–2322.

Mohamed, S., Z. Ghahramani, and K. A. Heller (2009). Bayesian exponential family pca. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (Eds.), *Advances in Neural Information Processing Systems 21*, pp. 1089–1096.

Murdoch, T. and A. Detsky (2013). The inevitable application of big data to health care. *JAMA 309*(13), 1351–1352.

Murray, J. (2018). Multiple imputation: a review of theoretical and practical findings. *Statistical Science*.

Narasimhan, B., L. Rubin, S. Gross, M. Bendersky, and P. W. Lavori (2017). Software for distributed computation on medical databases: A demonstration project. *Journal of Statistical Software 77*(13), 99–122.

Newman, D., S. Hettich, C. Blake, and C. Merz (1998). Uci repository of machine learning databases. *Irvine, CA: University of California, Department of Information and Computer Science.*.

Pagès, J. (2004). Analyse factorielle de donnÃ©es mixtes. *Revue de Statistique Appliquée 52*(4), 93–111.

Pagès, J. (2015). *Multiple factor analysis by example using R*. Chapman & Hall/CRC the R series (CRC Press). Taylor & Francis Group.

Pannekoek, J. and A. van Strien (2001). Trim 3 manual (trends & indices for monitoring data). *Statistics Netherlands*.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2*(11), 559–572.

Peres-Neto, P. R., S. Dray, and C. J. F. t. Braak (2016). Linking trait variation to the environment: critical issues with community-weighted mean correlation resolved by the fourth-corner approach. *Ecography*, n/a–n/a.

Quartagno, M. and J. Carpenter (2016). Multiple imputation for ipd meta-analysis: allowing for heterogeneity and studies with missing covariates. *Statistics in Medicine 35 (17)*, 2938–2954.

Quartagno, M. and J. Carpenter (2017). *jomo: A package for Multilevel Joint Modelling Multiple Imputation*.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Recht, B. (2011, December). A simpler approach to matrix completion. *J. Mach. Learn. Res. 12*, 3413–3430.

Recht, B., M. Fazel, and P. Parrilo (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review 52*(3), 471–501.

Resche-Rigon, M. and I. R. White (2016). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research*.

Robin, G., H.-T. Wai, J. Josse, O. Klopp, and E. Moulines (2018). Low-rank interaction with sparse additive effects model for large data frames. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 31*, pp. 5496–5506. Curran Associates, Inc.

Rohde, A. and A. B. Tsybakov (2011, 04). Estimation of high-dimensional low-rank matrices. *Ann. Statist. 39*(2), 887–930.

Rossaint, R., B. Bouillon, V. Cerny, T. J. Coats, J. Duranteau, E. Fernández-Mondéjar, D. Filipescu, B. J. Hunt, R. Komadina, G. Nardi, E. A. M. Neugebauer, Y. Ozier, L. Riddez, A. Schultz, J.-L. Vincent, and D. R. Spahn (2016, Apr). The european guideline on management of major bleeding and coagulopathy following trauma: fourth edition. *Critical Care 20*(1), 100.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Salmon, J., Z. Harmany, C. Deledalle, and R. Willett (2014). Poisson noise reduction with non-local pca. *Journal of Mathematical Imaging and Vision 48*(2), 279–294.

Sayoud, M., H. Salhi, B. Chalabi, A. Allali, M. Dakki, A. Qninba, M. E. Agbani, H. Azafzaf, C. Feltrup-Azafzaf, H. Dlensi, N. Hamouda, W. A. L. Ibrahim, H. Asran, A. A. Elnoor, H. Ibrahim, K. Etayeb, E. Bouras, W. Bashaimam, A. Berbash,

C. Deschamps, J. Mondain-Monval, A. Brochet, S. Véran, and P. D. du Rau (2017). The first coordinated trans-north african mid-winter waterbird census: The contribution of the international waterbird census to the conservation of waterbirds and wetlands at a biogeographical level. *Biological Conservation 206*, 11 – 20.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall/CRC.

Singh, A. P. and G. J. Gordon (2008). A unified view of matrix factorization models. In *ECML/PKDD*.

Srebro, N. (2004). *Learning with Matrix Factorizations*. Ph. D. thesis, Massachusetts Institute of Technology.

Srebro, N. and T. Jaakkola (2003). Weighted low-rank approximations. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pp. 720–727. AAAI Press.

Stekhoven, D. and P. Bühlmann (2012). Missforest - nonparametric missing value imputation for mixed-type data. *Bioinformatics 28*, 113–118.

Talagrand, M. (1996, January). A new look at independence. *Ann. Probab. 24*(1), 1–34.

Taylor, J. and R. Tibshirani (2018). Post-selection inference for -penalized likelihood models. *Canadian Journal of Statistics 46*(1), 41–61.

ter Braak, C. J., P. Peres-Neto, and S. Dray (2017, January). A critical issue in model-based inference for studying trait-based community assembly and a solution. *PeerJ 5*, e2885.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Timmerman, M. E. (2006). Multilevel component analysis. *British Journal of Mathematical and Statistical Psychology 59*(2), 301–320.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics 12*(4), 389–434.

Tseng, P. and S. Yun (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Math. Program. 117*(1-2, Ser. B), 387–423.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation* (1st ed.). Springer Publishing Company, Incorporated.

Udell, M., C. Horn, R. Zadeh, and S. Boyd (2016). Generalized low rank models. *Foundations and Trends in Machine Learning 9*(1).

Udell, M. and A. Townsend (2018). Why are big data matrices approximately low rank? *IAM Mathematics of Data Science (SIMODS)*.

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC, Boca Raton.

van Buuren, S. and K. Groothuis-Oudshoorn (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software, Articles 45*(3), 1–67.

Verbanck, M., J. Josse, and F. Husson (2013). Regularised PCA to denoise and visualise data. *Statistics and Computing*, 1–16.

Xu, H., C. Caramanis, and S. Sanghavi (2010). Robust pca via outlier pursuit. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, NIPS'10, USA, pp. 2496–2504. Curran Associates Inc.

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics 15 (2)*, 265–286.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Méthodes de rang faible pour l'analyse de données multi-source, hétérogènes et incomplètes

Mots Clefs : abondance d'espèces, complétion de matrices, données hétérogènes, famille exponentielle, modèles de rang faible

Résumé :
Dans les applications modernes des statistiques et de l'apprentissage, il est courant que les données récoltées présentent un certain nombre d'imperfections. En particulier, les données sont souvent *hétérogènes*, c'est-à-dires qu'elles contiennent à la fois des informations quantitatives et qualitatives, *incomplètes*, lorsque certaines informations sont inaccessibles ou corrompues, et *multi-sources*, c'est-à-dire qu'elles résultent de l'aggrégation de plusieurs jeux de données indépendants. Dans cette thèse, nous développons plusieurs méthodes pour l'analyse de données hétérogènes, incomplètes et multi-sources. Nous nous attachons à étudier tous les aspects de ces méthodes, en fournissant des études théoriques précises, des implémentations disponibles au public, ainsi que des évaluations empiriques. En particulier, nous considérons en détail deux applications issues de l'écologie pour la première et de la médecine pour la seconde.

Title : Low-rank methods for multi-source, heterogeneous and incomplete data

Keys words : exponential family models, low-rank models, matrix completion, species abundance data

Abstract :
In modern applications of statistics and machine learning, one often encounters many data imperfections. In particular, data are often *heterogeneous*, i.e. combine quantitative and qualitative information, *incomplete*, with missing values caused by machine failures or by the nonresponse phenomenon, and *multi-source*, when the data result from the compounding of diverse sources. In this dissertation, we develop several methods for the analysis of multi-source, heterogeneous and incomplete data. We provide a complete framework, and study all the aspects of the different methods, with thorough theoretical studies, open source implementations, and empirical evaluations. We study in details two particular applications from ecology and medical sciences.

Université Paris-Saclay, Espace Technologique, Bat. Discovery - RD 128 - 2e ét., 91190 Saint-Aubin