Approches pénalisées pour les analyses en sous-groupes : application en épidémiologie

Nadim Ballout

THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de :

l'Université Claude Bernard Lyon 1

Ecole Doctorale ED 205

Interdisciplinaire Sciences-Santé

Spécialité de doctorat : Biostatistiques

Discipline : Epidémiologie

Soutenue publiquement le 22/06/2020, par :

Nadim Ballout

Approches pénalisées pour les analyses en sous-groupes : Application en épidémiologie

---

Devant le jury composé de :

Anne-Laure FOUGERES	**Présidente du jury**
Professeure, Université Claude Bernard Lyon 1	
Sophie LAMBERT-LACROIX	**Rapporteure**
Professeure, Université de Grenoble	
Joseph SALMON	**Rapporteur**
Professeur, Université de Montpellier	
Marta AVALOS	**Examinatrice**
Maître de Conférences, Université de Bordeaux	
Jean-Louis MARTIN	**Examinateur**
Chargé de Recherche, IFSTTAR	
Vivian VIALLON	**Directeur de thèse**
Maître de Conférences, IARC	

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. Frédéric FLEURY**
Président du Conseil Académique	M. Hamda BEN HADID
Vice-président du Conseil d'Administration	M. Didier REVEL
Vice-président du Conseil Formation et Vie Universitaire	M. Philippe CHEVALIER
Vice-président de la Commission Recherche	M. Fabrice VALLEE
Directrice Générale des Services	M. Damien VERHAEGHE

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Doyen : M. Gilles RODE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Doyenne : Mme Carole BURILLON
Faculté d'Odontologie	Doyenne : Mme Dominique SEUX
Institut des Sciences Pharmaceutiques et Biologiques	Directrice : Mme Christine VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. Xavier PERROT
Département de formation et Centre de Recherche en Biologie Humaine	Directrice: Mme la Professeure Anne-Marie SCHOTT

COMPOSANTES & DEPARTEMENTS DE SCIENCES & TECHNOLOGIE

UFR Biosciences	Directrice : Mme Kathrin GIESELER
Département Génie Electrique et des Procédés (GEP)	Directrice : Mme Rosaria FERRIGNO
Département Informatique	Directeur : M. Behzad SHARIAT
Département Mécanique	Directeur : M. Marc BUFFAT
UFR – Faculté des sciences	Administrateur provisoire : M. Bruno ANDRIOLETTI
UFR (STAPS)	Directeur : M. Yannick VANPOULLE
Observatoire de Lyon	Directeur : Mme Isabelle DANIEL
Ecole Polytechnique Universitaire Lyon 1	Directeur : M. Emmanuel PERRIN
Ecole Supérieure de Chimie, Physique, Electronique (CPE Lyon)	Directeur : M. Bernard BIGOT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. Christophe VITON
Institut de Science Financière et d'Assurances	Directeur : M. Nicolas LEBOISNE
ESPE	Administrateur provisoire : M. Pierre CHAREYRON

Acknowledgements

C'est avec une vive émotion que j'écris ces quelques lignes afin de remercier toutes les personnes qui ont participé, de près ou de loin, aux résultats de ce travail de thèse.

Je voudrais dans un premier temps remercier mon directeur de thèse, Vivian VIALLON, pour sa gentillesse, sa confiance placée en moi, pour toutes les connaissances que j'ai pu acquérir grâce à lui au long de ce périple, et bien plus.

Merci à Sylviane LAFONT, Franck PICARD, et Julien JACQUES d'avoir accepté de composer mon Comité de suivi de thèse, et d'avoir ainsi suivi mon travail.

Je remercie également l'équipe UMRESTTE, notamment Amina NDIAYE pour son aide particulière. A noter que cette thèse a été financée par l'IFSTTAR. Un grand merci pour la mise à disposition du matériel informatique de qualité.

Je ne pourrais continuer ce paragraphe sans remercier les membres du jury de cette thèse qui ont accepté de juger mon travail : Marta AVALOS, Anne-Laure FOUGERES et Jean-Louis MARTIN, ainsi que Sophie LAMBERT-LACROIX et Joseph SALMON, rapporteurs, à qui je souhaite bon courage pour la lecture de ce manuscrit. Merci d'avance pour vos retours, remarques, et suggestions.

Merci aussi aux collègues pour la bonne ambiance, les bons moments partagés, et surtout les parties de tarot ! Une pensée chaleureuse pour Anne-Marie BIGOT, sa gentillesse légendaire, et son accompagnement administratif.

Alors que j'espère avoir remercié la majorité des personnes impliquée dans ma thèse au niveau de la sphère professionnelle, j'aimerais exprimer toute ma gratitude pour une collègue, qui a franchi la limite professionnelle/personnelle en devenant bien plus qu'une collègue : une vraie amie. Merci Clémence pour ton soutien à chaque instant et ton aide très spéciale.

J'en profite également pour remercier finalement la totalité de mes amis en France ou au Liban, pour tous leurs encouragements et les bons moments passés avec eux au cours de cette thèse.

Je ne pourrais clore ces remerciements sans avoir une pensée très émue pour ma famille (loin des yeux mais près du coeur) à qui je dédie ce travail. Tout d'abord à mes parents (Wafaa et Hassan) sans qui rien n'aurait été possible, merci d'avoir toujours cru en moi. Vous pouvez être fier de votre fils ! A mon frère (Wassim), ma sœur (Souheir) et leurs belles petites familles, leur soutien inconditionnel et sans faille dans les moments

difficiles, surtout les appels vidéo avec les êtres les plus chers à mes yeux (mes nièces Elena et Marina). Je vous aime.

Résumé

Dans un contexte où les cancers et l'insécurité routière font partie des principales causes de décès en France et dans le monde, chercher à en étudier les risques et aider à la prise en charge des malades et des victimes constituent un enjeu majeur de santé publique. Les études épidémiologiques mises en place pour répondre à ces besoins requièrent la disponibilité de nombreuses données qui deviennent de plus en plus détaillées et complexes. Certaines des méthodes statistiques utilisées classiquement ne satisfont pas pleinement aux exigences imposées par la taille et les caractéristiques de ces bases de données. L'objectif de ce travail de thèse est donc de développer des méthodes statistiques mieux adaptées, motivées par deux applications particulières : 1) la description des associations entre lésions chez les victimes d'accident de la route en fonction du type d'usager ; et 2) l'étude du rôle de certains métabolites dans le développement du cancer du sein, en fonction du sous-type de cancer du sein.

D'un point de vue méthodologique, les analyses stratifiées, ou en sous-groupes, constituent le coeur de nos recherches. En notant K le nombre de sous-groupes considérés, l'inférence statistique dans le contexte de ces analyses en sous-groupes revient en général à l'estimation de K vecteurs de paramètres, un vecteur par sous-groupe. Or, on s'attend généralement à une certaine homogénéité entre les K vrais vecteurs de paramètres. Nos méthodes reposent sur des pénalités de type fused lasso ou data shared lasso, et permettent de tirer profit de cette homogénéité pour réduire la complexité de la tâche d'apprentissage et améliorer la performance statistique de l'estimation. Par ailleurs, elles permettent l'identification des hétérogénéités éventuelles parmi les K vecteurs.

Dans le projet concernant la description des associations entre lésions chez les victimes d'accident de la route, nous nous sommes placés dans le cadre de l'estimation de modèles graphiques binaires stratifiés. Nous avons développé deux méthodes d'estimation basées chacune sur des régressions logistiques multiples en utilisant soit une pénalité de type fused lasso généralisé soit une pénalité de type data shared lasso. Dans le second projet, nous nous sommes placés dans le cadre général des études cas-témoins (appariées ou non), lorsque plusieurs sous-types de malades existent. Dans le cas des données appariées, nous avons étendu le data shared lasso au modèle de régression logistique conditionnelle, et avons montré la supériorité de l'approche par rapport à d'autres stratégies plus classiques. Dans le cas des données non-appariées, nous avons travaillé sous

le modèle de régression logistique multinomial, pour lequel deux formulations pénalisées par la norme ℓ_1 ont été proposées dans la littérature. Nous montrons que l'une de ses formulations correspond en fait à la version data shared lasso de l'autre : nos résultats nous permettent ainsi de comparer formellement les deux formulations, et fournir des recommandations sur le choix de la formulation à utiliser en pratique.
Globalement, nos résultats confirment que les méthodes tirant profit de l'homogénéité entre les K vecteurs, telles que celles reposant sur la pénalité data shared lasso, conduisent à des améliorations substantielles en termes d'efficacité d'estimation, lorsque cette homogénéité existe. Elles ciblent en effet une paramétrisation plus parcimonieuse lorsque des similarités existent entre les sous-groupes. De plus, leur implémentation est relativement aisée, et en tout cas comparable à celle de méthodes plus classiques. Nous avons développé des codes permettant leur implémentation sous le logiciel R, qui sont accessible via la plateforme Github. Nous recommandons leur utilisation, en complément des approches plus classiques.

Mots-clés : Analyse stratifiée ; Modèles graphiques ; Régression logistique multiple ; Régression logistique multinomiale ; Régression logistique conditionnelle ; DataShared Lasso ; Sécurité routière ; Cancer du sein.

Abstract

In a context where cancers and road safety are among the main causes of death in France and worldwide, seeking to study the risks and helping to provide care for patients and victims is a major public health issue. Epidemiological studies responding to these challenges require the availability of a large amount of data more and more detailed and complex. Some of statistical methods traditionally used do not fully meet the requirements imposed by the size and characteristics of these databases. The aims of this thesis is therefore to develop better adapted statistical methods, motivated by two particular applications : 1) the description of associations between injuries in road accident victims according to the type of user ; and 2) the study of certain metabolites role in the development of breast cancer, according to the subtype of breast cancer. From a methodological point of view, stratified (or subgroup) analyses are our research focus. If we note K the number of subgroups, statistical inference in the context of these subgroup analyses generally amounts to the estimation of K parameter vectors, i.e. one vector per subgroup. However, homogeneity is generally expected between the true K parameter vectors. Our methods are based on fused lasso or data shared lasso penalties, and take advantage of this homogeneity to reduce the complexity of the learning task and improve the statistical performance of the estimation. In addition, they allow the identification of possible heterogeneity among the K vectors. In the project about the description of associations between injuries among road accident victims, we worked in the context of the estimation of stratified binary graphical models. We developed two estimation methods, each based on multiple logistic regressions using either a generalized fused lasso penalty, or a data shared lasso penalty. In the second project, we worked in the general framework of case-control studies (matched or unmatched), when there are several disease subtypes. For matched designs, we extended data shared lasso to the conditional logistic regression model, and showed the superiority of the approach compared to other classical strategies. For unmatched designs, we worked under the multinomial logistic regression model, for which two formulations penalized by the ℓ_1 norm have been proposed in the literature. We show that one of these formulations actually corresponds to the data shared lasso version of the other : our results thus allow us to formally compare the two formulations, and provide recommendations on the choice of the formulation to be used in practice. Overall, our results confirm that methods which take advantage of

the homogeneity between the K vectors, such as those based on the data shared lasso penalty, lead to substantial improvements in terms of estimation efficiency, when this homogeneity exists. Indeed, they target a more parsimonious parameterization when similarities exist between subgroups. Moreover, their implementation is relatively easy, and comparable to more traditional methods. We have developed codes allowing their implementation under the R software, which are accessible via the Github platform. We recommend their use, in addition to more classical approaches.

Key words : Stratified analysis ; Graphical models ; Multiple logistic regressions ; Multinomial logistic regression ; Conditional logistic regression ; Datashared Lasso ; Road safety ; Breast cancer.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

1.1 Introduction générale

1.1.1 Contexte applicatif général

Mes travaux de thèse concernent principalement le développement de méthodes statistiques motivées par des applications en épidémiologie du cancer et du risque en matière de sécurité routière.

Depuis 2004, le cancer est la première cause de mortalité prématurée en France, devant les maladies cardiovasculaires (URL). A l'échelle mondiale, le Centre international de recherche sur le cancer (CIRC) estime à 9,5 millions de décès par cancer (tous sexes et âges confondus), et à 18,1 millions de cas incidents de cancer en 2018 (WHO, URL). D'après les données Eurostat (URL), plus de 1 320 000 personnes (soit 261 pour 100 000 habitants après standardisation) sont décédées des suites d'un cancer en 2015, représentant 22% des décès chez les femmes (principalement par cancers du sein et du poumon) et 29% chez les hommes (cancer du poumon et cancer colorectal) (URL). Selon l'Institut national du cancer (INCa), le nombre de nouveaux cas de cancer pour l'année 2018 en France métropolitaine a été estimé à 382 000 (204 600 chez l'homme et 177 400 chez la femme) (URL).

L'insécurité routière constitue un autre problème majeur de santé publique. Chaque année, on compte plus de 3 000 décès et 300 000 blessés parmi les victimes d'accidents de la route en France - les populations jeunes en payant le plus fort tribut (URL). Après avoir

sensiblement diminué sur la période 2008-2013, la mortalité par accident de la route est relativement stable ces dernières années. D'après l'Observatoire national interministériel de la sécurité routière (ONISR), 3 488 personnes sont décédées en 2018 en France entière, 3 684 en 2017, et 3 738 en 2016 (URL). Le nombre de blessés par accident de la route estimé est également stable : 73 253 en 2018, et 76 840 en 2017. Cependant, d'après une étude sur le nombre de blessés toutes gravités confondues, ces derniers chiffres pourraient être assez largement sous-évalués (Amoros et al., 2008). Les coûts engendrés par l'insécurité routière (coûts de la prise prise en charge, hospitalisation, rééducation, et coûts liés au déficit de productivité) constituent une charge importante de l'économie de l'Etat. Un rapport de l'ONISR fait état du coût des accidents corporels en France métropolitaine qui s'élèverait à 38,3 milliards d'euros, soit 2,2% du PIB (Observatoire national Interministériel de la sécurité routière 2017 : URL). Par ailleurs, un rapport sur la situation de la sécurité routière dans le monde publié en 2018 évoque que les "traumatismes sont actuellement la principale cause de décès chez les enfants et les jeunes adultes de 5 à 29 ans, mettant en évidence la nécessité de modifier le programme actuel pour la santé de l'enfant et de l'adolescent" (WHO, Rapport de situation sur la sécurité routière dans le monde 2018 : URL).

A l'échelle de l'Union Européenne, un rapport publié en 2016 fait état de plus de 26 000 décès liés aux accidents de la route en 2015, représentant près de 5,1 décès sur 100 000 habitants après standardisation (URL). D'après ce même rapport, dans l'ensemble des États membres de l'UE, c'est en France, en Allemagne et en Italie, que le nombre des victimes de la route a été le plus élevé en 2015.

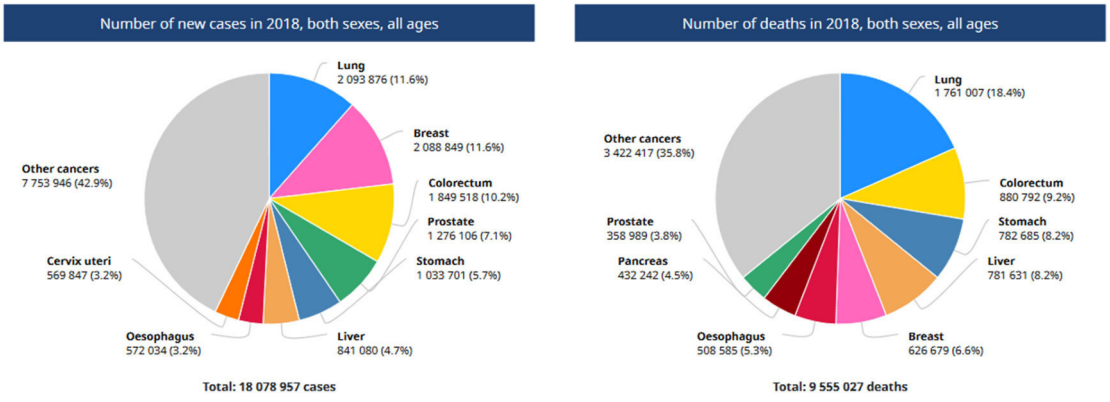


FIGURE 1.1: Nombre estimé de nouveaux cas et de décès en 2018 de tous types de cancers dans le monde pour les deux sexes et tous âges. *Source : Globocan 2018, IARC, OMS*

L'ensemble de ces chiffres souligne l'importance du cancer et de l'insécurité routière en terme de santé publique, et donc l'importance des études épidémiologiques visant à étudier le risque de cancer d'une part, et le risque en matière de sécurité routière d'autre part, dans la population. L'Organisation mondiale de la santé (OMS) définit en 1968 l'épidémiologie comme "l'étude de la distribution des maladies et des invalidités dans les populations humaines, ainsi que des influences qui déterminent cette distribution". Selon Jenicek et Cléroux, l'épidémiologie consiste en un "raisonnement et une méthode appliqués à la description des phénomènes de santé, à l'explication de leur étiologie et à la recherche des méthodes d'intervention les plus efficaces" (Jenicek and Cléroux, 1994). Cette discipline va donc au-delà du simple cadre de l'étude des épidémies et des maladies contagieuses. Elle s'applique également à étudier des phénomènes de santé plus complexes qui peuvent tenir compte du caractère multifactoriel des phénomènes étudiés (Gori, 1989).

Les études épidémiologiques dites observationnelles se divisent en deux grandes catégories : les études descriptives, et les études étiologiques (Bouyer and Cordier, 2003). Les enquêtes descriptives (par la mise en place d'études de variation, registres et sondages) visent à décrire l'état de santé d'une population : fréquence, incidence, répartition d'une maladie, d'un accident (de la circulation par exemple), gravité d'une lésion, etc. (Bouyer and Cordier, 2003). Cette étape descriptive permet d'avoir un aperçu quant aux besoins des populations en termes de prévention et de surveillance sur une base régionale ou nationale (Thacker and Stroup, 1998). Elle permet d'estimer le fardeau global lié à certaines maladies, en anglais le "burden of disease", à travers des critères tels que le nombre d'années de vie perdues dû à un état de santé particulier, etc. (Prüss-Üstün et al., 2003). Les études étiologiques visent quant à elles à décrire les déterminants (ou causes) de la maladie ou de l'état de santé étudié. Elles reposent sur des analyses statistiques d'association (Bouyer and Cordier, 2003).

Ces deux types d'étude reposent sur l'analyse statistique de données dites observationnelles (par opposition aux données expérimentales ou interventionnelles sur lesquelles reposent la plupart des essais cliniques par exemple). Les données aujourd'hui disponibles sont de plus en plus détaillées et complexes. Les travaux que j'ai effectués au cours de mes trois années de doctorat m'ont permis de développer deux méthodes statistiques. La première était motivée par une application en épidémiologie descriptive en sécurité routière et visait à décrire précisément les lésions subies par les victimes d'accident de

la route. La deuxième était motivée par une application en épidémiologie étiologique du cancer, et visait à étudier le rôle carcinogène de certains métabolites.

1.1.2 Description des lésions subies par les victimes d'accident de la route

En matière d'épidémiologie descriptive en sécurité routière, un outil essentiel en France est le Registre du Rhône, qui recense et décrit de manière continue les dommages corporels des victimes d'accidents de la route survenus dans le département du Rhône depuis 1995. L'inclusion des victimes dans ce Registre repose sur le lieu de l'accident - celui-ci impliquant au moins un véhicule en mouvement (patins et planches à roulettes inclus). Cette base recense approximativement 100 000 victimes d'accident de la route. Pour chacune d'entre elles, un recueil complet des bilans lésionnels, ainsi que la description des circonstances de l'accident sont enregistrés dans la base. Qualifié par le Comité national des Registres, son exhaustivité a été estimée à 73% et 87% pour les blessés mineurs/modérés, et les blessés graves survivants respectivement. L'échelle AIS (Abbreviated Injury Scale) (Committee on Medical Aspects of Automotive Safety, 1971) a été utilisée pour décrire le niveau d'atteinte de l'intégrité corporelle, qui caractérise l'événement de santé à considérer. Cette échelle est la principale échelle de gravité utilisée en accidentologie de la circulation et a été établie à partir des années 70 par l'Association for Advancement of Automotive Medicine(AAAM). Elle a ensuite été remaniée de multiples fois. La version 1998 de l'AIS a été traduite en français par un travail collaboratif entre l'Inrets, Institut national de recherche sur les transports et leur sécurité (Mireille Chiron, Amina N'Diaye), le CEESAR, Centre européen d'études de sécurité et d'analyse des risques, (Hervé Guillemot), et l'Institut de veille sanitaire, InVS (Bertrand Thélot). Cette traduction a fait l'objet d'un rapport publié par l'InVS en octobre 2004, tandis qu'une dernière version est disponible depuis mars 2013 (Committee on Medical Aspects of Automotive Safety, 1971).

La classification de l'Abbreviated Injury Scale 90 (AIS 90) a été utilisée pour coder chaque blessure que la victime a subie, décrivant ainsi son tableau lésionnel. Cette classification décrit le type et l'emplacement de la lésion en utilisant six chiffres. Le premier chiffre

identifie la région du corps [R] (tête, visage, colonne vertébrale, etc.), la seconde identifie la structure anatomique (vaisseaux, nerfs, etc.), les troisième et quatrième chiffres identifient la structure anatomique spécifique, ou la nature de la lésion lorsqu'une zone entière est atteinte [S] (colonne vertébrale, contusion, brûlure, etc.) et les cinquième et sixième précisent le type de lésion [N] (fractures, rupture, lacération, etc.). Voir quelques exemples de codage AIS dans le tableau 1.1.

TABLE 1.1: Exemples de codage AIS

R	T	S	N	Description
3	5	02	00	Cou/Squelette/Hyoïde/Fracture
4	2	10	04	Thorax/Vaisseaux/Artère pulmonaire/plaie (perforation)
5	4	14	10	Abdomen/Organes internes/Jéjunum-iléon (intestin grêle)/ Contusion (hématome)

Les données du Registre ont été largement utilisées pour décrire les prévalences, et leurs évolutions temporelles, des lésions subies par les victimes d'accident de la route. La description complète et détaillée de leurs lésions (zones corporelles et gravité) est essentielle, notamment pour la prise en charge des victimes (Lehmann et al., 2007).

Jusqu'à présent, les études descriptives menées sur le Registre du Rhône ont porté sur des lésions prises isolément, ou sur des regroupements entre lésions : estimation du nombre de blessés graves (Amoros et al., 2019), évolution des traumatisés crâniens (Koura et al., 2014), etc.

Cependant, le plus souvent, les victimes d'accidents de la route sont des polytraumatisés. Une description plus fine des tableaux lésionnels des victimes repose ainsi sur la description détaillée des associations éventuelles entre lésions. L'identification de telles associations (e.g. est-ce que les victimes souffrant de lésions aux membres inférieurs souffrent typiquement plus souvent d'autres lésions également) revêt deux intérêts principaux. Le premier concerne l'évaluation des besoins en termes de services de soins : une bonne connaissance des tableaux lésionnels, des victimes dans leur ensemble, est primordiale pour adapter au mieux leur accueil dans les services d'urgence. Un autre intérêt concerne le diagnostic. Si l'on identifie une association positive élevée entre une lésion externe particulière facile à diagnostiquer, et une lésion interne donnée plus difficile à diagnostiquer, alors une attention particulière sera accordée aux victimes souffrant de cette lésion externe, sachant que leur risque de souffrir également de cette lésion interne est plus élevé. Pour décrire la distribution des tableaux lésionnels, et notamment les associations entre lésions, nous avons opté pour une modélisation via les modèles graphiques

binaires. Selon le codage AIS, le jeu de données original de ce travail de thèse portant sur les accidents de la route contenait potentiellement 1 348 codes distincts correspondant à 1 348 lésions distinctes. Chaque lésion peut être modélisée par une variable aléatoire binaire qui équivaut à 1 si la victime souffre de cette lésion et 0 sinon. Cependant, comme plusieurs lésions ont des prévalences très faibles, certaines ont été regroupées. Après avoir converti le code AIS en code CIM-10 (Classification Internationale des Maladies et des maladies apparentées, 10ème révision), des regroupements ont été faits selon une adaptation du modèle EUROCOST (Lyons et al., 2006), pour finalement travailler avec 36 groupes de lésions. Ceci nous a ainsi mené à considérer 36 variables binaires, $U_1, \ldots, U_p$, où $U_j = 1$, si et seulement si le tableau lésionnel original de la victime contient au moins une blessure appartenant au j-ème groupe du modèle EUROCOST. Notons également que ces 36 groupes de lésions peuvent être classés en 6 catégories, correspondant approximativement aux zones corporelles (voir Table 2.2, pour plus de détails). Chaque victime est alors représentée par un vecteur $\mathbf{U} = (U_1, ..., U_p)^T \in \{0,1\}^p$ de p variables binaires, dont chaque composante j indique si la victime souffrait de la lésion j ou pas. Dans ce contexte, un modèle couramment utilisé est le modèle binaire exponentiel quadratique, également connu sous le nom de modèle d'Ising (Cox and Wermuth, 1994; Höfling and Tibshirani, 2009; Banerjee et al., 2008; Ravikumar et al., 2010). Le modèle d'Ising suppose l'existence d'un vecteur de paramètre $\mathbf{\Theta}^* = ((\theta^*_j)_{1\leq j\leq p}, (\theta^*_{j,\ell})_{1\leq j<\ell\leq p})^T$ dans $\mathbb{R}^{p(p+1)/2}$, tel que pour tout vecteur $\mathbf{u} = (u_1, ..., u_p) \in \{0,1\}^p$, la loi jointe du vecteur $\{\mathbf{U} = \mathbf{u}\}$ est donnée par

$$\mathbb{P}_{\mathbf{\Theta}^*}(\mathbf{U} = \mathbf{u}) = \exp\Big\{ \sum_{j=1}^{p} \theta^*_j u_j + \sum_{j=1}^{p-1} \sum_{\ell=j+1}^{p} \theta^*_{j,\ell} u_j u_\ell - A(\mathbf{\Theta}^*) \Big\}. \tag{1.1}$$

La log *partition function* $A : \mathbb{R}^p \to \mathbb{R}$ est un terme de normalisation qui garantit que $\sum_{\mathbf{u}\in\{0,1\}^p} \mathbb{P}_{\mathbf{\Theta}}(\mathbf{U} = \mathbf{u}) = 1$ pour tout $\mathbf{\Theta} \in \mathbb{R}^{p(p+1)/2}$, et se définit par

$$A(\mathbf{\Theta}) = \log \sum_{\mathbf{u}\in\{0,1\}^p} \exp\left(\sum_{j=1}^{p} \theta_j u_j + \sum_{j=1}^{p-1} \sum_{\ell=j+1}^{p} \theta_{j,\ell} u_j u_\ell \right). \tag{1.2}$$

Ce modèle repose donc sur l'estimation de $p(p+1)/2$ paramètres. Dans notre contexte applicatif, le nombre de lésions différentes considérées p vaut 36,et le nombre de paramètres total à estimer vaut donc 36*(36+1)/2, soit 666 paramètres. Nous nous retrouvons face

à un problème dit de grande dimension. Pour ce type de problème, les approches d'estimation "classiques" sont le plus souvent inopérantes, et on leur préfère des variantes pénalisées (voir paragraphe 1.2 ci-dessous pour une illustration dans le cas simple de la régression).

1.1.3 Etude du rôle carcinogène de certains métabolites

Plusieurs grandes cohortes ont été mises en place dans le monde afin d'étudier les causes et les facteurs de risques des cancers. En Europe, la cohorte European Prospective Investigation into Cancer and Nutrition (EPIC) menée par le CIRC est l'une des plus importantes études de cohortes au monde, avec plus de 521 000 participants recrutés dans 10 pays européens et suivis depuis plus de 15 ans (IARC, URL). Cette étude vise à étudier de manière prospective l'étiologie du cancer par rapport à l'alimentation, au mode de vie, aux facteurs environnementaux, ainsi qu'aux facteurs biologiques - des échantillons sanguins ayant été prélevés pour chaque participant au moment de leur inclusion dans l'étude. (description détaillée : Riboli et al., 2002).

En particulier, la recherche en épidémiologie du cancer s'intéresse particulièrement aux métabolites, et à leur effet carcinogène éventuel. Les niveaux de concentration sanguine (par exemple) des métabolites sont, en effet, à la fois le résultat des facteurs génétiques et environnementaux (incluant le style de vie) d'un individu. Ils pourraient donc expliquer les effets carcinogènes d'exposition comme le tabac, l'alcool, la pollution atmosphérique, ou encore ceux de certains gènes.

Sur la cohorte EPIC, plusieurs études métabolomiques cas-témoins nichées dans la cohorte ont été conduites (pour lesquelles les niveaux de concentration plasmatiques de centaines, voire de milliers, de métabolites sont mesurés) (Carayol et al., 2017; His, 2019; Bachlechner, 2016). Pour rappel, une étude cas-témoins consiste à recruter des sujets présentant l'événement de santé étudié (les cas), et parallèlement des personnes comparables mais ne présentant pas l'événement de santé étudié (les témoins). L'étude de cohorte consiste quant à elle en un suivi longitudinal, à l'échelle individuelle, d'un groupe de sujets. Ce type d'étude repose sur la collecte d'informations concernant des caractéristiques et des expositions des sujets à différents moments. L'objectif est de comparer la survenue d'une pathologie dans ce groupe de sujets en fonction de leur exposition à

un facteur présumé causal pour cette pathologie. Dans le cas d'une étude cas-témoins nichée dans une cohorte, les cas et les témoins sont recrutés à partir d'une cohorte déjà existante, et pour lesquels les données ont déjà été enregistrées (par exemple en recrutant tous les sujets dont la pathologie d'étude a été déclarée dans la cohorte, ainsi que les sujets semblables aux cas, mais dont la pathologie n'a pas été déclarée).

L'objectif de ces études est d'identifier les métabolites susceptibles d'être associés au risque de cancer. Là encore, les méthodes statistiques classiques se heurtent au problème des données de grande dimension, pour lequel les approches pénalisées peuvent fournir une solution.

1.2 Méthodes pénalisées

Pour simplifier l'exposé, nous nous plaçons dans le cadre simple du modèle de régression linéaire. La plupart de nos arguments s'étendent naturellement à d'autres modèles de régression paramétriques, notamment les modèles de régression linéaire généralisés, les modèles de régression utilisés dans l'analyse de survie (par exemple, le modèle de Cox), etc.

Modèle de régression linéaire Pour tout entier $m \geq 1$, notons $[m]$ l'ensemble $\{1, ..., m\}$. Soit $\mathbf{X} \in \mathbb{R}^{n*p}$, la matrice des données (déterministe) représentant les n observations $\mathbf{x}_i = (x_{i,1}, ..., x_{i,p}) \in \mathbb{R}^p$ du vecteur de p covariables, pour $i \in [n]$. Soit $\mathbf{X}_j = (x_{1,j}, ..., x_{n,j}) \in \mathbb{R}^n$, la j-ème colonne de $\mathbf{X}$, correspondant aux n observations de la j-ème covariable, pour $j \in [p]$. Soit $\mathbf{Y} = (y_1, ..., y_n)^T \in \mathbb{R}^n$, le vecteur réponse associé, tel que

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^* + \epsilon = \sum_{j=1}^{p} \beta_j^* \mathbf{X}_j + \epsilon \tag{1.3}$$

Dans cette expression, $\boldsymbol{\beta}^* = (\beta_1^*, ..., \beta_p^*) \in \mathbb{R}^p$ est le vecteur (inconnu) des coefficients de la régression, où β_j^* traduit l'influence de la covariable $\mathbf{X}_j$ sur $\mathbf{Y}$. Généralement, nous supposons que les composantes du vecteurs $\epsilon = (\epsilon_1, ..., \epsilon_n) \in \mathbb{R}^n$ sont indépendantes et identiquement distribuées $(i.i.d)$ selon une loi normale $N(0, \sigma^2)$ avec $\sigma > 0$ fixe mais inconnu.

Moindres carrés ordinaire Une méthode classique pour estimer $\boldsymbol{\beta}^*$ est la méthode des moindres carrés ordinaires (MCO). Elle consiste à minimiser la somme des carrés des résidus. L'estimateur $\hat{\boldsymbol{\beta}}^{MCO}$ de $\boldsymbol{\beta}^*$ est défini par

$$\hat{\boldsymbol{\beta}}^{MCO} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2,$$

où pour $a \in \mathbb{R}^n$, $||a||_2^2 = \sqrt{\sum_{i=1}^n a_i^2}$. L'unicité de l'estimateur MCO est établi lorsque la matrice $\mathbf{X}$ est de rang p. Dans ce cas, il s'écrit sous la forme

$$\hat{\boldsymbol{\beta}}^{MCO} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Les propriétés théoriques de cet estimateur sont bien connues. En particulier, lorsque $rg(\mathbf{X}) = p$ (cas où $p \leq n$), son erreur de prédiction quadratique moyenne est de l'ordre de

$$\frac{||\mathbf{X}(\hat{\boldsymbol{\beta}}^{MCO} - \boldsymbol{\beta}^*)||_2^2}{n} = \mathcal{O}_{\mathbb{P}}\left(\frac{p}{n}\right).$$

D'après ce résultat, lorsque p est fixe, l'erreur de prédiction quadratique moyenne tend vers 0 à la vitesse n^{-1} avec $n \to \infty$. Toutefois, d'après ce résultat, on peut voir que l'estimateur des MCO souffre du fléau de la dimension : par exemple si $p = n^\alpha$, avec $0 < \alpha < 1$, la vitesse vers laquelle l'erreur de prédiction moyenne tend vers 0 est non pas n^{-1}, mais $n^{-(1-\alpha)}$.

Néanmoins, quand $p > n$ - soit peu d'individus, mais beaucoup d'informations sur chacun - la matrice $\mathbf{X}^T\mathbf{X}$ n'est plus inversible. Une infinité de solutions est alors donnée pour le problème des MCO. La non-unicité de la solution $\boldsymbol{\beta}^{MCO}$ induit une réelle difficulté sur l'analyse et l'interprétation des solutions.

En règle générale, on ne peut pas améliorer les MCO, sauf en tirant profit d'informations préalables (ou d'hypothèses) : par exemple, la parcimonie du vecteur de paramètres.

Différentes méthodes pénalisées sont proposées pour traiter les problèmes de l'estimateur MCO. Elles introduisent des estimateurs pénalisés de la forme

$$\hat{\boldsymbol{\beta}}^{pen} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + pen(\boldsymbol{\beta}), \tag{1.4}$$

où $pen(.)$ est une fonction réelle de $\boldsymbol{\beta} \in \mathbb{R}^p$. Différents choix de la fonction de pénalisation $pen(.)$ sont possibles, en fonction de l'information préalable disponible (ou hypothèse).

Parcimonie/-Sparsity Dans la plupart des applications en grande dimension, il est légitime de supposer que seul un sous-ensemble de covariables est réellement associé à la variable réponse $\mathbf{Y}$. Cela revient à supposer que le vecteur de coefficients $\boldsymbol{\beta}^*$ est parcimonieux ("sparse" en anglais), i.e. que peu de ses composantes $\beta_j^*, j \in \{1, ..., p\}$ sont différentes de zéro. On définit le support $S(\beta^*)$ de $\boldsymbol{\beta}^*$ par les indices des coefficients non-nuls de $\boldsymbol{\beta}^*$:

$$S(\beta^*) = \{j : \beta_j^* \neq 0\}.$$

Si $\beta_j^* = 0$, alors la variable $\mathbf{X}_j$ n'explique pas la réponse $\mathbf{Y}$. Trouver $S(\beta^*)$, le support de β^* est donc équivalent à effectuer de la sélection de variables. Dans la suite, nous nous intéressons à l'estimation de $\boldsymbol{\beta}^*$ et à la sélection des β_j^* non-nuls.

1.2.1 Méthodes pénalisées dans le cas parcimonieux

Norme ℓ_0 Pour répondre à ce problème, la façon la plus naturelle pour obtenir des solutions parcimonieuses est d'utiliser la pénalité basée sur la norme ℓ_0 dans l'équation (1.4), c'est-à-dire $pen^{\ell_0}(\boldsymbol{\beta}) = \lambda||\boldsymbol{\beta}||_0 = \lambda \sum_{j=1}^p \mathbb{I}(\beta_j \neq 0)$ et $\mathbb{I}(.)$ est la fonction indicatrice. Mais l'estimateur correpondant est impossible à calculer en temps polynomial (NP-dur à calculer, Natarajan, 1995). Une autre approche qui encourage la sparsité du vecteur estimé est le Lasso, qui repose sur une pénalisation par la norme ℓ_1 : $pen^{lasso}(\boldsymbol{\beta}) = \lambda||\boldsymbol{\beta}||_1 = \lambda \sum_{j=1}^p |\beta_j|$.

Lasso : Least Absolute Shrinkage and Selection Operator L'estimateur lasso introduit par Tibshirani (1996) est défini par :

$$\hat{\boldsymbol{\beta}}_\lambda^{lasso} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1,$$

pour $\lambda \geq 0$. La formule précédente équivaut à une autre définition :

$$\hat{\boldsymbol{\beta}}_t^{lasso} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2 \text{ sous contrainte} \sum_{j=1}^p |\beta_j| \leq t \tag{1.5}$$

où $t \geq 0$. Contrairement à la régression ridge, qui elle repose sur une pénalisation par la norme ℓ_2, le lasso opère typiquement une sélection de variables, en plus de l'estimation des coefficients. Le terme de pénalité $||\boldsymbol{\beta}||_1$ rétrécit certaines estimations de $\hat{\boldsymbol{\beta}}_\lambda^{lasso}$

vers zéro et en met d'autres égaux à zéro. La force de cette pénalité est controlée par λ (ou t). Pour $\lambda = 0$ (i.e. $t \to \infty$), le lasso correpond à une regression lineaire classique ($\hat{\boldsymbol{\beta}}_0^{lasso} = \hat{\boldsymbol{\beta}}^{MCO}$). En revanche, pour $\lambda \to \infty$ (i.e. $t = 0$), tous les estimations deviennent nulles ($\hat{\boldsymbol{\beta}}_{\lambda\to\infty}^{lasso} = \mathbf{0}_p$). Le paramètre λ contrôle le niveau de parcimonie de la solution dans le sens où, la solution a tendance à être d'autant plus parcimonieuse que λ est grand [ou que t est petit] . Donc pour $\lambda > 0$, le lasso produit typiquement des solutions parcimonieuses, i.e. le vecteur $\hat{\boldsymbol{\beta}}_\lambda^{lasso}$ est sparse, i.e. $card\{S(\hat{\boldsymbol{\beta}}_\lambda^{lasso})\} << p$.

Le Lasso est une méthode relativement populaire en vertu des propriétés théoriques des estimateurs qu'elle renvoie (voir le paragraphe suivant), et de la convexité du problème d'optimisation sur lequel elle repose. Même si en général il n'existe pas de solution explicite, des algorithmes rapides existent pour résoudre ce problème tels que le lars (Efron et al., 2004) ou coordinate descent (Friedman et al., 2010).

Sous des hypothèses portant sur la matrice de design X et pour un choix approprié du paramètre de pénalité λ, il a été établi qu'avec une grande probabilité, le lasso retrouve exactement le vrai support (i.e. le vrai signal du support) (Zhao and Yu, 2006; Wainwright, 2009). Le lasso est alors dit consistant en sélection de variables, ou sparsistent. La condition d'irreprésentabilité est une condition suffisante pour obtenir la consistance des variables sélectionnées. Notons $S^* = S(\boldsymbol{\beta}^*)$ et S_c^* son complémentaire. On note $\mathbf{X}_S^*$ la matrice de dimensions $n \times |S^*|$ qui correspond aux colonnes de la matrice $\mathbf{X}$ dont l'indice appartient a l'ensemble S^*. L'hypothèse d'irrepresentabilité est satisfaite si :

— pour $\gamma \in]0, 1]$:

$$||\mathbf{X}_{S_c^*}^T \mathbf{X}_{S^*} (\mathbf{X}_{S^*}^T \mathbf{X}_{S^*})^{-1}||_\infty \leq (1 - \gamma)$$

où $||.||_\infty$ représente la norme L_∞ ; pour tout $M \in \mathbb{R}^{m\times n}$, cette norme est définie par $||M||_\infty = \underset{i=1,\ldots,m}{max} \sum_{j=1}^n |M_{i,j}|$.

— pour $C_{min} > 0$:

$$\Lambda_{min}(\frac{1}{n}\mathbf{X}_{S^*}^T \mathbf{X}_{S^*}) \leq C_{min}$$

où $\Lambda_{min}(M)$ représente la valeur propre minimale de la matrice M.

En d'autres termes, la condition d'irreprésentabilité stipule que le modèle restreint à S^* est identifiable et que les colonnes de S_c ne sont pas trop alignées avec celles de S^*.

Sous ces hypothèses, il a été établi qu'avec une grande probabilité :

— le lasso a une solution unique $\hat{\boldsymbol{\beta}}^{lasso} \in \mathbb{R}^p$ avec $S(\hat{\boldsymbol{\beta}}^{lasso}_{\lambda}) \subseteq S(\boldsymbol{\beta}^*)$ et satisfait la limite l_∞

$$||\hat{\boldsymbol{\beta}}^{lasso}_{S^*} - \boldsymbol{\beta}^*_{S^*}||_\infty \leq \underbrace{\left[||(\mathbf{X}^T_{S^*}\mathbf{X}_{S^*}/n)^{-1}||_\infty + \frac{4\sigma}{\sqrt{C_{min}}}\right]}_{g(\lambda_n)}$$

— Si $\underset{j \in S^*}{min}|\beta^*_j| > g(\lambda_n)$, alors le lasso retrouve exactement le vrai signal du support (i.e. $sign(\hat{\boldsymbol{\beta}}^{lasso}) = sign(\boldsymbol{\beta}^*)$)

Voir Wainwright (2009), pour plus de détails.

Sous des hypothèses sur la matrice de design $\mathbf{X}$, on peut montrer (Bickel et al., 2009) que l'erreur de prédiction quadratique moyenne est oraculaire, de l'ordre

$$\frac{||\mathbf{X}(\hat{\boldsymbol{\beta}}^{lasso} - \boldsymbol{\beta}^*)||_2^2}{n} = \mathcal{O}_{\mathbb{P}}\left(|S^*|\frac{log(p)}{n}\right).$$

Pour récapituler, le lasso produit des solutions parcimonieuses (sparse), combine de bonnes propriétés numériques (convexité du problème) et, sous certaines hypothèses, de bonnes propriétés théoriques (consistance en sélection de variables, erreur de prédiction oraculaire).

1.3 Le cas des données stratifiées

1.3.1 Motivations

Les analyses en sous-groupes sont fréquemment réalisées en complément des analyses principales en épidémiologie. Les sous-groupes peuvent être définis selon un critère (sexe, âge, type d'usager, sous-types de maladies, etc.), à partir de données initiales de patients/participants/victimes, et permettent l'étude d'un évènement de santé au sein de chaque sous-groupe. L'intérêt est souvent de mettre en évidence une interaction entre des facteurs de risques de l'évènement d'intérêt et la variable critère utilisée pour définir les sous-groupes. Dans le cadre des accidents de la route, les caractéristiques de l'accident jouent un rôle important dans l'ensemble des lésions subies par les victimes. Dans ce sens, les tableaux lésionnels sont susceptibles de varier en fonction du type d'usager de la route : piétons, occupants d'une voiture, motards, cyclistes, etc. Il convient donc

d'étudier les associations entre les lésions en fonction du type d'usager de la route, ce qui correspond donc à étudier des associations de lésions à travers plusieurs sous-groupes prédéfinis de victimes. De même, dans le cadre du cancer du sein, et de manière plus générale en recherche clinique, l'efficacité d'un traitement est étudiée initialement sur la population générale. Cependant, des effets différents peuvent être révélés parmi des sous-groupes (par exemple : sous-types de diabétiques). Dans chacune des deux situations, les analyses en sous-groupes reviennent à estimer K fois plus de paramètres, où K est le nombre de groupes. Cependant, une homogénéité entre les sous-groupes est le plus souvent attendue : certains facteurs de risque sont susceptibles d'être partagés par certains sous-groupes, et ces facteurs de risque communs peuvent avoir le même niveau d'association avec l'évènement d'intérêt dans différents sous-groupes.

De fait, nous allons pouvoir tirer profit de cette homogénéité par la prise en compte adéquate de cette parcimonie structurée particulière (Bach et al., 2012) pour réduire la complexité de la tâche d'apprentissage, et améliorer l'estimation (Viallon et al., 2016).

Lorsque le vecteur de paramètres à estimer est de grande dimension (comme cela a été abordé ci-dessus), il est nécessaire de tirer profit de toute l'information disponible a priori, telle que la sparsité du vecteur de paramètres $\boldsymbol{\beta}^*$. Dans le cadre des données stratifiées, ou en sous-groupes, où K vecteurs de paramètres sont à estimer $(\boldsymbol{\beta}^{(1)*}, ..., \boldsymbol{\beta}^{(K)*})$, on peut également chercher à tirer profit de l'information (ou l'hypothèse) concernant l'homogénéité possible entre ces vecteurs des paramètres.

Par ailleurs, nous souhaitons être capable d'identifier les hétérogénéités éventuelles entre ces sous- groupes, comme par exemple identifier des effets du tabac qui peuvent s'avérer différents entre les hommes et les femmes (intérêt "clinique").

Dans ce travail de thèse, nous utiliserons la pénalité introduite pour le DataShared Lasso pour tenir compte de l'homogénéité attendue entre les K vecteurs de paramètre à estimer, sans pour autant masquer les heterogénéités éventuelles existant entre ces vecteurs (Gross and Tibshirani, 2016; Ollier and Viallon, 2017).

1.3.2 DataShared Lasso

1.3.2.1 Notations

Soit un n-échantillon, $n \geq 1$, tel que l'observation $i \in [n]$ correspond au triplet (Y_i, x_i, Z_i) où $Y_i \in \mathbb{R}$ est la variable d'intérêt, $x_i \in \mathbb{R}^p$ le vecteur des covariables, et $Z_i \in [K]$ la

variable catégorielle décrivant la strate d'appartenance de l'observation i. Soit $n_k = \sum_{i\in[n]} \mathbb{I}(Z_i = k)$, le nombre d'observations de la strate k, si bien que $n = \sum_{k\in[K]} n_k$. Pour tout $k \in [K]$, soit $\mathbf{X}^{(k)} \in \mathbb{R}^{n_k \times p}$, la matrice des données (déterministe) représentant les n_k observations $\mathbf{x}_i = (x_{i,1}^{(k)}, ..., x_{i,p}^{(k)}) \in \mathbb{R}^p$ du vecteur de p covariables, pour $i \in [n_k]$. Soit $\mathbf{X}_j^{(k)} = (x_{1,j}^{(k)}, ..., x_{n,j}^{(k)}) \in \mathbb{R}_k^n$, la j-ème colonne de $\mathbf{X}^{(k)}$, correspondant aux n_k observations de la j-ème covariable, pour $j \in [p]$. Soit $\mathbf{Y}^{(k)} = (y_1^{(k)}, ..., y_{n_k}^{(k)})^T \in \mathbb{R}^{n_k}$ est le vecteur réponse associé, tel que

$$\mathbf{Y}^{(k)} = \mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)*} + \epsilon^{(k)} = \sum_{j=1}^{p} \beta_j^{(k)*}\mathbf{X}_j^{(k)} + \epsilon^{(k)} \qquad \text{pour tout } k \in [K]. \tag{1.6}$$

Dans cette expression, $\boldsymbol{\beta}^{(k)*} = (\beta_1^{(k)*}, ..., \beta_p^{(k)*}) \in \mathbb{R}^p$ est le vecteur (inconnu) des coefficients de la régression, où $\beta_j^{(k)*}$ traduit l'influence de la covariable $\mathbf{X}_j$ sur $\mathbf{Y}$ dans la strate k. Généralement, nous supposons que les composantes du vecteur $\epsilon^{(k)} = (\epsilon_1^{(k)}, ..., \epsilon_{n_k}^{(k)}) \in \mathbb{R}^{n_k}$ sont indépendantes et identiquement distribuées ($i.i.d$) selon une loi normale $N(0, \sigma^2)$ avec $\sigma > 0$ fixe mais inconnu.

1.3.2.2 DataShared Lasso

La méthode DataShared Lasso est basée sur la décomposition sur-paramétrée suivante du paramètre $\beta_j^{(k)*}$

$$\beta_j^{(k)*} = \mu_j^* + \gamma_j^{(k)*}, \tag{1.7}$$

pour tout $j \in [p]$ and $k \in [K]$. Ici, μ_j^* peut être considéré comme le paramètre global de la covariable j et est commun à toutes les strates, tandis que γ_j^* capte la variation du paramètre de la strate k autour de μ_j^*.

Même si la décomposition (1.7) est sur-paramétrée, en reprenant les idées du lasso qui repose sur la pénalisation du vecteur de paramètres par la norme ℓ_1, les estimations de μ_j^* et $\gamma_j^{(k)*}$ pour $k \in [K]$ et $j \in [p]$, peuvent être dérivées en minimisant le critère suivant

$$(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\gamma}}^{(1)}, ..., \hat{\boldsymbol{\gamma}}^{(K)}) \in \underset{\boldsymbol{\mu}, \boldsymbol{\gamma}^{(1)}, ..., \boldsymbol{\gamma}^{(K)}}{\operatorname{argmin}} \sum_{k=1}^{K} ||\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}(\boldsymbol{\mu} + \boldsymbol{\gamma}^{(k)})||_2^2 + \lambda_1 ||\boldsymbol{\mu}||_1 + \sum_{k=1}^{K} \lambda_2^{(k)} ||\boldsymbol{\gamma}^{(k)}||_1, \tag{1.8}$$

pour des valeurs appropriées des paramètres de régularisation $\lambda_1 \geq 0$ et $\lambda_2^{(k)} \geq 0$ pour tout $k \in [K]$. DataShared Lasso revient à définir les estimations $(\hat{\boldsymbol{\beta}}^{DSL(1)}, ..., \hat{\boldsymbol{\beta}}^{DSL(K)})$ de la forme $\hat{\boldsymbol{\beta}}^{DSL(k)} = \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\gamma}}^{(k)}$. La pénalité sur $||\boldsymbol{\mu}||_1$ encourage la sparsité du vecteur de paramètres globaux, tandis que celle sur $||\boldsymbol{\gamma}^{(k)}||_1$ encourage l'homogénéité entre les vecteurs $\hat{\boldsymbol{\beta}}^{DSL(k)}$.

La minimisation du critère (1.8) est équivalente à la minimisation du critère suivant

$$\sum_{k=1}^{K} ||\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)}||_2^2 + \lambda_1 \left(||\boldsymbol{\mu}||_1 + \sum_{k=1}^{K} \frac{\lambda_2^{(k)}}{\lambda_1} ||\boldsymbol{\beta}^{(k)} - \boldsymbol{\mu}||_1 \right) \tag{1.9}$$

avec $\boldsymbol{\beta}^{(k)} \in \mathbb{R}^p$ pour tout $k \in [K]$ et $\boldsymbol{\mu} \in \mathbb{R}^p$. Il est ainsi facile de voir que $\hat{\mu}_j$ est une version *shrunkée* et pondérée de la médiane de $(\hat{\beta}_j^{(1)}, ..., \hat{\beta}_j^{(K)})$ qu'on note $\text{WSmedian}_{\boldsymbol{\tau}}(\hat{\beta}_{1,j}, ...\hat{\beta}_{K,j})$ avec $\boldsymbol{\tau} = (\tau_1, ..., \tau_K)$ et $\tau_k = \lambda_2^{(k)}/\lambda_1$; en effet, on a $\hat{\mu}_j \in \underset{b}{\operatorname{argmin}}(|b| + \sum_{k=1}^{K} \tau_k |\hat{\beta}_j^{(k)} - b|)$. Donc, le paramètre global estimé de la covariable j correspond à une version *shrunkée* et pondérée de la médiane des paramètres estimés de la covariable j à travers les strates. En conséquence, les estimations $(\hat{\boldsymbol{\beta}}^{DSL(1)}, ..., \hat{\boldsymbol{\beta}}^{DSL(K)})$ sont encouragées à être proches de leur médiane *shrunkée* et pondérée $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, ..., \hat{\mu}_p)$ au sens de la norme ℓ_1.
On note C le nombre de paramètres non nuls dans le modèle. La complexité de DataShared Lasso est définie par $C^{DSL} = ||\boldsymbol{\mu}^*||_0 + \sum_k ||\boldsymbol{\beta}^{(k)*} - \boldsymbol{\mu}^*||_0$ (voir Figure 1.2).

Comme nous le montrons ci-dessous, la méthode DataShared Lasso peut être considérée comme une généralisation de plusieurs approches plus classiques basées sur d'autres paramétrisations du modèle, qui correspondent à des contraintes particulières sous la décomposition (1.7).

1.3.2.3 Pooled Lasso

Une première méthode classique consiste à négliger l'information associée aux strates et de travailler sous l'hypothèse $\boldsymbol{\beta}^{(1)*} = ... = \boldsymbol{\beta}^{(K)*} = \boldsymbol{\beta}^*$. Ainsi Pooled Lasso revient à définir les estimations $(\hat{\boldsymbol{\beta}}^{Po(1)}, ..., \hat{\boldsymbol{\beta}}^{Po(K)})$ de la forme $\hat{\boldsymbol{\beta}}^{Po(k)} = \hat{\boldsymbol{\beta}}^{Po}$ pout tout $k \in [K]$

comme solutions minimisant le critère suivant

$$\hat{\boldsymbol{\beta}}^{Po} \in \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{k=1}^{K} ||\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1, \tag{1.10}$$

pour une valeur appropriée du paramètre de régularisation $\lambda \geq 0$. Le principal défaut de cette méthode réside dans le fait qu'elle n'identifie pas les hétérogénéités éventuelles entre les vecteurs $\boldsymbol{\beta}^{(k)}$. La complexité de Pooled Lasso est définie par $C^{Po} = \sum_j ||\beta^*||_0$ (voir Figure 1.2). Cette méthode peut être considéré comme un cas particulier de DataShared Lasso quand on impose la contrainte $\gamma_j^{(k)*} = 0$ pour tout $j \in [p]$ et $k \in [K]$, qui implique $\beta_j^{(k)*} = \mu_j^*$ $j \in [p]$ et $k \in [K]$.

1.3.2.4 Indep Lasso

La méthode Indep Lasso est une autre stratégie classique, qui consiste à estimer indépendamment les vecteurs $\boldsymbol{\beta}^{(1)*}, ..., \boldsymbol{\beta}^{(K)*}$. Plus précisément, la sélection des covariables associées à la variable d'intérêt sur chaque strate et l'estimation de leurs effets sont résolus par un Lasso simple appliqué sur chaque strate. Ainsi Indep Lasso revient à définir les estimations $(\hat{\boldsymbol{\beta}}^{In(1)}, ..., \hat{\boldsymbol{\beta}}^{In(K)})$ comme solutions minimisant le critère suivant

$$(\hat{\boldsymbol{\beta}}^{In(1)}, ..., \hat{\boldsymbol{\beta}}^{In(K)}) \in \underset{\boldsymbol{\beta}^{(1)},...,\boldsymbol{\beta}^{(K)}}{\operatorname{argmin}} \sum_{k=1}^{K} \left(||\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)}||_2^2 + \lambda^{(k)}||\boldsymbol{\beta}^{(k)}||_1 \right), \tag{1.11}$$

pour des valeurs appropriées des paramètres de régularisation $\lambda^{(k)} \geq 0$ pour tout $k \in [K]$. Le principal défaut de cette méthode réside dans le fait qu'elle ne tient pas compte de l'homogénéité potentielle entre les vecteurs $\boldsymbol{\beta}^{(k)}$. De plus, pour tout $k_1 \neq k_2$ et $j \in [p]$ fixé, on ne pourrait pas interpréter les différences observées entre $\hat{\beta}_j^{In(k_1)}$ et $\hat{\beta}_j^{In(k_2)}$ puisque ces valeurs sont différentes par construction. La complexité d'Indep Lasso est définie par $C^{In} = \sum_k ||\boldsymbol{\beta}^{(k)}||_0$ (voir Figure 1.2). Cette méthode peut être considérée comme un cas particulier de DataShared Lasso quand on impose la contrainte $\mu_j^* = 0$ pour tout $j \in [p]$, qui implique $\beta_j^{(k)*} = \gamma_j^{(k)*}$ pour tout $j \in [p]$ et $k \in [K]$.

1.3.2.5 Ref Lasso

La méthode Ref Lasso est une autre stratégie classique qui consiste à sélectionner une strate de référence $r \in [K]$, a priori, puis à décomposer les paramètres $\beta_j^{(k)*}$ comme suit

$$\beta_j^{(k)*} = \beta_j^{(r)*} + \delta_j^{(k)*}, \tag{1.12}$$

avec $\delta_j^{(r)*} = 0$ pour tout $j \in [p]$ et $k \in [K]$ (Gertheiss and Tutz, 2012). Pour simplifier, supposons que la première strate soit choisie comme référence, $r = 1$. Ainsi Ref Lasso revient à définir les estimations $(\hat{\boldsymbol{\beta}}^{Ref(1)}, ..., \hat{\boldsymbol{\beta}}^{Ref(K)})$ de la forme $\hat{\boldsymbol{\beta}}^{Ref(k)} = \hat{\boldsymbol{\beta}}^{Ref(1)} + \hat{\boldsymbol{\delta}}^{(k)}$ comme solutions minimisant le critère suivant

$$\begin{aligned}(\hat{\boldsymbol{\beta}}^{Ref(1)}, \hat{\boldsymbol{\delta}}^{(2)}, ..., \hat{\boldsymbol{\delta}}^{(K)}) \in \underset{\boldsymbol{\beta}^{(1)}, \boldsymbol{\delta}^{(2)}, ..., \boldsymbol{\delta}^{(K)}}{\operatorname{argmin}} & \sum_{k=1}^{K} ||\mathbf{Y}^{(k)} - \mathbf{X}^{(k)}(\boldsymbol{\beta}^{(1)} + \boldsymbol{\delta}^{(k)})||_2^2 \\ & + \lambda_1 ||\boldsymbol{\beta}^{(1)}||_1 + \sum_{k=2}^{K} \lambda_2^{(k)} ||\boldsymbol{\delta}^{(k)}||_1\end{aligned} \tag{1.13}$$

pour des valeurs appropriées des paramètres de régularisation λ_1 et $\lambda_2^{(k)}$. Cette méthode tient en partie compte de l'homogénéité éventuelle entre les vecteurs $\boldsymbol{\beta}^{(k)}$. En effet, seules les différences entre les paramètres de la strate de référence $r = 1$ et les paramètres des autres strates sont pénalisées. Donc pour tout $k_1 \neq 1$ et $k_2 \neq 1$, on ne peut pas interpréter les différences éventuelles entre $\hat{\beta}_j^{Ref(k_1)}$ et $\hat{\beta}_j^{Ref(k_2)}$ pout tout $j \in [p]$. La complexité de Ref Lasso est définie par $C^{Ref} = ||\boldsymbol{\beta}^{(r)*}||_0 + \sum_k ||\boldsymbol{\beta}^{(k)*} - \boldsymbol{\beta}^{(r)*}||_0$. Donc la complexité de cette méthode dépend du choix arbitraire de la strate de référence r (voir Figure 1.2). C^{Ref} est minimale si $\beta_j^{(r)*} \in \text{mode}(0, \beta_j^{(1)*}, ..., \beta_j^{(K)*})$, pour tout $j \in [p]$. Trouver une strate r répondant à cette condition est rare en pratique. Pour tout $j \in [p]$, soit r_j^* la strate de référence optimale spécifique à la covariable j. La version optimale de Ref Lasso (OptRef Lasso) consiste donc à utiliser la paramétrisation $\beta_j^{(k)*} = \beta_j^{(r_j)*} + \delta_j^{(k)*}$, pour tout $k \neq r_j^*$. Il existe toujours $r_j^* \in [K]$ tel que $\beta_j^{(r_j)*} \in \text{mode}(0, \beta_j^{(1)*}, ..., \beta_j^{(K)*})$ mais qui ne sont jamais accessibles en pratique.

La méthode Ref Lasso peut être considérée comme un cas particulier de DataShared Lasso quand on impose la contrainte $\gamma_j^{(r)*} = 0$ pour tout $j \in [p]$, qui implique $\mu_j^* = \beta_j^{(r)*}$ et $\gamma_j^{(k)*} = \beta_j^{(k)*} - \beta_j^{(r)*}$ pour tout $j \in [p]$ et $k \in [K] \backslash r$.

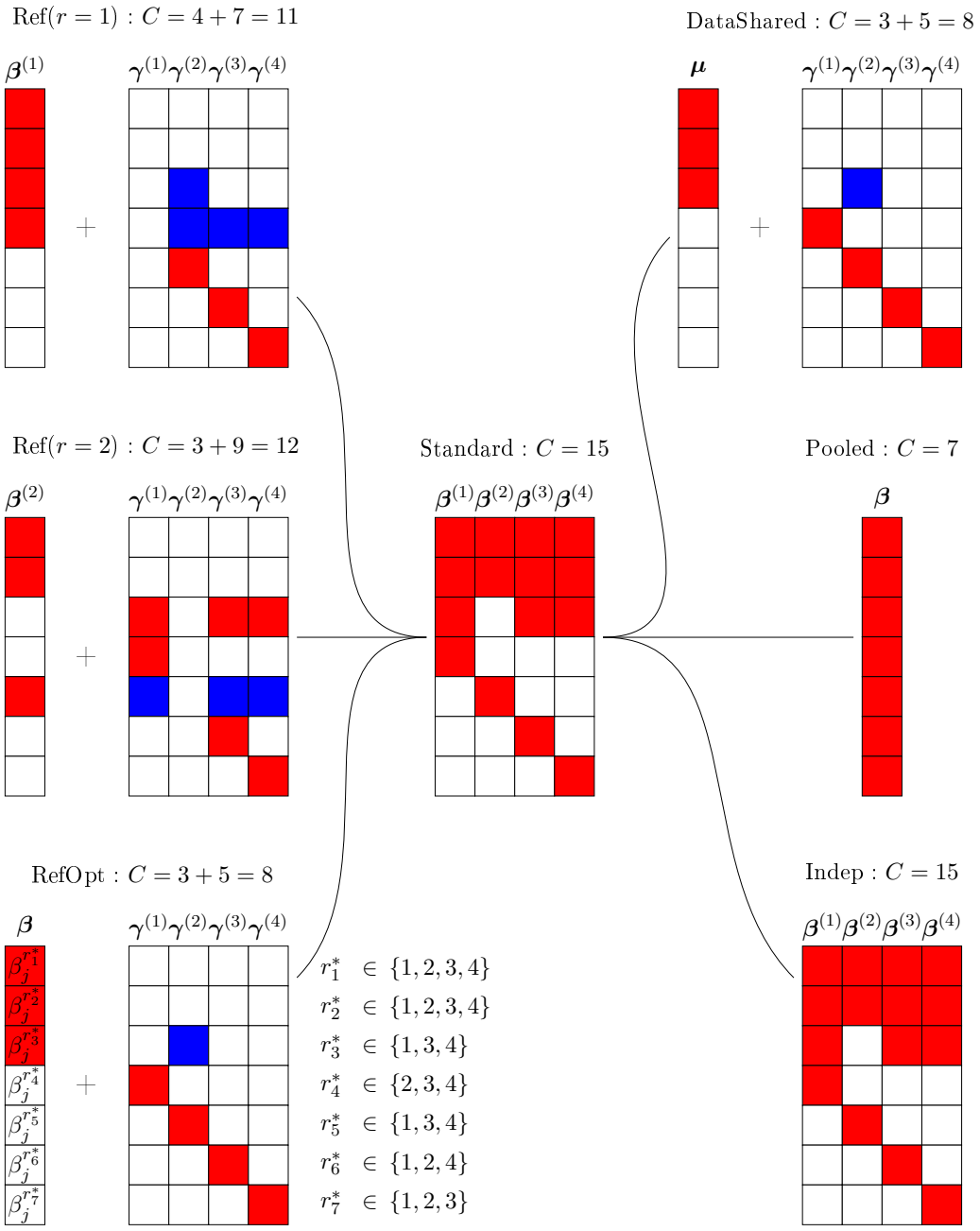


FIGURE 1.2 Une représentation graphique qui montre la paramétrisation et la complexité (notée C) obtenues par les méthodes DataShared Lasso, Pooled Lasso, Indep Lasso, Ref Lasso et OptRef Lasso, sur un exemple simple pour $p = 7$ et $K = 4$. Dans chaque matrice et vecteur, les entrées rouges correspondent à une valeur $\beta \in \mathbb{R}^*$, les entrées bleues correspondent à la valeur $-\beta$ et les entrées blanches à la valeur 0. La méthode Ref Lasso est représentée par deux paramétrisations, la première [resp. seconde] est obtenue en choisissant la strate 1 [resp. 2] comme strate de référence. Cet exemple illustre comment la complexité de la méthode Ref Lasso est affectée par le choix de la strate de référence.

1.3.2.6 Proprietés de DataShared Lasso

Dans le cadre de modèles linéaires stratifiés, il a été démontré que DataShared Lasso visait la même paramétrisation que OptRef Lasso (Ollier and Viallon, 2017). Pour chaque

covariable, DataShared Lasso permet d'identifier les strates sur lesquelles les paramètres diffèrent de ceux de la strate de référence optimale, sous des hypothèses presque identiques à celles requises par OptRef Lasso (qui est infaisable en pratique). Il a en outre été démontré que DataShared Lasso était aussi performant que OptRef Lasso, et que, tant sur le plan théorique qu'empirique, elle surpassait les trois approches les plus courantes (Pooled Lasso, Indep Lasso et Ref Lasso) (Ollier and Viallon, 2017). D'autre part, Ollier and Viallon (2017) ont établi la sparsistency de DataShared Lasso sous certaines hypothèses techniques dans le cas des modèles de régression linéaire.

D'autre part, d'un point de vue pratique, une propriété intéressante de DataShared Lasso est qu'elle peut être écrite comme un lasso pondéré après une transformation simple des données originales. On note $\boldsymbol{\mathcal{Y}} = (\mathbf{Y}^{(1)^T}, ..., \mathbf{Y}^{(K)^T}) \in \mathbb{R}^n$ le vecteur des n observations de la variable réponse sur l'ensemble des strates. Le critère à minimiser s'écrit alors sous la forme suivante :

$$||\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}\boldsymbol{\phi}||_2^2 + \lambda||\boldsymbol{\phi}||_1, \tag{1.14}$$

avec,

$$\boldsymbol{\mathcal{X}} = \begin{pmatrix} \mathbf{X}^{(1)} & \frac{\mathbf{X}^{(1)}}{\tau_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}^{(K)} & 0 & \cdots & \frac{\mathbf{X}^{(K)}}{\tau_K} \end{pmatrix} \in \mathbb{R}^{n \times (K+1)p} \text{ et } \boldsymbol{\phi} = \begin{pmatrix} \boldsymbol{\mu} \\ \tau_1 \boldsymbol{\gamma}^{(1)} \\ \vdots \\ \tau_K \boldsymbol{\gamma}^{(K)} \end{pmatrix} \in \mathbb{R}^{(K+1)p},$$

pour des valeurs données $\tau_k \geq 0, k \in [K]$.

Il est donc facile à implémenter sous une variété de modèles de régression, puisque des algorithmes rapides et très efficaces sont maintenant disponibles pour résoudre le problème d'optimization du lasso sous de nombreux modèles de régression : par exemple le paquet glmnet est maintenant disponible sur R, Matlab et Python et utilise plusieurs astuces pour rendre l'implémentation extrêmement rapide, telles que des règles fortes pour éliminer les prédicteurs et les problèmes connexes dans la régression lasso (Tibshirani et al., 2012; El Ghaoui et al., 2012). Cette propriété de réecriture comme un lasso s'étend naturellement à l'ensemble des modèles généralisés, aux modèles de Cox, etc.

1.3.3 Comparaison avec des méthodes alternatives

1.3.3.1 Group Lasso

Dans ce contexte, le sous-groupe est constitué de sous-ensembles d'échantillons et non de sous-groupes de variables. Définissons $\boldsymbol{\beta}_j = (\beta_j^{(1)}, ..., \beta_j^{(K)})$ et

$$\tilde{\mathbf{X}}_j = \begin{pmatrix} \mathbf{X}_j^{(1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{X}_j^{(K)} \end{pmatrix} \in \mathbb{R}^{n \times K},$$

pour tout $j \in [p]$. Les estimations $(\hat{\boldsymbol{\beta}}_1, ..., \hat{\boldsymbol{\beta}}_p)$ peuvent être dérivées en minimisant le critère suivant

$$(\hat{\boldsymbol{\beta}}_1, ..., \hat{\boldsymbol{\beta}}_j) \in \underset{\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_p}{\operatorname{argmin}} ||\mathbf{Y} - \sum_{j=1}^{p} \tilde{\mathbf{X}}_j \boldsymbol{\beta}_j||_2^2 + \lambda \sum_{j=1}^{p} ||\boldsymbol{\beta}_j||_2, \tag{1.15}$$

pour une valeur appropriée du paramètre de régularisation $\lambda \geq 0$. Ainsi Group Lasso revient à définir les estimations $(\hat{\boldsymbol{\beta}}^{Gr(1)}, ..., \hat{\boldsymbol{\beta}}^{Gr(K)})$ de la forme $\hat{\boldsymbol{\beta}}^{Gr(k)} = (\hat{\beta}_1^{(k)}, ..., \hat{\beta}_p^{(k)})$, pour tout $k \in [K]$. Le terme $\sum_{j=1}^{p} ||\boldsymbol{\beta}_j||_2$ encourage la parcimonie sur les sous-groupes, et non sur les covariables. Cependant, le Group Lasso n'est pas bien adapté à l'identification des hétérogénéités.

1.3.3.2 Fused Lasso généralisé

La méthode Fused Lasso généralisé consiste à estimer les vecteurs $\boldsymbol{\beta}^{(1)*}, ..., \boldsymbol{\beta}^{(K)*}$ en pénalisant toutes les différences entre les paramètres des différentes strates. Ainsi Fused Lasso revient à définir les estimations $(\hat{\boldsymbol{\beta}}^{Fus(1)}, ..., \hat{\boldsymbol{\beta}}^{Fus(K)})$ comme solutions minimisant le critère suivant

$$(\hat{\boldsymbol{\beta}}^{Fus(1)}, ..., \hat{\boldsymbol{\beta}}^{Fus(K)}) \in \underset{\boldsymbol{\beta}^{(1)}, ..., \boldsymbol{\beta}^{(K)}}{\operatorname{argmin}} \sum_{k=1}^{K} ||\mathbf{Y}^{(k)} - \mathbf{X}^{(k)} \boldsymbol{\beta}^{(k)}||_2^2 + \lambda_1 \sum_{k=1}^{K} ||\boldsymbol{\beta}^{(k)}||_1 + \lambda_2 \sum_{\substack{(k_1, k_2) \in [K]^2 \\ k_1 < k_2}} ||\boldsymbol{\beta}^{(k_1)} - \boldsymbol{\beta}^{(k_2)}||_1, \tag{1.16}$$

pour des valeurs appropriées des paramètres de régularisation $\lambda_1 \geq 0$ et $\lambda_2 \geq 0$. Le terme $\sum_k ||\boldsymbol{\beta}^{(k)}||_1$ encourage les solutions $\hat{\boldsymbol{\beta}}^{Fus(k)}$ à être sparse et le terme $\sum_{k_1<k_2} ||\boldsymbol{\beta}^{(k_1)} - \boldsymbol{\beta}^{(k_2)}||_1$ encourage l'homogénéité des vecteurs solutions $\hat{\boldsymbol{\beta}}^{Fus(k)}$: on peut alors interpréter les différences observées entre $\hat{\beta}_j^{Fus(k_1)}$ et $\hat{\beta}_j^{Fus(k_2)}$ pour deux strates $k_1 \neq k_2$ et $j \in [P]$ fixé.

Si Kp est fixe, la méthode Fused Lasso renvoie des estimateurs asymptotiquement optimaux, sous des hypothèses assez générales (Voir Gertheiss and Tutz, 2012; Viallon et al., 2016). Toutefois, lorsque nous sommes dans un cadre non-asymptotique, l'approche et les propriétés qui lui appartiennent n'ont pour le moment pas été décrites. Du fait de ces limites, nous nous sommes intéressés à l'utilisation de la méthode DataShared Lasso qui montre beaucoup d'avantages à la fois sur les plans théoriques et pratiques, par sa facilité de mise en œuvre et ses propriétés adaptées dans le cadre des données stratifiées.

1.4 Contributions

Dans cette thèse, j'ai étendu le DataShared Lasso dans deux directions particulières : dans le cadre de l'estimation de modèles graphiques binaires sur données stratifiées (travail motivé par la description des tableaux lésionnels des victimes d'accident de la route), et dans le cadre de l'analyse de données cas-témoins en présence de plusieurs sous-types de maladie (travail motivé par l'analyse des études métabolomiques disponibles dans la cohorte EPIC).

1.4.1 Modèles graphiques binaires : données stratifiées

Les modèles graphiques sont utilisés dans de nombreuses applications telles que le diagnostic médical, la sécurité informatique, etc. De plus en plus souvent, l'estimation de tels modèles doit être effectuée sur plusieurs strates prédéfinies de l'ensemble de la population. Les modèles graphiques se sont avérés utiles pour modéliser la distribution conjointe de p variables données, et ils fournissent une représentation graphique des dépendances conditionnelles entre elles (Lauritzen, 1996). Un modèle graphique est un graphique non dirigé qui consiste en un ensemble de p nœuds correspondant aux p variables, avec un ensemble d'arêtes joignant certaines paires de nœuds. Plus précisément, l'arête est absente

entre deux nœuds, si et seulement si les deux variables correspondantes sont indépendantes conditionnellement aux autres variables. Dans la situation des accidentés de la route, chaque lésion peut être modélisée par une variable aléatoire binaire qui équivaut à 1 si la victime souffre de cette lésion et 0 sinon. La description du tableau lésionnel des victimes revient ensuite à décrire la distribution conjointe des p variables binaires, où p est la cardinalité de l'ensemble des lésions possibles. L'identification de la structure du modèle graphique binaire se réduit à la détermination de paramètres $(\theta^*_{j,\ell})_{1\leq j<\ell\leq p}$ non-nuls dans le modèle d'Ising (1.1), puis à un problème de sélection de modèle. Cependant, lorsque p est supérieur à 20 ou 30 variables, l'inférence sous le modèle d'Ising est difficile à cause de la complexité de la fonction de log-partition (1.2). En particulier, l'estimation du maximum de vraisemblance ne peut généralement pas être effectuée. Diverses solutions sont apparues dans la littérature. Wainwright et al. (2007) ont proposé d'utiliser les modèles de régressions logistiques pénalisées par la norme ℓ_1 multiple, étendant l'approche développée par Meinshausen and Bühlmann (2006) dans le cas gaussien. Suivant la terminologie adoptée dans Wang et al. (2009), nous allons nous référer à cette approche comme SepLogit. Höfling and Tibshirani (2009) ont considéré une variante basée sur la pseudo vraisemblance pénalisée par la norme ℓ_1. Banerjee et al. (2008) ont dérivé une approximation gaussienne de la log-vraisemblance d'Ising, tandis que Yang and Ravikumar (2011) ont utilisé une inférence variationnelle basée sur des approximations alternatives de la fonction de log-partition. Ces approches ont été comparées empiriquement dans Viallon et al. (2014). Selon les designs qu'ils ont étudié, toutes ces méthodes ont donné des résultats similaires, et raisonnablement bons, pour identifier la structure d'un modèle graphique binaire simple.

En particulier, l'approche SepLogit repose sur l'observation suivante (Viallon et al., 2014). Pour tout vecteur $\mathbf{u} \in \{0,1\}^p$ et tout $j \in [p]$, soit $\mathbf{u}_{-j} \in \{0,1\}^{p-1}$ le vecteur correspondant au vecteur $\mathbf{u}$ après élimination de la j-ème composante. Sous le modèle (1.1), on a pour tout $j \in [p]$,

$$\text{logit}\{\mathbb{P}_{\boldsymbol{\theta}^*}(U_j = 1|\mathbf{U}_{-j} = \mathbf{u}_{-j})\} = \theta^*_j + \sum_{\ell\neq j} \theta^*_{j,\ell} u_\ell. \tag{1.17}$$

Les paramètres peuvent donc être estimés via p régressions logistiques. Dans les modèles d'Ising en général, tout comme dans leur application sur les tableaux lésionnels, les paramètres $\theta^*_{j,\ell}$ pour $j \neq \ell$ revêtent un intérêt particulier. Ils correspondent aux log odds-ratios conditionnels entre la lésion j et la lésion ℓ. Ils indiquent ainsi si ces deux

lésions sont préférentiellement associées dans les tableaux lésionnels. Leur connaissance permet une représentation graphique des associations entre lésions à travers un modèle graphique, qui décrit la loi de probabilité du vecteur $\mathbf{U}$. Soit $\mathcal{G} = (V, E)$ le graphe correspondant, où V est l'ensemble des p sommets correspondant aux p composantes de $\mathbf{U}$ (les p lésions) et l'ensemble d'arêtes $E \subseteq \{(j, \ell) \in V^2 : j < \ell\}$ décrit les relations d'indépendance conditionnelle parmi ces composantes. Plus précisément, l'arête (j, ℓ) entre les lésions U_j et U_ℓ est absente si et seulement si U_j et U_ℓ sont indépendantes conditionnellement aux autres variables, et donc si et seulement si $\theta^*_{j,\ell} = 0$.

Sous l'approche SepLogit, la sélection des paramètres $\theta^*_{j,\ell}$ non nuls peut être effectuée par des régressions logistiques lasso, c'est-à-dire pénalisées par la norme ℓ_1 des paramètres. Plus formellement, en notant $L_j(\boldsymbol{\theta}_j)$ la log-vraisemblance sous le modèle de régression logistique (1.17) pour la j-ème variable, avec $\boldsymbol{\theta}_j = (\theta_j, \boldsymbol{\theta}_{j,\ell}) \in \mathbb{R}^p$, l'estimation des paramètres $\boldsymbol{\theta}^*_j = (\theta^*_j, (\boldsymbol{\theta}^*_{j,\ell})) \in \mathbb{R}^p$ du modèle (1.1), correspondant à la j-ème lésion, est obtenue par minimisation de la fonction suivante :

$$- L_j(\boldsymbol{\theta}_j) + \lambda \|\boldsymbol{\theta}_j\|_1, \tag{1.18}$$

pour une valeur du paramètre de régularisation $\lambda \geq 0$ appropriée. En ce qui concerne l'estimation de modèles graphiques sur des données stratifiées, on se retrouve alors face à un problème d'estimation de K modèles graphiques, où K est le nombre de strates. Quelques approches ont été proposées, principalement pour les modèles graphiques gaussiens. Par exemple, Danaher et al. (2014) ont développé une approche basée sur une pénalité fused lasso. Hallac et al. (2017) ont développé une approche similaire dans le cas particulier où les strates correspondent aux timestamps. Pour les modèles graphiques binaires, Guo et al. (2015) ont basé leur approche sur la pseudo-vraisemblance et sur une décomposition multiplicative des coefficients. Toutefois, comme nous le verrons plus loin, cette approche n'est pas adaptée à l'identification des hétérogénéités qui peuvent exister entre les structures de dépendance conditionnelle des différentes strates. Afin d'identifier les différences entre ces K modèles, on peut adapter l'approche SepLogit lasso (1.18) ci-dessus pour encourager les paramètres des K modèles graphiques à être proches les uns des autres. De cette façon, nous pourrons plus facilement interpréter les différences détectées, le cas échéant. Pour ce faire, on peut reprendre les idées de l'approche DataShared Lasso (Ollier and Viallon, 2017) ou encore du fused lasso généralisé (Viallon et al., 2016). L'extension

de DataShared Lasso et Fused Lasso dans ce cadre est un de mes projets présenté au chapitre 2.

1.4.2 Etude cas-temoins en présence de sous-types de maladie

L'analyse d'études cas-témoins avec plusieurs sous-types de cas est de plus en plus courante, par exemple, en épidémiologie du cancer. Pour de nombreuses maladies, principalement considérées auparavant comme une seule maladie (cancer du sein, cancer colorectal), plusieurs sous-types ont maintenant été reconnus. Ils peuvent être histologiques, comme pour le cancer du sein, ou anatomiques, comme pour le cancer colorectal. Même si des points communs peuvent exister entre ces sous-types, ils ont leurs propres spécificités en ce qui concerne à la fois le pronostic et l'étiologie. Par exemple, les épidémiologistes du cancer s'intéressent de plus en plus à l'identification de facteurs de risque spécifiques aux sous-type pour divers sites de cancer.

Pour les études cas-témoins non-appariées avec plusieurs sous-types de cas, le modèle de régression logistique multinomial (McCullagh and Nelder, 1989; Begg and Gray, 1984) est une extension naturelle du modèle de régression logistique binaire . Si $K-1$ désigne le nombre de sous-types pour un entier donné $K > 1$, l'inférence sous ce modèle consiste à estimer les $K-1$ vecteurs de paramètre de taille p, où p désigne le nombre de covariables (qui peut inclure des interactions, ainsi que le terme d'intercept). Par ailleurs, pour les études cas-contrôle appariées avec les $K-1$ sous-types de cas, l'échantillon total peut être décomposé en $K-1$ sous-échantillons, un pour chaque sous-type. En supposant, pour des raisons de simplicité, un plan d'appariement 1 :1, chaque sous-échantillon est constitué de paires composées d'un cas pour un sous-type particulier, et d'un contrôle apparié. Ensuite, chaque sous-échantillon peut être analysé séparément, par exemple en appliquant un modèle de régression logistique conditionnelle sparse (Avalos et al., 2015). Encore une fois, l'analyse globale se résume à l'estimation de $K-1$ vecteurs de paramètre de taille p.

Comme il existe des points communs entre les sous-types de cas, un certain niveau d'homogénéité est attendu entre ces $K-1$ vecteurs de paramètre, à la fois dans les cas appariés et non appariés. Tenir correctement compte de cette homogénéité est essentiel pour réduire la dimensionnalité et améliorer l'efficacité de l'estimation. Dans le cas apparié, les $K-1$ modèles de régression logistique conditionnelle sparse doivent être estimés sur des $K-1$ sous-groupes, ces sous-groupes étant définis en fonction du sous-type de

cas de chaque paire.

Des extensions de DataShared Lasso sur les analyses des études cas-témoins lorsque plusieurs sous-types de cas sont présents constituent la deuxième partie de ce travail de thèse, et notamment sur les données portant sur l'étude du cancer du sein, dans laquelle nous sommes en présence de plusieurs sous-types de cas (différents sous-types de cancer du sein). Ainsi, les sous-types de cas forment des strates. Pour répondre à l'étude de cette situation - ou à n'importe quelle étude cas-témoins comportant des données stratifiées – l'extension de DataShared Lasso est basée sur des modèles de régression logistique conditionnelle.

Une autre extension de DataShared Lasso sur le cas des données non-appariées est basée sur des modèles de régressions logistiques multinomiales sparse. En effet, deux formulations de modèles de régression logistique multinomiale sparse ont été proposées dans la littérature. Une première, que nous appellerons standard, repose sur la sélection d'une catégorie de référence et sur l'estimation de $(K-1)$ paramètre vecteurs (Begg and Gray, 1984). De manière alternative, nous pouvons également adopter une formulation plus symétrique du modèle, dans laquelle aucune catégorie de référence ne doit être sélectionnée et les K vecteurs de paramètre doivent être estimés (Friedman et al., 2010). L'estimation non pénalisée est impossible sous ce modèle surparamétré à cause d'un manque évident d'identifiabilité. Cependant, une estimation pénalisée par la norme ℓ_1 peut être réalisée, telle que implémentée dans le très répandu package glmnet sur R. À notre connaissance, il n'existe dans la littérature aucune recommandation claire sur la manière de choisir entre les deux formulations de modèles de régression logistique multinomiale sparse. Nous établirons formellement que les stratégies pénalisées par la norme ℓ_1 associées aux deux formulations diffèrent par la manière dont elles rendent compte de l'homogénéité potentielle parmi les vecteurs de paramètres à estimer. Plus précisément, nous montrons que les estimations pénalisées par la norme ℓ_1 dérivées sous la formulation symétrique coïncident avec les estimations dérivées sous la formulation standard lors de l'utilisation d'une pénalité DataShared Lasso.

Chapitre 2

Structure estimation of binary graphical models on stratified data : application to the description of injury tables for victims of road accidents.

Ballout N, Viallon V. Structure estimation of binary graphical models on stratified data : Application to the description of injury tables for victims of road accidents. Stat Med. 2019 ;38(14) :2680-2703. doi :10.1002/sim.8138. Ballout and Viallon (2019).

Abstract

Graphical models are used in many applications such as medical diagnostics, computer security, etc. Increasingly often, the estimation of such models has to be performed on several predefined strata of the whole population. For instance, in epidemiology and clinical research, strata are often defined according to age, gender, treatment or disease type, etc. In this article, we propose new approaches dedicated to the estimation of binary graphical models on such strata. These approaches are implemented by combining well-known methods that have been developed in the context of a single binary graphical model, with penalties encouraging structured sparsity, that have recently been shown to be appropriate when dealing with stratified data. Empirical comparisons on synthetic data highlight that our approaches generally outperform its competitors. We present an application of the approach to study associations among the injuries suffered by victims of road accidents according to road user type.

Key words :Graphical models ; Ising models ; Multiple logistic regressions ; Penalization ; Stratified analysis ; Structured sparsity.

2.1 Introduction

In this article, we consider the estimation of the conditional dependence structure among a set of binary variables across several predefined sub-groups, or strata, of a population. As an illustration, we will consider the description of the injury tables suffered by victims of road accidents. This description is key to the quantification of the needs in terms of care services and thus, in a long term perspective, to improve the care of the victims. Fine description of the associations among injuries could further turn out to be useful for diagnostic purposes : if a given external, hence easy to diagnose, injury is strongly positively associated with some internal and harder to diagnose injury, then special attention would be given to a victim suffering from the external injury as this victim would be more likely to suffer from the internal one as well. For all these reasons, clinicians ask for statistical tools that can accurately summarize injury tables of victims, as well as the associations among injuries. Of course, the characteristics of the accident play an important role in the set of injuries suffered by the victims. In particular, injury tables are likely to vary according to the road user type of the victim : pedestrian, car occupant, motorized two-wheeler user, cyclist, etc. Therefore, associations among injuries have to be studied according to road user types, that is, across several predefined strata of the population of victims.

Graphical models have proven valuable when modeling the joint distribution of p given variables, and they provide a graphical representation of the conditional dependences between them (Lauritzen, 1996). A graphical model is a non-directed graph that consists of a set of p nodes corresponding to the p variables, along with a set of edges joining some pairs of nodes. More precisely, an edge is absent between two nodes if and only if the two corresponding variables are conditionally independent given the other variables. Regarding the application we have in mind, each injury can be modeled by a binary random variable that equals 1 if the victim suffers from this injury and 0 otherwise. The description of the injury tables of victims then reduces to the description of the joint distribution of p binary variables, where p is the cardinality of the set of all possible injuries. When working with binary variables, the quadratic exponential binary model, also known as the Ising model, is commonly used (Besag, 1974; Cox and Wermuth, 1994). Identifying the structure of the binary graphical model reduces to the determination of non-null parameters in the Ising model (see Section 2.2.1 below for details), and then to a model selection problem. However, when p is larger than 20 or 30 variables, inference

under the Ising model is challenging because of the intractability of the log-partition function. In particular, maximum likelihood estimation can generally not be performed. Various solutions have arisen in the literature. Wainwright et al. (2007) proposed to use multiple ℓ_1-penalized logistic regressions, extending the approach developed by Meinshausen and Bühlmann (2006) in the Gaussian case. Following the terminology adopted in Wang et al. (2009), we will refer to this approach as SepLogit. Höfling and Tibshirani (2009) considered a variant based on ℓ_1-penalized pseudo-likelihood. Banerjee et al. (2008) derived a Gaussian approximation of the Ising log-likelihood, while Yang and Ravikumar (2011) used variational inference based on alternative approximations of the log-partition function. These approaches have been empirically compared in Viallon et al. (2014). Under the designs they considered, all these methods performed similarly, and reasonably well.

When models have to be estimated on several predefined strata of a population, the general objective is to take advantage of the potential homogeneity among the corresponding models, while not masking the heterogeneities. Several authors have studied the estimation of regression models on such stratified data. Viallon et al. (2016) as well as (Gertheiss and Tutz, 2010, 2012) studied generalized fused lasso estimates under generalized linear models. Gross and Tibshirani (2016) and Ollier and Viallon (2017) independently developed an alternative approach referred to as DataShared Lasso. It can be seen as an extension of the common strategy which consists in first selecting a reference stratum and then adding interaction terms between each covariate and the indicators of the remaining strata. By considering an appropriate overparametrization, DataShared Lasso bypasses the arbitrary choice of the reference stratum and mimics the strategy based on an optimal and covariate-specific choice of the reference stratum. Ollier and Viallon (2017) established the sparsistency of DataShared Lasso under some technical assumptions in the case of linear regression models. In particular, for each covariate, DataShared Lasso is able to identify the strata on which the parameters differ from those of the optimal reference stratum under nearly the same assumptions as those required by the optimal (and infeasible in practice) strategy. From a practical perspective, DataShared Lasso can be written as a weighted lasso on a simple transformation of the original data. It is therefore easy to implement under a variety of regression models, since very efficient lasso solvers are now available under many regression models : for instance the glmnet package is now available in R, Matlab and Python and uses several tricks to make the implementation extremely fast, such as strong rules for discarding predictors (Tibshirani

et al., 2012; El Ghaoui et al., 2012).

As for the estimation of graphical models on stratified data, a few approaches have been proposed, mostly for Gaussian graphical models. For example, Danaher et al. (2014) developed an approach based on a fused lasso penalty. Hallac et al. (2017) developed a similar approach in the special case where strata correspond to timestamps. For binary graphical models, Guo et al. (2015) based their approach on the pseudo-likelihood and a multiplicative decomposition of the coefficients. However, as will be made clearer below, this approach is not tailored for the identification of the heterogeneities that may exist between the conditional dependence structures of the different strata. In this article, we propose two versions of the approach of Danaher et al. (2014) which was originally developed for Gaussian graphical models. Our approaches combine the SepLogit method with either a generalized fused lasso penalty or, even more simply, a DataShared Lasso penalty. In Section 2.2, we briefly recall some basics about the Ising model and the SepLogit method. Then, we describe our proposals and formally explain why they are better suited for the identification of heterogeneities than other approaches, including Guo et al's approach. In Section 3.4, we present results from an empirical study, which establish that our approaches outperform its competitors under the settings we consider. Our empirical results further suggest that our two proposals generally perform similarly, while the version based on DataShared Lasso being much cheaper regarding computational time. Section 2.4 presents the application of our approaches to describe injury tables on a French registry of victims of road accidents. Possible extensions are discussed in Section 2.5.

2.2 Methods

2.2.1 The Ising Model

The injury table of a victim can be modeled as a realization of the random variable $\mathbf{U} = (U_1, ..., U_p)^T \in \{0,1\}^p$ where U_j indicates the presence of the injury j in the considered injury table and p is the cardinality of the set of all possible injuries. The description of the injury tables then reduces to the estimation of the joint distribution of $\mathbf{U}$, given an n-sample $\mathbf{U}_1, ..., \mathbf{U}_n$ of independent and identically distributed (i.i.d.) replicas of $\mathbf{U}$. In this context, a commonly used model is the quadratic exponential

binary model, also known as the Ising model (Cox and Wermuth, 1994; Höfling and Tibshirani, 2009; Banerjee et al., 2008; Ravikumar et al., 2010). The Ising model assumes the existence of a parameter vector $\boldsymbol{\Theta}^* = ((\theta^*_j)_{1\leq j\leq p}, (\theta^*_{j,\ell})_{1\leq j<\ell\leq p})^T$ in $\mathbb{R}^{p(p+1)/2}$ such that for any vector $\mathbf{u} = (u_1, ..., u_p) \in \{0,1\}^p$, the probability of the event $\{\mathbf{U} = \mathbf{u}\}$ is given by

$$\mathbb{P}_{\boldsymbol{\Theta}^*}(\mathbf{U} = \mathbf{u}) = \exp\Big\{ \sum_{j=1}^{p} \theta^*_j u_j + \sum_{j=1}^{p-1} \sum_{\ell=j+1}^{p} \theta^*_{j,\ell} u_j u_\ell - A(\boldsymbol{\Theta}^*) \Big\}. \tag{2.1}$$

In other words, probabilities $\mathbb{P}_{\boldsymbol{\Theta}^*}(\mathbf{U} = \mathbf{u})$ are assumed to depend on main effects and first-order interactions only, and the Ising model can be seen as a special case of the more general model presented in Section A.2 in the Appendix.

The so-called log *partition function* $A : \mathbb{R}^p \to \mathbb{R}$, is a normalization term ensuring that $\sum_{\mathbf{u}\in\{0,1\}^p} \mathbb{P}_{\boldsymbol{\Theta}}(\mathbf{U} = \mathbf{u}) = 1$ for every $\boldsymbol{\Theta} \in \mathbb{R}^{p(p+1)/2}$, and is defined as

$$A(\boldsymbol{\Theta}) = \log \sum_{\mathbf{u}\in\{0,1\}^p} \exp\left(\sum_{j=1}^{p} \theta_j u_j + \sum_{j=1}^{p-1} \sum_{\ell=j+1}^{p} \theta_{j,\ell} u_j u_\ell \right). \tag{2.2}$$

For every $j > \ell$, let $\theta^*_{j,\ell} = \theta^*_{\ell,j}$. For every $\mathbf{u} = (u_1, ..., u_p) \in \{0,1\}^p$ and every $j \in [p]$, where $[p]$ is the set of values $\{1, ..., p\}$, further denote by $\mathbf{u}_{-j} = (u_1, ..., u_{j-1}, u_{j+1}, ..., u_p)^T \in \{0,1\}^{p-1}$ the vector obtained after the elimination of the jth component of $\mathbf{u}$. Under (2.1), we have, for every $j \in [p]$,

$$\text{logit}\{\mathbb{P}_{\boldsymbol{\Theta}^*}(U_j = 1 | \mathbf{U}_{-j} = \mathbf{u}_{-j})\} = \theta^*_j + \sum_{\ell\neq j} \theta^*_{j,\ell} u_\ell = \theta^*_j + \boldsymbol{\theta}^{*T}_{-j} \mathbf{u}_{-j}, \tag{2.3}$$

with $\boldsymbol{\theta}^*_j = (\theta^*_j, \theta^*_{j,1}, .., \theta^*_{j,j-1}, \theta^*_{j,j+1}, ..., \theta^*_{j,p})^T$ and $\boldsymbol{\theta}^*_{-j}$ corresponds to the vector $\boldsymbol{\theta}^*_j$ after the elimination of the constant term θ^*_j. Therefore, parameters $\theta^*_{j,\ell}$ correspond to conditional log odds-ratios and conditional independence between U_j and U_ℓ is equivalent to $\theta^*_{j,\ell} = 0$. The Ising model is naturally associated to a graphical model, that is a non-directed graph $\mathcal{G} = (V, E)$. The p vertices of set V correspond to the p components of $\mathbf{U}$. The set of edges $E \subseteq \{(j,\ell) \in V^2 : j < \ell\}$ describes the conditional independence relationships among these components. More precisely, the edge (j,ℓ) between nodes j and ℓ is absent if and only if U_j and U_ℓ are independent conditionally on the other variables, that is if and only if $\theta^*_{j,\ell} = 0$. Therefore, the identification of the edge set, or the structure of the graph, reduces to the identification of the zeros in the vector $\boldsymbol{\Theta}^*$. However, the estimation of $\boldsymbol{\Theta}^*$ and the selection of the non-zero elements of $\boldsymbol{\Theta}^*$ under

the Ising model is not straightforward because of the form of the log-partition function. Defined as a sum over 2^p terms, this function can not be computed in a reasonable time for $p \geq 20$ so, for instance, maximum likelihood estimation can not be performed. One popular strategy to get around this problem has been proposed by Wainwright, (Wainwright et al., 2007) and will be referred to as SepLogit in the sequel. From (2.3), parameters of model (2.1) can be estimated through p logistic regression models. More precisely, denote by $L(\boldsymbol{\theta}_j; \boldsymbol{\mathcal{U}}_j, \boldsymbol{\mathcal{U}}_{-j})$ the log-likelihood under the logistic regression model (2.3). Here $\boldsymbol{\mathcal{U}}_j = (U_{j,1}, \ldots, U_{j,n}) \in \mathbb{R}^n$ contains the n observations of variable U_j, $\boldsymbol{\mathcal{U}}_{-j} \in \mathbb{R}^{n \times (p-1)}$ is the matrix containing the observations of the $p-1$ remaining variables, and $\boldsymbol{\theta}_j^T = (\theta_j, \boldsymbol{\theta}_{-j}^T) = (\theta_j, \theta_{j,1}, ..., \theta_{j,j-1}, \theta_{j,j+1}, ..., \theta_{j,p}) \in \mathbb{R}^p$ is the vector of parameters, over which optimization is performed to return (penalized) maximum likelihood estimates. Under SepLogit, estimation of $\boldsymbol{\theta}_j^* = (\theta_j^*, \boldsymbol{\theta}_{-j}^{*T})^T$ and selection of the non-zero values in $\boldsymbol{\theta}_{-j}^*$ are both achieved by minimizing the following lasso criterion, for an appropriate value of the regularization parameter $\lambda_j \geq 0$,

$$- L(\boldsymbol{\theta}_j; \boldsymbol{\mathcal{U}}_j, \boldsymbol{\mathcal{U}}_{-j}) + \lambda_j \|\boldsymbol{\theta}_{-j}\|_1. \tag{2.4}$$

Here, $\|\boldsymbol{\theta}_{-j}\|_1 = \sum_{\ell \neq j} |\theta_{j,\ell}|$ is the L_1-norm of $\boldsymbol{\theta}_{-j}$. For every $j = 1, ..., p$, we denote by $\hat{\boldsymbol{\theta}}_j$ any solution minimizing criterion (2.4). From these p vectors two estimates of the parameter $\theta_{j,\ell}^*$ are obtained for every $(j, \ell) \in [p]^2$: $\hat{\theta}_{j,\ell}$ from the vector $\hat{\boldsymbol{\theta}}_j$ and $\hat{\theta}_{\ell,j}$ from the vector $\hat{\boldsymbol{\theta}}_\ell$, with $\hat{\theta}_{j,\ell} \neq \hat{\theta}_{\ell,j}$ in general. Of course, it is even possible that $\hat{\theta}_{j,\ell} = 0$ and $\hat{\theta}_{\ell,j} \neq 0$ for example. This asymmetry issue is resolved by considering either the SepLogit AND or the SepLogit OR strategy. According to the SepLogit AND strategy the edge (j, ℓ) is present in the edge set E if both $\hat{\theta}_{j,\ell}$ and $\hat{\theta}_{\ell,j}$ are non-zero. According to SepLogit OR the edge (j, ℓ) is present in the set E if either $\hat{\theta}_{j,\ell}$ or $\hat{\theta}_{\ell,j}$ is non-zero.

2.2.2 Estimation of binary graphical models on K strata : "standard" approaches

In our illustrating example, clinicians expect the injury tables to depend on the road user type (car occupants, pedestrians, ...). Therefore, associations among injuries suffered by victims of road accidents should be studied according to road user type. In other words, K binary graphical models have to be estimated, where K is the number of road user types. In this context, we assume the existence of K vectors $\boldsymbol{\Theta}^{(k)*} =$

$((\theta_j^{(k)*})_{1\leq j\leq p}, (\theta_{j,\ell}^{(k)*})_{1\leq j<\ell\leq p})^T$ in $\mathbb{R}^{p(p+1)/2}$, for $k = 1, \ldots, K$, such that the probability of observing the injury table $\{\mathbf{U} = \mathbf{u}\}$ in the k-th stratum is given by

$$\mathbb{P}_{\mathbf{\Theta}^{(k)*}}(\mathbf{U} = \mathbf{u}) = \exp\Big\{ \sum_{j=1}^{p} \theta_j^{(k)*} u_j + \sum_{j=1}^{p-1} \sum_{\ell=j+1}^{p} \theta_{j,\ell}^{(k)*} u_j u_\ell - A(\mathbf{\Theta}^{(k)*}) \Big\}. \tag{2.5}$$

Of course, vectors $(\mathbf{\Theta}^{(k)*})_{k\in[K]}$ can be estimated separately, by applying the SepLogit method on each stratum independently. More precisely, set $\boldsymbol{\theta}_j^{(k)*} = (\theta_j^{(k)*}, \boldsymbol{\theta}_{-j}^{(k)*T})^T = (\theta_j^{(k)*}, \theta_{j,1}^{(k)*}, .., \theta_{j,j-1}^{(k)*}, \theta_{j,j+1}^{(k)*}, ..., \theta_{j,p}^{(k)*})^T \in \mathbb{R}^p$. Estimates of these vectors returned by what we will refer to as Indep-SepLogit are defined as minimizers of the following criterion, for appropriate tuning parameters $\lambda_j^{(k)} \geq 0$,

$$\sum_{k=1}^{K} \Big(-L(\boldsymbol{\theta}_j^{(k)}; \boldsymbol{\mathcal{U}}_j^{(k)}, \boldsymbol{\mathcal{U}}_{-j}^{(k)}) + \lambda_j^{(k)} \|\boldsymbol{\theta}_{-j}^{(k)}\|_1 \Big), \tag{2.6}$$

where $\boldsymbol{\mathcal{U}}_j^{(k)} \in \mathbb{R}^{n_k}$ and $\boldsymbol{\mathcal{U}}_{-j}^{(k)} \in \mathbb{R}^{n_k \times (p-1)}$ contain the observations of variable j and the $p-1$ remaining variables respectively, for the victims belonging to stratum k. Here n_k is the number of observations belonging to the k-th stratum. However, Indep-SepLogit does not account for the potential homogeneity among the vectors $\mathbf{\Theta}^{(k)*}$, $k \in [K]$. Indeed, even if associations between injuries may vary according to road user type, we still expect that $\theta_{j,\ell}^{(k_1)*} = \theta_{j,\ell}^{(k_2)*}$ for some $(k_1, k_2) \in [K]^2$ and some $(j, \ell) \in [p]^2$. By not accounting for this expected homogeneity, Indep-SepLogit has two undesirable properties. First, the returned estimates are of unnecessarily high dimension and so typically have poor performance. Second, when homogeneity is expected, the identification of heterogeneities is generally of particular interest. In our example for instance, clinicians are interested in identifying which associations of injuries are more likely for each road user type ; automobile manufacturers may further be interested in the associations of injuries that are more likely for car occupants, etc. However, differences between estimates $\hat{\theta}_{j,\ell}^{(k_1)}$ and $\hat{\theta}_{j,\ell}^{(k_2)}$ returned by Indep-SepLogit can not be interpreted as true differences, since $\hat{\theta}_{j,\ell}^{(k_1)}$ and $\hat{\theta}_{j,\ell}^{(k_2)}$ are different by construction, as soon as these two quantities are non-null.

Another standard strategy in epidemiology consists in first selecting a reference stratum m, *a priori*, and then decomposing the parameters of (2.6) according to the equation $\boldsymbol{\theta}_j^{(k)*} = \boldsymbol{\theta}_j^{(m)*} + \boldsymbol{\gamma}_j^{(k)*}$, for all $k \in [K]$, with $\boldsymbol{\gamma}_j^{(m)*} = \mathbf{0}_p$ (Gertheiss and Tutz, 2012). More precisely, estimates of $\boldsymbol{\theta}_j^{(m)*}$ and the $\boldsymbol{\gamma}_j^{(k)*}$'s returned by what we will refer to as Ref-SepLogit are defined as minimizers of the following criterion, for appropriate values

of the tuning parameters $\lambda_{j,1} \geq 0$ and $\lambda_{j,2}^{(k)} \geq 0$,

$$\sum_{k=1}^{K} -L((\boldsymbol{\theta}_j^{(m)} + \boldsymbol{\gamma}_j^{(k)}); \boldsymbol{\mathcal{U}}_j^{(k)}, (\boldsymbol{\mathcal{U}}_{-j}^{(k)})) + \lambda_{j,1} \|\boldsymbol{\theta}_{-j}^{(m)}\|_1 + \sum_{k \neq m} \lambda_{j,2}^{(k)} \|\boldsymbol{\gamma}_j^{(k)}\|_1, \tag{2.7}$$

where $\boldsymbol{\theta}_{-j}^{(m)}$ corresponds to the vector $\boldsymbol{\theta}_j^{(m)}$ after the elimination of the constant term $\theta_j^{(m)}$. Here, we will use the particular value $\lambda_{j,2}^{(k)} = \lambda_{j,2}\omega_k$ with $\omega_k = \sqrt{K(n_k/n)}$, as suggested in Ollier and Viallon (2017). Ref-SepLogit can be seen as a special case of the approach proposed by Cheng et al. (2014) when covariates reduce to indicators of the strata. In equation (2.7), the penalty term $\|\boldsymbol{\theta}_{-j}^{(m)}\|_1$ reflects the expected sparsity in the vector $\boldsymbol{\theta}_j^{(m)*}$, while the penalty term $\sum_{k \neq m} \lambda_{j,2}^{(k)} \|\boldsymbol{\gamma}_j^{(k)}\|_1 = \sum_{k \neq m} \lambda_{j,2}^{(k)} |\boldsymbol{\theta}_j^{(k)} - \boldsymbol{\theta}_j^{(m)}\|_1$ reflects the expected similarity between vectors $\boldsymbol{\theta}_j^{(m)*}$ and $\boldsymbol{\theta}_j^{(k)*}$, for all k. More precisely, the latter penalty shrinks estimates $\hat{\boldsymbol{\theta}}_j^{(k)}$ towards $\hat{\boldsymbol{\theta}}_j^{(m)}$, and encourages equality of these two vectors. However, only the differences between components of $\hat{\boldsymbol{\theta}}_j^{(k)}$ and $\hat{\boldsymbol{\theta}}_j^{(m)}$ can be interpreted as true differences. More precisely, for all $k_1 \neq m$ and $k_2 \neq m$ differences between estimates $\hat{\theta}_{j,\ell}^{(k_1)}$ and $\hat{\theta}_{j,\ell}^{(k_2)}$ returned by Ref-SepLogit can not be interpreted as true differences since their difference is not penalized, and their values are different by construction as long as they are different from $\hat{\theta}_{j,\ell}^{(m)}$. In order to account for the expected homogeneity among vectors $(\boldsymbol{\Theta}^{(k)*})_{k \in [K]}$, Guo et al. (2015) proposed an alternative approach based on the multiplicative decomposition $\theta_{j,\ell}^{(k)*} = \phi_{j,\ell}^* \gamma_{j,\ell}^{(k)*}$. Here, for all $j < \ell$, $\phi_{j,\ell}^*$ is common to all K strata and controls the occurrence of common links shared across strata, while $\gamma_{j,\ell}^{(k)*}$ is an individual factor specific to the k-th stratum, $k \in [K]$. The approach proposed by Guo et al. (2015) relies on the use of the pseudo-likelihood, which is another solution to get around the asymmetry issue of SepLogit. Moreover, the domain of parameters $\phi_{j,\ell}$ is restricted to $\mathbb{R}_+$ to avoid sign ambiguities between $\phi_{j,\ell}$ and $\gamma_{j,\ell}^{(k)}$. More importantly, their approach relies on a penalty of the form $\eta_1 \sum_{j<\ell} \phi_{j,\ell} + \eta_2 \sum_{j<\ell} \sum_{k=1}^{K} |\gamma_{j,\ell}^{(k)}|$, as proposed by Zhou and Zhu (2010) under linear regression models. Keeping in mind that $\phi_{j,\ell} \geq 0$, the first term of the penalty shrinks estimates of $\phi_{j,\ell}^*$ towards 0, while the second term shrinks estimates of $\gamma_{j,\ell}^{(k)*}$ toward 0. Therefore, the way this penalty couples the estimations across the K strata is as follows : if $\hat{\phi}_{j,\ell} = 0$ then $\hat{\theta}_{j,\ell}^{(k)} = 0$ for all $k \in [K]$, and hence there is no link between nodes j and ℓ in any of the K graphs. On the other hand, if $\hat{\phi}_{j,\ell} \neq 0$, then some of the $\hat{\gamma}_{j,\ell}^{(k)}$ and hence some of the $\hat{\theta}_{j,\ell}^{(k)}$ still have the possibility of being zero, for some $k \in [K]$. However, this approach is not well suited for the application we have in mind. Indeed, differences between non-zero estimates $\hat{\theta}_{j,\ell}^{(k_1)}$

and $\hat{\theta}_{j,\ell}^{(k_2)}$ cannot be interpreted as true differences since $\hat{\theta}_{j,\ell}^{(k_1)}$ and $\hat{\theta}_{j,\ell}^{(k_2)}$ are different by construction, as long as they are both non-null.

2.2.3 Joint estimation of binary graphical models on K strata : our proposal

To fully account for the expected homogeneity across the K graphs and be able to interpret differences between estimates of, say, $\hat{\theta}_{j,\ell}^{(k_1)}$ and $\hat{\theta}_{j,\ell}^{(k_2)}$, our proposal relies on the combination of SepLogit and penalties used in the context of regression modeling on stratified data. In the following two paragraphs, the principle of the estimation of the vectors $(\boldsymbol{\theta}_j^{(k)*})_{k\in[K]}$, for a given $j \in [p]$ is presented, using either a fused penalty or a DataShared Lasso penalty (Gross and Tibshirani, 2016; Ollier and Viallon, 2017). Next, we propose two strategies to combine the estimates obtained for all $j \in [p]$, extending the ideas of SepLogit AND and SepLogit OR described above.

2.2.3.1 Fused-SepLogit.

Our first proposal follows the ideas developed in Danaher et al. (2014) under Gaussian graphical models as well as those of Gertheiss and Tutz (2010, 2012) and Viallon et al. (2016) under generalized linear regression models. It relies on a generalized fused lasso penalty. More precisely, for all $j \in [p]$, the method we will refer to as Fused-SepLogit returns estimates of $\boldsymbol{\theta}_j^{(k)*}$, for $k \in [K]$, defined as minimizers of the following criterion, for appropriate values of the tuning parameters $\lambda_{j,1} \geq 0$ and $\lambda_{j,2} \geq 0$,

$$\sum_{k=1}^{K} \Big(-L(\boldsymbol{\theta}_j^{(k)}; \boldsymbol{\mathcal{U}}_j^{(k)}, \boldsymbol{\mathcal{U}}_{-j}^{(k)}) + \lambda_{j,1} \|\boldsymbol{\theta}_j^{(k)}\|_1 \Big) + \lambda_{j,2} \sum_{k_1<k_2} \|\boldsymbol{\theta}_j^{(k_1)} - \boldsymbol{\theta}_j^{(k_2)}\|_1. \tag{2.8}$$

The fused-like penalty term $\|\boldsymbol{\theta}_j^{(k_1)} - \boldsymbol{\theta}_j^{(k_2)}\|_1$ shrinks estimates $\hat{\boldsymbol{\theta}}_j^{(k_1)}$ and $\hat{\boldsymbol{\theta}}_j^{(k_2)}$ towards each other, and therefore encourages equality of these two vectors. Accordingly, differences between components of $\hat{\boldsymbol{\theta}}_j^{(k_1)}$ and $\hat{\boldsymbol{\theta}}_j^{(k_2)}$ can be interpreted as true differences, and the expected homogeneity is likely to be fully accounted for. More precisely, estimates derived from the adaptive version of (2.8) have been shown to enjoy an asymptotic oracle property, in the fixed Kp case ; see Gertheiss and Tutz (2012) and Viallon et al. (2016)).

2.2.3.2 DataShared-SepLogit.

Our second proposal can be seen as an improved version of Ref-SepLogit. It consists in extending a method that was independently developed in Gross and Tibshirani (2016) and Ollier and Viallon (2017) under generalized linear models (see also Ollier and Viallon, 2014). Applied to our context, it first relies on the following additive decomposition of $\boldsymbol{\theta}_j^{(k)*}$

$$\boldsymbol{\theta}_j^{(k)*} = \boldsymbol{\mu}_j^* + \boldsymbol{\gamma}_j^{(k)*}, \quad \text{for each } k \in [K] \tag{2.9}$$

where $\boldsymbol{\mu}_j^* = (\mu_j^*, \boldsymbol{\mu}_{-j}^{*T})^T = (\mu_j^*, \mu_{j,1}^*, ..., \mu_{j,j-1}^*, \mu_{j,j+1}^*, ..., \mu_{j,p}^*)^T \in \mathbb{R}^p$ "morally" contains what is common between the K strata, while $\boldsymbol{\gamma}_j^{(k)*} = (\gamma_j^{(k)*}, \gamma_{j,1}^{(k)*}, ..., \gamma_{j,j-1}^{(k)*}, \gamma_{j,j+1}^{(k)*}, ..., \gamma_{j,p}^{(k)*})^T \in \mathbb{R}^p$ for $k \in [K]$ captures the variation in stratum k around $\boldsymbol{\mu}_j^*$. Estimates of $\boldsymbol{\mu}_j^*$ and the $\boldsymbol{\gamma}_j^{(k)*}$'s are then derived as minimizers of the following criterion, for appropriate values of $\lambda_{j,1} \geq 0$ and $\lambda_{j,2}^{(k)} \geq 0$,

$$\sum_{k=1}^{K} -L((\boldsymbol{\mu}_j + \boldsymbol{\gamma}_j^{(k)}); \boldsymbol{\mathcal{U}}_j^{(k)}, (\boldsymbol{\mathcal{U}}_{-j}^{(k)})) + \lambda_{j,1}\|\boldsymbol{\mu}_{-j}\|_1 + \sum_{k=1}^{K} \lambda_{j,2}^{(k)} \|\boldsymbol{\gamma}_j^{(k)}\|_1. \tag{2.10}$$

where $\boldsymbol{\mu}_{-j}$ corresponds to the vector $\boldsymbol{\mu}_j$ after the elimination of the constant term μ_j. Again, we will use the particular value $\lambda_{j,2}^{(k)} = \lambda_{j,2}\omega_k$ with $\omega_k = \sqrt{K(n_k/n)}$, as suggested in Ollier and Viallon (2017). It is easily shown that, at optimum, we have $\hat{\mu}_{j,\ell} = \text{argmin}_{m \in \mathbb{R}}(\lambda_{j,1}|m| + \sum_k \lambda_{j,2}^{(k)}|\hat{\theta}_{j,\ell}^{(k)} - m|)$ for all $\ell > 1$. Therefore, among the infinitely many decompositions of the form (2.9), DataShared Lasso targets the one such that $\hat{\mu}_{j,\ell}$ is a weighted and shrunk towards 0 version of the median of the set of estimates $(\hat{\theta}_{j,\ell}^{(1)}, \ldots, \hat{\theta}_{j,\ell}^{(K)})$. In other words, the penalty term $\left(\lambda_{j,1}\|\boldsymbol{\mu}_{-j}\|_1 + \sum_{k=1}^K \lambda_{j,2}^{(k)}\|\boldsymbol{\gamma}_j^{(k)}\|_1\right) = \left(\lambda_{j,1}\|\boldsymbol{\mu}_{-j}\|_1 + \sum_{k=1}^K \lambda_{j,2}^{(k)}\|\boldsymbol{\theta}_j^{(k)} - \boldsymbol{\mu}_j\|_1\right)$ shrinks the estimators $\hat{\theta}_{j,\ell}^{(k)}$, $k \in [K]$, towards their "weighted and shrunk towards 0" median. For the constant terms, we have $\hat{\mu}_j = \text{argmin}_{m \in \mathbb{R}} \sum_k \lambda_{j,2}^{(k)}|\hat{\theta}_j^{(k)} - m|$ and $\hat{\mu}_j$ is then simply a weighted median of the set of values $(\hat{\theta}_j^{(1)}, \ldots, \hat{\theta}_j^{(K)})$ (see Section A.1.1 in the Appendix for more details). Ollier and Viallon (2017) compared the performance of DataShared Lasso and Ref-Lasso, both formally and empirically. By bypassing the arbitrary choice of the reference stratum, DataShared Lasso was shown to achieve performance similar to the strategy that would consist in applying Ref-Lasso with an optimal reference stratum (in our context here, this unknown optimal reference stratum would typically vary for each association j, ℓ). In addition, DataShared Lasso can be rewritten as a standard lasso after a simple transformation of

the original data (see Section A.1.2 in the Appendix). Therefore, its implementation is very easy and DataShared Lasso has roughly the same computational cost as Ref-Lasso. We refer to Ollier and Viallon (2017) and Gross and Tibshirani (2016).

2.2.3.3 Combining the $(\hat{\boldsymbol{\theta}}_j^{(k)})_{j\in[p],k\in[K]}$ to derive the K estimated graphs.

For every $j<\ell$, the approaches based on SepLogit (Indep-SepLogit, Ref-SepLogit, Fused-SepLogit and DataShared-SepLogit) all return two vectors of estimates for $(\theta_{j,\ell}^{(1)*},...,\theta_{j,\ell}^{(K)*})$: $(\hat{\theta}_{j,\ell}^{(1)},...,\hat{\theta}_{j,\ell}^{(K)})$ and $(\hat{\theta}_{\ell,j}^{(1)},...,\hat{\theta}_{\ell,j}^{(K)})$. Of course, we may still have, for some k,ℓ,j, $\hat{\theta}_{j,\ell}^{(k)}\neq\hat{\theta}_{\ell,j}^{(k)}$, or even $\hat{\theta}_{j,\ell}^{(k)}=0$ while $\hat{\theta}_{\ell,j}^{(k)}\neq 0$ for instance. But we may also have other asymmetry issues here. For instance, we may have a fully homogeneous vector $(\hat{\theta}_{j,\ell}^{(1)},...,\hat{\theta}_{j,\ell}^{(K)})$, that is, a vector whose components are all equal suggesting that the association between variables U_ℓ and U_j does not vary across the strata, while $(\hat{\theta}_{\ell,j}^{(1)},...,\hat{\theta}_{\ell,j}^{(K)})$ exhibits some heterogeneities, suggesting that the association between variables U_ℓ and U_j does vary across the strata.

To get around these asymmetry issues, we propose two adaptations of SepLogit AND and SepLogit OR. When only one graph has to be estimated on a single stratum (or on the population as a whole), the complexity of the estimated graph can be defined as the number of edges of this graph. Then SepLogit AND [resp. OR] returns the graph with lowest [resp. highest] complexity given the parameters $\hat{\theta}_{j,\ell}$ and $\hat{\theta}_{\ell,j}$, for $(j,\ell)\in[p]^2$. When a collection of K graphs is estimated over K strata, we propose two strategies, SepLogit MIN and SepLogit MAX, returning two collections of K graphs, with lowest and highest complexities respectively. We will define the complexity of a collection of K estimated graphs based on the number of heterogeneities among the sets $(\hat{\theta}_{j,\ell}^{(1)},...,\hat{\theta}_{j,\ell}^{(K)})\in\mathbb{R}^K$ (or $(\hat{\theta}_{\ell,j}^{(1)},...,\hat{\theta}_{\ell,j}^{(K)})\in\mathbb{R}^K$), for $(j,\ell)\in[p]^2$. More precisely, for any vector $\boldsymbol{S}=(s_1,...,s_K)\in\mathbb{R}^K$, we define its complexity $comp(\boldsymbol{S})$ as the number of distinct non-null values among $(s_1,...,s_K)$. Then, for every $j<\ell$, denote by $\boldsymbol{S}_j$ the vector $(\hat{\theta}_{j,\ell}^{(1)},...,\hat{\theta}_{j,\ell}^{(K)})$ and by $\boldsymbol{S}_\ell$ the vector $(\hat{\theta}_{\ell,j}^{(1)},...,\hat{\theta}_{\ell,j}^{(K)})$ returned by Fused-SepLogit, DataShared-SepLogit, Ref-SepLogit or Indep-SepLogit. Now, let $\boldsymbol{S}_{j,\ell}^{\min}$ [resp. $\boldsymbol{S}_{j,\ell}^{\max}$] denote the vector among $(\boldsymbol{S}_j,\boldsymbol{S}_\ell)$ with lowest [resp. highest] complexity, as measured by function $comp$. In other words, if $comp(\boldsymbol{S}_j)<comp(\boldsymbol{S}_\ell)$ then $\boldsymbol{S}_{j,\ell}^{\min}=\boldsymbol{S}_j$ and $\boldsymbol{S}_{j,\ell}^{\max}=\boldsymbol{S}_\ell$, while $\boldsymbol{S}_{j,\ell}^{\min}=\boldsymbol{S}_\ell$ and $\boldsymbol{S}_{j,\ell}^{\max}=\boldsymbol{S}_j$ otherwise. Finally, the MIN [resp. MAX] strategy returns the collection of graphs constructed from vectors $\boldsymbol{S}_{j,\ell}^{\min}$ [resp. $\boldsymbol{S}_{j,\ell}^{\max}$] obtained for all $j<\ell$, which is the collection with lowest [resp.

highest] possible complexity given initial estimates $(\hat{\theta}^{(1)}_{j,\ell}, ..., \hat{\theta}^{(K)}_{j,\ell})$ and $(\hat{\theta}^{(1)}_{\ell,j}, ..., \hat{\theta}^{(K)}_{\ell,j})$, for $(j, \ell) \in [p]^2$.

2.3 Simulation study

We empirically compared the approaches presented above on synthetic data, following the simulation framework developed in Guo et al. (2015). Indep-SepLogit, Ref-SepLogit and DataShared-SepLogit were implemented using the glmnet package, while Guo et al's approach and Fused-SepLogit were implemented using the BMN and FusedLasso packages respectively. Note that the BMN and FusedLasso packages are not maintained on the CRAN anymore, but are still available from the archives. For the sake of brevity, results are only presented for the SepLogit MIN strategy presented above. Selection of tuning parameters was performed using the BIC. Following ideas introduced in Efron et al. (2004) (see also Viallon et al., 2016; Ollier and Viallon, 2014), a two-step BIC approach was also considered and yielded very similar performance (results not shown).

2.3.1 Data generation

The IsingSampler package of R was used to generate the data, given matrices $\mathbf{\Theta}^{(k)*} = (\theta^{(k)*}_{j,\ell})_{p*p}$. Following the framework considered by Guo et al. (2015), these matrices were defined as $\mathbf{\Theta}^{(k)*} = \boldsymbol{\mu}^* + \mathbf{\Psi}^{(k)*}$, where $\boldsymbol{\mu}^* = (\mu^*_{j,\ell})_{p*p}$ represents the common structure across all strata and $\mathbf{\Psi}^{(k)*} = (\psi^{(k)*}_{j,\ell})_{p*p}$ represents the structure specific to stratum k, for $k \in [K]$.

For the common part, we again followed the framework of Guo et al. (2015) and considered three types of graphs which are the chain graph, the 3-nearest neighbor graph and the scale-free graph. See Figure 2.1 and Guo et al. (2015) for more details. As for the specific part, non-zero values of each $\mathbf{\Psi}^{(k)*}$ were randomly generated on the set $[-1,-0.5] \cup [0.5, 1]$. The number of non-zero values in each $\mathbf{\Psi}^{(k)*}$ was the same for every k and depended on a parameter $\rho \in [0, 1]$. This parameter corresponds to the ratio between the number of individual edges and the number of common edges. Therefore, this ratio represents the level of heterogeneity between the different strata. In particular, if $\rho = 0$, the structures are identical over the K strata. Five values of ρ were considered : 0, 0.25, 0.5, 0.75 and 1.

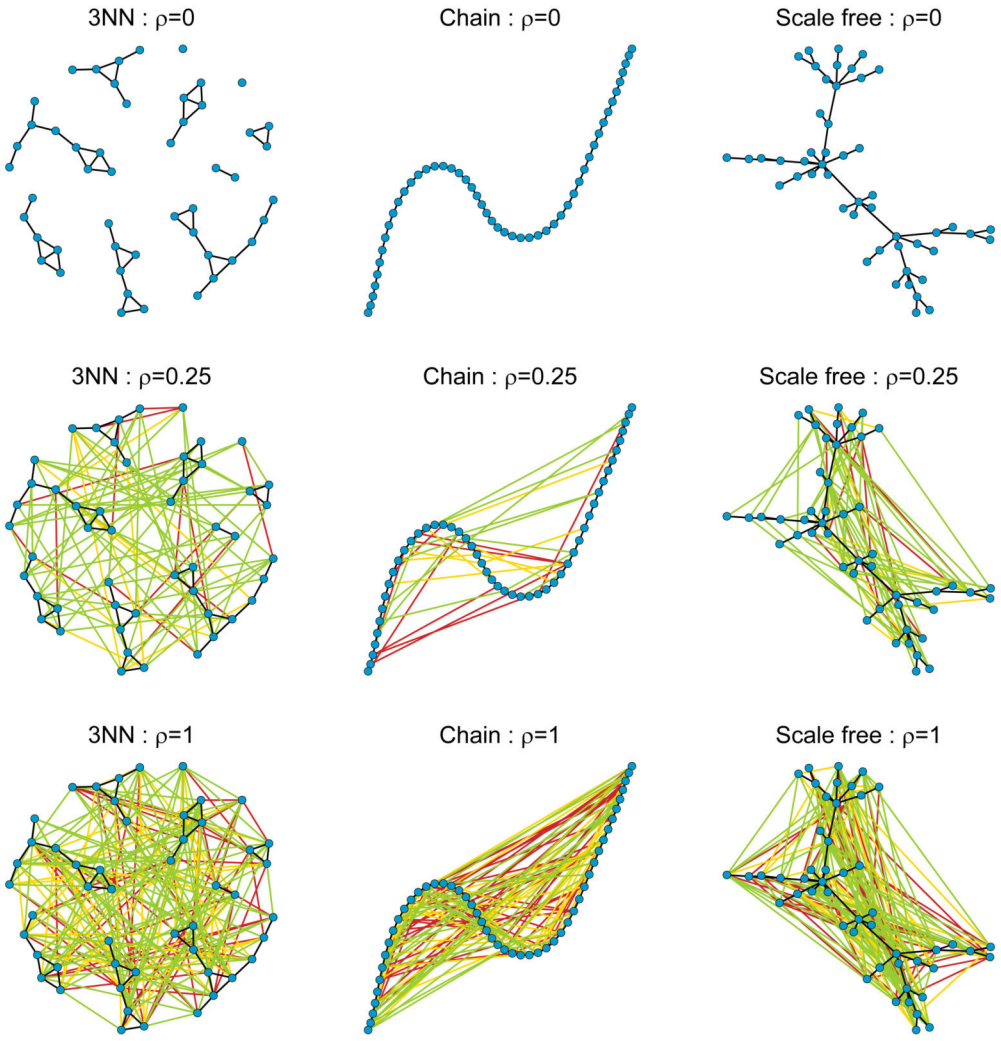


FIGURE 2.1: A graphical representation for the three types of network of structuring with a ratio ρ equal to 0, 0.25 and 1, with $p = 50$ and $K = 3$. The black edges represent the common structure and the red, blue and green edges represent the structures specific to each stratum.

For each common structure and ratio ρ, we considered 50 replicates of data consisting of 500 observations in each stratum, with $(p = 10, K = 3)$ in a first simulation study. In a second study, we considered 50 replicates of data consisting of 666, 1143, 1920 and 3060 observations in 1st, 2nd, 3rd and 4th stratum respectively, with $(p = 36, K = 4)$. These choices were motivated by the dimension in our leading example (see Section 2.4). Note, however, that Fused-SepLogit and Guo et al. (2015) approach were only considered in the first simulation study to save computational time.

2.3.2 Evaluation criteria

Estimates $\hat{\mathbf{\Theta}}^{(k)} = (\hat{\theta}^{(k)}_{j,\ell})_{p*p}$ returned by the various methods were computed and compared to $\mathbf{\Theta}^{(k)*}$ on each simulated data. Two criteria were computed and averaged over the 50 replicates of each of the considered simulation design. The first one, Acc.S, measures the accuracy regarding the support of each matrix $\mathbf{\Theta}^{(k)*}$, that is the indices of the non-zero off-diagonal elements of matrices $\mathbf{\Theta}^{(k)*}$, $k \in [K]$. More precisely, Acc.S is defined as

$$\text{Acc.S} = \frac{1}{K} \sum_{k \in [K]} \left(\frac{\sum_{j>\ell} \left(\mathbb{1}[\theta^{(k)*}_{j,\ell} \neq 0 \ , \ \hat{\theta}^{(k)}_{j,\ell} \neq 0] + \mathbb{1}[\theta^{(k)*}_{j,\ell} = 0 \ , \ \hat{\theta}^{(k)}_{j,\ell} = 0] \right)}{p(p-1)/2} \right).$$

We also evaluated the performance of each method regarding the identification of the heterogeneity between matrices $\mathbf{\Theta}^{(k)*}$, $k \in [K]$. Here, we report results for Acc.H, which is defined as follows

$$\text{Acc.H} = \frac{\sum_{j>\ell} \left(\mathbb{1}[Z^*_{j,\ell} \neq 0 \ , \ \hat{Z}_{j,\ell} \neq 0] + \mathbb{1}[Z^*_{j,\ell} = 0 \ , \ \hat{Z}_{j,\ell} = 0] \right)}{p(p-1)/2},$$

where

$$Z^*_{j,\ell} = \begin{cases} 0 & \text{if } \ \theta^{(1)^*}_{j,\ell} = \theta^{(2)^*}_{j,\ell} = ... = \theta^{(K)^*}_{j,\ell} \\ 1 & \text{otherwise} \end{cases} \quad \text{and} \quad \hat{Z}_{j,\ell} = \begin{cases} 0 & \text{if } \ \hat{\theta}^{(1)}_{j,\ell} = \hat{\theta}^{(2)}_{j,\ell} = ... = \hat{\theta}^{(K)}_{j,\ell} \\ 1 & \text{otherwise.} \end{cases}$$

In other words, $Z^*_{j,\ell}$ [resp. $\hat{Z}_{j,\ell}$] is a binary variable which equals 1 if the association between variables j and ℓ varies across the K strata, under the true model [resp. as identified by the considered method], and Acc.H corresponds to the accuracy regarding the support recovery of $Z^*_{j,\ell}$. Other criteria were considered, such that the Rand index, and lead to similar results (not shown).

TABLE 2.1: Averages of the computational times (in minutes) needed to estimate the K models (computed over 50*3*5=750 runs, where 50 is the number of replicates, 3 is the number of designs and 5 the number of ρ values). The column 1st [resp. 2nd] corresponds to the first [resp. second] simulation study.

Method \ Simulation study	**1st**	**2nd**
Indep-SepLogit	10^{-3}	0.15
Ref-SepLogit	1	60
DataShared-SepLogit	1	60
Fused-SepLogit	30	x
Guo et al's	180	x

2.3.3 Results

2.3.3.1 First simulation study with $p = 10$, $K = 3$ and balanced strata sizes.

In this first situation, where all strata share the same size, the reference stratum for Ref-SepLogit was selected randomly. Results are presented on Figure 2.2. Overall, the common structure type only marginally affects the comparisons between the five considered methods. In other respect, and as expected, the performance of Indep-SepLogit regarding Acc.S is independent of the level of heterogeneity ρ, while DataShared-SepLogit, Fused-SepLogit and Guo et al's method globally outperform Indep-SepLogit, especially when ρ is low (i.e., when homogeneity is high). These three methods perform similarly regarding Acc.S, with a slight advantage for Fused-SepLogit and DataShared-SepLogit for $\rho = 0$ and an even slighter advantage for Guo et al's method for $\rho \geq 0.75$. Interestingly, and still considering Acc.S, Ref-SepLogit performs nearly as well as DataShared-SepLogit, Fused-SepLogit and Guo et al's approach when homogeneity is high (and ρ is small), but notably worse for $\rho \geq 0.5$. As for Acc.H, Fused-SepLogit and DataShared-SepLogit perform similarly and outperform Ref-SepLogit (especially for high values of ρ) and noticeably outperform Indep-SepLogit and Guo et al's method. In addition, Guo et al's method and Indep-SepLogit perform similarly regarding this criterion confirming that Guo et al's method is not well suited for the identification of heterogeneities. The average computational times needed to estimate the K models for each method are presented in Table 2.1. Overall, DataShared-SepLogit is 180 times faster than Guo et al's approach, 30 times faster than Fused-SepLogit and, and shares the same computational cost as Ref-SepLogit. In other words, DataShared-SepLogit appears as the best compromise between estimation performance and computational time.

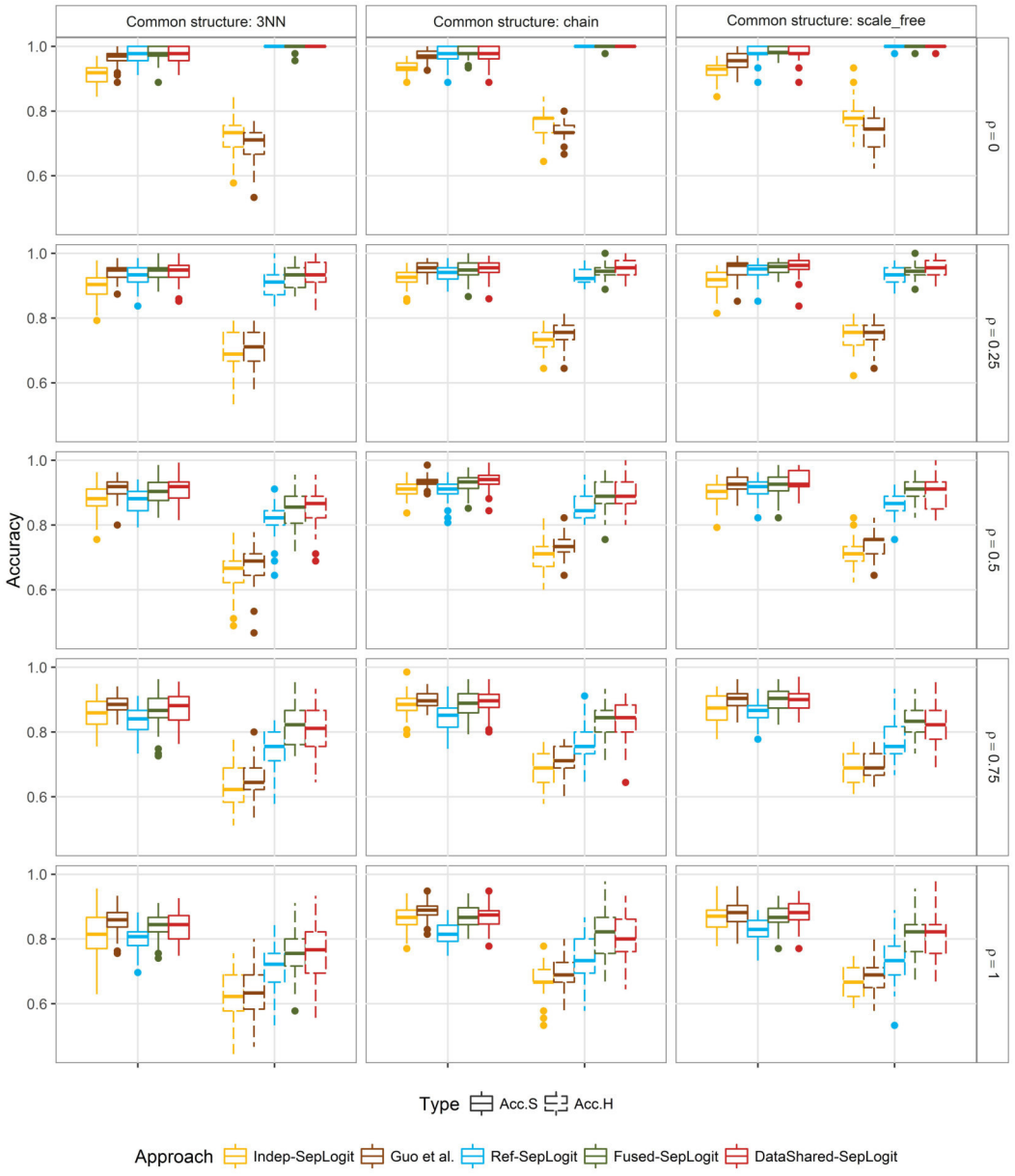


FIGURE 2.2: Boxplots for the values of Acc.S and Acc.H obtained for each method on the 50 replicates of each simulation design in the first simulation study.

2.3.3.2 Second simulation study with $p = 36$, $K = 4$ and an unbalanced strata sizes.

Motivated by the dimensions observed in our illustrative example, we further considered the situation where $K = 4$, $p = 36$ and with n_k equal to the number of observations in the k-th stratum of our illustrative example divided by 10. The reference stratum for

Ref-SepLogit was set to the largest one here, which is a common choice in practice. To save computational time, only results for Indep-SepLogit, Ref-SepLogit and DataShared-SepLogit were considered. We also considered adaptive versions of Ref-SepLogit and DataShared-SepLogit, with adaptive weights derived from initial unpenalized maximum likelihood estimates ; see, e.g., Ollier and Viallon (2014) for details on the adaptive version of DataShared Lasso. The results are presented in Figure 2.3. Overall, they are consistent with those obtained in the first simulation study : DataShared-SepLogit outperforms both Indep-SepLogit and Ref-SepLogit, regarding the identification of the heterogeneities (as measured by Acc.H), as well as support recovery for low values of ρ, but also for higher values of ρ when using the the adaptive version. Globally, adaptive versions of DataShared-SepLogit and Ref-SepLogit outperfoms their "standard" counterparts, regarding both Acc.S and Acc.H, especially for high values of ρ.

2.4 Application

Data description (Registry data) The Rhône registry contains data describing all the victims of road traffic accidents that have occurred in the Rhône since 1996. The Rhône is a French Department (1,600,000 inhabitants) whose main urban center is Lyon. The Rhône registry is officially recognized by the National Registry Committee and is managed by the ARVAC (Association pour le Registre des Victimes d'Accidents de la route). Data collection began on January, 1st 1995. Up to now, the data from 1996 to 2014 have been fully computerized and validated. This data contains the identification of the victims (name, sex, date of birth), the accident characteristics (date, time, place, type of vehicle, etc.) as well as the complete injury tables of victims. The injury table of each victim describes each injury he suffered, coded according to the **A**bbreviated **I**njury **S**cale **90** classification. This classification describes the type and the location of the injury using six digits written as 12(34)(56). The first digit identifies the body region **[R]** (Head, Face, Lower extremity, Spine, etc.), the second one identifies the type of the anatomic structure **[T]** (Vessels, Nerves, etc.), the third and fourth digits identify the specific anatomic structure, or the nature of the injury when an entire area is reached, **[S]** (Cervical Column, Back Column, Contusion, Burn, etc.) and the fifth and sixth ones specify the type of injury **[N]** (Fractures, rupture, laceration, etc.). Using this coding, the original dataset contained 1348 distinct codes corresponding to 1348 distinct injuries.

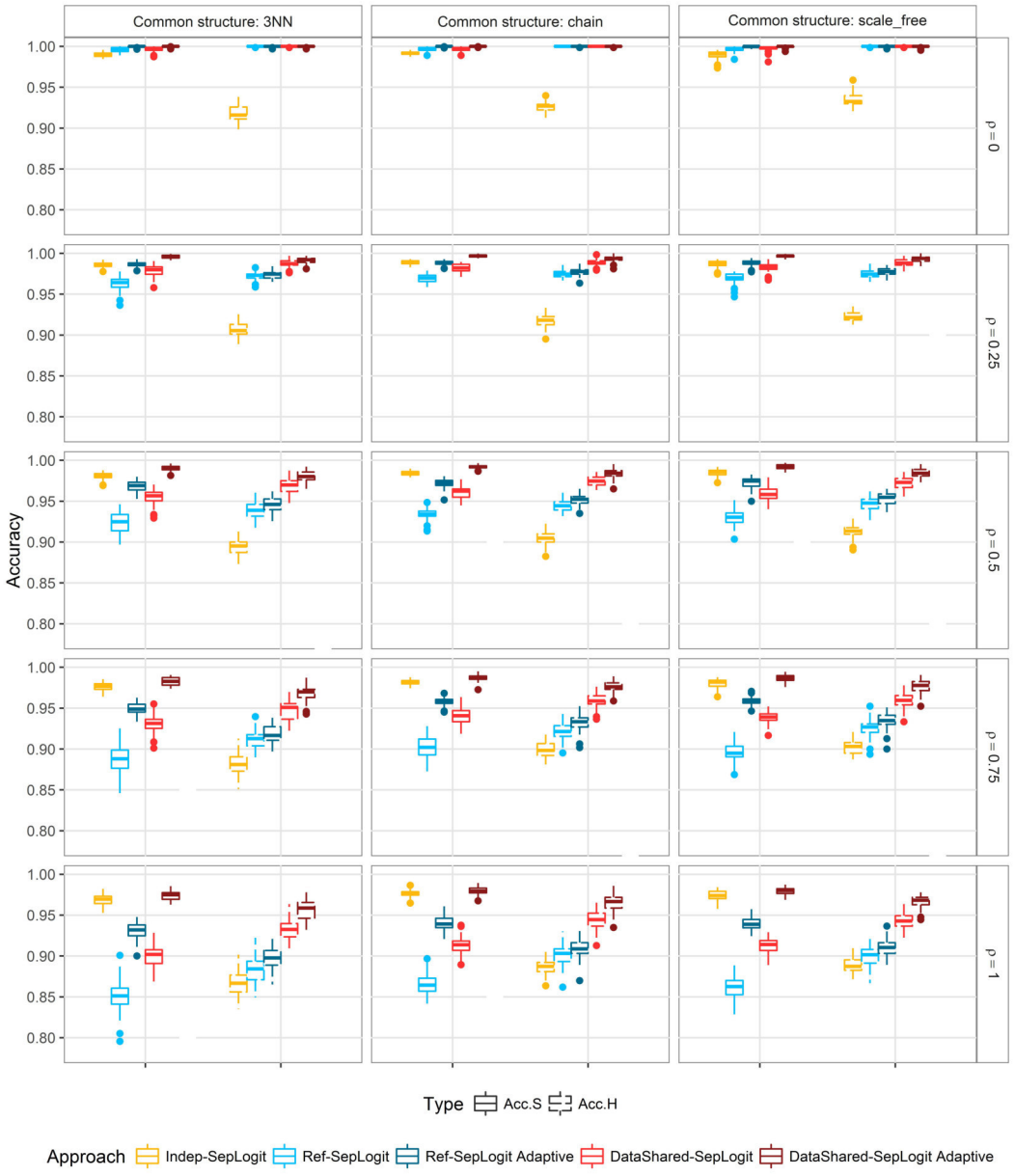


FIGURE 2.3: Boxplots for the values of Acc.S and Acc.H obtained for each method on the 50 replicates of each simulation design in the second simulation study.

However, because most of these injuries have very low prevalences, we decided to group some of them. After converting the AIS codes to ICD-10 codes (International Statistical Classification of Diseases and Related Health Problems, 10-th Revision), we used the grouping proposed in the EUROCOST model (see Lyons et al., 2006, Table 1). We further decided to split the "Internal-organ injury" group into two groups, corresponding to "Internal-organ injury thorax" and "Internal-organ injury abdomen". We finally worked with 36 groups of injuries. Formally, this led us to consider 36 binary $(0,1)$-variables $U_1, \ldots, U_p$, where $U_j = 1$ if and only if the original injury table of the victim contains

at least one injury falling into the j-th group of the EUROCOST model. Further note that these 36 groups of injuries can be classified into 6 classes, roughly corresponding to body areas (see Table 2.2). From now on the 36 groups of injuries will simply be referred to as injuries.

In this application, we used the data collected over the last 10 available years (2005 to 2014). We considered four strata defined according to the road user type of the victim : pedestrians, cyclists, motorized two-wheelers users and car occupants. Overall, this data contained 67,894 victims and 109,793 injuries. See Table 2.3 for more details. In Figure 2.4, estimates of the prevalence of each injury on the different strata are presented. Colors correspond to the class of each injury and the length of each bar corresponds to the prevalence of each injury. Some differences can be noticed across the strata. For instance, compared to other road users, car occupants are more likely to suffer from whiplash and spine injuries, while they are less likely to suffer from injuries of the upper and lower extremities. On the other hand, motorized two wheelers users are less likely to suffer from concussion and wound face, as expected since their head and face are supposedly protected by their helmet.

Stratified graphical model estimation We applied the adaptive versions of Fused-SepLogit, DataShared-SepLogit, Ref-SepLogit and Indep-SepLogit, using SepLogit MIN strategy. When applying Ref-SepLogit, the reference stratum was set to the largest one, which corresponds to car occupants.

First consider the results returned by DataShared-SepLogit. Figure 2.5 presents a simplified version of the estimated structure of the graphical models on each of the 4 strata : to improve legibility, only edges corresponding to estimated conditional odds-ratios greater than or equal to 2 are presented (see Table 2.4 for more details, and the Discussion for our motivation to focus on these associations only). This figure was generated making use of the chordDiagram function of the circlize R package. Nodes (injuries) are represented around a circle to facilitate the comparison of the structures across the strata.

TABLE 2.2: Descriptions, labels and classes of injuries

Description	**Label**	**Class**
Concussion	Concussion	Head-Face
Other skull-brain injury	Skull-Injury	Head-Face
Open wound of head	Wound-Head	Head-Face
Eye injury	Eye-Injury	Head-Face
Fracture of facial bones	Fracture-Face	Head-Face
Open wound of face	Wound-Face	Head-Face
Fracture/dislocations/sprain/ strain of vertebral/spine	Spine	Spine
Whiplash injury/neck sprain/ distortion of cervical spine	Whiplash	Spine
Spinal cord injury	Spinal-Cord	Spine
Internal-organ injuries /Thorax	Internal-Thorax	Thorax-Abdomen
Internal-organ injuries /Abdomen	Internal-Abdomen	Thorax-Abdomen
Fracture of rib/sternum	Rib-Frac	Thorax-Abdomen
Fracture of clavicle or scapula	Clavic-Frac	Upper Extremity
Fracture of upper arm	UpArm-Frac	Upper Extremity
Fracture of elbow or forearm	ForeArm-Frac	Upper Extremity
Fracture of wrist	Wrist-Frac	Upper Extremity
Fracture of hand or fingers	Hand-Frac	Upper Extremity
Dislocation/sprain/ strain of shoulder or elbow	UpArm-Disloc	Upper Extremity
Dislocation/sprain/ strain of hand or fingers	Hand-Disloc	Upper Extremity
Injury to nerves of upper extremity	UpArm-Nerves	Upper Extremity
Complex soft tissue injury of upper extremity	UpExtrem-Injury	Upper Extremity
Fracture of pelvis	Pelvis-Frac	Lower Extremity
Fracture of hip	Hip-Frac	Lower Extremity
Fracture of femoral shaft	Femur-Frac	Lower Extremity
Fracture of knee or lower leg	LowLeg-Frac	Lower Extremity
Fracture of ankle	Ankle-Frac	Lower Extremity
Fracture of foot(excludes ankle)	Foot-Frac	Lower Extremity
Dislocation/sprain/strain of knee	Knee-Disloc	Lower Extremity
Dislocation/sprain/strain of ankle or foot	Ankle-Disloc	Lower Extremity
Dislocation/sprain/strain of hip	Hip-Disloc	Lower Extremity
Injury to nerves of lower extremity	LowExtrem-Nerves	Lower Extremity
Complex soft tissue injury of lower extremity	LowExtrem-Injury	Lower Extremity
Superficial injury (including contusions and bruises)	Contusions	Others
Open wounds	OpenWound	Others
Mild burns	Burns	Others
Other and unspecified injury	Unspecif	Others

TABLE 2.3: Number of victims and injuries in each stratum

	Pedestrians	**Cyclists**	**Motorized T-W Users**	**Car Occupants**
Victims	6663	11431	19204	30596
Injuries : before grouping	15348	20703	41779	55966
Injuries : after grouping	11865	17404	32932	47592

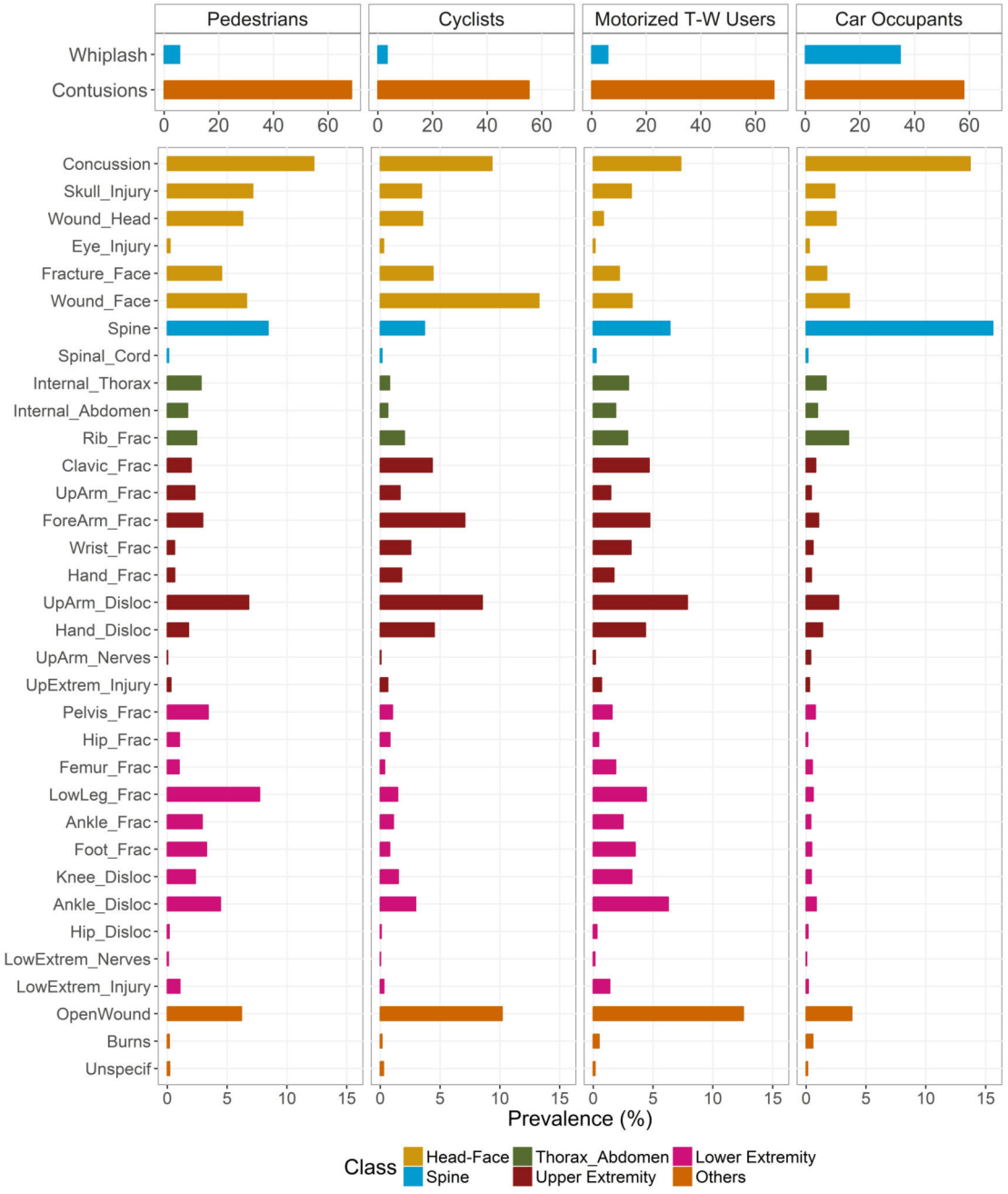


FIGURE 2.4: Injury prevalences in each stratum.

Node colors correspond to the class they belong to (see Table 2.2), while edge colors correspond to the classes of the two connected injuries. If these two injuries belong to distinct classes, then the edge is grey, otherwise the edge shares the same color as the two injuries. The edge width is related to the value of the corresponding conditional odds-ratio. Lastly, the size of a node corresponds to the sum of the edge widths over the edges involving this node.

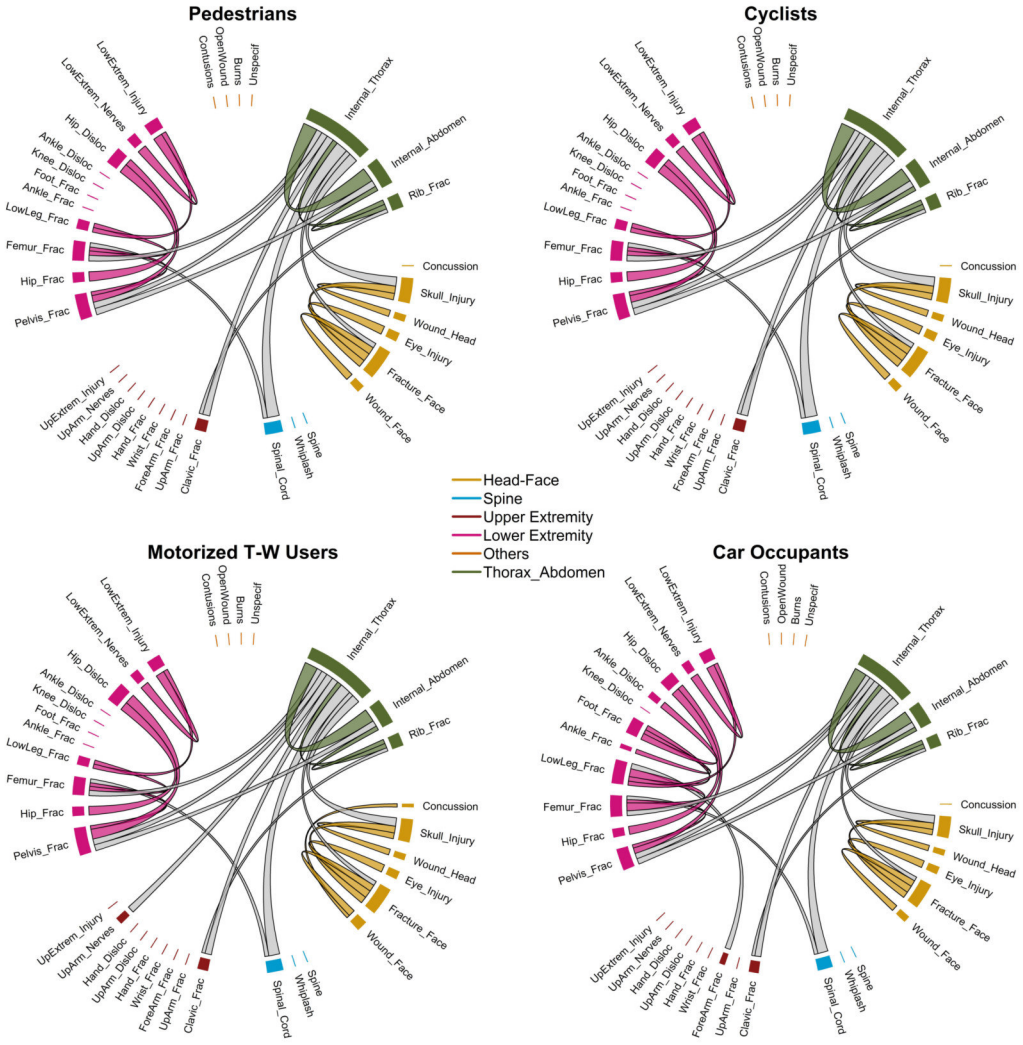


FIGURE 2.5: Application of the DataShared-SepLogit Adaptive MIN approach on the Rhône Registry Data to describe the injury tables of victims of road accidents, according to road user type : pedestrians, cyclists, motorized T-W and car occupants. Only edges corresponding to conditional odds-ratios greater than or equal to 2 are represented.

Overall, the estimated structures are very similar over the four strata, and most edges connect injuries belonging to the same class. In particular, associations between injuries

TABLE 2.4: Number of associations in each stratum

OR \ Stratum	Pedestrians	Cyclists	Motorized T-W Users	Car Occupants
$\neq 1$	241	265	244	248
> 1	40	38	43	42
≥ 2	22	22	24	26

of the Head-Face class are very similar for motorized two-wheelers users and the other users. This can be considered as unexpected at first sight, given the protective effect of helmet : as mentioned above, Figure 2.4 shows that injuries are less frequent in the Head-Face area for motorized two-wheelers users. However, our findings suggest that motorized two-wheelers users suffering from one injury in the Head-Face area are likely to suffer from other injuries in the Head-Face area as well, just as any other road user. Even if a lack of power is a possible interpretation, another simple interpretation is that the helmet was not able to protect these users well enough, either because of the violence of the impact, or because the helmet felt down, or because the user simply did not have a helmet : when the helmet loses its protective effect, motorized two-wheelers users are just like any other road users, hence similar associations are observed between injuries of the Head-Face class in all strata.

Lastly, the main difference among the structures estimated on the 4 strata concerns associations between injuries of the lower extremities, which are more numerous for car occupants, while prevalences of these injuries are lower for these road users (see Figure 2.4). Here again, there is a simple explanation : for car occupants to suffer from such injuries, substantial deformations of the car body or collisions with the dashboard are quite likely, which typically generate multiple injuries.

Comparison of methods Results obtained with Fused-SepLogit were very similar to those obtained with DataShared-SepLogit and are therefore not presented here for the sake of brevity. On the other hand, Indep-SepLogit (see Figure A.2 in the Appendix) returns structures with many more heterogeneities while Ref-SepLogit returns structures with fewer heteregeneities (see Figure A.1 in the Appendix). For an alternative graphical representation, see also Figures A.4, A.5 and A.6 in the Appendix. These results were expected, given the results of our simulation study. Indeed, assuming the model obtained by DataShared-SepLogit is close to the true one, car occupants correspond to the worst reference stratum to use in combination with Ref-SepLogit since it leads to the highest

level of heterogeneity, hence the highest complexity, and sparsistency is not guaranteed ; see Ollier and Viallon (2017) for more details. In other words, it is likely that applying Ref-SepLogit after choosing car occupants as the reference stratum, though quite natural, is not optimal in this application. Here again, DataShared-SepLogit by-passes the arbitrary choice of the reference stratum, and is likely to perform better than Ref-SepLogit for nearly the same computational cost. In other respect, the heterogeneities returned by Indep-SepLogit are not interpretable and are likely mostly noise.

2.5 Discussion

In this article, we described two methods based on multiple penalized logistic regressions to jointly estimate binary graphical models on several pre-defined strata. By appropriately penalizing heterogeneities across the corresponding structures, the proposed methods take benefit of the potential homogeneity among these structures, and allow the interpretation of the identified heterogeneities. Focusing on the identification of heterogeneities, we observed better performance for our methods compared to the naive strategy consisting in estimating the structure on each stratum independently, the strategy proposed by Guo et al. (2015), and the standard strategy that consists in estimating the structures after choosing an arbitrary reference stratum. We shall stress that DataShared-SepLogit is similar to Ref-SepLogit regarding computational time, and is then computationally more efficient than Fused-SepLogit. On the other hand, DataShared-SepLogit is similar to Fused-SepLogit, and generally much better than Ref-SepLogit, regarding the accuracy for support recovery and the identification of heterogeneities.

We may further mention that methods based on group lasso penalties (Ma and Michailidis, 2016), were not considered in this work because they are not well suited for the identification of heterogeneities. On the other hand, an interesting lead for future work could consist in extending the approach of Tao et al. (2016) which uses ℓ_0-like penalties to encourage homogeneity when estimating multiple Gaussian graphical models.

Now, focusing on our illustrative application, most of the observed associations were common to all user types, while heterogeneities mostly concerned car occupants. We have to insist on the fact that only positive associations were presented here (Figure 2.5) actually presents associations corresponding to conditional odds-ratio greater than

or equal to two only ; see Figure A.3 in the Appendix for models with the whole list of positive associations). The main reason why negative associations are less interesting in this particular application has to do with the way clinicians record injuries for each victim : they are likely overlook some injuries when more severe injuries are present, and most negative associations may simply be due to this reporting bias. In addition, because all observations correspond to victims of road accidents, our data set consists of individuals who suffer from at least one injury. Therefore, our sample is biased compared to the whole population ; the resulting collider bias (Hernán et al., 2004) typically makes causal interpretation of negative association hazardous. Even if no causal interpretation can be given to the identified associations, the description of these associations from the injury tables of victims of road accidents is still relevant, since this sub-population of victims of road accidents is the one that clinicians have to take care of. In future work, finer groupings of the injuries may be used to improve clinical interpretability, along with other definitions for the strata, which may include the severity of the crash, etc.

Although we focused on the standard quadratic exponential binary model, all the methods we presented here can be extended to estimate stratified binary graphical models with higher-order interactions (see Section A.2 in the Appendix). However, the computational burden may rapidly become an issue when considering more complex models. In addition, the graphical representation and interpretation of these general models is less straightforward.

Lastly, in applications where selection bias is absent, combining the ideas described here with those presented in Champion et al. (2017) (see also van de Geer and Bühlmann (2013) for the ℓ_0 version) could lead to a powerful approach to estimate causal DAGs on stratified data.

Chapitre 3

Sparse estimation for case-control studies with multiple disease subtypes.

Ballout N, Garcia C, Viallon V. Sparse estimation for case-control studies with multiple disease subtypes. Biostatistics. Biostatistics 2020. doi.org :10.1093/biostatistics/kxz063. Ballout et al. (2020b).

Abstract

The analysis of case-control studies with several disease subtypes is increasingly common, e.g. in cancer epidemiology. For matched designs, a natural strategy is based on a stratified conditional logistic regression model. Then, to account for the potential homogeneity among disease subtypes, we adapt the ideas of DataShared Lasso, which has been recently proposed for the estimation of stratified regression models. For unmatched designs, we compare two standard methods based on L_1-norm penalized multinomial logistic regression. We describe formal connections between these two approaches, from which practical guidance can be derived. We show that one of these approaches, which is based on a symmetric formulation of the multinomial logistic regression model, actually reduces to a DataShared Lasso version of the other. Consequently, the relative performance of the two approaches critically depends on the level of homogeneity that exists among disease subtypes : more precisely, when homogeneity is moderate to high, the non-symmetric formulation with controls as the reference is not recommended. Empirical results obtained from synthetic data are presented, which confirm the benefit of

properly accounting for potential homogeneity under both matched and unmatched designs, in terms of estimation and prediction accuracy, variable selection and identification of heterogeneities. We also present preliminary results from the analysis of a case-control study nested within the EPIC cohort, where the objective is to identify metabolites associated with the occurrence of subtypes of breast cancer.

Key words : Conditional logistic regression ; Multinomial logistic regression ; Lasso ; Sparsity ; Structured sparsity.

3.1 Introduction

The rise of -omics and other high-dimensional data in medical science gives researchers access to numerous features that may predict outcomes of interest, like cancer development. However, this relatively cheap source of information comes at a price : the curse of dimensionality makes multivariate modeling of such data impossible without further assumptions. In other words, some prior piece of information has to be properly accounted for to reduce dimensionality and accurately estimate high-dimensional multivariate models. The prior information about the sparsity of the parameter vector is one common assumption for the parametric regression models. The use of L_1-norm regularized approaches, such as the Lasso (Tibshirani, 1996), has been shown to yield optimal sparse estimates when the true vector is sparse, under technical assumptions on the design matrix (Wainwright, 2009; Bach, 2010; Bickel et al., 2009). As a result, L_1-penalized logistic models are now standard tools when studying risk factors of a disease in a high-dimensional setting (Park and Hastie, 2007; Wu et al., 2009) .

For many diseases that were primarily considered as one single disease (breast cancer, colorectal cancer), several subtypes have now been recognized. They can either be histological, as for breast cancer, or anatomical, as for colorectal cancer. Even if commonalities may exist among these subtypes, they have their own specificities regarding both prognosis and etiology. For example, the cancer epidemiology community is now increasingly concerned with the identification of subtype specific risk factors for various cancer sites. One illustrating example is presented in Section 3.5, where the objective is the identification of metabolites associated with breast cancer subtypes, based on a matched case-control study nested in the EPIC (European Prospective Investigation into Cancer and nutrition) cohort study.

Formally, let $K-1$ denote the number of case/disease subtypes, for some $K > 1$. In matched case-control studies, and assuming for simplicity a 1 :1 matching, each case has his own control. Then, the overall sample can naturally be divided into $K-1$ subsamples. Each subsample can be analyzed separately using, e.g., a conditional logistic regression model. On the other hand, for unmatched studies with multiple subtypes, controls are "shared" for all case subtypes, and the sample can not be split according to disease subtype. The analysis of such data typically relies on a multinomial logistic regression model (McCullagh and Nelder, 1989; Begg and Gray, 1984).

Under both matched and unmatched settings, the inference boils down to the estimation of $K-1$ parameter vectors. But, as mentioned above, commonalities are generally expected among disease subtypes. More precisely, some risk factors are likely to be shared by some subtypes, and these shared risk factors may have the same level of association across various subtypes. Then, the $K-1$ parameter vectors are expected to show some level of homogeneity, in the sense that some zeros are likely to be in the same positions, and that some non-zero values are likely identical across subtypes. Properly accounting for this particular structured sparsity (Bach et al., 2012) is key to reduce the complexity of the inference task and improve estimation efficiency (Viallon et al., 2016). Recently, DataShared Lasso has been introduced as a way to account for the expected homogeneity among the $K-1$ parameter vectors to be estimated under stratified regression models (Gross and Tibshirani, 2016; Ollier and Viallon, 2017).

In this article, we will show how the ideas of DataShared Lasso can be applied to analyze both matched and unmatched case-control studies with multiple disease subtypes. In Section 3.2, we consider stratified sparse conditional logistic models under matched designs, for which DataShared Lasso is naturally appealing. Section 3.3 is devoted to the unmatched setting and sparse multinomial logistic regression models, for which the link with DataShared Lasso is less obvious at first sight. Two formulations of sparse multinomial logistic regression models exist in the literature (Krishnapuram et al., 2005 ; Friedman et al., 2010), without clear guidance on how to chose between them. We will formally establish that one of these two formulations corresponds to a DataShared Lasso version of the other. In Section 3.4, we present results from a simulation study. Under both the matched and unmatched settings, our results illustrate the superiority of DataShared Lasso compared to its competitors when homogeneity exists among the parameter vectors to be estimated, in terms of prediction and estimation accuracy, as well as support recovery (i.e., the ability to identify the position of the non-zero entries of these vectors) and identification of heterogeneities among these vectors. Section 3.5 is devoted to our illustrative example. Concluding remarks are given in Section 3.6.

3.2 Matched case-control studies with multiple subtypes of cases and stratified conditional logistic models

Conditional logistic regression is a standard tool for the analysis of matched case-control studies when a single type of disease is considered (Pearce, 2016; Rothman et al., 2008). Here, we show how the ideas of DataShared Lasso can be applied to handle the situation where $K-1$ disease subtypes are present, for some given integer $K > 1$.

3.2.1 Setting

Consider a matched case-control study where information about subtype is available for each case. We denote the number of subtypes by $K-1$, for some given integer $K > 1$. For simplicity, we further assume a 1 :1 matched case-control design, and we denote by $m \geq 1$ the total number of pairs of individuals. Because each case has its own control, the total sample can be divided into $K-1$ subsamples. For any $k \in \{1, .., K-1\}$, the k-th subsample $\mathcal{M}_k$ is made of the m_k pairs composed by each case of subtype k and its matched control.

For any $\ell \in \{1, \ldots, m_k\}$, we let $\mathbf{x}^{(k)}_{\ell,case}$ and $\mathbf{x}^{(k)}_{\ell,control}$ denote the vectors of covariates (of length p) for the case and the control, respectively, in the ℓ-th matched pair of $\mathcal{M}_k$. We then have $Y^{(k)}_{\ell,case} = 1$ and $Y^{(k)}_{\ell,control} = 0$, which represent the disease indicators for the two individuals composing this matched pair. The association between covariates and disease subtype k can be studied by applying a conditional logistic regression model restricted to observations in $\mathcal{M}_k$. Under this model, we assume the existence of a vector $\boldsymbol{\delta}^*_k \in \mathbb{R}^p$ of true values of parameters such that the probability that the case is the one observed in pair ℓ, given that a case is observed in pair ℓ, is (Greenland, 2000)

$$\Pr(Y^{(k)}_{\ell,case} = 1 | Y^{(k)}_{\ell,case} + Y^{(k)}_{\ell,control} = 1, \mathbf{x}^{(k)}_{\ell,case}, \mathbf{x}^{(k)}_{\ell,control}) = \frac{\exp(\boldsymbol{\delta}^{*T}_k \mathbf{x}^{(k)}_{\ell,case})}{\exp(\boldsymbol{\delta}^{*T}_k \mathbf{x}^{(k)}_{\ell,case}) + \exp(\boldsymbol{\delta}^{*T}_k \mathbf{x}^{(k)}_{\ell,control})}. \tag{3.1}$$

Introduce $\mathbf{1}_{m_k} = (1, \ldots, 1)^T \in \mathbb{R}^{m_k}$ and let $\boldsymbol{\Delta}^{(k)}$ denote the $m_k \times p$ matrix whose ℓ-th row equals $(\mathbf{x}^{(k)}_{\ell,control} - \mathbf{x}^{(k)}_{\ell,case})$, for $\ell \in \{1, .., m_k\}$. Vector $\boldsymbol{\delta}^*_k$ can be estimated by maximizing the log conditional likelihood $L^{(cond)}_k$ restricted to pairs in $\mathcal{M}_k$, which is

defined for any vector $\boldsymbol{\delta}_k \in \mathbb{R}^p$ as

$$L_k^{(cond)}(\boldsymbol{\delta}_k) = -\sum_{\ell=1}^{m_k} \log[1 + \exp\{\boldsymbol{\delta}_k^T(\mathbf{x}_{\ell,control}^{(k)} - \mathbf{x}_{\ell,case}^{(k)})\}]$$
$$= -[\log\{\mathbf{1}_{m_k} + \exp(\boldsymbol{\Delta}^{(k)}\boldsymbol{\delta}_k)\}]^T\mathbf{1}_{m_k}. \tag{3.2}$$

Equivalently, vectors $\boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_{K-1}^*$ can be estimated simultaneously by maximizing the following global criterion over $\boldsymbol{\delta} = (\boldsymbol{\delta}_1^T, \ldots, \boldsymbol{\delta}_{K-1}^T)^T$,

$$L^{(cond)}(\boldsymbol{\Delta}_{In}, \boldsymbol{\delta}) = \sum_{k=1}^{K-1} L_k^{(cond)}(\boldsymbol{\delta}_k) = -[\log\{\mathbf{1}_m + \exp(\boldsymbol{\Delta}_{In}\boldsymbol{\delta})\}]^T\mathbf{1}_m, \tag{3.3}$$

with

$$\boldsymbol{\Delta}_{In} = \begin{pmatrix} \boldsymbol{\Delta}^{(1)} & \ldots & \mathbf{0}_{m_1,p} \\ \vdots & \ddots & \vdots \\ \mathbf{0}_{m_{K-1},p} & \ldots & \boldsymbol{\Delta}^{(K-1)} \end{pmatrix}.$$

For future use, observe that function $L^{(cond)}$ is defined for any pair $(\boldsymbol{\Delta}, \boldsymbol{\delta})$ with $\boldsymbol{\Delta} \in \mathbb{R}^{m\times d}$ and $\boldsymbol{\delta} \in \mathbb{R}^d$, for any integer $d \geq 1$. Moreover, estimation of the $K-1$ vectors $\boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_{K-1}^*$, is performed simultaneously but still independently when maximizing the above criterion. Coupling the estimation of the $K-1$ vectors, that is making the estimation of each vector to depend on each other, is deemed necessary to allow the estimates to share the same similarities as $\boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_{K-1}^*$ when such similarities exist. This can be achieved by using appropriate penalties, such as the one employed in DataShared Lasso presented below.

3.2.2 DataShared Lasso

DataShared Lasso was introduced by Gross and Tibshirani (2016) and Ollier and Viallon (2017) in the context of stratified regression models, as a way to account for the expected homogeneities among the parameter vectors to be estimated. The key to the approach is a reparametrization of the model. More precisely, instead of the original parametrization based on $\delta_{k,j}^*$, for $k \in \{1, ..., K-1\}$ and $j \in \{1, ..., p\}$, DataShared Lasso is based on the following over-parametrized decomposition

$$\delta_{k,j}^* = \mu_j^* + \gamma_{k,j}^*. \tag{3.4}$$

Here μ_j^* can be seen as the "global" parameter for covariate j and is common to all subtypes, while $\gamma_{k,j}^*$ captures the variation of the parameter for subtype k around this global parameter. As will be shown in Section 3.2.3, DataShared Lasso can be seen as a generalization of several more standard L_1-penalized approaches based on other parametrizations of the model, which correspond to particular constraints in decomposition (3.4).

Even if decomposition (3.4) is over-parametrized, estimates of μ_j^* and $\gamma_{k,j}^*$ for $k \in \{1, ..., K-1\}$ and $j \in \{1, ..., p\}$ can be derived by maximizing the following L_1-penalized criterion over $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$ and the $\boldsymbol{\gamma}_k$'s, with $\boldsymbol{\gamma}_k = (\gamma_{k,1}, \ldots, \gamma_{k,p})$,

$$\sum_{k=1}^{K-1} L_k^{(cond)}(\boldsymbol{\mu} + \boldsymbol{\gamma}_k) - \lambda(\|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K-1} \|\boldsymbol{\gamma}_k\|_1). \tag{3.5}$$

As usual, appropriate values of the tuning parameter λ can be obtained in practice by cross-validation (Bühlmann and Geer, 2011) or through the maximization of BIC-like criteria (Schwarz, 1978). We will refer to this approach as CondLogist_DataSharedLasso. The L_1-norm penalty $\|\boldsymbol{\mu}\|_1$ encourages sparsity of the vector of global parameters, while the $\|\boldsymbol{\gamma}_k\|_1$'s encourage homogeneity among vectors $\widehat{\boldsymbol{\delta}}_k$ defined as $\widehat{\boldsymbol{\delta}}_k = \widehat{\boldsymbol{\mu}} + \widehat{\boldsymbol{\gamma}}_k$, for $k \in \{1, ..., K-1\}$. Moreover, Gross and Tibshirani (2016) and Ollier and Viallon (2017) showed that optimal parameters especially satisfy

$$\widehat{\mu}_j = \underset{m}{\text{argmin}}\{|m| + \sum_{k=1}^{K-1} |\widehat{\delta}_{k,j} - m|\} = \text{median}(\widehat{\delta}_{1,j}, \ldots, \widehat{\delta}_{K-1,j}, 0).$$

In words, the estimated global parameter for covariate j corresponds to a shrunk version of the median of the estimated parameters for covariate j across disease subtypes. As a result, estimates $(\widehat{\boldsymbol{\delta}}_1, \ldots, \widehat{\boldsymbol{\delta}}_{K-1})$ produced by CondLogist_DataSharedLasso, are encouraged to be close to their shrunk median $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_p)$ in the L_1-norm sense, hence similar.

We shall stress that the penalty $\sum_{k=1}^{K-1} \|\boldsymbol{\gamma}_k\|_1$ can be generalized to $\sum_{k=1}^{K-1} \tau_k \|\boldsymbol{\gamma}_k\|_1$, for some $(\tau_k)_{k \geq 1}$, e.g., to penalize more heavily terms $\|\boldsymbol{\gamma}_k\|_1$ associated with larger sample size m_k. For simplicity, we focus on the case $\tau_k = 1$ here, and refer to Gross and Tibshirani (2016) and Ollier and Viallon (2017) for more details on the general case.

3.2.3 Implementation and relationship with more standard strategies

A first nice property of DataShared Lasso is that it can be written as a simple lasso, which makes it readily implementable. In particular, the DataShared Lasso criterion can be rewritten as

$$\sum_{k=1}^{K-1} L_k^{(cond)}(\boldsymbol{\mu}+\boldsymbol{\gamma}_k) - \lambda(\|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K-1}\|\boldsymbol{\gamma}_k\|_1) = -[\log\{\mathbf{1}_m + \exp(\boldsymbol{\Delta}_{DS}\boldsymbol{\Gamma})\}]^T\mathbf{1}_m - \lambda\|\boldsymbol{\Gamma}\|_1 \tag{3.6}$$

with $\boldsymbol{\Gamma} = (\boldsymbol{\mu}^T, \boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_{K-1}^T)^T \in \mathbb{R}^{K\times p}$ and

$$\boldsymbol{\Delta}_{DS} = \begin{pmatrix} \boldsymbol{\Delta}^{(1)} & \boldsymbol{\Delta}^{(1)} & \mathbf{0}_{m_1,p} & \ldots & \mathbf{0}_{m_1,p} \\ \boldsymbol{\Delta}^{(2)} & \mathbf{0}_{m_2,p} & \boldsymbol{\Delta}^{(2)} & \ldots & \mathbf{0}_{m_2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Delta}^{(K-1)} & \mathbf{0}_{m_{K-1},p} & \mathbf{0}_{m_{K-1},p} & \ldots & \boldsymbol{\Delta}^{(K-1)} \end{pmatrix}.$$

Here, $\boldsymbol{\Delta}^{(k)}$ still denotes the $m_k \times p$ matrix whose ℓ-th row equals $(\mathbf{x}_{\ell,control}^{(k)} - \mathbf{x}_{\ell,case}^{(k)})$. Criterion (3.6) corresponds to an L_1-penalized version of the log-likelihood (3.2) or (3.3), with design matrix $\boldsymbol{\Delta}_{DS}$ instead of $\boldsymbol{\Delta}^{(k)}$ or $\boldsymbol{\Delta}_{In}$. In other words, any solver for the L_1-penalized conditional logistic regression model can be used to implement CondLogist_DataSharedLasso. For instance, the cLogitLasso (Avalos et al., 2015) and cLogitL1 (Reid and Tibshirani, 2014) packages are available for R users.

In addition, this new writing of the DataShared Lasso criterion highlights its connection with three more standard approaches based on other reparametrizations of the model, and which correspond to particular constraints in decomposition (3.4). These standard approaches consist in maximizing a criterion similar to (3.6) above with $\boldsymbol{\Delta}_{DS}$ replaced, in turn, by $\boldsymbol{\Delta}_{In}$ given above, and $\boldsymbol{\Delta}_{Po}$ and $\boldsymbol{\Delta}_{Re}^{(1)}$ given by

$$\boldsymbol{\Delta}_{Po} = \begin{pmatrix} \boldsymbol{\Delta}^{(1)} \\ \boldsymbol{\Delta}^{(2)} \\ \vdots \\ \boldsymbol{\Delta}^{(K-1)} \end{pmatrix}, \quad \boldsymbol{\Delta}_{Re}^{(1)} = \begin{pmatrix} \boldsymbol{\Delta}^{(1)} & \mathbf{0}_{m_1,p} & \ldots & \mathbf{0}_{m_1,p} \\ \boldsymbol{\Delta}^{(2)} & \boldsymbol{\Delta}^{(2)} & \ldots & \mathbf{0}_{m_2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Delta}^{(K-1)} & \mathbf{0}_{m_{K-1},p} & \ldots & \boldsymbol{\Delta}^{(K-1)} \end{pmatrix}.$$

First, consider the constraint $\mu_j^* = 0$ for all $j \in \{1, ..., p\}$ in decomposition (3.4). In this case, the reparametrization is simply a change of notation compared with the original

parametrization : $\delta^*_{k,j} = \gamma^*_{k,j}$. The constraint $\mu^*_j = 0$ for all $j \in \{1, ..., p\}$ can be imposed in criterion (3.6) simply by eliminating the first block of p columns in $\boldsymbol{\Delta}_{DS}$, that is by replacing $\boldsymbol{\Delta}_{DS}$ by $\boldsymbol{\Delta}_{In}$. As detailed in Appendix B.1 in the Supplementary Materials, under this additional contraint, CondLogist_DataSharedLasso reduces to the simple approach which consists in running one L_1-penalized conditional logistic regression (that is an L_1-penalized version of criterion (3.2)) on each subsample $\mathcal{M}_k$ independently, and so we refer to this approach as CondLogist_IndepLasso. Second, consider the constraint $\gamma^*_{k,j} = 0$ for all $k \in \{1, ..., K-1\}$ and $j \in \{1, ..., p\}$. In this case, $\delta^*_{k,j} = \mu^*_j$ for all k : working under this constraint corresponds to assuming that vectors $\boldsymbol{\delta}^*_1, \ldots, \boldsymbol{\delta}^*_K$ are all equal to a common vector, $\boldsymbol{\mu}^*$. This vector $\boldsymbol{\mu}^* \in \mathbb{R}^p$ can again be estimated by maximizing the same criterion as (3.6), this time after eliminating the $K-1$ last blocks of p columns in $\boldsymbol{\Delta}_{DS}$, that is after replacing $\boldsymbol{\Delta}_{DS}$ by $\boldsymbol{\Delta}_{Po}$. This corresponds to pooling all the subsamples together, and we will refer to this approach as CondLogist_PooledLasso. Finally, consider the constraint $\gamma^*_{1,j} = 0$ for all $j \in \{1, ..., p\}$. In this case, we have $\mu^*_j = \delta^*_{1,j}$ and $\gamma^*_{k,j} = \delta^*_{k,j} - \delta^*_{1,j}$ for all $j \in \{1, ..., p\}$ and $k > 1$. The $(K-1) \times p$ parameters $\mu^*_j (= \delta^*_{1,j})$ and $\gamma^*_{k,j}$ for $j \in \{1, ..., p\}$ and $k \geq 2$, can be estimated by maximizing the same criterion as (3.6), after eliminating the second block of p columns in $\boldsymbol{\Delta}_{DS}$, that is after replacing $\boldsymbol{\Delta}_{DS}$ by $\boldsymbol{\Delta}^{(1)}_{Re}$. This corresponds to working under the decomposition $\boldsymbol{\delta}^*_k = \boldsymbol{\delta}^*_1 + \boldsymbol{\gamma}^*_k$ for $k \geq 2$. In other words, this corresponds to considering the first subtype as the reference subtype, while parameter $\gamma^*_{k,j}$, for $j \in \{1, ..., p\}$ and $k \geq 2$, captures how the association of covariate j and subtype k differs from that of covariate j and subtype 1. We will refer to this approach as CondLogist_RefLasso. Of course, any subtype r can be considered as the reference, not necessarily the first one.

Each of these three more standard approaches, CondLogist_IndepLasso, CondLogist_PooledLasso and CondLogist_RefLasso, can therefore be regarded as one particular constrained version of CondLogist_DataSharedLasso, where the additional constraint makes decomposition (3.4) identifiable. However, the flexibility of the over-parametrization on which CondLogist_DataSharedLasso relies makes the approach generally better than the other three, as we now explain. First, the parametrization used in CondLogist_PooledLasso is not flexible enough to account for subtype specificities, and then results in biased estimates unless all vectors $\boldsymbol{\delta}^*_k$ are equal. On the other hand, the parametrizations used in CondLogist_IndepLasso and CondLogist_RefLasso are flexible enough to avoid such a

bias. But, as detailed in Ollier and Viallon (2017), these parametrizations are still suboptimal, because they generally involve unnecessarily large numbers of non-zero true parameters. As a matter of fact, the optimal parametrization of the form (3.4) is such that $\|\boldsymbol{\mu}^*\|_0 + \sum_k \|\boldsymbol{\delta}_k^* - \boldsymbol{\mu}^*\|_0$ is minimized, with $\|\cdot\|_0$ standing for the L_0 pseudo-norm. The optimal choice for μ_j^* is therefore $\delta_{r_j,j}^*$ for any $r_j \in \{1, ..., K-1\}$ such that $\delta_{r_j,j}^*$ is the mode of the collection of values $(\delta_{1,j}^*, \ldots, \delta_{K-1,j}^*, 0)$. In other words, the optimal parametrization of the form (3.4) relies on optimal covariate-specific references. The corresponding optimal version of CondLogist_RefLasso, applied with such optimal covariate-specific references, can of course not be implemented in practice because these optimal covariate-specific references are unknown. But, in the setting of stratified linear models, the DataShared Lasso strategy was shown to target the same parametrization as this optimal version of CondLogist_RefLasso (Ollier and Viallon, 2017). It was further shown to perform as well as this optimal version of CondLogist_RefLasso, and to outperform the three more standard approaches, both theoretically and empirically (Ollier and Viallon, 2017). Results from our simulation study presented in Section 3.4 will confirm those described in Ollier and Viallon (2017) under linear regression models. In particular, the strategy based on the DataShared Lasso penalty usually better accounts for homogeneity than the other three appraoches, which translates into better estimation and prediction accuracy, overall support recovery and identification of heterogeneities.

3.3 Unmatched case-control studies with multiple subtypes of cases and sparse multinomial logistic models

We now turn our attention to the unmatched setting. When $K-1$ subtypes of cases are present for some given integer $K > 1$, the outcome Y can be modeled as a categorical variable, taking values in $\{1, ..., K\}$. Hereafter, we will assume that $Y = K$ for controls, while $Y = k$ for cases of subtype k, for any $k \in \{1, ..., K-1\}$. When no natural order exists among the categories of Y, the multinomial logistic regression model is a natural extension of the standard logistic regression model. Below, we will recall some basics about the multinomial logistic regression model. We will first introduce the L_1-penalized approach based on the symmetric formulation of the model, as implemented in the popular `glmnet` R package (Friedman et al., 2010). We will then show that it corresponds to a DataShared Lasso version of the more standard formulation, which relies on the initial

choice of a reference category. For ease of notation, we will mostly focus on models with no intercept. Our presentation would mainly be the same if intercepts were considered, except that intercept terms are generally not penalized, and L_1-norms $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$ would be replaced by $\sum_{j=2}^{p} |\beta_j|$ if β_1 corresponds to the intercept. See the last paragraph in Section 3.3.1 for additional details.

3.3.1 The multinomial logistic regression model

For any collection of vectors $(\mathbf{u}_1, \ldots, \mathbf{u}_K) \in \mathbb{R}^{p \times K}$, any $k \in \{1, ..., K\}$, and any $\mathbf{x}_0 \in \mathbb{R}^p$ for some $p \geq 1$, introduce the function $p_k(\mathbf{x}_0; \mathbf{u}_1, \ldots, \mathbf{u}_K) = \exp(\mathbf{x}_0^T \mathbf{u}_k) / \{\sum_{\ell=1}^{K} \exp(\mathbf{x}_0^T \mathbf{u}_\ell)\}$. In its symmetric formulation, the multinomial logistic regression model assumes the existence of K vectors $(\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*) \in \mathbb{R}^{p \times K}$ of true values of parameters such that

$$\Pr(Y = k | \mathbf{x} = \mathbf{x}_0) = \frac{\exp(\mathbf{x}_0^T \boldsymbol{\beta}_k^*)}{\sum_{\ell=1}^{K} \exp(\mathbf{x}_0^T \boldsymbol{\beta}_\ell^*)} = p_k(\mathbf{x}_0; \boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*), \tag{3.7}$$

for any value $\mathbf{x}_0 \in \mathbb{R}^p$ of the covariate vector. Because $\sum_{k=1}^{K} \Pr(Y = k | \mathbf{x} = \mathbf{x}_0) = 1$ for any $\mathbf{x}_0 \in \mathbb{R}^p$, this formulation is over-parametrized and vectors $\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*$ in Equation (3.7) are defined up to a constant only. Indeed, for any $\boldsymbol{\nu} \in \mathbb{R}^p$ and any $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) \in \mathbb{R}^{p \times K}$, $p_k(\mathbf{x}_0; \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) = p_k(\mathbf{x}_0; \boldsymbol{\beta}_1 + \boldsymbol{\nu}, \ldots, \boldsymbol{\beta}_K + \boldsymbol{\nu})$. In other words, if model (3.7) holds with vectors $(\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*)$, then it holds with vectors $(\boldsymbol{\beta}_1^* + \boldsymbol{\nu}, \ldots, \boldsymbol{\beta}_K^* + \boldsymbol{\nu})$ for any $\boldsymbol{\nu} \in \mathbb{R}^p$ as well. For future use, note that it especially holds with vectors $(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_K^*, \ldots, \boldsymbol{\beta}_{K-1}^* - \boldsymbol{\beta}_K^*, \mathbf{0}_p)$, which corresponds to the particular choice $\boldsymbol{\nu} = -\boldsymbol{\beta}_K^*$. Because of this lack of identifiability, standard maximum likelihood estimation based on this parametrization can not be used to derive estimates of $\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*$, and constrained or penalized versions of the likelihood have to be used instead. In particular, the `glmnet` R package (Friedman et al., 2010) produces estimates defined as maximizers of the L_1-penalized version of the log-likelihood

$$L(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) - \lambda \sum_{k=1}^{K} \|\boldsymbol{\beta}_k\|_1 = \frac{1}{n} \sum_{i=1}^{n} \log\{p_{y_i}(\mathbf{x}_i; \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)\} - \lambda \sum_{k=1}^{K} \|\boldsymbol{\beta}_k\|_1 \tag{3.8}$$

for an appropriate value of the regularization parameter λ. We will refer to this approach as MultinomLogist_SymLasso. It works under the implicit assumption that (at least) one of the infinitely many collections of vectors $\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*$ satisfying (3.7) is sparse, and looks for the "sparsest", or more precisely, the one with lowest $\sum_k \|\boldsymbol{\beta}_k^*\|_1$. In particular,

Friedman et al. (2010) show that maximizers $\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K$ of criterion (3.8) are such that

$$\text{median}(\widehat{\beta}_{1,j}, \ldots, \widehat{\beta}_{K,j}) = 0, \quad \text{for all } j \in \{1, ..., p\}. \tag{3.9}$$

Equation (3.9) establishes that the L_1-norm penalization solves the lack of identifiability for each covariate by targeting a collection of vectors $\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K$ such that, for each covariate, the median of its parameters across the K categories is null. As mentioned above, when intercepts are considered, they are generally not penalized, in which case the lack of identifiability remains for them. In `glmnet`, this is resolved by mean centering, which corresponds to imposing the constraint $\sum_{k=1}^{K} \hat{\beta}_{k,1} = 0$ (Friedman et al., 2010), with $\hat{\beta}_{k,1}$ standing for the intercept estimate for the k-th category.

3.3.2 Relationship with DataShared Lasso

Now, let us turn our attention to the "standard" formulation of the multinomial logistic regression model, which resolves the lack of identifiability of the symmetric one by first selecting a reference category, typically K. Then, this formulation assumes the existence of $K-1$ parameter vectors, say $\boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_{K-1}^*$, such that $\Pr(Y = k|\mathbf{x} = \mathbf{x}_0) = p_k(\mathbf{x}_0, \boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_{K-1}^*, \mathbf{0}_p)$. The two formulations – symmetric and standard – are strictly equivalent. Indeed, and as mentioned above, for any $\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*$ satisfying the symmetric formulation of the model, vectors $\boldsymbol{\delta}_1^*, \ldots \boldsymbol{\delta}_{K-1}^*$ defined as $\boldsymbol{\delta}_k^* = \boldsymbol{\beta}_k^* - \boldsymbol{\beta}_K^*$ for $k \in \{1, ..., K-1\}$ satisfy the standard one. When the dimension p of the covariates is large, the expected sparsity within vectors $(\boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_{K-1}^*)$ can be accounted for by looking for estimates maximizing an L_1-penalized log-likelihood (Krishnapuram et al., 2005)

$$\frac{1}{n} \sum_{i=1}^{n} \log\{p_{y_i}(\mathbf{x}_i; \boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{K-1}, 0)\} - \lambda \sum_{k=1}^{K-1} \|\boldsymbol{\delta}_k\|_1.$$

We will refer to this approach as MultinomLogist_StdLasso. The ideas of DataShared Lasso can further be applied to account for the homogeneity among vectors $(\boldsymbol{\delta}_k^*)$'s when the subtypes are expected to share commonalities. Considering as in Section 3.2 the decomposition $\boldsymbol{\delta}_k = \boldsymbol{\mu} + \boldsymbol{\gamma}_k$ for $k \in \{1, ..., K-1\}$, the method we will refer to as

MultinomLogist_StdDataSharedLasso then simply consists in maximizing the criterion

$$\frac{1}{n}\sum_{i=1}^{n}\log\{p_{y_i}(\mathbf{x}_i;\boldsymbol{\mu}+\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\mu}+\boldsymbol{\gamma}_{K-1},\mathbf{0}_p)\}-\lambda\Big(\|\boldsymbol{\mu}\|_1+\sum_{k=1}^{K-1}\|\boldsymbol{\gamma}_k\|_1\Big).$$

Interestingly, this criterion is exactly the same as the one in Equation (3.8), after using the change of variable $\boldsymbol{\mu}=-\boldsymbol{\beta}_K$ and $\boldsymbol{\gamma}_k=\boldsymbol{\beta}_k$ for all $k<K$; see Appendix B.2 in the Supplementary Materials for the detailed derivation of this result. This equality formally establishes that working under the symmetric formulation (3.7) with an L_1-norm penalty, as in the `glmnet` R package, exactly corresponds to working under the more standard formulation with a DataShared Lasso penalty to encourage homogeneity among vectors $(\boldsymbol{\delta}_1^*,\ldots,\boldsymbol{\delta}_{K-1}^*)$. More precisely, the estimates $(\widehat{\boldsymbol{\beta}}_1,\ldots,\widehat{\boldsymbol{\beta}}_K)$ and $(\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\gamma}}_1,\ldots,\widehat{\boldsymbol{\gamma}}_{K-1})$ produced by MultinomLogist_SymLasso and MultinomLogist_StdDataSharedLasso, respectively, are such that $\widehat{\boldsymbol{\mu}}=-\widehat{\boldsymbol{\beta}}_K$ and $\widehat{\boldsymbol{\beta}}_k=\widehat{\boldsymbol{\gamma}}_k$ for all $k\in\{1,...,K-1\}$.

This equivalence between MultinomLogist_SymLasso and MultinomLogist_StdDataSharedLasso further allows the derivation of guidance on whether to use MultinomLogist_SymLasso or MultinomLogist_StdLasso in practice : by by-passing the arbitrary choice of the reference category, MultinomLogist_SymLasso will typically target a sparser parametrization than MultinomLogist_StdLasso if disease subtypes share commonalities, and is then expected to produce better estimates. MultinomLogist_StdDataSharedLasso can be seen as a way to compensate any suboptimal choice of the reference category in MultinomLogist_StdLasso. Although different at first sight, MultinomLogist_SymLasso and MultinomLogist_StdDataSharedLasso produce the same estimates and we will simply refer to any of them as MultinomLogist_SymLasso in the rest of our article.

3.4 Simulation study

3.4.1 Evaluation criteria

To compare the performance of the considered approaches (under both the matched and unmatched settings), several criteria are evaluated. Given estimates $\hat{\boldsymbol{\delta}}_1,\ldots,\hat{\boldsymbol{\delta}}_{K-1}$ of true vectors of parameters $\boldsymbol{\delta}_1^*,\ldots,\boldsymbol{\delta}_{K-1}^*$ (under the unmatched setting, they correspond to

vectors involved in the standard formulation with controls as the reference category), a first common criterion when evaluating L_1-penalized approaches is the accuracy with respect to support recovery, which measures the ability to correctly identify patterns of null, positive and negative entries in the vector of parameters to be estimated. In our context, we consider the following criterion :

$$\text{Sgn\_Accuracy} = \frac{\sum_{k=1}^{K-1}\sum_{j=1}^{p}\left(\mathbb{1}[\text{sgn}(\delta^*_{k,j}) = \text{sgn}(\hat{\delta}_{k,j})] - \mathbb{1}[\text{sgn}(\delta^*_{k,j}) \times \text{sgn}(\hat{\delta}_{k,j}) = -1]\right)}{(K-1)p},$$

where $\text{sgn}(x) = +1$ if $x > 0$, $\text{sgn}(x) = -1$ if $x < 0$ and $\text{sgn}(x) = 0$ if $x = 0$. This criterion is a slight modification of the standard accuracy (Metz, 1978; Viallon et al., 2016), where the term $-\mathbb{1}[\text{sgn}(\delta^*_{k,j}) \times \text{sgn}(\hat{\delta}_{k,j}) = -1]$ is included to penalize approaches that tend to produce positive [resp. negative] estimate while the true value is negative [resp. positive], since this is particularly unwanted in practice. Good approaches are expected to have a high Sgn_Accuracy.

In our framework, vectors $\boldsymbol{\delta}^*_1, ..., \boldsymbol{\delta}^*_{K-1}$ are not only expected to be sparse. They may also have some zeros in the same positions, and some non-zero entries may be equal for different subtypes. Estimates $\hat{\boldsymbol{\delta}}_1, \ldots, \hat{\boldsymbol{\delta}}_{K-1}$ should share the same structure to be able to identify heterogeneities. For any $j \in \{1, ..., p\}$, a good approach should then be able to produce estimates $\hat{\delta}_{1,j}, \ldots, \hat{\delta}_{K-1,j}$ such that, for any $(k_1 \neq k_2) \in \{1, ..., K-1\}^2$, $\hat{\delta}_{k_1,j} = \hat{\delta}_{k_2,j}$ if and only if $\delta^*_{k_1,j} = \delta^*_{k_2,j}$. One standard criterion to evaluate this capacity is the Rand Index (Rand, 1971), which is defined in our context as

$$\text{RandIndex} = \frac{\sum_{j=1}^{p}\sum_{k1=1}^{K-2}\sum_{k2>k1}^{K-1}\left(\mathbf{1}[\delta^*_{k1,j} = \delta^*_{k2,j}, \hat{\delta}_{k1,j} = \hat{\delta}_{k2,j}] + \mathbf{1}[\delta^*_{k1,j} \neq \delta^*_{k2,j}, \hat{\delta}_{k1,j} \neq \hat{\delta}_{k2,j}]\right)}{p(K-2)!}.$$

Again, good approaches are expected to have a high RandIndex.

We also evaluate the approaches with respect to estimation error and prediction accuracy. As for the estimation error, we used the following criterion, which should be as low as possible

$$\text{Est\_Error} = \sum_{k=1}^{K-1} \frac{\|\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}^*_k\|_2^2}{\|\boldsymbol{\delta}^*_k\|_2^2}.$$

As for the prediction accuracy, we computed an AUC-like criterion, which was adapted to our matched and unmatched settings. Under both settings, our AUC compares predicted probabilities with observed outcomes on an independent test sample of size $n^{(test)} =$

10,000. In the matched setting, our AUC is defined as the weighted average of the AUCs computed in each subsample $\mathcal{M}_k^{(test)}$. In the unmatched setting, we adapted the one class versus all other classes approach (Provost and Domingos, 2000; Fawcett, 2006) ; see Appendix B.3 in the Supplementary Material for details on this adaptation. In either setting, good approaches are expected to have a high AUC.

3.4.2 The matched setting

We performed a simulation study to assess the performance of DataShared Lasso in the context of matched case-control studies. We compared CondLogist_DataSharedLasso with CondLogist_IndepLasso, CondLogist_PooledLasso, and CondLogist_RefLasso. For the latter, the first subtype was selected as the reference. We set the number of covariates to $p = 100$, and the number of disease subtypes to $(K - 1) = 6$. We further set the number of pairs of observations in each subsample to $m_1 = 200$, $m_2 = 100$ and $m_k = 50$ for $k = 3, \ldots, 6$, so that the total number of observations was $n = 1000$. In this "high"-dimensional setting, we implemented a cross-validation technique in the spirit of the one-step lasso (Bühlmann and Meier, 2008) to select the optimal regularization parameters and obtain the final parameter estimates.

Here, we briefly describe the simulation designs we considered. Additional details are provided in Appendix B.4 of the Supplementary Materials. Four configurations corresponding to four levels of homogeneity among vectors $\boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_6^*$ were considered : full homogeneity (the 6 vectors are equal), low heterogeneity ($\boldsymbol{\delta}_2^*, \ldots, \boldsymbol{\delta}_6^*$ are equal, and $\boldsymbol{\delta}_1^*$ is different from them), moderate heterogeneity ($\boldsymbol{\delta}_4^*, \boldsymbol{\delta}_5^*$, and $\boldsymbol{\delta}_6^*$ are equal, while the three other vectors are different from them, and from each other) and full heterogeneity (the 6 vectors have nothing in common). We shall stress that under the low and moderate heterogeneity configurations, the first subtype is the worst choice for the reference used in CondLogist_RefLasso, in the sense that $\|\boldsymbol{\delta}_r^*\|_0 + \sum_{k \neq r} \|\boldsymbol{\delta}_k^* - \boldsymbol{\delta}_r^*\|_0$ is maximized for $r = 1$. The comparison between the performance of CondLogist_RefLasso and CondLogist_DataSharedLasso will allow the assessment of the impact of a suboptimal choice for the reference when applying CondLogist_RefLasso.

To illustrate the relative performance of the approaches as a function of signal strength, we made it vary through a parameter $\delta \in \{0.1, 0.25, 0.5, 0.75\}$, which determines the magnitude of the non-zero true parameters of our generating model, and is related to

the log-odds-ratio for an increase of one standard-deviation of the corresponding covariates ; see Appendix B.4 in the Supplementary Materials for more details. Under each of the four configurations, and for each of the four signal strengths, we generated 200 samples under model (3.1). Figure 3.1 presents the criteria averaged over these 200 replicates, along with the 95% confidence intervals, for each of the four methods we compared here, that is CondLogist_DataSharedLasso, CondLogist_IndepLasso, CondLogist_PooledLasso, and CondLogist_RefLasso. Boxplots showing the distribution of the criteria over the 200 replicates for each method, under each configuration and for each signal strength, are presented on Figure B.1 of the Supplementary Materials. First consider the case of full homogeneity. Because all the vectors $\boldsymbol{\delta}_k^*$ are equal, the optimal strategy is of course CondLogist_PooledLasso, which is based on a parametrization with p_0 non-zero parameters (where p_0 is the number of non-zero parameters in each vector $\boldsymbol{\delta}_k^*$; we have $p_0 = 10$ in our simulation study). On the other hand, because CondLogist_IndepLasso is based on a parametrization with $(K-1)p_0$ non-zero parameters, it performs poorly compared to CondLogist_PooledLasso in this configuration, in terms of the four criteria we considered : because it is unable to account for homogeneity, estimates produced by CondLogist_IndepLasso are fully heterogeneous (its RandIndex is very low, as expected), hence with a large variance, and performs poorly in terms of estimation and prediction accuracy, and also support recovery (because it is unable to borrow strength from the various subtypes). On the other hand, both CondLogist_DataSharedLasso and CondLogist_RefLasso account for homogeneity, and perform nearly as well as CondLogist_PooledLasso in terms of each of our criteria under this configuration of full homogeneity. It is noteworthy that in this particular case, any subtype is an optimal reference in CondLogist_RefLasso ($\|\boldsymbol{\delta}_r^*\|_0 + \sum_{k\neq r} \|\boldsymbol{\delta}_k^* - \boldsymbol{\delta}_r^*\|_0 = p_0$ for any r), which explains why CondLogist_DataSharedLasso and CondLogist_RefLasso perform similarly in this case. Next, in the case of low heterogeneity, CondLogist_PooledLasso produces biased estimates and is not optimal since vectors $\boldsymbol{\delta}_k^*$ are not all equal anymore. Interestingly, CondLogist_RefLasso does not outperform CondLogist_PooledLasso in this case, and these two approaches actually produce very similar estimates under this configuration. This is due to the particular choice for the reference subtype in CondLogist_RefLasso : when $\boldsymbol{\delta}_2^* = \ldots = \boldsymbol{\delta}_6^*$, and $\boldsymbol{\delta}_1^*$ is different from them, the penalty term $\sum_k \|\boldsymbol{\delta}_k - \boldsymbol{\delta}_1\|_1$ generally prevents the approach to identify these heterogeneities. As a matter of fact, any other choice for the reference would have led to better performance for CondLogist_RefLasso. As mentioned above, CondLogist_DataSharedLasso bypasses the arbitrary choice of the

reference, and mimics the optimal version of CondLogist_RefLasso applied with an optimal, possibly covariate-specific, reference. Under this low heterogeneity configuration, CondLogist_DataSharedLasso allows the identification of heterogeneities (its RandIndex is higher than that of CondLogist_RefLasso and CondLogist_PooledLasso), and substantially outperforms the other approaches with respect to all criteria. As the level of heterogeneity increases, the complexity of the estimation task increases, and the performance of CondLogist_DataSharedLasso tends to that of CondLogist_IndepLasso. But, as long as some level of homogeneity is present (moderate heterogeneity configuration), CondLogist_DataSharedLasso outperforms the other approaches. Under the full heterogeneity configuration, CondLogist_DataSharedLasso still performs on average as well as CondLogist_IndepLasso, which is the optimal strategy in this case, while CondLogist_PooledLasso, and to a lesser extent CondLogist_RefLasso, perform worse.

Overall, our results illustrate that the performance of CondLogist_IndepLasso does not depend on the level of heterogeneity, in terms of support recovery, prediction accuracy and estimation accuracy. In the total absence of homogeneity, this performance is optimal. But, as the level of homogeneity increases, methods that account for homogeneity can target better (*i.e.*, sparser) parametrizations, and yield substantial improvements in terms of estimation performance. Among the four approaches we compared here, CondLogist_DataSharedLasso appears as the best approach to account for homogeneity when it is present. In addition, it performs as well as CondLogist_IndepLasso on average when no homogeneity is present at all.

3.4.3 The unmatched setting

We further performed a simulation study in the unmatched setting to illustrate the relative performance of MultinomLogist_StdLasso and MultinomLogist_SymLasso (the later being the same as MultinomLogist_StdDataSharedLasso), depending on the level of homogeneity among vectors $\boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_{K-1}^*$ of the standard formulation. We again set $K-1 = 6$ disease subtypes, and considered four configurations : full homogeneity, low heterogeneity, moderate heterogeneity and full heterogeneity. To save computational time, a low-dimensional setting with $n = 1000$ and $p = 20$ was considered here. To generate the data, we adapted the framework described in Section 3.4.2 to the unmatched setting. We used intercept terms, $(\delta_{1,0}, \ldots, \delta_{K-1,0})$, chosen in such a way that $\Pr(Y = K) = 0.5$

and $\Pr(Y = k)$ ranged from 0.05 to 0.2 for $k \in \{1, ..., K-1\}$. In this low-dimensional setting, regularization parameters were selected as minimizers of the BIC after adapting the Lasso-OLS hybrid ideas to our context (Efron et al., 2004), in the same way as in Viallon et al. (2016).

Figure 3.2 presents the criteria averaged over 200 replicates, along with their 95% confidence intervals. Boxplots summarizing the full distribution of the criteria over the 200 replicates are presented in Figure B.2 of the the Supplementary Materials. Overall, the conclusions drawn from the comparison between MultinomLogist_SymLasso and MultinomLogist_StdLasso in this unmatched setting are consistent with those drawn when comparing CondLogist_DataSharedLasso with CondLogist_IndepLasso in the matched setting. More precisely, the two methods perform similarly in case of full heterogeneity, but the performance of MultinomLogist_SymLasso improves as the level of homogeneity increases, while that of MultinomLogist_StdLasso remains roughly unchanged. In particular, MultinomLogist_SymLasso substantially outperforms MultinomLogist_StdLasso with respect to all criteria in the case of full homogeneity. This was expected since the number of non-zero parameters to be estimated under the standard formulation is $(K-1)p_0$ (where p_0 is the number of non-zero parameters in each $\boldsymbol{\delta}_k^*$; this was set to $p_0 = 10$ in our simulation study), while MultinomLogist_SymLasso (or equivalently MultinomLogist_StdDataSharedLasso) is able to target a parametrization with only p_0 non-zero parameters in the case of full homogeneity; see Appendix B.6 in the Supplementary Materials, for more details. Just as in the matched setting, our results confirm that using DataShared Lasso (or, equivalently, the symmetric formulation in this unmatched setting) allows the homogeneity to be accounted for when present, which translates into better estimation and prediction accuracy, support recovery and identification of heterogeneties.

3.5 Application

3.5.1 Data description

The European Prospective Investigation into Cancer and Nutrition (EPIC) study is an ongoing multicenter prospective study aiming to investigate prospectively the etiology of cancer in relation to diet, lifestyle and environmental factors. Its design has been

previously described in detail (Riboli et al., 2002). From 1992 to 2000, a total of 521,324 participants were recruited across 10 European countries. Among these participants, 246,000 women, aged from 35 to 70 years, provided a blood sample at inclusion. Here, we present preliminary results from the analysis of a case-control study nested in EPIC, whose main objective was to assess the association between metabolites and the risk of subtypes of breast cancer for women older than 50 : 1415 cases of breast cancer were included, along with 1415 matched controls (using incidence density sampling). We shall stress that the methods presented in Section 3.2 can be applied when the case-control study is nested within a cohort, as is the case here. This is because the analysis of the k-th disease subtype still relies on a conditional logistic regression model with parameter $\boldsymbol{\delta}_k^*$, which measures the level of association between the covariates and disease subtype k.

For these 2830 individuals, plasma samples collected at inclusion in the study were analyzed by mass spectrometry (AbsoluteIDQ p180 Kit) allowing the measurement of the concentrations of 127 metabolites. These concentrations were log-transformed to reduce skewness. We considered six histological subtypes of breast cancer, based on the presence/absence of hormone receptors : HER2-enriched (100 pairs of observations), triple negative (134 pairs), Luminal A PR- (164 pairs), Luminal A PR+ (820 pairs), Luminal B PR- (58 pairs) and Luminal B PR+ (139 pairs).

3.5.2 Results

Figure 3.3 provides a graphical representation of the log odds-ratio estimates $\hat{\boldsymbol{\delta}}_1, \ldots, \hat{\boldsymbol{\delta}}_6$ produced by each of the four methods for the 6 subtypes of breast cancer. For 79 out of the 127 measured metabolites, all methods produced a zero estimate for all subtypes. These "constantly" null estimates are not reported on Figure 3.3 to improve legibility. Also, the remaining 48 metabolites were anonymized as the biological interpretation of the results is out of the scope of this preliminary analysis. When analyzing such data, most practitioners would start by pooling all subtypes together (that is, ignoring subtypes) to identify metabolites associated with breast cancer as a whole. In this application, CondLogist_PooledLasso does identify several metabolites associated with breast cancer, which naturally raises the question of whether these identified metabolites (and maybe other ones as well) may be more specifically associated with particular subtypes.

The independent analyses of each subtype, as implemented in CondLogist_IndepLasso, identifies many metabolites associated with the Luminal A PR+ subtype, and fewer metabolites for the other subtypes. In particular, no metabolite is identified for the Luminal B PR- and the HER2-enriched subtypes. Moreover, very few metabolites were found to be associated with more than one subtype : to name a few exceptions, M96 appeared to be associated with both Luminal A PR+ and Luminal A PR-, and M28 with Luminal A PR+ and Triple Negative. Clearly, this heterogeneity across the subtypes can be the result of a combination of : (*i*) true heterogeneities, (*ii*) lack of power for some subtypes (many metabolites are identified in the case of Luminal A PR+, which is the most frequent subtype, while no metabolite is identified for Luminal B PR- or the HER2-enriched which are the two least frequent subtypes), and (*iii*) sample variability combined with correlations among the metabolites. Indeed, if two metabolites are strongly correlated, CondLogist_IndepLasso will typically identify one or the other on two different samples even if these samples are drawn from the same population (that is, in the absence of true heterogeneity between the two samples). In other words, and just as in subgroup analyses (Wang et al., 2007), it is hazardous to claim and interpret heterogeneities on the basis of the independent analyses of subtypes. Because heterogeneities are penalized when using CondLogist_DataSharedLasso (and, in a less optimal way, when using CondLogist_RefLasso), heterogeneities identified by CondLogist_DataSharedLasso are supported by the data, and are more likely true ones. In the present application, CondLogist_DataSharedLasso produces estimates that are quite similar to those produced by CondLogist_PooledLasso, suggesting that the data does not support departure from homogeneity in the levels of association between most metabolites and breast cancer across subtypes. A few heterogeneities are identified though, suggesting that some metabolites might be more specifically associated with the Luminal A PR+ subtype (M18, M27, M42, M43, M63 but also M111 whose association with other subtypes exist, but is stronger with Luminal A PR+), or Luminal A PR- (M96). The comparison with the results produced by CondLogist_RefLasso is also instructive, in particular the estimates produced for M18 and M63. Because Luminal A PR+ was chosen as the reference when applying CondLogist_RefLasso, it is here unable to identify any heterogeneity for this particular subtype, which is consistent with the results of our simulation study under the low heterogeneity configuration.

3.6 Discussion

In this article, we considered the analysis of high-dimensional case-control studies, when several disease subtypes exist, under both unmatched and matched settings. In the latter case, our analysis further covers matched case-control studies nested within a cohort. We have shown that estimation and prediction accuracy, support recovery and the ability to identify heterogeneities across subtypes, could all be substantially improved when commonalities exist among subtypes, provided methods that properly account for these commonalities, e.g. those based on the DataShared Lasso penalty, are used. Our findings are in line with the empirical and theoretical results of Ollier and Viallon (2017) in the case of stratified linear regression models, as well as the empirical results of Ballout and Viallon (2019) for stratified binary graphical models.

Under matched designs, the original parametrization relies on $K-1$ vectors $\boldsymbol{\delta}_k^*$, which represent the log odds-ratios that compare each of the $K-1$ disease subtypes with controls. Based on an over-parametrized reparametrization, CondLogist_DataSharedLasso is able to target a sparser parametrization when commonalities exist among subtypes, which can yield substantial improvements in terms of estimation efficiency. In the absence of commonalities, it still performs as well as the standard, independent analysis of each subtype. Under unmatched designs, the standard formulation of multinomial logistic regression models relies on the same parametrization, involving $K-1$ vectors $\boldsymbol{\delta}_k^*$ that compare each disease subtype with controls. We formally established that applying the ideas of DataShared Lasso along with this parametrization was actually equivalent to applying a standard lasso on the symmetric formulation of the model. This symmetric formulation relies on an over-parametrized parametrization with K vectors $\boldsymbol{\beta}_k^*$, and takes advantage of the fact that controls do not necessarily have to be considered as the reference category in unmatched settings. Again, the resulting parametrization can be much sparser than the standard one, and yields generally better estimation efficiency, especially when the level of homogeneity among subtypes is high.

The methods we presented to account for potential commonalities are simple to implement under both designs. Under matched designs, CondLogist_DataSharedLasso is as easy to implement as CondLogist_RefLasso or CondLogist_IndepLasso. Under unmatched designs, MultinomLogist_SymLasso (which is equivalent to MultinomLogist_StdDataSharedLasso) is implemented in the `glmnet` R package. Given the simplicity

of their implementation and the possibly substantial gain in terms of estimation performance, we strongly encourage the use of these approaches when analyzing case-control studies with several disease subtypes.

As pointed out in our application to the EPIC data, the methods that account for potential commonalities are especially useful to claim and interpret heterogeneities across subtypes, contrary to methods that do not account for them. An interesting extension would concern the derivation of valid p-values or confidence intervals for the nonzero parameters identified by CondLogist_DataSharedLasso or MultinomLogist_SymLasso, in particular those corresponding to heterogeneities across subtypes. Given the connection of DataShared Lasso with the lasso (see, e.g., Equation (3.6) under matched designs), this post-selection inference could be derived by extending strategies proposed for lasso estimates (Lee et al., 2016). In other respects, when the identification of heterogeneities is of primary interest, study design is an important step to ensure balanced sample sizes across subtypes (which was not the case in our application to the EPIC data).

The estimation of several parameter vectors considered here is closely related to multi-task learning (Evgeniou and Pontil, 2004), for which a number of other structured sparsity inducing norms have been proposed in the literature, including the group lasso and generalized fused lasso (Lounici et al., 2011; Viallon et al., 2016). We shall first mention that the group lasso is not well suited for the identification of heterogeneities. On the other hand, the generalized fused lasso has shown good properties in the context of stratified regression models, both under generalized linear models (Viallon et al., 2016), survival models (Sennhenn-Reulen and Kneib, 2016) and binary graphical models (Ballout and Viallon, 2019). Its extension to conditional logistic regression models or multinomial logistic models constitutes another interesting lead for future work.

3.7 Software

Software in the form of R codes is available on `Github`. The link to the codes in the matched setting is `https://github.com/NadimBLT/SL1CLR`. The link to the codes in the unmatched setting is `https://github.com/NadimBLT/L1MLR`.

3.8 Supplementary Materials

Supplementary materials can be found in Appendix B.

Acknowledgments

This work was partially supported by the French National Cancer Institute (L'Institut National du Cancer ; INCA) (grant number 2015-166 ; PI : S. Rinaldi). The authors are grateful to the Principal Investigators of each of the EPIC centres for sharing the data of our illustrative example.

Where authors are identified as personnel of the International Agency for Research on Cancer / World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer / World Health Organization.

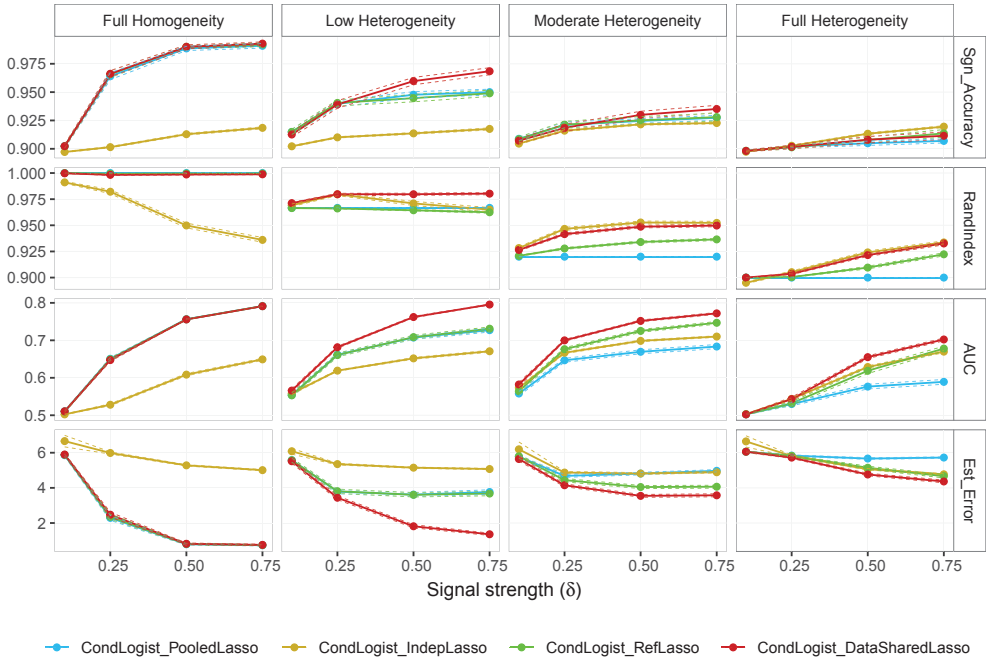


FIGURE 3.1: Results of the simulation study in the matched setting. Solid lines correspond to averages over the 200 replicates, while 95% confidence intervals appear as dotted lines.

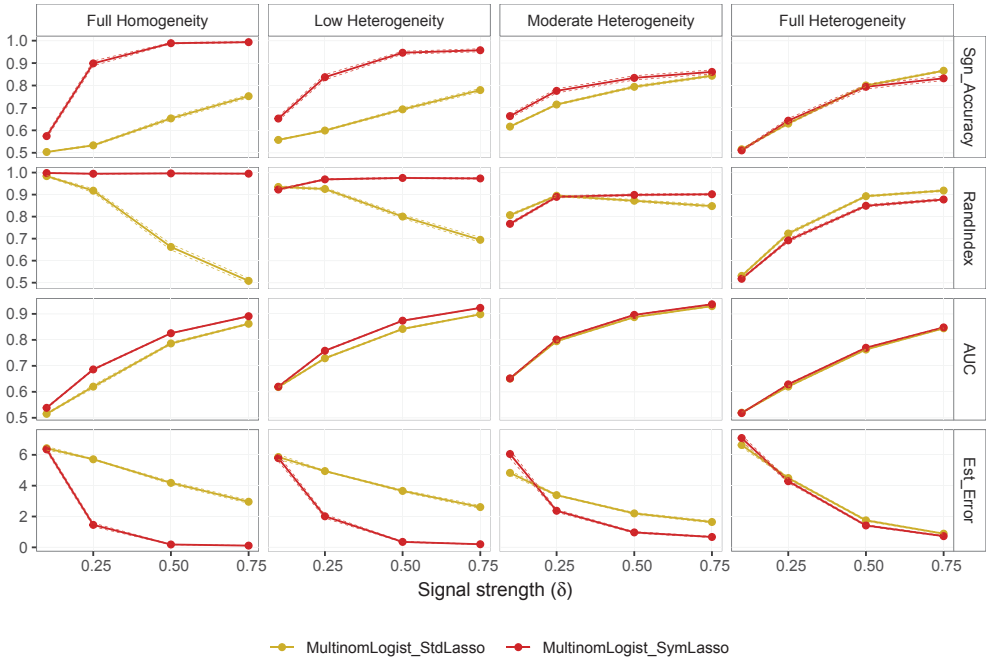


FIGURE 3.2: Results of the simulation study in the unmatched setting. Solid lines correspond to averages over the 200 replicates, while 95% confidence intervals appear as dotted lines

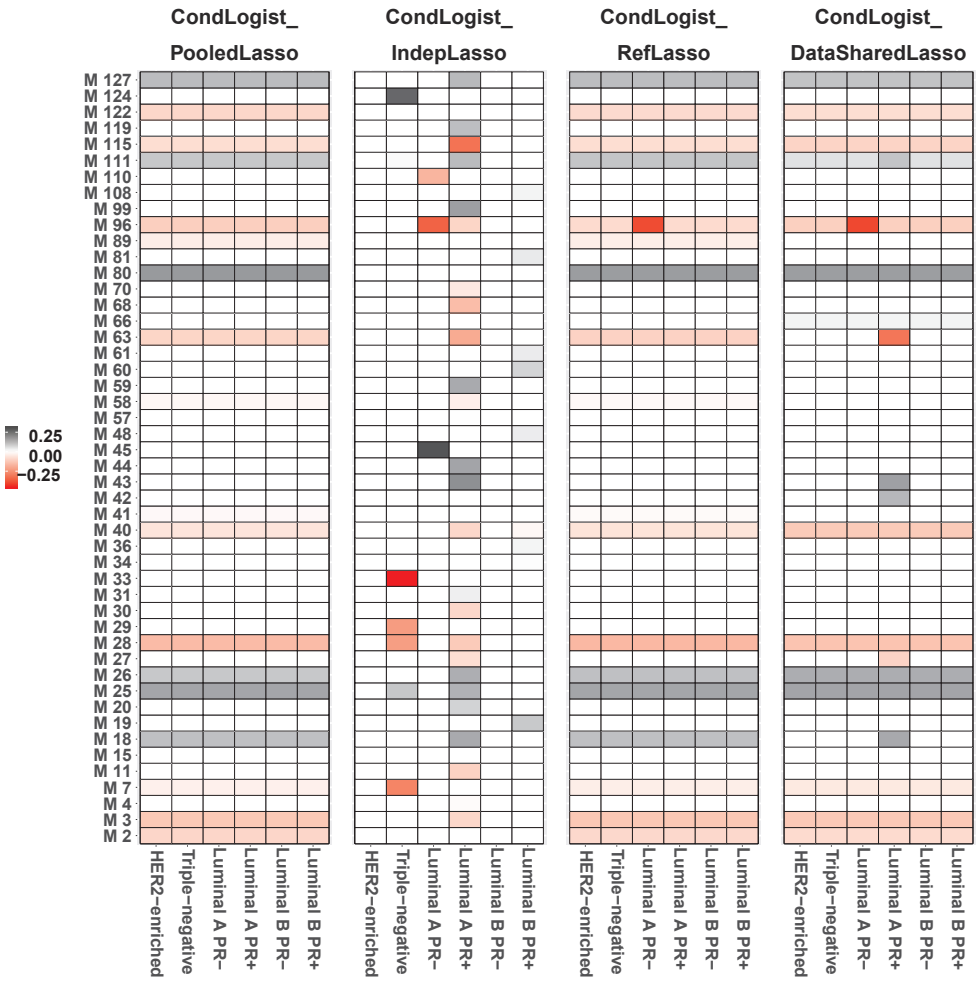


FIGURE 3.3: Preliminary results from the analysis of the matched case-control study nested in EPIC. Six breast cancer histological subtypes are considered : HER-enriched, Triple Negative, Luminal A PR-, Luminal A PR+, Luminal B PR- and Luminal B PR+. Results obtained after the application of four different methods (CondLogist_PooledLasso, CondLogist_IndepLasso, CondLogist_RefLasso, and CondLogist_DataSharedLasso) are presented. For CondLogist_RefLasso, the Luminal A PR+ subtype was selected as the reference. For each method, estimates of $\boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_6^*$ are combined in a matrix, with 6 columns (one for each subtype) and 48 rows (out of the 127 original metabolites, the 79 metabolites for which the four methods produced a zero estimate for all 6 subtypes were eliminated from the plot). In each of the four matrices, each entry represents the estimated level of association between one metabolite and one particular breast cancer subtype. White entries correspond to null associations, grey entries indicate positive associations, while red entries indicate negative association; see the scale on the left of the figure. For example, CondLogist_IndepLasso identifies a strongly inverse association between metabolite M33 and Triple-Negative breast cancer.

Chapitre 4

Discussion

Dans ce travail de thèse, deux méthodes basées sur des régressions logistiques multiples pénalisées ont tout d'abord été décrites pour estimer conjointement des modèles graphiques binaires sur plusieurs strates prédéfinies. Ensuite, l'analyse d'études cas-témoins sur des données de grandes dimensions ont été considérés lorsque plusieurs sous-types de maladies existent, dans des contextes non appariés et appariés.

Dans chacun des travaux, le travail a consisté à appliquer des pénalités sur les hétérogénéités entre les structures correspondantes, permettant aux méthodes proposées dans cette thèse de tirer parti de l'homogénéité potentielle entre ces structures, et facilitant l'interprétation des hétérogénéités identifiées. En mettant l'accent sur l'identification des hétérogénéités, nous avons observé une meilleure performance de nos méthodes par rapport à la stratégie classique qui consiste à estimer les structures séparément et à la stratégie standard qui elle-même consiste à estimer les structures après avoir choisi une strate de référence arbitraire.

La méthode DataShared Lasso permet de tenir compte de l'homogénéité attendue entre les K vecteurs à estimer selon des modèles de régression stratifiés (Gross and Tibshirani, 2016; Ollier and Viallon, 2017). La méthode DataSharedLasso est ainsi capable de cibler une parametrisation plus sparse lorsqu'il existe des points communs entre sous-groupes, ce qui peut entraîner des améliorations substantielles en termes d'efficacité d'estimation. Nos conclusions sont conformes aux résultats empiriques et théoriques d'Ollier and Viallon (2017) dans le cas des modèles de régression linéaire stratifiée.

Dans l'ensemble, nos résultats montrent que la performance d'Indep Lasso ne dépend

pas du niveau d'hétérogénéité, et que cette performance est optimale en l'absence d'homogénéité. Cependant, plus le niveau d'homogénéité augmente, plus les méthodes qui tiennent compte de l'homogénéité peuvent cibler de meilleures (i.e. sparser) paramétrisations et produire des améliorations importantes en termes de performance d'estimation : la méthode DataShared Lasso semble ainsi être une bonne approche pour tenir compte de l'homogénéité quand elle est présente. De plus, cette dernière méthode fonctionne aussi bien que la méthode Indep Lasso lorsqu'il n'y a aucune homogénéité. Par ailleurs, ce travail de thèse a permis de montrer que la méthode DataShared Lasso est plus performante que la méthode Ref Lasso, notamment lorsque le niveau d'hétérogénéité est élevé. De plus, ce travail de thèse a permis de montrer que la méthode Ref Lasso masque les hétérogénéités entre les strates, et que ceci est aggravé lorsque la strate de référence présente également de nombreuses hétérogénéités.
Une propriété intéressante de la méthode DataShared Lasso est qu'elle peut être écrite comme un lasso pondéré par une transformation simple des données originales. Elle est donc facile à implémenter sous une variété de modèles de régression, puisque des algorithmes rapides et très efficaces sont maintenant disponibles pour résoudre le problème d'optimization du lasso sous de nombreux modèles de régression : par exemple le paquet glmnet est maintenant disponible sur R, Matlab et Python. Par ailleurs, ce package glmnet permet l'utilisation des matrices sparses, conduisant ainsi à des gains de calculs substantiels.

L'estimation de plusieurs vecteurs de paramètres considérés dans ces travaux est étroitement liée à l'apprentissage multitâche (Evgeniou and Pontil, 2004), pour lequel un certain nombre de pénalités structurées induisant la sparsité ont été proposées dans la littérature, comme par exemple le group lasso (Lounici et al., 2011) qui n'a pas été considéré dans ce travail puisqu'il n'est pas adapté à l'identification des hétérogénéités, ou le fused lasso (Viallon et al., 2016) qui a été utilisé dans le premier travail de cette thèse portant sur les modèles graphiques binaires. Cette dernière méthode, le fused lasso, a notamment montré de bonnes propriétés dans le contexte des modèles de régression stratifiée, à la fois sous les modèles linéaires généralisés (Viallon et al., 2016) et les modèles de survie (Sennhenn-Reulen and Kneib, 2016).
Comme nous l'avons vu dans les résultats du premier travail de cette thèse, les estimations de DataShared Lasso et de Fused Lasso sont similaires. En général, dans un cadre

asymptotique, les estimations de Fused Lasso sont optimales et devraient être préférées à DataShared Lasso, sous des hypothèses assez générales. En effet, la méthode DataShared Lasso permet de donner une description partielle de la structure des paramètres pour une covariable donnée, alors que Fused Lasso permet, en principe, de donner une description complète de cette structure - sous des conditions très fortes (voir Ollier and Viallon, 2017), grâce à la pénalisation des $K(K-1)/2$ différences $||\boldsymbol{\beta}^{(k_1)} - \boldsymbol{\beta}^{(k_2)}||_1$ pour tout $k_1 > k_2$ qui connecte complètement les paramètres de différentes strates pour chaque covariable $j \in [p]$. DataShared Lasso pénalise les différences entre les paramètres $\beta_j^{(k)}$ pour $k \in [K]$ et leur médiane shrunkée et pondérée μ_j pour tout $j \in [p]$ par le terme $||\boldsymbol{\beta}^{(k)} - \boldsymbol{\mu}||_1$. Donc DataShared Lasso connecte les parmètres $\beta_j^{(k)*}$ par μ_j^*, pour tout $j \in [p]$. Sous cette connexion de paramètres, pour tout $k_1 \neq k_2$, DataShared Lasso nous permet d'interpréter les différences entre $\hat{\beta}_j^{(k_1)}$ et $\hat{\beta}_j^{(k_2)}$, tel que $\hat{\beta}_j^{(k)} = \hat{\mu}_j + \hat{\gamma}_j^{(k)}$ pour tout $k \in [K]$, si au moins une de ces deux estimations est égale à $\hat{\mu}_j$ (i.e. si au moins une des estimations $\hat{\gamma}_j^{(k_1)}$ et $\hat{\gamma}_j^{(k_2)}$ est égale à 0). Plus précisément, pour tout $k_1 \neq k_2$, seules les différences entre $\hat{\beta}_j^{(k_1)}$ et $\hat{\beta}_j^{(k_2)}$ telles que $\hat{\beta}_j^{(k_1)} = \hat{\beta}_j^{(k_2)} = \hat{\mu}_j$ (i.e. $\hat{\gamma}_j^{(k_1)} = \hat{\gamma}_j^{(k_2)} = 0$) ou telles que $\hat{\beta}_j^{(k_1)} = \hat{\mu}_j$ et $\hat{\beta}_j^{(k_2)} \neq \hat{\mu}_j$ (i.e. $\hat{\gamma}_j^{(k_1)} = 0$ et $\hat{\gamma}_j^{(k_2)} \neq 0$) sont interprétables, et pour tout $k_1 \neq k_2$ tel que $\hat{\beta}_j^{(k_1)} \neq \hat{\mu}_j$ et $\hat{\beta}_j^{(k_2)} \neq \hat{\mu}_j$ (i.e. $\hat{\gamma}_j^{(k_1)} \neq 0$ et $\hat{\gamma}_j^{(k_2)} \neq 0$) les différences entre $\hat{\beta}_j^{(k_1)}$ et $\hat{\beta}_j^{(k_2)}$ ne sont pas interprétables. La même analyse des connexions et interprétations des différences peut être envisagée pour la méthode Ref Lasso en remplaçant μ_j^* par $\beta_j^{(r)*}$ et $\hat{\mu}_j$ par $\hat{\beta}_j^{(r)}$. Voir les Figures 4.1 et 4.2 pour une illustration graphique qui illustre comment les méthodes Fused Lasso, Ref Lasso et DataShared Lasso connectent les paramètres, et décrivent la structure des paramètres pour une covariable dans deux exemples. Voir plus bas, une extension de DataShared Lasso qui permettrait d'identifier la partition complète des paramètres pour une covariable donnée.

	Fused	DataShared
Résultats asymptotiques	✓	✓
Résultats non-asymptotiques	-	✓
Implémentation	Difficile	Très facile
Description de la structure des paramètres pour une covariable	Complète (Sous des conditions très fortes)	partielle

Le fait que la méthode Ref Lasso masque les hétérogénéités entre les strates - surtout quand la strate de référence présente de nombreuses hétérogénéités - a été confirmé dans les résultats de l'application aux données servant à décrire les tableaux lésionnels. En

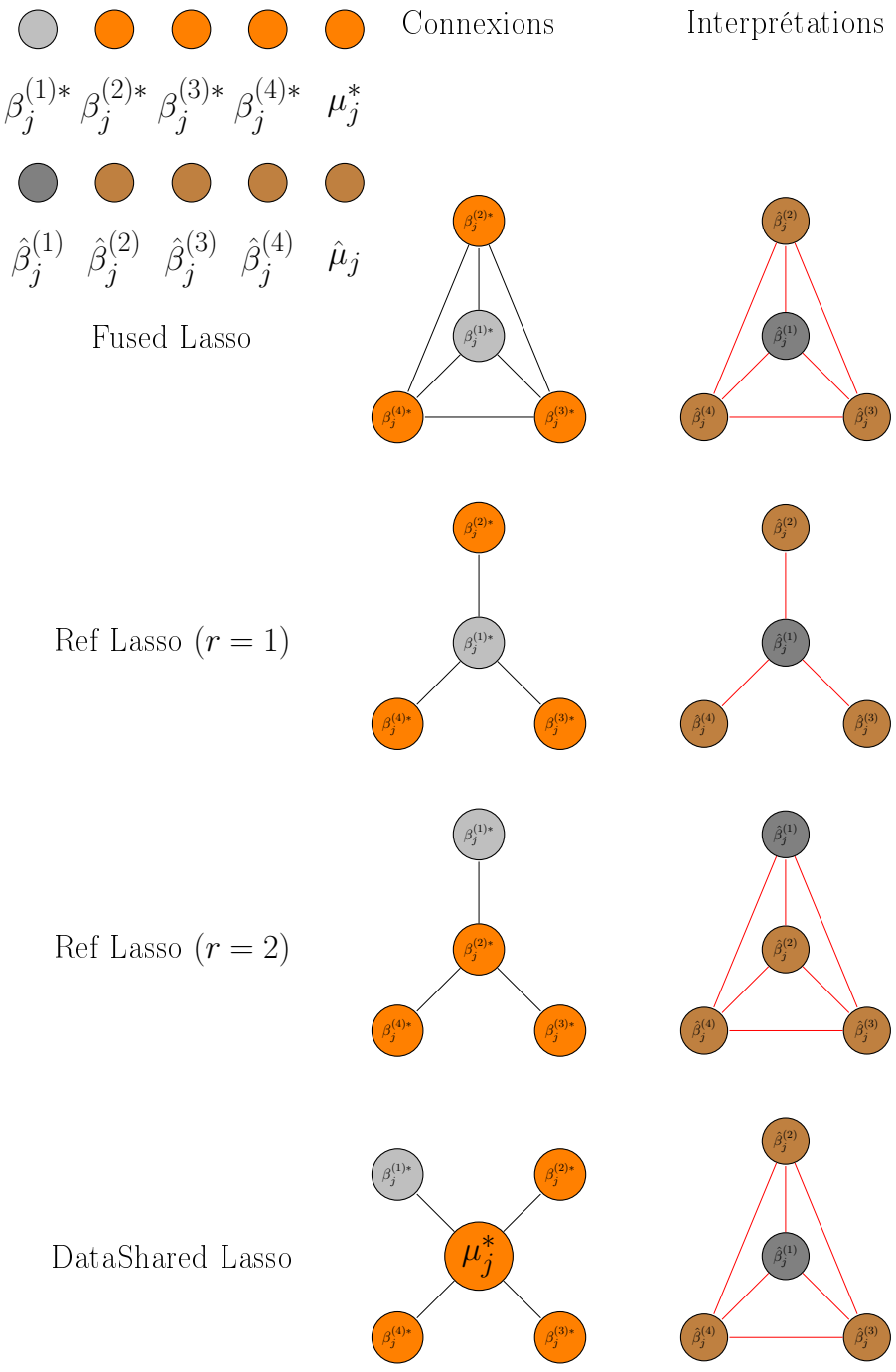


FIGURE 4.1: Une illustration graphique qui illustre comment pour une covariable $j \in [p]$, les méthodes Fused Lasso, Ref Lasso et DataShared Lasso connectent les paramètres de manière générale (les connexions sont représentées par les arêtes noires) et décrivent la structure des paramètres sur un exemple particulier et simple : $K = 4$, $\beta_j^{(1)*} = \beta_1 \in \mathbb{R}$ et $\beta_j^{(2)*} = \beta_j^{(3)*} = \beta_j^{(4)*} = \mu_j^* = \beta_2 \in \mathbb{R}$ tel que $\beta_2 \neq \beta_1$. Nous supposons que toutes les méthodes renvoient les mêmes estimations : $\hat{\beta}_j^{(1)} = \beta_3 \in \mathbb{R}$ et $\hat{\beta}_j^{(2)} = \hat{\beta}_j^{(3)} = \hat{\beta}_j^{(4)} = \beta_4 \in \mathbb{R}$ tel que $\beta_3 \neq \beta_4$ et nous suppsosons que $\hat{\mu}_j = \beta_4$. La capacité des méthodes à interpréter les différences entre les estimations est représentée par les arêtes rouges. Plus précisément, s'il y a une arête rouge entre deux estimations, cela signifie que nous pouvons identifier (interpréter) si ces deux estimations sont identiques ou non.

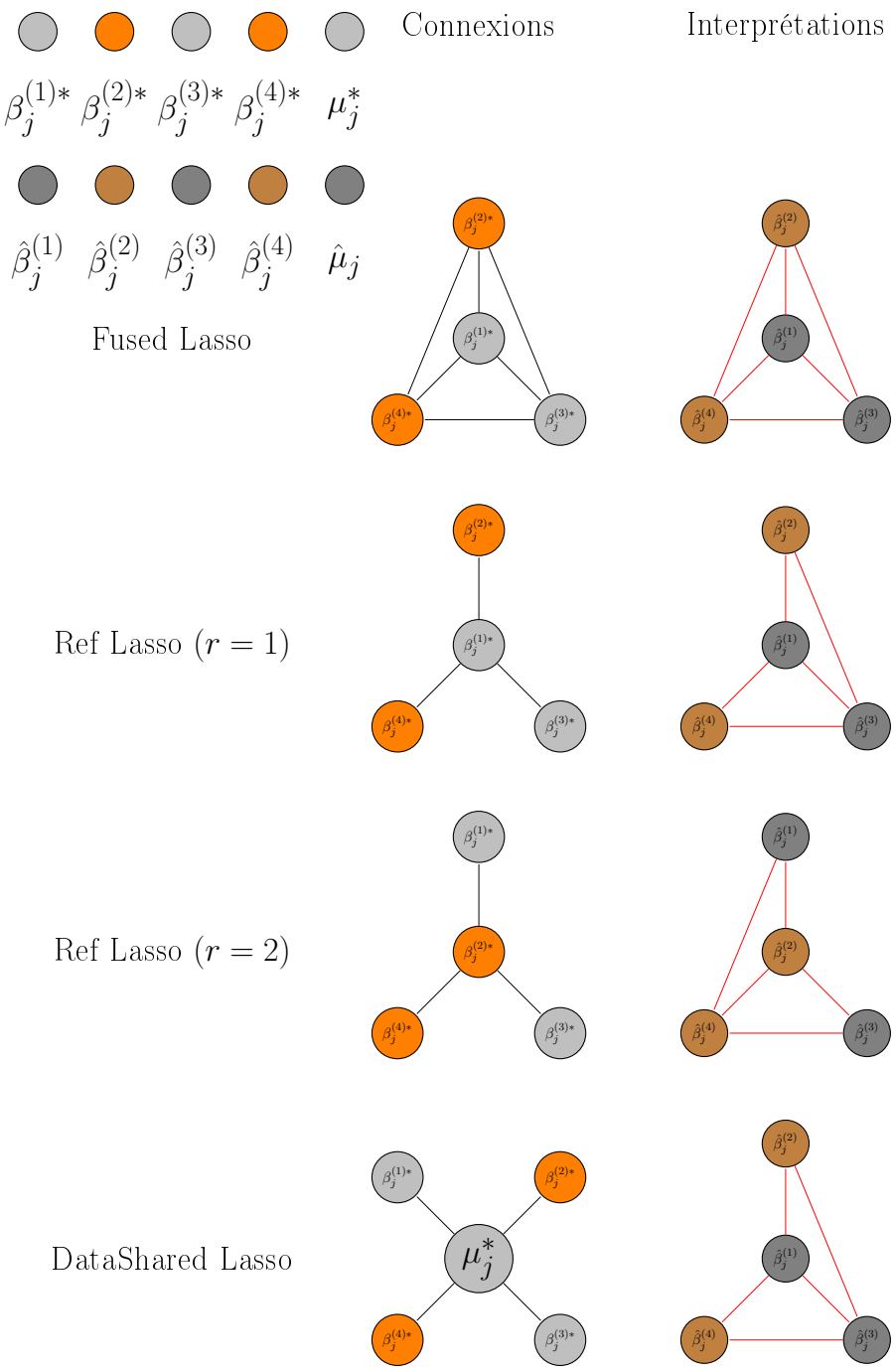


FIGURE 4.2: Une illustration graphique qui illustre comment pour une covariable $j \in [p]$, les méthodes Fused Lasso, Ref Lasso et DataShared Lasso connectent les paramètres de manière générale (les connexions sont représentées par les arêtes noires) et décrivent la structure des paramètres sur un exemple particulier et simple : $K = 4$, $\beta_j^{(1)*} = \beta_j^{(3)*} = \mu_j^* = \beta_1 \in \mathbb{R}$ et $\beta_j^{(2)*} = \beta_j^{(4)*} = \beta_2 \in \mathbb{R}$ tel que $\beta_2 \neq \beta_1$. Nous supposons que toutes les méthodes renvoient les mêmes estimations : $\hat{\beta}_j^{(1)} = \hat{\beta}_j^{(3)} = \beta_3 \in \mathbb{R}$ et $\hat{\beta}_j^{(2)} = \hat{\beta}_j^{(4)} = \beta_4 \in \mathbb{R}$ tel que $\beta_3 \neq \beta_4$ et nous suppsosons que $\hat{\mu}_j = \beta_3$.La capacité des méthodes à interpréter les différences entre les estimations est représentée par les arêtes rouges. Plus précisément, s'il y a une arête rouge entre deux estimations, cela signifie que nous pouvons identifier (interpréter) si ces deux estimations sont identiques ou non.

l'occurrence, les occupants de voiture - étant la plus grande strate - constituent naturellement la strate de référence, et présentent de nombreuses hétérogénéités (par rapport aux autres strates). En d'autres termes, il est vraisemblable que le fait d'appliquer la méthode Ref-SepLogit - après avoir choisi les occupants de voiture comme strate de référence - ne soit pas optimale dans cette application. Cependant, la méthode DataShared-SepLogit contourne le choix arbitraire de la strate de référence, et est de fait susceptible de fonctionner mieux que Ref-SepLogit pour un coût de calcul quasiment identique. En terme de comparaison, les temps de calcul sont similaires pour les méthodes DataShared-SepLogit et Ref-SepLogit - plus efficaces que Fused-SepLogit ; tandis que la précision de la sélection de variable et l'identification des hétérogénéités sont similaires pour les méthodes DataShared-SepLogit et Fused-SepLogit (et notamment dans le cadre des modèles graphiques binaires) – généralement meilleures que Ref-SepLogit. De plus, la méthode DataShared-SepLogit, impliquant uniquement des régressions lasso logistiques, est très facile à mettre en œuvre, et cette méthode est implémenté dans le package glmnet ce qui facilite son utilisation.

La plupart des associations de lésions observées dans l'application illustrative (voir chapitre 2) étaient communes à tous les types d'utilisateurs, alors que les hétérogénéités concernaient davantage les occupants de voitures - à noter que seules les associations positives ont été présentées dans ce travail de thèse (Figure 2.5) (associations avec des $OR \geq 2$). En effet, les associations négatives sont moins intéressantes dans cette application illustrative dû au fait que les cliniciens enregistrent les lésions pour chaque victime - ils négligent probablement certaines lésions lorsque des lésions plus graves sont présentes, et la plupart des associations négatives pourraient être liées à ce biais. De plus, les observations de la première partie de ce travail de thèse correspondent à des victimes d'accidents de la route, soit un ensemble de données ne se composant que de personnes qui souffrent d'au moins une lésion. Par conséquent, l'échantillon sur lequel repose notre travail est biaisé par rapport à l'ensemble de la population : un biais de collision rend généralement dangereuse l'interprétation causale de l'association négative. Même si aucune interprétation causale ne peut être donnée aux associations identifiées, la description de ces associations à partir des tableaux lésionnels des victimes d'accidents de la route reste pertinente, puisque cette sous-population de victimes d'accidents de la route est celle que les cliniciens doivent prendre en charge.

Les résultats du deuxième travail composant cette thèse sont conformes aux résultats empiriques et théoriques de la méthode DataShared Lasso dans le cas des modèles de régression linéaire stratifiés, ainsi qu'aux résultats empiriques exposés dans la première partie de ce travail (voir chapitre 2) pour les modèles graphiques binaires stratifiés.

Les analyses ont notamment intégré les études cas-témoins appariées nichées dans une cohorte portant sur le cancer. Nous avons montré que l'efficacité de l'estimation et de prédiction, la sélection de variables et la capacité d'identifier les hétérogénéités entre les sous-types de maladie pourraient toutes être considérablement améliorées lorsqu'il existe des points communs entre les sous-types de cas, à condition que des méthodes qui tiennent dûment compte de ces points communs soient utilisées (par exemple celles fondées sur la pénalité DataShared Lasso).

Dans le cas des designs appariés, la paramétrisation originale repose sur les $K-1$ vecteurs $\boldsymbol{\delta}_k^*$, qui représentent les odds ratio logarithmiques comparant chacun des $K-1$ sous-types de maladie aux témoins. Sur la base d'une reparamétrisation sur-paramétrée, CondLogist_DataSharedLasso est capable de cibler une paramétrisation plus parcimonieuse lorsqu'il existe des points communs entre sous-types, ce qui peut entraîner des améliorations importantes en termes d'efficacité d'estimation. En l'absence de points communs entre les sous-types de cas, cette méthode fonctionne aussi bien que l'analyse standard et indépendante de chaque sous-type de cas. Dans le cas de modèles non appariés, la formulation standard des modèles de régression logistique multinomiale repose sur la même paramétrisation, impliquant $K-1$ vecteurs $\boldsymbol{\delta}_k^*$ qui comparent chaque sous-type de maladie aux témoins. Nous avons formellement établi que l'application des idées de DataShared Lasso avec cette paramétrisation équivalait en fait à l'application d'un lasso standard sur la formulation symétrique du modèle. Cette formulation symétrique repose sur une paramétrisation sur-paramétré avec K vecteurs $\boldsymbol{\beta}_k^*$, et profite du fait que les témoins ne doivent pas nécessairement être considérés comme la catégorie de référence dans des réglages non appariés. Encore une fois, la paramétrisation qui en résulte peut être beaucoup plus sparse que la paramétrisation standard et donne généralement une meilleure efficacité d'estimation, surtout lorsque le niveau d'homogénéité entre les

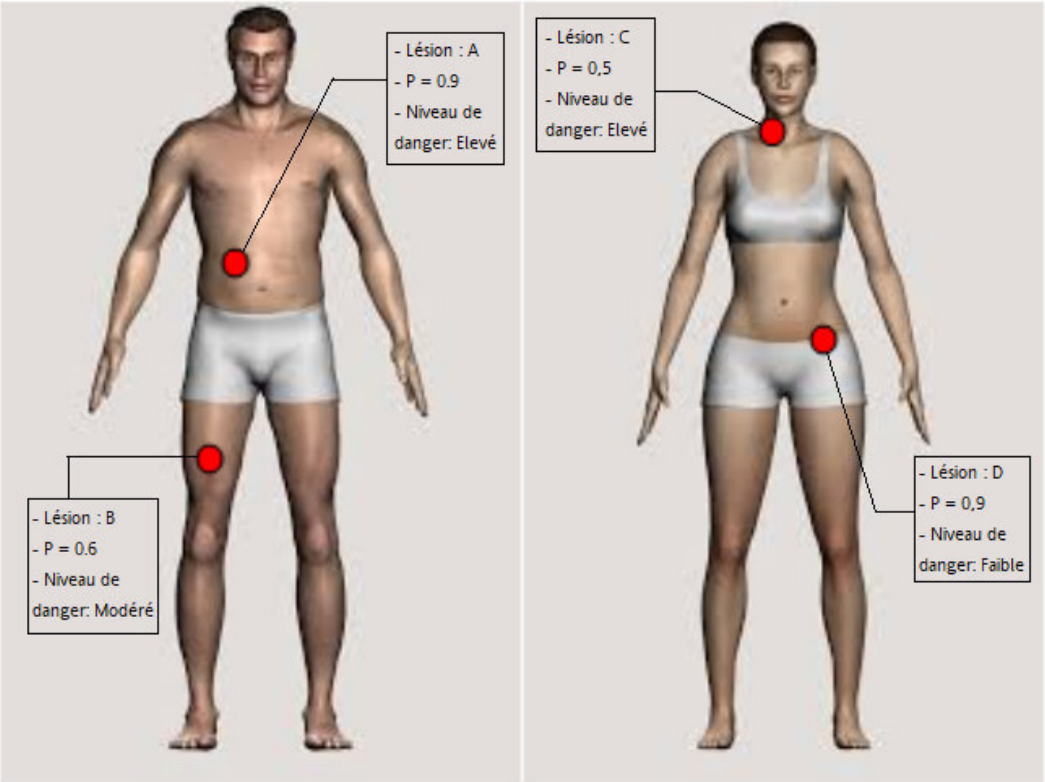
sous-types est élevé.

Perspectives Dans le cadre des victimes d'accidents de la route, des regroupements plus fins des lésions pourraient être utilisés pour améliorer l'interprétation clinique, ou alors des définitions différentes pourraient être données aux strates (par exemple, gravité de l'accident). Dans le cadre clinique, ce travail de thèse trouve également son intérêt chez les cliniciens et notamment pour certaines lésions qui peuvent passer inaperçues chez des patients dans un état grave et pour lesquels l'enjeu vital prime sur l'enjeu fonctionnel. Ces associations peuvent avoir un intérêt particulier lorsque l'on s'intéresse à une même région corporelle : par exemple, contusion pulmonaire et hémorragies intra thoraciques. Ces derniers sont certes plus vitales mais la contusion peut entraîner une infection pulmonaire, voire, à long terme, un tableau d'insuffisance respiratoire chronique.
Actuellement, une collaboration avec des cliniciens permet une meilleure compréhension des besoins en terme d'outils statistiques nécessaires : ils ont besoin de pouvoir déceler une lésion "illisible" grâce à la présence d'une lésion "lisible", et ceci dans les plus brefs délais. Pour répondre à ces besoins, nous avons vu que les calculs d'associations entre regroupements de lésions apportaient de nombreuses informations sur la présence de lésions. Cependant, de nombreux facteurs interviennent lors des accidents, et les associations entre lésions diffèrent (par exemple, type d'usagers, caractéristiques d'accident, etc.). Pour rappel, les résultats des méthodes proposées dans ce manuscrit ont été appliqués sur des groupes de lésions ($p = 36$), présentés par type d'usager et lorsque l'association entre les lésions était relativement élevée ($OR \geq 2$). Il est clair que l'interprétation et la présentation graphique de ses résultats (ou la présentation de ses résultats en général) n'est pas assez facile. Or, d'après les cliniciens, il semblerait encore plus intéressant de pouvoir étudier les associations entre des lésions précises (et non des regroupements de lésions ou des regroupements de lésions plus fins et plus détaillés), et avec des associations plus faibles ($OR > 1$), ce qui paraît complexe étant donné le nombre de lésions ($p = 1348$), et le nombre d'associations potentielles. De plus, des interactions d'ordre 2 ou 3 peuvent également intervenir, ce qui rend la tâche encore plus complexe. De fait, une des perspectives de ce travail de thèse est de développer un outil, qui selon les caractéristiques imposées par l'accident, n'afficherait que les lésions associées avec la lésion A-(lisible), par exemple, les lésions B, C et D-(illisible). Pour y parvenir, nous devons

développer des modèles statistiques dans lesquels un maximum de caractéristiques des victimes d'accident sont introduites, et ainsi estimer les associations entre les lésions, et donc la probabilité de présence d'une lésion "illisible", avec beaucoup de précision. Plus précisément, nous devons nous intéresser à l'estimation d'une probabilité conditionnelle, du type $\mathbb{P}(Y = 1|\mathbf{X})$ où Y est une variable binaire qui indique la présence de la lésion d'intérêt et $\mathbf{X} \in \mathbb{R}^p$ un ensemble de prédicteurs qui décrivent la présence d'autres lésions, les caractéristiques de l'accident, celles de la victime, etc. L'estimation de la probabilité $\mathbb{P}(Y = 1|\mathbf{X})$ peut être faite à l'aide d'approches statistiques analogues, qui constituent le champ de la classification supervisée, sur données de grande dimension. Parmi ces approches, on peut évoquer les modèles de régression logistique pénalisée (comme le lasso évoqué ci-dessus par exemple), le boosting, les forêts aléatoires, les réseaux de neurones (deep learning), etc... (Hastie et al., 2001). On pourra également s'intéresser aux approches qui permettent d'agréger plusieurs estimations de la probabilité $\mathbb{P}(Y = 1|\mathbf{X})$, comme par exemple le super-learner de (van der Laan Mark J. et al., 2007). Si les performances prédictives de ces modèles sont assez bonnes, ils pourront être développés sous forme d'application de type Shiny, qui peut être sur mobile ou tablette directement destinée aux cliniciens. L'idée serait de permettre aux cliniciens d'enregistrer sur cette application les données d'entrée : les caractéristiques de l'accident, celles de la victime et la lésion ou les lésions apparentes sur pictogrammes ou diagnostiquées par l'équipe médicale lors de son arrivée sur le lieux de l'accident, pour finalement n'avoir que les lésions probables associées, et ceci en quelques clics : une représentation graphique en 3D d'un être humain pourrait simplifier la visualisation de l'accidenté, permettant par exemple de visualiser en rouge les zones ou les lésions probablement atteintes, avec une probabilité de prédiction (voir Figure 4.3).

Le lien important entre lasso et DataShared Lasso nous permet d'étendre les extensions déjà réalisées à partir de la méthode lasso sur la méthode DataShared Lasso, et donc sur les méthodes proposées dans ce travail qui sont basées sur la méthode DataShared Lasso. Une extension intéressante des méthodes présentées sur les données stratifiées consisterait à dériver des p-values valides ou des intervalles de confiance pour les paramètres non nuls identifiés par DataShared-SepLogit, CondLogist_DataSharedLasso ou MultinomLogist_SymLasso, en particulier ceux correspondant aux hétérogénéités entre sous-types. Étant donné le lien entre le lasso et DataShared Lasso, cette inférence post-sélection pourrait être obtenue en étendant les stratégies proposées pour les estimations

FIGURE 4.3: Exemple de représentation graphique simple d'un être humain qui pourrait simplifier la visualisation de l'accidenté, où nous visualisons en rouge les lésions probablement atteintes, avec une probabilité de prédiction et un niveau de danger.



du lasso (Lockhart et al., 2014) ou (Lee et al., 2016) . Ensuite, comme évoqué ci-dessus, les connexions des paramètres qui sont faites par DataShared Lasso ne conduisent qu'à une description partielle de la structure des paramètres pour une covariable. Une extension intéressante du DataShared Lasso (notée DataShared Lasso Iterated) dont nous présenterons ci-dessous le principe pourrait être capable de compléter cette description.

Pour simplifier la présentation de DataShared Lasso Iterated, nous allons nous concentrer sur une seule covariable ($p = 1$) et considérer l'exemple présenté dans la Figure 4.2, où $\beta^{(1)*} = \beta^{(3)*} = \beta_1 \in \mathbb{R}$ et $\beta^{(2)*} = \beta^{(4)*} = \beta_2$ tel que $\beta_2 \neq \beta_1 \in \mathbb{R}$. Par cette Figure, nous pouvons voir que nous ne pouvons pas interpréter la différence entre $\hat{\beta}^{(2)}$ et $\hat{\beta}^{(4)}$ s'il y a une différence. Pour y parvenir, nous devons donc coupler l'estimation de ces deux paramètres. Dans cet exemple, supposons que DataShared Lasso nous permette d'identifier que $\hat{\beta}^{(1)} = \hat{\beta}^{(3)} = \hat{\mu}$ (i.e. $\hat{\gamma}^{(1)} = \hat{\gamma}^{(3)} = 0$) et que $\hat{\beta}^{(2)} \neq \hat{\mu}$ et $\hat{\beta}^{(4)} \neq \hat{\mu}$ (i.e. $\hat{\gamma}^{(2)} \neq 0$ et $\hat{\gamma}^{(4)} \neq 0$). L'idée est alors de faire une autre étape en décomposant μ par $\mu^{(1,3)}$ pour $k \in \{1, 3\}$, et par $\mu^{(2,4)}$ pour $k \in \{2, 4\}$. En faisant cela, nous connectons

les paramètres $\beta^{(1)*}$ et $\beta^{(3)*}$ par $\mu^{(1,3)*}$ et les paramètres $\beta^{(2)*}$ et $\beta^{(4)*}$ par $\mu^{(2,4)*}$ (voir Figure 4.4). Plus précisément, dans la première itération du DataShared Lasso Iterated, nous faisons un DataShared Lasso standard et donc nous minimisons le critère $||\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}^{Iter_1}\boldsymbol{\phi}^{Iter_1}||_2^2 + \lambda||\boldsymbol{\phi}^{Iter_1}||_1$, avec,

$$\boldsymbol{\mathcal{X}}^{Iter_1} = \begin{pmatrix} \mathbf{X}^{(1)} & \mathbf{X}^{(1)} & 0 & 0 & 0 \\ \mathbf{X}^{(2)} & 0 & \mathbf{X}^{(2)} & 0 & 0 \\ \mathbf{X}^{(3)} & 0 & 0 & \mathbf{X}^{(3)} & 0 \\ \mathbf{X}^{(4)} & 0 & 0 & 0 & \mathbf{X}^{(4)} \end{pmatrix} \in \mathbb{R}^{n*5} \text{ et } \boldsymbol{\phi}^{Iter_1} = \begin{pmatrix} \mu \\ \gamma^{(1)} \\ \gamma^{(2)} \\ \gamma^{(3)} \\ \gamma^{(4)} \end{pmatrix} \in \mathbb{R}^5.$$

Par cette itération, nous identifions (typiquement) que les estimations $\hat{\beta}^{(1)}$ et $\hat{\beta}^{(3)}$ sont identiques et qu'ils sont différents de $\hat{\beta}^{(2)}$ et $\hat{\beta}^{(4)}$. Ensuite, dans la deuxième itération, nous minimisons le critère $||\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}^{Iter_2}\boldsymbol{\phi}^{Iter_2}||_2^2 + \lambda||\boldsymbol{\phi}^{Iter_2}||_1$, après avoir défini la matrice de design et le vecteur de paramètres comme suit

$$\boldsymbol{\mathcal{X}}^{Iter_2} = \begin{pmatrix} \mathbf{X}^{(1)} & 0 & 0 & 0 \\ \mathbf{X}^{(3)} & 0 & 0 & 0 \\ 0 & \mathbf{X}^{(2)} & \mathbf{X}^{(2)} & 0 \\ 0 & \mathbf{X}^{(4)} & 0 & \mathbf{X}^{(4)} \end{pmatrix} \in \mathbb{R}^{n*4} \text{ et } \boldsymbol{\phi}^{Iter_2} = \begin{pmatrix} \mu^{(1,3)} \\ \mu^{(2,4)} \\ \gamma^{(2)} \\ \gamma^{(4)} \end{pmatrix} \in \mathbb{R}^4.$$

Par cette deuxième itération, nous pouvons donc identifier si les estimations $\hat{\beta}^{(2)}$ et $\hat{\beta}^{(4)}$ sont identiques ou non. Nous sommes alors en mesure d'obtenir des estimations du type $\hat{\gamma}^{(2)} = \hat{\gamma}^{(4)} = 0$, et donc identifier que $\hat{\beta}^{(2)}$ et $\hat{\beta}^{(4)}$ sont identiques. Donc DataShared Lasso Iterated permet, en principe, de donner une description complète de la structure des paramètres (voir Figure 4.4).

Enfin, un travail est en cours de finalisation en collaboration avec Lola Etievant et Vivian Viallon, reposant sur une technique de cross-validation dans l'esprit de one-step lasso (Bühlmann and Meier, 2008) visant à sélectionner le paramètre de régularisation. Cette technique a déjà été implémentée pour les modèles de régression logistique conditionnelle.

Le one-step lasso est une procédure en deux étapes et peut être considérée comme une extension du lasso adaptatif. Il consiste à utiliser le lasso d'une première étape comme estimateur initial. Bühlmann and Meier (2008) suggèrent d'utiliser la validation croisée

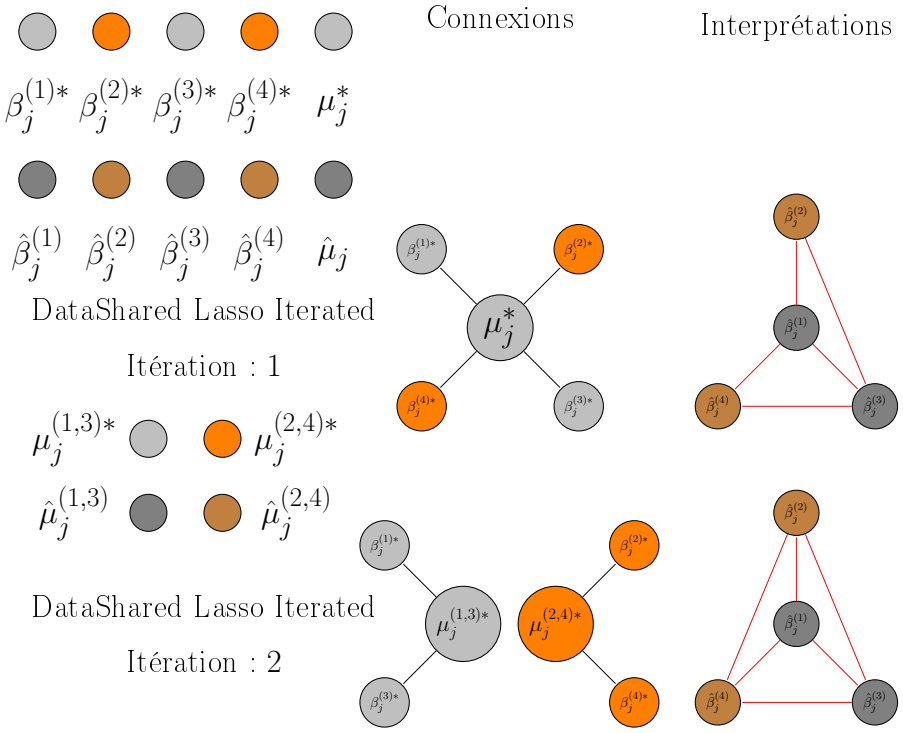


FIGURE 4.4: Une illustration graphique qui illustre comment pour une covariable $j \in [p]$, la méthode DataShared Lasso Iterated connectent les paramètres (les connexions sont représentées par les arêtes noires) et décrit la structure des paramètres sur un exemple particulier et simple : $K = 4$, $\beta_j^{(1)*} = \beta_j^{(3)*} = \mu_j^* = \mu_j^{(1,3)*} = \beta_1 \in \mathbb{R}$ et $\beta_j^{(2)*} = \beta_j^{(4)*} = \mu_j^{(2,4)*} = \beta_2 \in \mathbb{R}$ tel que $\beta_2 \neq \beta_1$. Nous supposons que DataShared Lasso Iterated renvoie les estimations : $\hat{\beta}_j^{(1)} = \hat{\beta}_j^{(3)} = \hat{\mu}_j = \hat{\mu}_j^{(1,3)} = \beta_3 \in \mathbb{R}$ et $\hat{\beta}_j^{(2)} = \hat{\beta}_j^{(4)} = \hat{\mu}_j^{(2,4)} = \beta_4 \in \mathbb{R}$ tel que $\beta_3 \neq \beta_4$. La capacité de cette méthode à interpréter les différences entre les estimations est représentée par les arêtes rouges. Plus précisément, s'il y a une arête rouge entre deux estimations, cela signifie que nous pouvons identifier (interpréter) si ces deux estimations sont identiques ou non.

pour sélectionner le paramètre de régularisation optimal λ qui reflète l'erreur de prédiction optimale. Ce choix est populaire dans les données de grandes dimension. Dans le second projet, nous avons identifié un défaut de la validation croisée pour le one-step lasso. En fait, dans one-step lasso et à chaque étape, nous sélectionnons le paramètre de régularisation par l'approhe de la validation croisée. Dans cette approche, nous calculons l'erreur de validation croisée qui est une estimation de l'erreur de prédiction réelle, puis nous sélectionnons le paramètre de régularisation qui a l'erreur de validation croisée minimale. La validation croisée est une technique pour évaluer les modèles prédictifs en partitionnant l'ensemble de données original D en train D^{train} et test $D^{test} (= D \setminus D^{train})$ sous-ensembles, où D^{train} sert à former le modèle et D^{test} est un ensemble indépendant pour évaluer le modèle. Le one-step lasso ou le lasso adaptatif est incompatible avec

l'idée de validation croisée, car dans la deuxième étape, nous utilisons des poids qui ont été calculés sur les données totales D, donc dans la technique de validation croisée, l'ensemble de test D^{test} n'est plus totalement indépendant.
Une adaptation de la cross-validation, qui peut être vue comme une cross-validation emboitée, et qui corrige ce défaut, permet une meilleure sélection du paramètre de régularisation pour le one-step lasso et le lasso adaptatif. Une version de l'article correspondant est disponible sur arXiv (Ballout et al., 2020a).

Annexe A

Structure estimation of binary graphical models on stratified data : application to the description of injury tables for victims of road accidents : APPENDICES

Figure A.1 [resp. A.2] presents the results obtained by applying Ref-SepLogit Adaptive MIN with car occupants as the reference stratum [resp. Indep-SepLogit Adaptive MIN]. Only edges corresponding to conditional odds-ratios greater than or equal to 2 are presented to make comparison with the results of Figure 2.5 in the Main text easier.

In Figure A.3, we present another version of the structure of the graphical models estimated by DataShared-Seplogit Adaptive MIN, on each of the four strata. Here, only positive conditional associations, that is edges corresponding to estimated conditional odds-ratios greater than one, are presented. To better visualize the differences between graphs, Figure A.4, A.5, A.6 and A.7 present an alternative version of the graphs presented in Figure 2.5 of the Main text, A.1, A.2 and A.3. In this version, we first present the common graph, for which each edge is defined as the median of the four corresponding edges estimated over the four strata. We then present the differences between each stratum-specific graph and this common graph. The chordDiagram function of the circlize R package was used to generate these figures. Colors of the nodes (injuries) correspond to the class they belong to (see Table 2.2). The line type of each edge is related to the sign of the corresponding conditional odds-ratio. The size of each node in the common structure [resp. specific structure for k-th stratum] is related to the prevalences of injuries computed on the whole population of victims [resp. on victims that are in the k-th stratum].

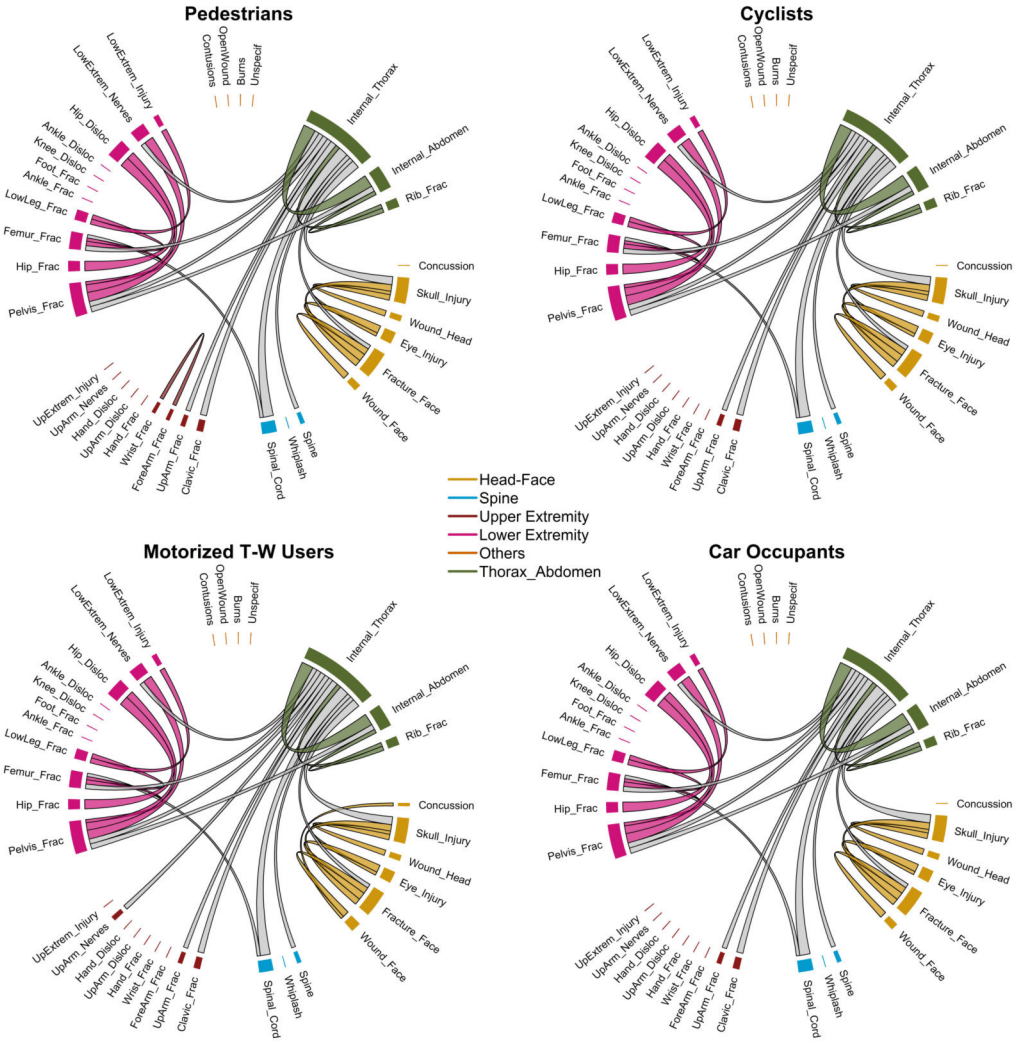


FIGURE A.1: Application of the Ref-SepLogit Adaptive MIN approach on the Rhône Registry Data to describe the injury tables of victims of road accidents, according to road user type : pedestrians, cyclists, motorized T-W and car occupants. Only edges corresponding to conditional odds-ratios greater than or equal to 2 are represented. The reference stratum was set to "car occupants".

A.1 DataShared-SepLogit

A.1.1 Identifiability

As mentioned in the main text, DataShared-SepLogit is primarily based on the following additive decomposition

$$\boldsymbol{\theta}_j^{(k)*} = \boldsymbol{\mu}_j^* + \boldsymbol{\gamma}_j^{(k)*}, \quad \text{for each } k \in [K].$$

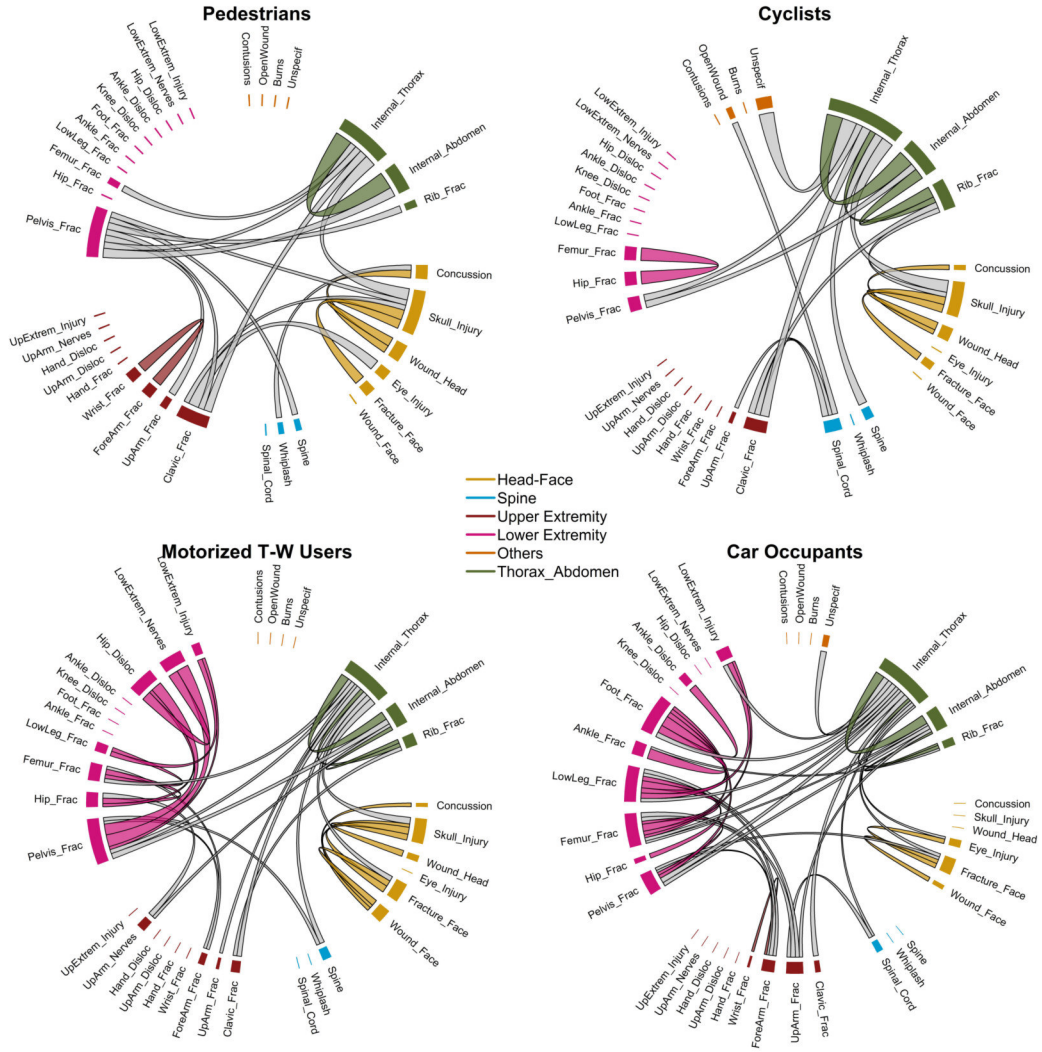


FIGURE A.2: Application of the Indep-SepLogit Adaptive MIN approach on the Rhône Registry Data to describe the injury tables of victims of road accidents, according to road user type : pedestrians, cyclists, motorized T-W and car occupants. Only edges corresponding to conditional odds-ratios greater than or equal to 2 are represented.

Of course, there are infinitely many decompositions of this form. But, first observe that the minimization of criterion

$$\sum_{k=1}^{K} -L((\boldsymbol{\mu}_j + \boldsymbol{\gamma}_j^{(k)}); \boldsymbol{\mathcal{U}}_j^{(k)}, (\boldsymbol{\mathcal{U}}_{-j}^{(k)})) + \lambda_{j,1} \|\boldsymbol{\mu}_{-j}\|_1 + \sum_{k=1}^{K} \lambda_{j,2}^{(k)} \|\boldsymbol{\gamma}_j^{(k)}\|_1,$$

over $\boldsymbol{\mu}_j$ and the $\boldsymbol{\gamma}_j^{(k)}$'s is equivalent to the minimization of the following one,

$$\sum_{k=1}^{K} -L(\boldsymbol{\theta}_j^{(k)}; \boldsymbol{\mathcal{U}}_j^{(k)}, (\boldsymbol{\mathcal{U}}_{-j}^{(k)})) + \lambda_{j,1} \left(\|\boldsymbol{\mu}_{-j}\|_1 + \sum_{k=1}^{K} \frac{\lambda_{j,2}^{(k)}}{\lambda_{j,1}} \|\boldsymbol{\theta}_j^{(k)} - \boldsymbol{\mu}_j\|_1 \right),$$

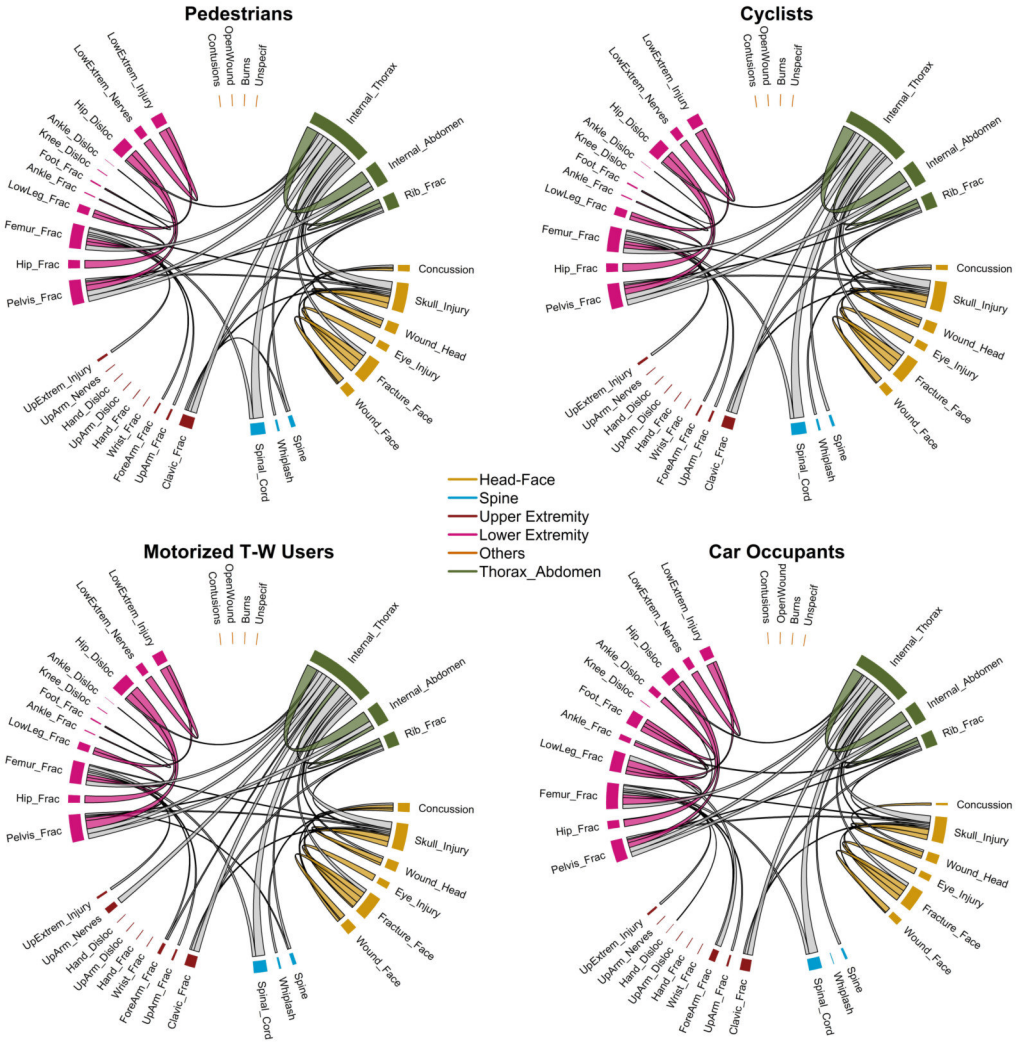


FIGURE A.3: Application of the DataShared-SepLogit Adaptive MIN approach on the Rhône Registry Data to describe the injury tables of victims of road accidents, according to road user type : pedestrians, cyclists, motorized T-W and car occupants. All positive conditional associations are represented.

over $\boldsymbol{\mu}_j$ and the $\boldsymbol{\theta}_j^{(k)}$'s. Next, by setting $\tau = (\tau_1, ..., \tau_K)$ with $\tau_k = \lambda_{j,2}^{(k)}/\lambda_{j,1}$, and defining the shrunk and weighted version of the median of $(\alpha_1, ..., \alpha_K)$ as $WSmedian(\alpha_1, ..., \alpha_K; \tau) = \underset{\alpha}{argmin}(|\alpha| + \sum_{k=1}^{K} \tau_k |\alpha_k - \alpha|)$, it is easy to see that $\hat{\mu}_{j,\ell} \in WSmedian(\hat{\theta}_{j,\ell}^{(1)}, ..., \hat{\theta}_{j,\ell}^{(K)})$. In other words, for any particular values of $\tau_k = \lambda_{j,2}^{(k)}/\lambda_{j,1}$, the penalty term $\lambda_{j,1}\left(\|\boldsymbol{\mu}_{-j}\|_1 + \sum_{k=1}^{K} \tau_k \|\boldsymbol{\gamma}_j^{(k)}\|_1\right) = \lambda_{j,1}\left(\|\boldsymbol{\mu}_{-j}\|_1 + \sum_{k=1}^{K} \tau_k \|\boldsymbol{\theta}_j^{(k)} - \boldsymbol{\mu}_j\|_1\right)$ shrinks the estimates $\hat{\theta}_{j,\ell}^{(k)}$, $k \in [K]$, towards their "weighted and shrunk towards 0" median $\hat{\mu}_{j,\ell} = WSmedian(\hat{\theta}_{j,\ell}^{(1)}, ..., \hat{\theta}_{j,\ell}^{(K)})$. For the constant terms, we have $\hat{\mu}_j = \text{argmin}_{m \in \mathbb{R}} \sum_k \lambda_{j,2}^{(k)} |\hat{\theta}_j^{(k)} - m|$ and $\hat{\mu}_j$ is then simply a weighted median of the set of values $(\hat{\theta}_j^{(1)}, \ldots, \hat{\theta}_j^{(K)})$. We refer to Ollier and Viallon (2017) and Gross and Tibshirani (2016) for more details.

A.1.2 Implementation

Another interesting property of DataShared Lasso is that it can be rewritten as a simple lasso, which makes its implementation straightforward. Denote by $Y_j = (\boldsymbol{\mathcal{U}}_j^{(1)}, ..., \boldsymbol{\mathcal{U}}_j^{(K)})$ the vector containing the n observations of the response variable and let $\tilde{\boldsymbol{\mathcal{U}}}_{-j}^{(k)} = (\mathbf{1}_{n_k}, \boldsymbol{\mathcal{U}}_{-j}^{(k)}) \in \mathbb{R}^{n_k * p}$, where $\mathbf{1}_{n_k}$ is the vectors of size n_k with components all equal 1. The criterion to be minimized can then be written as

$$-L(\Phi_j; \mathbf{Y}_j, \mathbf{X}_j) + \lambda_j \|\Phi_{-j}\|_1$$

with,

$$\mathbf{X}_j = \begin{pmatrix} \boldsymbol{\mathcal{U}}_{-j}^{(1)} & \frac{1}{\tau_1}\tilde{\boldsymbol{\mathcal{U}}}_{-j}^{(1)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\mathcal{U}}_{-j}^{(K)} & 0 & \cdots & \frac{1}{\tau_K}\tilde{\boldsymbol{\mathcal{U}}}_{-j}^{(K)} \end{pmatrix} \in \mathbb{R}^{n*((K+1)p-1)} \text{ and}$$

$$\Phi_j = \begin{pmatrix} \boldsymbol{\mu}_j \\ \tau_1 \boldsymbol{\gamma}_j^{(1)} \\ \vdots \\ \tau_K \boldsymbol{\gamma}_j^{(K)} \end{pmatrix} \in \mathbb{R}^{(K+1)p},$$

where Φ_{-j} corresponds to the vector Φ_j after the elimination of the constant term μ_j (i.e. $\Phi_{-j} = (\boldsymbol{\mu}_{-j}^T, \tau_1 \boldsymbol{\gamma}_j^{(1)T}, ..., \tau_K \boldsymbol{\gamma}_j^{(K)T})^T$). Thus, estimates $\hat{\boldsymbol{\mu}}_j$ and $\hat{\boldsymbol{\gamma}}_j^{(k)}$'s can be obtained as the solution of a simple lasso with response vector $\mathbf{Y}_j$ and design matrix $\mathbf{X}_j$.

A.2 Extension to more general binary graphical models

Under more general binary graphical models, we assume the existence of a parameter vector $\boldsymbol{\Theta}^* = ((\theta_j^*)_{1\leq j\leq p}, (\theta_{j,\ell}^*)_{1\leq j<\ell\leq p}, (\theta_{j,\ell,m}^*)_{1\leq j<\ell<m\leq p}, ..., (\theta_{1,2,...,p}^*))^T$ in $\mathbb{R}^{(2^p-1)}$ such that for any vector $\mathbf{u} = (u_1, ..., u_p) \in \{0,1\}^p$, the probability of the event $\{\mathbf{U} = \mathbf{u}\}$ is given by

$$\mathbb{P}_{\boldsymbol{\Theta}^*}(\mathbf{U} = \mathbf{u}) =$$

$$\exp\Big\{ \sum_{j=1}^{p} \theta_j^* u_j + \sum_{j=1}^{p-1} \sum_{\ell=j+1}^{p} \theta_{j,\ell}^* u_j u_\ell + \sum_{j=1}^{p-2} \sum_{\ell=j+1}^{p-1} \sum_{m=\ell+1}^{p} \theta_{j,\ell,m}^* u_j u_\ell u_m + ... + \theta_{1,2,...,p}^* u_1 u_2 ... u_p - A(\boldsymbol{\Theta}^*) \Big\}. \tag{A.1}$$

We can then still use the ideas of SepLogit under such models. For simplicity, let us consider the model which neglects third and higher-order interactions. This model assumes the existence of a parameter vector $\boldsymbol{\Theta}^* = ((\theta^*_j)_{1\leq j\leq p}, (\theta^*_{j,\ell})_{1\leq j<\ell\leq p}, (\theta^*_{j,\ell,m})_{1\leq j<\ell<m\leq p})^T$ in $\mathbb{R}^{p(p^2+5)/6}$ such that for any vector $\mathbf{u} = (u_1, ..., u_p) \in \{0,1\}^p$, the probability of the event $\{\mathbf{U} = \mathbf{u}\}$ is given by

$$\mathbb{P}_{\boldsymbol{\Theta}^*}(\mathbf{U} = \mathbf{u}) = \exp\Big\{ \sum_{j=1}^{p} \theta^*_j u_j + \sum_{j=1}^{p-1} \sum_{\ell=j+1}^{p} \theta^*_{j,\ell} u_j u_\ell + \sum_{j=1}^{p-2} \sum_{\ell=j+1}^{p-1} \sum_{m=\ell+1}^{p} \theta^*_{j,\ell,m} u_j u_\ell u_m - A(\boldsymbol{\Theta}^*) \Big\}. \tag{A.2}$$

In this case, the log *partition function* A is defined as

$$A(\boldsymbol{\Theta}) = \log \sum_{\mathbf{u}\in\{0,1\}^p} \exp\left(\sum_{j=1}^{p} \theta_j u_j + \sum_{j=1}^{p-1} \sum_{\ell=j+1}^{p} \theta_{j,\ell} u_j u_\ell + \sum_{j=1}^{p-2} \sum_{\ell=j+1}^{p-1} \sum_{m=\ell+1}^{p} \theta_{j,\ell,m} u_j u_\ell u_m \right). \tag{A.3}$$

For every $1 \leq j < \ell < m \leq p$, let $\theta^*_{j,\ell} = \theta^*_{\ell,j}$ and $\theta^*_{j,\ell,m} = \theta^*_{\ell,j,m} = \theta^*_{m,j,\ell}$. Under (A.2), we have, for every $\mathbf{u} = (u_1, ..., u_p) \in \{0,1\}^p$ and every $j \in [p]$,

$$\text{logit}\{\mathbb{P}_{\boldsymbol{\Theta}^*}(U_j = 1 | \mathbf{U}_{-j} = \mathbf{u}_{-j})\} = \theta^*_j + \sum_{\ell\neq j} \theta^*_{j,\ell} u_\ell + \sum_{\ell\neq j} \sum_{m\neq j,\ell} \theta^*_{j,\ell,m} u_\ell u_m = \theta^*_j + \boldsymbol{\theta}^{*T}_{-j} \tilde{\mathbf{u}}_{-j}, \tag{A.4}$$

with $\tilde{\mathbf{u}}_{-j} = (u_1, ..., u_{j-1}, u_{j+1}, ..., u_p, u_1u_2, ..., u_1u_{j-1}, u_1u_{j+1}, ..., u_1u_p, ..., u_{p-1}u_p)^T \in \{0,1\}^{p(p+1)/2}$ and

$\boldsymbol{\theta}^*_{-j} = (\theta^*_{j,1}, .., \theta^*_{j,j-1}, \theta^*_{j,j+1}, ..., \theta^*_{j,p}, \theta^*_{j,1,2}, .., \theta^*_{j,1,j-1}, \theta^*_{j,1,j+1}, ..., \theta^*_{j,1,p}, ..., \theta^*_{j,p-1,p})^T$.

Parameter estimates can be obtained by performing p logistic regressions, which would return two estimates for each parameter $\theta^*_{j,\ell}$, and three estimates for each parameter $\theta^*_{j,\ell,m}$. More precisely, we obtain $\hat{\theta}_{j,\ell}$ and $\hat{\theta}_{j,\ell,m}$ when taking the j-th variable as the response variable (and the other variables as explanatory variables), we obtain $\hat{\theta}_{\ell,j}$ and $\hat{\theta}_{\ell,j,m}$ when taking the ℓ-th variable as the response variable, and we obtain $\hat{\theta}_{m,j,\ell}$ when taking the m-th variable as the response variable and the others as explanatory variables. In general, we have $\hat{\theta}_{j,\ell} \neq \hat{\theta}_{\ell,j}$ $\hat{\theta}_{j,\ell,m} \neq \hat{\theta}_{\ell,j,m} \neq \hat{\theta}_{m,j,\ell}$. These asymmetry issues can be solved by extending the ideas of SepLogit AND and SepLogit MIN for instance.

Then, all the methods presented in this article can be easily extended using the same criteria as the ones presented under the simple Ising model, but with the following quantities

$\boldsymbol{\theta}^*_{-j} = (\theta^*_{j,1}, .., \theta^*_{j,j-1}, \theta^*_{j,j+1}, ..., \theta^*_{j,p}, \theta^*_{j,1,2}, .., \theta^*_{j,1,j-1}, \theta^*_{j,1,j+1}, ..., \theta^*_{j,1,p}, ..., \theta^*_{j,p-1,p})^T \in \mathbb{R}^{p(p+1)/2}$

$\boldsymbol{\mathcal{U}}_{-j} = (\mathcal{U}_1, ..., \mathcal{U}_{j-1}, \mathcal{U}_{j+1}, ..., \mathcal{U}_p, \mathcal{U}_1\mathcal{U}_2, ..., \mathcal{U}_1\mathcal{U}_{j-1}, \mathcal{U}_1\mathcal{U}_{j+1}, ..., \mathcal{U}_1\mathcal{U}_p, ..., \mathcal{U}_{p-1}\mathcal{U}_p)^T \in \{0,1\}^{p(p+1)/2}$

$\boldsymbol{\theta}_{-j}^{(k)*} = (\theta_{j,1}^{(k)*}, .., \theta_{j,j-1}^{(k)*}, \theta_{j,j+1}^{(k)*}, ..., \theta_{j,p}^{(k)*}, \theta_{j,1,2}^{(k)*}, .., \theta_{j,1,j-1}^{(k)*}, \theta_{j,1,j+1}^{(k)*}, ..., \theta_{j,1,p}^{(k)*}, ..., \theta_{j,p-1,p}^{(k)*})^T \in \mathbb{R}^{p(p+1)/2}$
$\boldsymbol{\mathcal{U}}_{-j}^{(k)} = (\mathcal{U}_1^{(k)}, ..., \mathcal{U}_{j-1}^{(k)}, \mathcal{U}_{j+1}^{(k)}, ..., \mathcal{U}_p^{(k)}, \mathcal{U}_1^{(k)}\mathcal{U}_2^{(k)}, ..., \mathcal{U}_1^{(k)}\mathcal{U}_{j-1}^{(k)}, \mathcal{U}_1^{(k)}\mathcal{U}_{j+1}^{(k)}, ..., \mathcal{U}_1^{(k)}\mathcal{U}_p^{(k)}, ..., \mathcal{U}_{p-1}^{(k)}\mathcal{U}_p^{(k)})^T \in \{0,1\}^{p(p+1)/2}$.

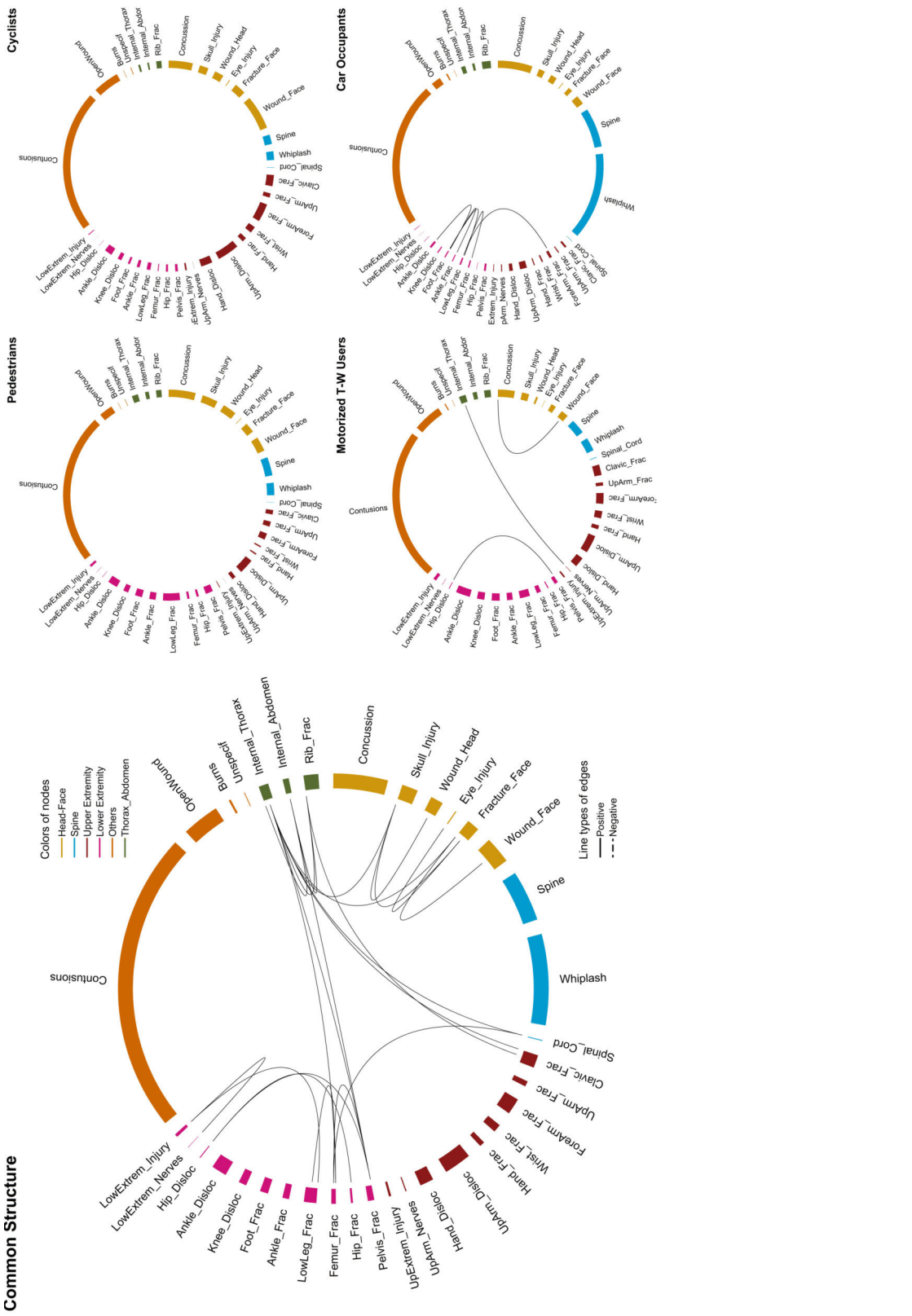


FIGURE A.4: Application of the DataShared-SepLogit Adaptive MIN approach on the Rhône Registry Data to describe the injury tables of victims of road accidents, according to road user type : pedestrians, cyclists, motorized T-W and car occupants. Only edges corresponding to conditional odds-ratios greater than or equal to 2 are represented.

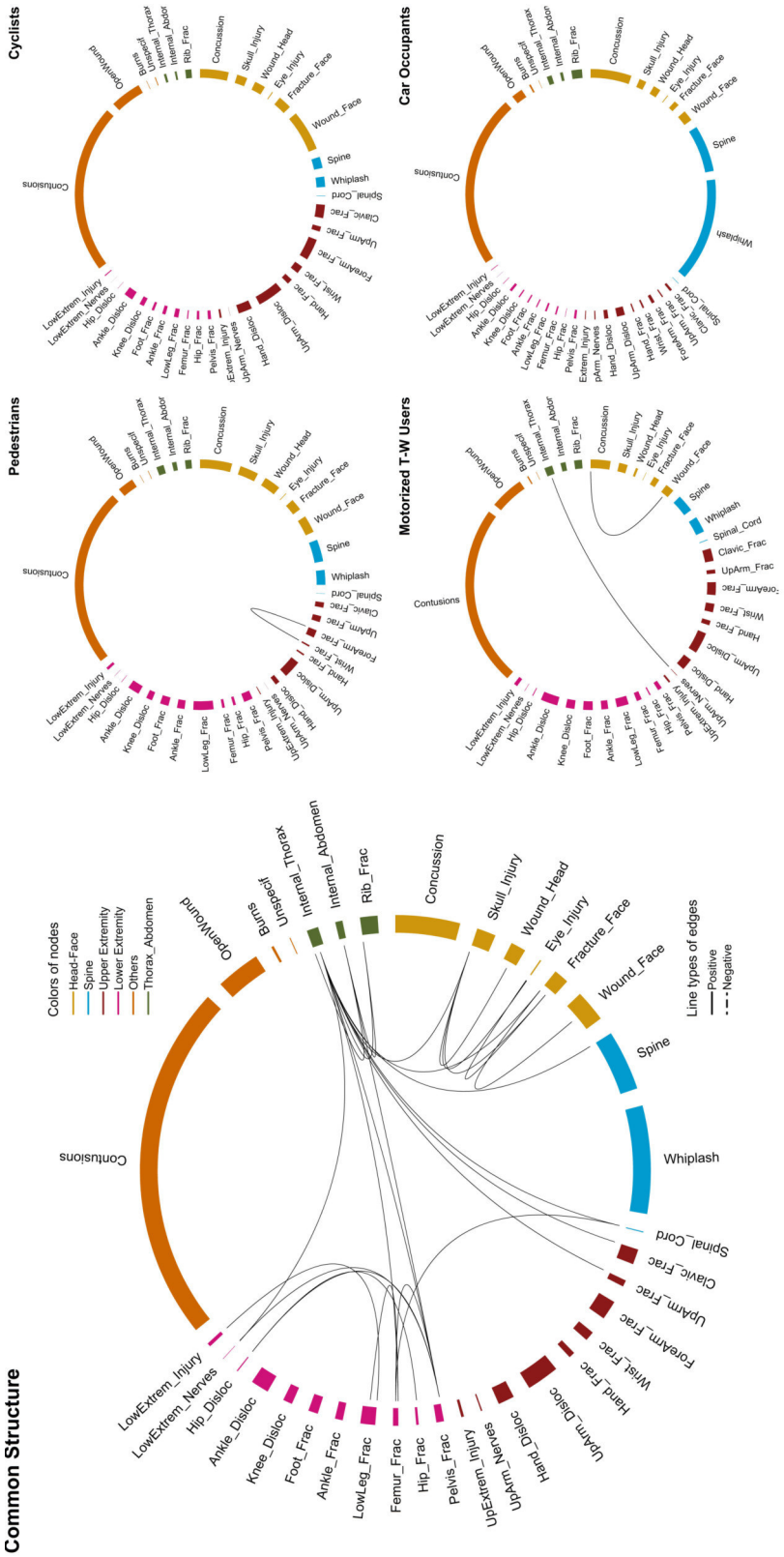


FIGURE A.5: Application of the Ref-SepLogit Adaptive MIN approach on the Rhône Registry Data to describe the injury tables of victims of road accidents, according to road user type : pedestrians, cyclists, motorized T-W and car occupants. Only edges corresponding to conditional odds-ratios greater than or equal to 2 are represented. The reference stratum was set to "car occupants".

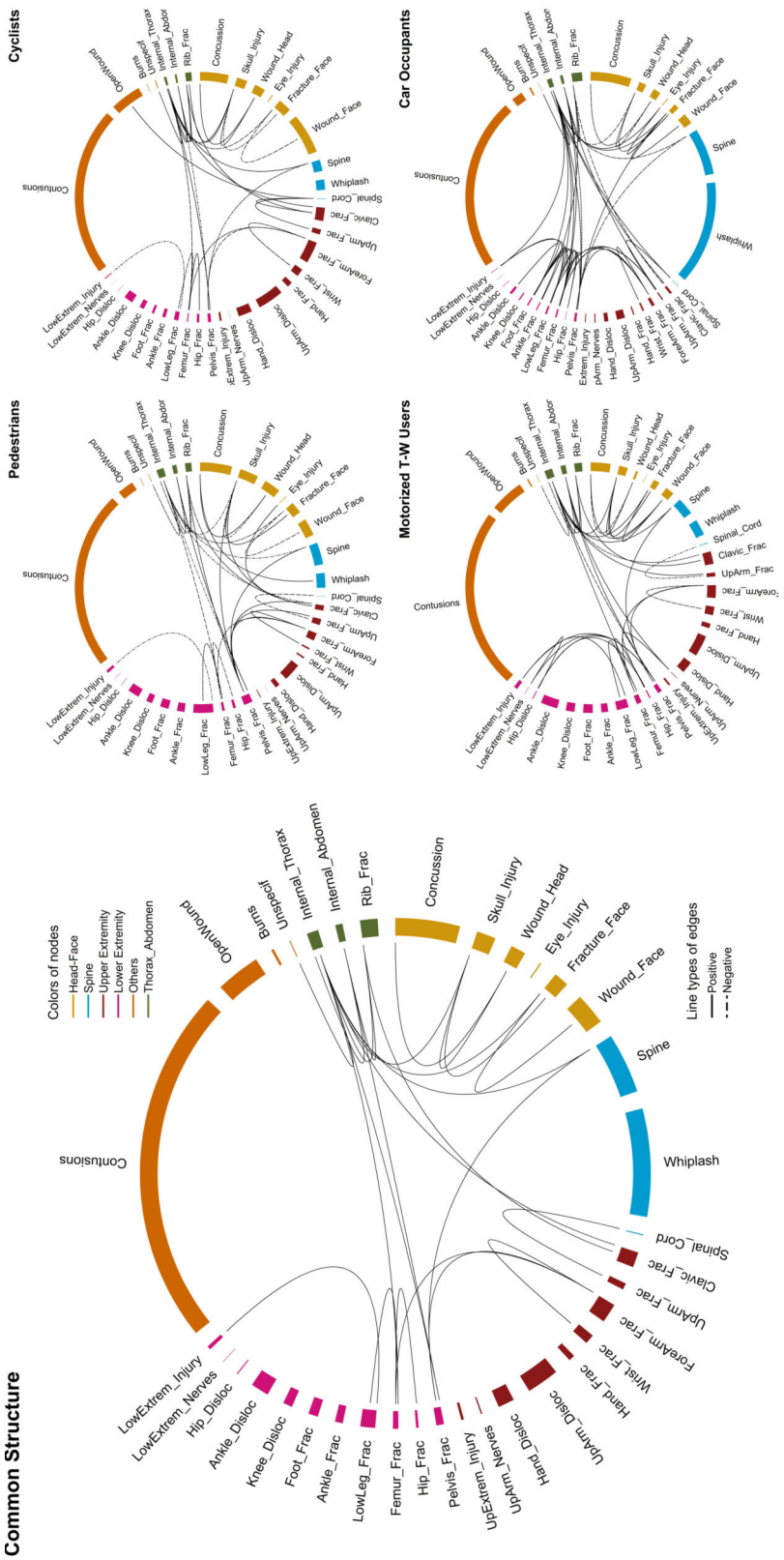


FIGURE A.6: Application of the Indep-SepLogit Adaptive MIN approach on the Rhône Registry Data to describe the injury tables of victims of road accidents, according to road user type : pedestrians, cyclists, motorized T-W and car occupants. Only edges corresponding to conditional odds-ratios greater than or equal to 2 are represented.

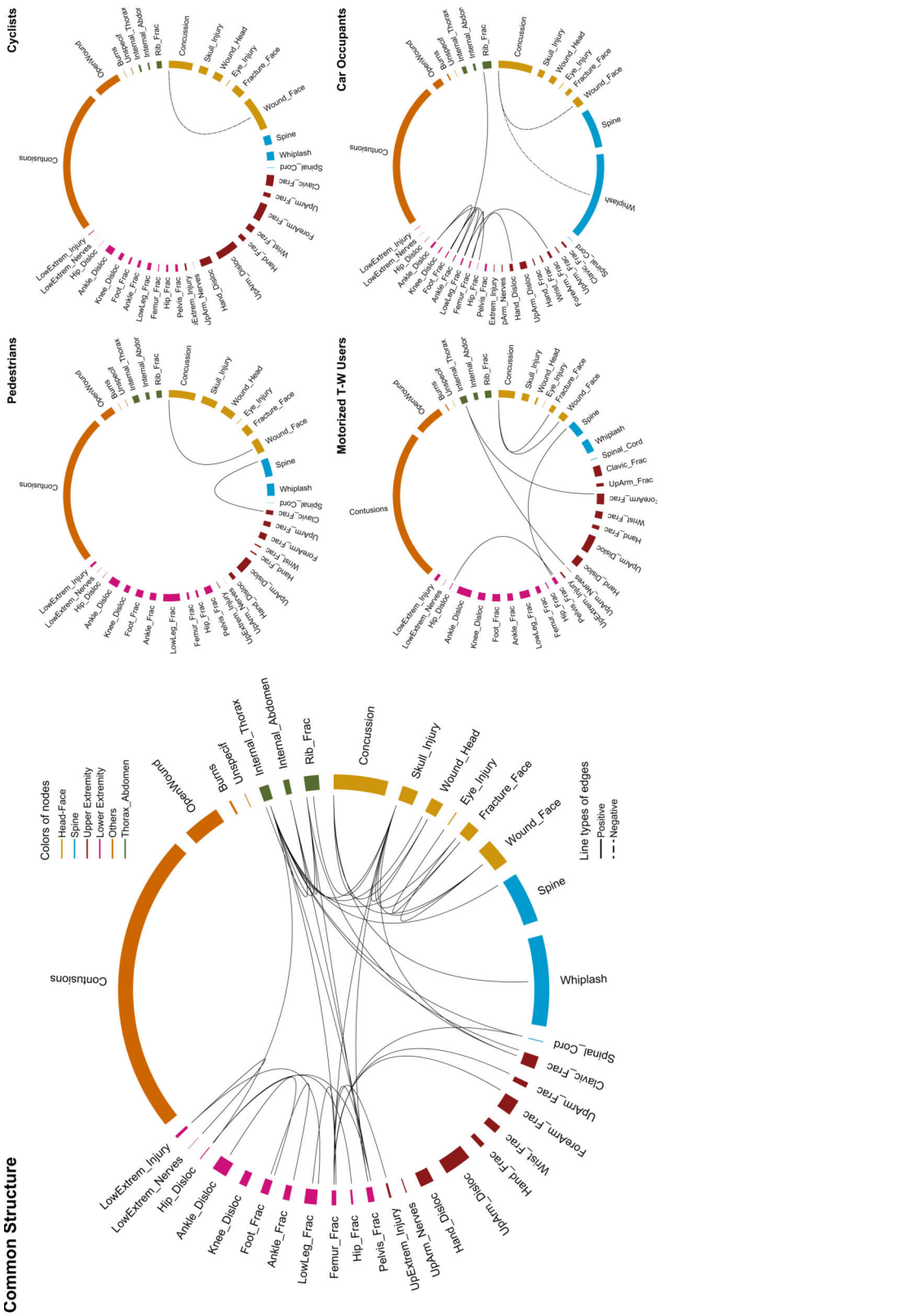


FIGURE A.7: Application of the DataShared-SepLogit Adaptive MIN approach on the Rhône Registry Data to describe the injury tables of victims of road accidents, according to road user type : pedestrians, cyclists, motorized T-W and car occupants. All positive conditional associations are represented.

Annexe B

Sparse estimation for case-control studies with multiple disease subtypes : Supplementary Materials

Additional technical details

B.1 Details on the "standard" L_1-penalized approaches presented in the matched design

Here, we provide additional details on the link between CondLogist_DataSharedLasso and the three more standard approaches presented in the matched design (CondLogist_IndepLasso, CondLogist_PooledLasso and CondLogist_RefLasso).

CondLogist_DataSharedLasso. First recall that estimates produced by CondLogist_DataSharedLasso are defined as $\hat{\boldsymbol{\delta}}_k = \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\gamma}}_k$ with

$$(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_{K-1}) = \text{argmax}_{(\boldsymbol{\mu}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{K-1}) \in \mathbb{R}^{p \times K}} \sum_{k=1}^{K-1} L_k^{(cond)}(\boldsymbol{\mu} + \boldsymbol{\gamma}_k) - \lambda(\|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K-1} \|\boldsymbol{\gamma}_k\|_1)$$

CondLogist_IndepLasso. This approach simply consists in working with the original parametrization and performing one L_1-penalized conditional logistic regression on

each subsample independently. Then, the most natural way of defining estimates returned by CondLogist_IndepLasso is

$$(\hat{\boldsymbol{\delta}}_1, \ldots, \hat{\boldsymbol{\delta}}_{K-1}) = \text{argmax}_{(\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_{K-1})\in\mathbb{R}^{p\times(K-1)}} \Big[\sum_{k=1}^{K-1} \big\{ L_k^{(cond)}(\boldsymbol{\delta}_k) - \lambda\|\boldsymbol{\delta}_k\|_1 \big\} \Big]$$

But, CondLogist_IndepLasso can also be seen as a special case of CondLogist_DataSharedLasso since we also have $(\hat{\boldsymbol{\delta}}_1, \ldots, \hat{\boldsymbol{\delta}}_{K-1}) = (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_{K-1})$, with

$$(\mathbf{0}_p, \hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_{K-1}) = \text{argmax}_{\substack{(\boldsymbol{\mu},\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_{K-1})\in\mathbb{R}^{p\times K} \\ \boldsymbol{\mu}=\mathbf{0}_p}} \Big[\sum_{k=1}^{K-1} \big\{ L_k^{(cond)}(\boldsymbol{\mu}+\boldsymbol{\gamma}_k) \big\} - \lambda(\|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K-1} \|\boldsymbol{\gamma}_k\|_1) \Big]$$

$$i.e., \quad (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_{K-1}) = \text{argmax}_{(\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_{K-1})\in\mathbb{R}^{p\times K}} \Big[\sum_{k=1}^{K-1} \big\{ L_k^{(cond)}(\boldsymbol{\gamma}_k) - \lambda\|\boldsymbol{\gamma}_k\|_1 \big\} \Big]$$

While estimated simultaneously, the $(K-1)$ parameter vectors are still estimated independently. This approach cannot take advantage of any potential similarity among $(\boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_K^*)$, and typically produces estimates with suboptimal properties when such similarity exists.

CondLogist_PooledLasso. This approach works under the (strong) assumption that all disease subtypes share the same parameter vector : $\boldsymbol{\delta}_1^* = \ldots = \boldsymbol{\delta}_{K-1}^* = \boldsymbol{\delta}^*$. Then, the most natural way of defining estimates returned by CondLogist_PooledLasso is

$$\hat{\boldsymbol{\delta}} = \text{argmax}_{\boldsymbol{\delta}\in\mathbb{R}^p} \Big[\sum_{k=1}^{K-1} L_k^{(cond)}(\boldsymbol{\delta}) - \lambda\|\boldsymbol{\delta}\|_1 \Big]$$

Again, there is a link between CondLogist_PooledLasso and CondLogist_DataSharedLasso since $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\mu}}$, with

$$(\hat{\boldsymbol{\mu}}, \mathbf{0}_p, \ldots, \mathbf{0}_p) = \text{argmax}_{\substack{(\boldsymbol{\mu},\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_{K-1})\in\mathbb{R}^{p\times K} \\ \boldsymbol{\gamma}_1=\ldots=\boldsymbol{\gamma}_{K-1}=\mathbf{0}_p}} \Big[\sum_{k=1}^{K-1} L_k^{(cond)}(\boldsymbol{\mu}+\boldsymbol{\gamma}_k) - \lambda(\|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K-1} \|\boldsymbol{\gamma}_k\|_1) \Big]$$

$$i.e., \quad \hat{\boldsymbol{\mu}} = \text{argmax}_{\hat{\boldsymbol{\mu}}\in\mathbb{R}^p} \Big[\sum_{k=1}^{K-1} L_k^{(cond)}(\boldsymbol{\mu}) - \lambda\|\boldsymbol{\mu}\|_1 \Big]$$

Because all $\boldsymbol{\delta}_k^*$'s are assumed to be equal, this approach obviously produces biased estimates when differences exist among the $\boldsymbol{\delta}_k^*$'s.

CondLogist_RefLasso. For simplicity, assume that the first disease subtype is chosen as the reference. Then, CondLogist_RefLasso works under the following reparametrization : $\boldsymbol{\delta}_k^* = \boldsymbol{\delta}_1^* + \boldsymbol{\gamma}_k^*$ for all $k \geq 2$. The most natural way of defining estimates returned by CondLogist_RefLasso is

$$(\hat{\boldsymbol{\delta}}_1, \hat{\boldsymbol{\gamma}}_2, \ldots, \hat{\boldsymbol{\gamma}}_{K-1}) = \text{argmax}_{(\boldsymbol{\delta}_1, \boldsymbol{\gamma}_2 \ldots, \boldsymbol{\gamma}_{K-1}) \in \mathbb{R}^{p \times (K-1)}} \Big[L_1^{(cond)}(\boldsymbol{\delta}_1) + \sum_{k=2}^{K-1} L_k^{(cond)}(\boldsymbol{\delta}_1 + \boldsymbol{\gamma}_k) - \lambda(\|\boldsymbol{\delta}_1\|_1 + \sum_{k=2}^{K-1} \|\boldsymbol{\gamma}_k\|_1) \Big]$$

But, we also have $(\hat{\boldsymbol{\delta}}_1, \ldots, \hat{\boldsymbol{\delta}}_{K-1}) = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}} + \widehat{\boldsymbol{\gamma}}_2, \ldots, \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\gamma}}_{K-1})$, with

$$(\hat{\boldsymbol{\mu}}, \mathbf{0}_p, \hat{\boldsymbol{\gamma}}_2, \ldots, \hat{\boldsymbol{\gamma}}_{K-1}) = \text{argmax}_{\substack{(\boldsymbol{\mu}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{K-1}) \in \mathbb{R}^{p \times K} \\ \boldsymbol{\gamma}_1 = \mathbf{0}_p}} \Big[\sum_{k=1}^{K-1} \big\{ L_k^{(cond)}(\boldsymbol{\mu} + \boldsymbol{\gamma}_k) \big\} - \lambda(\|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K-1} \|\boldsymbol{\gamma}_k\|_1) \Big]$$

$$i.e., \quad (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\gamma}}_2, \ldots, \hat{\boldsymbol{\gamma}}_{K-1}) = \text{argmax}_{(\boldsymbol{\mu}, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_{K-1}) \in \mathbb{R}^{p \times (K-1)}} \Big[L_1^{(cond)}(\boldsymbol{\mu}) + \sum_{k=2}^{K-1} L_k^{(cond)}(\boldsymbol{\mu} + \boldsymbol{\gamma}_k) - \lambda(\|\boldsymbol{\mu}\|_1 + \sum_{k=2}^{K-1} \|\boldsymbol{\gamma}_k\|_1) \Big]$$

More generally, when used with the r-th subtype as the reference, this approach encourages similarity between $\boldsymbol{\delta}_r^*$ and the remaining $K-2$ vectors. Consequently, its performance depends on the choice of the reference subtype. In particular, it will perform poorly if the true parameter vector pertaining to the chosen reference subtype is actually very different from the $K-2$ other ones. As explained in the main text, the optimal reference subtype is generally covariate-specific. CondLogist_DataSharedLasso bypasses the arbitrary choice of the reference subtype. Moreover, in the setting of stratified linear regression models, the DataShared Lasso strategy was shown to perform as well as the optimal (and non-implementable) strategy based on an a priori selection of optimal covariate-specific references.

B.2 Equivalence between MultinomLogist_(SymLasso and StdDataSharedLasso)

First observe that the contribution of an individual with covariate vector $\mathbf{x}_0$ to the likelihood of the symmetric formulation of the model (see Equation (3.8) of the Main Manuscript) is

$$\begin{aligned}
\prod_{k=1}^{K}\left\{p_k(\mathbf{x}_0;\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_K)\right\}^{\mathbf{1}(Y=k)} &= \prod_{k=1}^{K}\left\{\frac{\exp(\mathbf{x}_0^T\boldsymbol{\beta}_k)}{\sum_{\ell=1}^{K}\exp(\mathbf{x}_0^T\boldsymbol{\beta}_\ell)}\right\}^{\mathbf{1}(Y=k)} \\
&\overset{(*)}{=} \prod_{k=1}^{K-1}\left\{\frac{\exp(\mathbf{x}_0^T\boldsymbol{\gamma}_k)}{\exp(-\boldsymbol{\mu}^T\mathbf{x}_0)+\sum_{\ell=1}^{K-1}\exp(\mathbf{x}_0^T\boldsymbol{\gamma}_\ell)}\right\}^{\mathbf{1}(Y=k)} \times \left\{\frac{\exp(-\boldsymbol{\mu}^T\mathbf{x}_0)}{\exp(-\boldsymbol{\mu}^T\mathbf{x}_0)+\sum_{\ell=1}^{K-1}\exp(\mathbf{x}_0^T\boldsymbol{\gamma}_\ell)}\right\}^{\mathbf{1}(Y=K)} \\
&= \prod_{k=1}^{K-1}\left\{\frac{\exp(\mathbf{x}_0^T(\boldsymbol{\mu}+\boldsymbol{\gamma}_k))}{1+\sum_{\ell=1}^{K-1}\exp(\mathbf{x}_0^T(\boldsymbol{\mu}+\boldsymbol{\gamma}_\ell))}\right\}^{\mathbf{1}(Y=k)} \times \left\{\frac{1}{1+\sum_{\ell=1}^{K-1}\exp(\mathbf{x}_0^T(\boldsymbol{\mu}+\boldsymbol{\gamma}_\ell))}\right\}^{\mathbf{1}(Y=K)} \\
&= \prod_{k=1}^{K}\left\{p_k(\mathbf{x}_0;\boldsymbol{\mu}+\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\mu}+\boldsymbol{\gamma}_{K-1},\mathbf{0}_p)\right\}^{\mathbf{1}(Y=k)}
\end{aligned}$$

where we used the change of variable $\boldsymbol{\mu} = -\boldsymbol{\beta}_K$ and $\boldsymbol{\gamma}_k = \boldsymbol{\beta}_k$ for all $k < K$ to obtain the equality $(*)$. Using the same change of variable, we get $\sum_{k=1}^{K}\|\boldsymbol{\beta}_k\|_1 = \|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K-1}\|\boldsymbol{\gamma}_k\|_1$. Putting this all together, the L_1-penalized criterion (3.9) of the Main Manuscript equals, up to a change of variable,

$$\begin{aligned}
L(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_K) - \lambda\sum_{k=1}^{K}\|\boldsymbol{\beta}_k\|_1 &= \frac{1}{n}\sum_{i=1}^{n}\log\{p_{y_i}(\mathbf{x}_i;\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_K)\} - \lambda\sum_{k=1}^{K}\|\boldsymbol{\beta}_k\|_1 \\
&= \frac{1}{n}\sum_{i=1}^{n}\log\{p_{y_i}(\mathbf{x}_i;\boldsymbol{\mu}+\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\mu}+\boldsymbol{\gamma}_{K-1},\mathbf{0}_p)\} - \lambda\Big(\|\boldsymbol{\mu}\|_1 + \sum_{k=1}^{K-1}\|\boldsymbol{\gamma}_k\|_1\Big).
\end{aligned}$$

B.3 Additional details on the AUC criteria

In the unmatched setting, the AUC was computed as an adaptation of the one class versus all other classes approach (Provost and Domingos, 2000; Fawcett, 2006). More precisely, first remind that we generate data such that, e.g., $\boldsymbol{\delta}_4^* = \boldsymbol{\delta}_5^* = \boldsymbol{\delta}_6^*$ (under the configurations described as full homogeneity, low heterogeneity and moderate heterogeneity). Then, the

three classes 4, 5, 6 are undistinguishable under these configurations. More generally, the set of classes $\{1, ..., K\}$ can be partitioned into $G = \{g_1, ..., g_I\}$, where $I \leq K$ and $g_i \subset \{1, ..., K\}$, for all $i = \{1, \ldots, I\}$, such that $(\exists i \in \{1, \ldots, I\} : (k_1, k_2) \in g_i) \Leftrightarrow (\boldsymbol{\delta}^*_{k_1} = \boldsymbol{\delta}^*_{k_2})$; we set $\boldsymbol{\delta}^*_K = \mathbf{0}_p$ for the class corresponding to controls. For the sake of completeness, these partitions are as follows under the four configurations we considered

— Full homogeneity : $G = \{g_1, g_2\}$, $|G| = 2$, $g_1 = \{1, ..., 6\}$ and $g_2 = \{7\}$

— Low heterogeneity : $G = \{g_1, g_2, g_3\}$, $|G| = 3$, $g_1 = \{1\}$, $g_2 = \{2, ..., 6\}$ and $g_2 = \{7\}$

— Moderate heterogeneity : $G = \{g_1, g_2, g_3, g_4, g_5\}$, $|G| = 5$, $g_1 = \{1\}$, $g_2 = \{2\}$, $g_3 = \{3\}$, $g_4 = \{4, ..., 6\}$ and $g_5 = \{7\}$

— Full heterogeneity : $G = \{g_1, g_2, ..., g_7\}$, $|G| = 7$, for each $k \in \{1, ..., k\}$, $g_k = k$

Given such a partition G of $\{1, \ldots, K\}$, for any $g \subset G$, let $y_i^{(g)} = 1$ if $y_i \in g$ and 0 otherwise. Then set $n_g = \sum_{i=1}^n \mathbb{1}[y_i \in g]$, $T_0^{(g)} = \{i \in \{1, ..., n\} : y_i^{(g)} = 0\}$, $T_1^{(g)} = \{i \in \{1, ..., n\} : y_i^{(g)} = 1\}$, and

$$\hat{p}_i^{(g)} = \sum_{k \in g} \hat{p}_i^{(k)} = \sum_{k \in g} \frac{\exp\left(x_i^T \hat{\boldsymbol{\delta}}_k\right)}{\sum_{\ell=1}^K \exp\left(x_i^T \hat{\boldsymbol{\delta}}_\ell\right)}.$$

The AUC was then simply computed as

$$AUC = \sum_{g \subset G} \frac{n_g}{n} \frac{1}{|T_1^{(g)}| \times |T_0^{(g)}|} \sum_{i_1 \in T_1^{(g)}} \sum_{i_2 \in T_0^{(g)}} \mathbb{1}[\hat{p}_{i_1}^{(g)} > \hat{p}_{i_2}^{(g)}].$$

It is a weighted average of the one class versus all class AUC (Provost and Domingos, 2000; Fawcett, 2006), with classes replaced by the groups of classes in G.

B.4 Additional details on the simulation study

For any real numbers $a < b$, we denote the uniform distribution on $[a, b]$ by $\mathcal{U}_{[a,b]}$. For any $p \in [0, 1]$, we further denote the Bernoulli distribution with parameter p by $B(p)$. Parameters $\delta^*_{k,j}$ were generated as follows. One subset $J_1 \subset \{1, ..., p\}$ was first randomly selected, with $|J_1| = 10$. For $j \notin J_1$, we set $\delta^*_{k,j} = 0$ for all $k \in \{1, ..., K-1\}$. For $j \in J_1$, four configurations were considered, allowing the level of homogeneity among $(\delta^*_{1,j}, \ldots, \delta^*_{K-1,j})$ to vary. In the first configuration (full homogeneity), we set $\delta^*_{k,j} = (2\iota_j - 1)\delta$, for some $\delta > 0$ and with $\iota_j \sim B(1/2)$. In the second configuration (low

heterogeneity), for $j \in J_1$, we set $\delta^*_{k,j} = (2\iota_j - 1)\delta$ for $k \geq 2$ and $\delta^*_{1,j} = (2\iota_{k_j,j} - 1)\delta(1 + U_{k_j,j})$, with each $\iota_{k,j} \sim B(1/2)$ and $U_{k_j,j} \sim \mathcal{U}_{[\sqrt{K}/2,2\sqrt{K}]}$. Here, the limits $[\sqrt{K}/2, 2\sqrt{K}]$ were motivated by the non-asymptotic results obtained by Ollier and Viallon (2017) under stratified linear regression models (see the comment right after Theorem 1). In the third configuration (moderate heterogeneity), we set $\delta^*_{k,j} = (2\iota_j - 1)\delta$ for $k \notin \{1, 2, 3\}$ and $\delta^*_{k,j} = (2\iota_{k,j} - 1)\delta(1 + U_{k,j})$ for $k \in \{4, 5, 6\}$, with again $\iota_j \sim B(1/2)$, $\iota_{k,j} \sim B(1/2)$ and $U_{k,j} \sim \mathcal{U}_{[\sqrt{K}/2,2\sqrt{K}]}$. Finally, in the fourth configuration (full heterogeneity), we set $\delta^*_{k,j} = (2\iota_{k,j} - 1)\delta(1 + U_{k,j})$ for $k \in \{1, ..., K - 1\}$ with again $\iota_{k,j} \sim B(1/2)$ and $U_{k,j} \sim \mathcal{U}_{[\sqrt{K}/2,2\sqrt{K}]}$. In each configuration, parameter δ varied in $\{0.1, 0.25, 0.5, 0.75\}$ to study the impact of signal strength on the performance of the approaches ; these δ values correspond to log-odds-ratio for an increment of one standard deviation.

For each observation, covariates were generated under a multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}_p, \mathbf{\Sigma})$, where $\mathbf{\Sigma}_{i,j} = 0.3^{|i-j|}$. Pairs of observations were then created and randomly assigned to one stratum $\mathcal{M}_k$ in such a way that $m_1 = 200$, $m_2 = 100$ and $m_k = 50$ for $k = 3, \ldots, 6$. Within each pair ℓ of each stratum $\mathcal{M}_k$, the response variable $Y^{(k)}_{\ell,1}$ was then generated under model (3.1), that is, $Y^{(k)}_{\ell,1}$ was drawn from a Bernoulli distribution with parameter equal to

$$\frac{\exp(\boldsymbol{\delta}^{*T}_k \mathbf{x}^{(k)}_{\ell,1})}{\exp(\boldsymbol{\delta}^{*T}_k \mathbf{x}^{(k)}_{\ell,1}) + \exp(\boldsymbol{\delta}^{*T}_k \mathbf{x}^{(k)}_{\ell,2}))}.$$

Then, $Y^{(k)}_{\ell,2}$ was set to $1 - Y^{(k)}_{\ell,1}$. Denoting by *case* [resp. *control*] the index of the case [resp. control] in the ℓ-th generated pair, we then have

$$\Pr(Y^{(k)}_{\ell,case} = 1 | Y^{(k)}_{\ell,case} + Y^{(k)}_{\ell,control} = 1, \mathbf{x}^{(k)}_{\ell,case}, \mathbf{x}^{(k)}_{\ell,control}) = \frac{\exp(\boldsymbol{\delta}^{*T}_k \mathbf{x}^{(k)}_{\ell,case})}{\exp(\boldsymbol{\delta}^{*T}_k \mathbf{x}^{(k)}_{\ell,case}) + \exp(\boldsymbol{\delta}^{*T}_k \mathbf{x}^{(k)}_{\ell,control})},$$

and our data are indeed generated under model (3.1).

B.5 Additional results from the simulation study

Figure B.1 illustrates the distribution of the criteria whose averages are presented on Figure 3.1 (matched setting) in the main text.

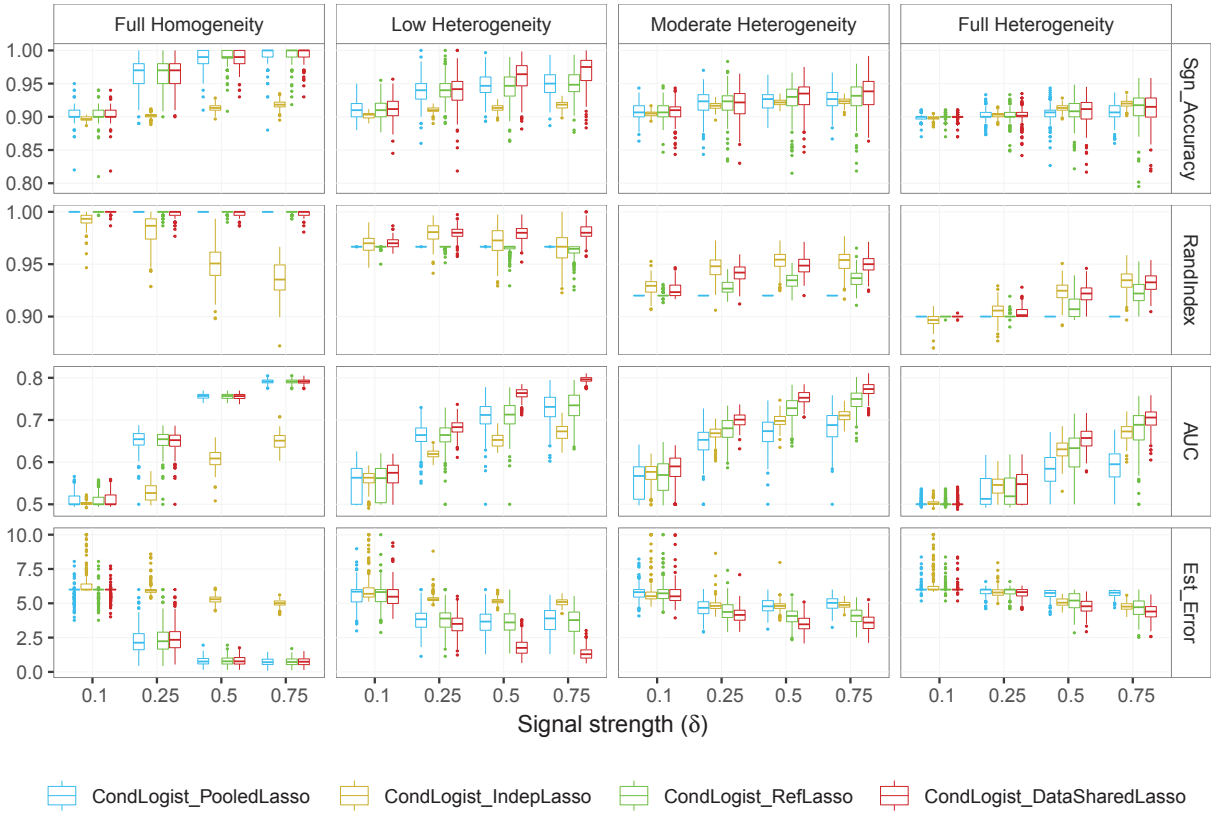


FIGURE B.1: Boxplots showing the distributions of the criteria for each of the four methods compared in the matched setting, over the 200 replicates of each considered configuration and signal strength.

Figure B.2 illustrates the distribution of the criteria whose averages are presented on Figure 3.2 (unmatched setting) in the main text.

B.6 The influence of the reference category when using MultinomLogist_StdLasso : a toy example

Consider a multinomial logistic regression model with $K-1$ disease subtypes, in the particular case of full homogeneity, that is when covariates associated with disease have the same level of association with all disease subtypes. For any $\boldsymbol{\beta}_1^*, ..., \boldsymbol{\beta}_K^*$ satisfying the symmetric formulation of the model, we would then have (i) $\boldsymbol{\beta}_1^* = \cdots = \boldsymbol{\beta}_{K-1}^*$, and (ii) $\boldsymbol{\beta}_K^* \neq \boldsymbol{\beta}_1^*$ (assuming as in the main text that $Y = K$ indicates controls, while $Y = k$ for $k \in \{1, ..., K-1\}$ indicates disease subtype k). Let p_0 denote the number of covariates associated with the disease, for some $1 \leq p_0 \leq p$. If the standard formulation of the model is used with controls as the reference, parameters to be estimated will be

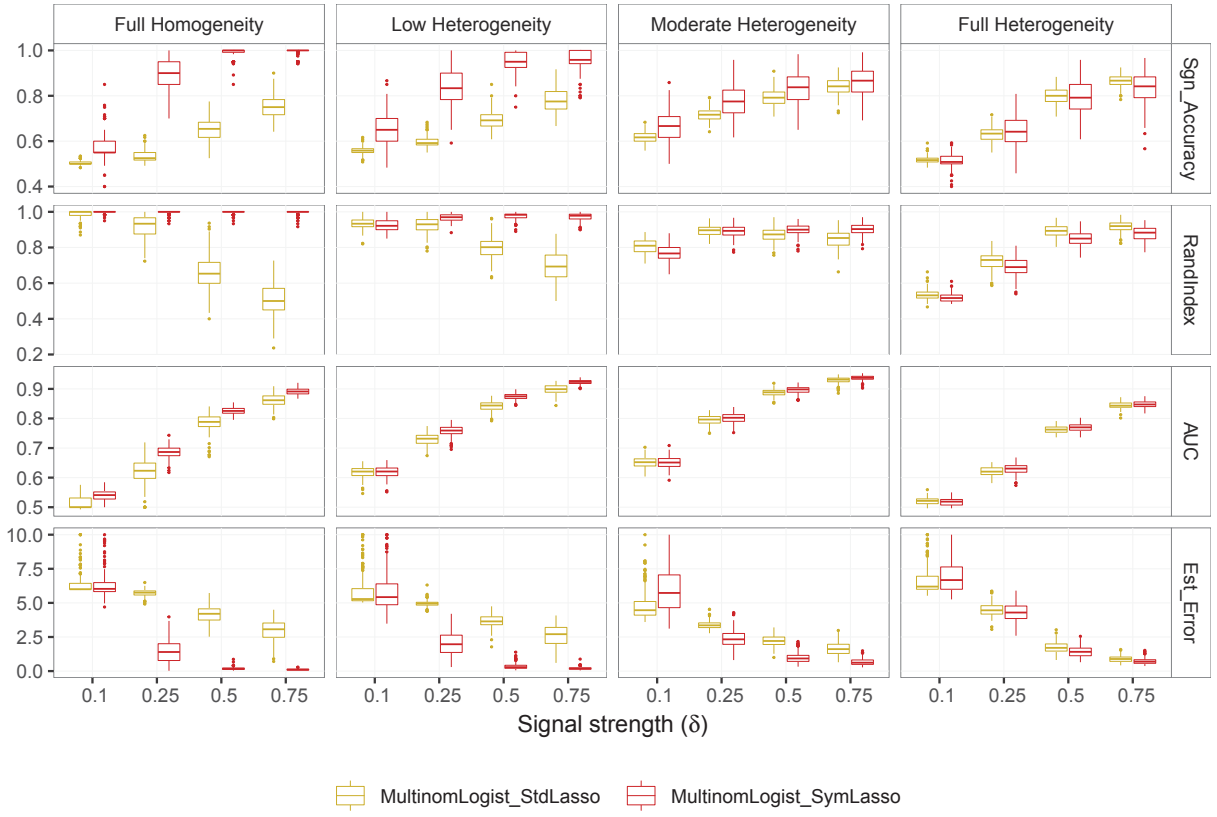


FIGURE B.2: Boxplots showing the distributions of the criteria for the two methods compared in the unmatched setting, over the 200 replicates of each considered configuration and signal strength.

$\boldsymbol{\delta}_1^*, \ldots, \boldsymbol{\delta}_{K-1}^*$, with $\boldsymbol{\delta}_k^* = \boldsymbol{\beta}_k^* - \boldsymbol{\beta}_K^*$ for $k = 1, \ldots, K-1$. Then, these $K-1$ vectors are all equal (since $\boldsymbol{\beta}_1^* = \ldots = \boldsymbol{\beta}_{K-1}^*$), and they all have p_0 non-zero components. The total number of non-zero parameters (hereafter referred to as model complexity) is therefore $(K-1)p_0$. Now, consider again the standard formulation, but this time using the first disease subtype as the reference. The parameters to be estimated would then be $\tilde{\boldsymbol{\delta}}_2^*, \ldots, \tilde{\boldsymbol{\delta}}_K^*$, with $\tilde{\boldsymbol{\delta}}_k^* = \boldsymbol{\beta}_k^* - \boldsymbol{\beta}_1^*$ for $k = 2, \ldots, K$. Because $\boldsymbol{\beta}_1^* = \ldots = \boldsymbol{\beta}_{K-1}^*$, the $K-2$ vectors $(\tilde{\boldsymbol{\delta}}_k^*)_{k=2,\ldots,K-1}$ are all null, while $\tilde{\boldsymbol{\delta}}_K^* = -\boldsymbol{\delta}_1^*$ has p_0 non-zero components. The model complexity is therefore p_0, which is much lower than $(K-1)p_0$. In this particular case, using subtype 1 (or any other subtype) as the reference category yields a parametrization which is much sparser than the parametrization we consider when using controls as the reference. Therefore, when applying MultinomLogist_StdLasso, controls represent the worst choice for the reference category in this situation (we recall that the choice of the reference category would not have any influence on the quality of the estimation in an unpenalized framework). Still considering this toy example, applying MultinomLogist_StdDataSharedLasso corrects any suboptimal choice for the reference category.

For example, whether the reference category is set to controls or to the first disease subtype, the model complexity is p_0 when applying MultinomLogist_StdDataSharedLasso, as illustrated in Figure B.3 below. Finally, MultinomLogist_SymLasso would target the sparsest collection of vectors $\boldsymbol{\beta}_1^*, ..., \boldsymbol{\beta}_K^*$ satisfying the symmetric formulation of the model. In the toy example considered here, this would be $(\mathbf{0}_p, \ldots, \mathbf{0}_p, \tilde{\boldsymbol{\delta}}_K^*)$, as illustrated in Figure B.3. The corresponding model complexity is again p_0.

Figure B.3 below gives a graphical representation of our toy example. It especially illustrates how model complexity (denoted by C) is affected by the choice of the reference category when working under the standard formulation of the multinomial logistic regression model, and how the decomposition targeted by DataShared Lasso corrects any sub-optimal choice for this reference category. For simplicity, Figure B.3 represents the case where $p_0 = p$, and where all covariates share the same level of association with the disease. In this case, a typical collection of vectors $\boldsymbol{\beta}_1^*, ..., \boldsymbol{\beta}_K^*$ satisfying the symmetric formulation of the model is such that, for all $j \in \{1, ..., p\}$, $\beta_{1,j}^* = \beta_1$ and $\beta_{K,j}^* = \beta_2$ for some $\beta_1, \beta_2 \neq 0$ such that $\beta_1 - \beta_2 \neq 0$.

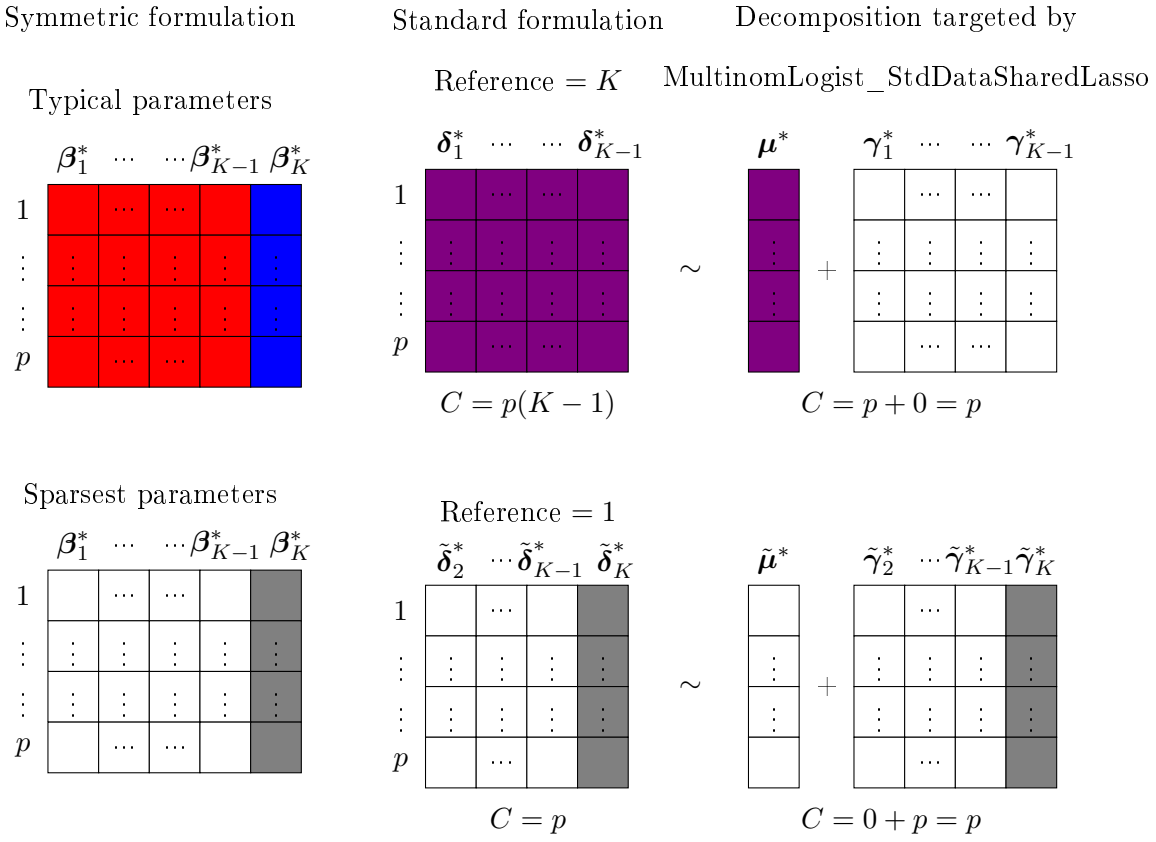


FIGURE B.3: Graphical representation of our toy example. In each matrix, red entries correspond to the "common" value β_1, blue entries correspond to the value β_2, purple entries to the value $(\beta_1 - \beta_2)$, gray entries to the value $-(\beta_1 - \beta_2)$ and white entries to the value 0. If MultinomLogist_StdLasso is applied after selecting the K-th category as the reference, model complexity is $C = (K-1)p$. The choice of the K-th categoty as the reference is clearly sub-optimal since selecting any other category as the reference, e.g. the first one, leads to a model complexity $C = p$. On the other hand, irrespective of the initial choice of the reference category, the complexity of the decomposition targeted by MultinomLogist_StdDataSharedLasso is optimal and equals p.

Bibliographie

Amoros, E., Lardy, A., Martin, J., Wu, D., and Viallon, V. (2019). Méthodologie redressement et extrapolation. l3 deliverable voiesur project. Technical report, ANR-11-VPTT-0007. Available at : https ://hal. archives-ouvertes. fr/hal ....

Amoros, E., Martin, J., and Laumon, B. (2008). Estimation de la morbidité routière, france. *Bulletin Épidémiologique Hebdomadaire, 19, Institut National de Veille Sanitaire*, pages 157–160.

Avalos, M., Pouyes, H., Grandvalet, Y., Orriols, L., and Lagarde, E. (2015). Sparse conditional logistic regression for analyzing large-scale matched data from epidemiological studies : a simple algorithm. *BMC bioinformatics*, 16(6) :S1.

Bach, F. (2010). Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4 :384–414.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2012). Structured sparsity through convex optimization. *Statistical Science*, 27(4) :450–468.

Bachlechner, U., F. A. S. A. e. a. (2016). Associations of anthropometric markers with serum metabolites using a targeted metabolomics approach : results of the epic-potsdam study. *Nutr & Diabetes*, 6 :e215.

Ballout, N., Etievant, L., and Viallon, V. (2020a). On the use of cross-validation for the calibration of the tuning parameter in the adaptive lasso. *ArXiv e-prints arXiv :2005.10119*.

Ballout, N., Garcia, C., and Viallon, V. (2020b). Sparse estimation for case–control studies with multiple disease subtypes. *Biostatistics*. kxz063.

Ballout, N. and Viallon, V. (2019). Structure estimation of binary graphical models on stratified data : Application to the description of injury tables for victims of road accidents. *Statistics in Medicine*, 38(14) :2680–2703.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar) :485–516.

Begg, C. B. and Gray, R. (1984). Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika*, 71(1) :11–18.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732.

Bouyer, J. and Cordier, S. Levallois, P. (2003). Epidémiologie. *Environnement et santé publique - Fondements et pratiques*, pages 89–118.

Bühlmann, P. and Geer, S. (2011). *Statistics for High-Dimensional Data : Method, Theory and Applications*. Springer Series in Statistics.

Bühlmann, P. and Meier, L. (2008). Discussion of "one-step sparse estimates in nonconcave penalized likelihood models" by h. zou and r. li. *Ann. Statist*, 36 :1534–1541.

Carayol, M., Leitzmann, M. F., Ferrari, P., Zamora-Ros, R., Achaintre, D., Stepien, M., Schmidt, J. A., Travis, R. C., Overvad, K., Tjønneland, A., Hansen, L., Kaaks, R., Kühn, T., Boeing, H., Bachlechner, U., Trichopoulou, A., Bamia, C., Palli, D., Agnoli, C., Tumino, R., Vineis, P., Panico, S., Quirós, J. R., Sánchez-Cantalejo, E., Huerta, J. M., Ardanaz, E., Arriola, L., Agudo, A., Nilsson, J., Melander, O., Bueno-de Mesquita, B., Peeters, P. H., Wareham, N., Khaw, K.-T., Jenab, M., Key, T. J., Scalbert, A., and Rinaldi, S. (2017). Blood metabolic signatures of body mass index : A targeted metabolomics study in the epic cohort. *Journal of Proteome Research*, 16(9) :3137–3146. PMID : 28758405.

Champion, M., Picheny, V., and Vignes, M. (2017). Inferring large graphs using ℓ_1-penalized likelihood. *Statistics and Computing*, pages 1–17.

Cheng, J., Levina, E., Wang, P., and Zhu, J. (2014). A sparse ising model with covariates. *Biometrics*, 70(4) :943–953.

Committee on Medical Aspects of Automotive Safety (1971). Rating the severity of tissue damage. i. the abbreviated scale. *Jama*, 215(2) :277–280.

Cox, D. and Wermuth, N. (1994). A note on the quadratic exponential binary distribution. *Biometrika*, 81(2) :403–408.

Danaher, P., Wang, P., and Witten, D. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 76(2) :373–397.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of statistics*, 32(2) :407–499.

El Ghaoui, L., Viallon, V., and Rabbani, T. (2012). Safe feature elimination in sparse supervised learning. *Pacific Journal of Optimization*, 8(4) :667–698.

Evgeniou, T. and Pontil, M. (2004). Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117. ACM.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27(8) :861 – 874. ROC Analysis in Pattern Recognition.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statist. Soft.*, 33(1) :1–22.

Gertheiss, J. and Tutz, G. (2010). Sparse modeling of categorial explanatory variables. *The Annals of Applied Statistics*, 4(4) :2150–2180.

Gertheiss, J. and Tutz, G. (2012). Regularization and model selection with categorial effect modifiers. *Statistica Sinica*, 22(3) :957–982.

Gori, G. B. (1989). Epidemiology and the concept of causation in multifactorial diseases. *Regulatory Toxicology and Pharmacology*, 9(3) :263 – 272.

Greenland, S. (2000). Small-sample bias and corrections for conditional maximum-likelihood odds-ratio estimators. *Biostatistics*, 1(1) :113–122.

Gross, S. M. and Tibshirani, R. (2016). Data shared lasso : A novel tool to discover uplift. *Computational statistics & data analysis*, 101 :226–235.

Guo, J., Cheng, J., Levina, E., Michailidis, G., and Zhu, J. (2015). Estimating heterogeneous graphical models for discrete data with an application to roll call voting. *The Annals of Applied Statistics*, 9(2) :821–848.

Hallac, D., Park, Y., Boyd, S., and Leskovec, J. (2017). Network inference via the time-varying graphical lasso. *CoRR*, abs/1703.01958.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Hernán, M., Hernández-Díaz, S., and Robins, J. (2004). A structural approach to selection bias. *Epidemiology*, 15(5) :615–625.

His, M., V. V. D. L. e. a. (2019). Prospective analysis of circulating metabolites and breast cancer in epic. *BMC Med*, 17.

Höfling, H. and Tibshirani, R. (2009). Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. *Journal of Machine Learning Research*, 10(Apr) :883–906.

Jenicek, M. and Cléroux, R. (1994). Epidémiologie clinique. *Population*, 49-1 :267.

Koura, K. G., Hours, M., Charnay, P., Tournier, C., Javouhey, E., and Luaute, J. (2014). Evolution de la qualité de vie après un traumatisme crânien par accident de la route. In *VIe Congrès International d'Épidémiologie-ADELF-EPITER*, pages 17–p.

Krishnapuram, B., Carin, L., Figueiredo, M. A., and Hartemink, A. J. (2005). Sparse multinomial logistic regression : Fast algorithms and generalization bounds. *IEEE transactions on pattern analysis and machine intelligence*, 27(6) :957–968.

Lauritzen, S. (1996). *Graphical models*, volume 17. Clarendon Press.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference with the lasso. *Ann. Statist.*, 44 :907–927.

Lehmann, R. K., Arthurs, Z. M., Cuadrado, D. G., Casey, L. E., Beekley, A. C., and Martin, M. J. (2007). Trauma team activation : simplified criteria safely reduces overtriage. *The American Journal of Surgery*, 193(5) :630 – 635. PAPERS FROM THE NORTH PACIFIC SURGICAL ASSOCIATION.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *Ann. Statist.*, 42(2) :413–468.

Lounici, K., Pontil, M., van de Geer, S., and Tsybakov, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, pages 2164–2204.

Lyons, R., Polinder, S., Larsen, C., Mulder, S., Meerding, W., Toet, H., Van Beeck, E., and the Eurocost Reference Group (2006). Methodological issues in comparing

injury incidence across countries. *International journal of injury control and safety promotion*, 13(2) :63–70.

Ma, J. and Michailidis, G. (2016). Joint structural estimation of multiple graphical models. *Journal of Machine Learning Research*, 17(166) :1–48.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC press.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462.

Metz, C. E. (1978). Basic principles of roc analysis. *Seminars in Nuclear Medicine*, 8(4) :283 – 298.

Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2) :227–234.

Ollier, E. and Viallon, V. (2014). Joint estimation of K related regression models with simple L_1-norm penalties. *ArXiv e-prints arXiv :1411.1594*.

Ollier, E. and Viallon, V. (2017). Regression modelling on stratified data with the lasso. *Biometrika*, 104(1) :83–96.

Park, M. Y. and Hastie, T. (2007). L_1-regularization path algorithm for generalized linear models. *J. Roy. Statist. Soc. Ser. B*, 69(4) :659–677.

Pearce, N. (2016). Analysis of matched case-control studies. *BMJ*, 352 :i969.

Provost, F. and Domingos, P. (2000). Well-trained pets : Improving probability estimation trees. CeDER Working Paper #IS-00-04, Stern School of Business, New York University, NY 10012.

Prüss-Üstün, A., Mathers, C., Corvalán, C., and Woodward, A. (2003). Introduction and methods : assessing the environmental burden of disease at national and local levels. *WHO Environmental Burden of Disease Series, No. 1*, pages 61–82.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336) :846–850.

Ravikumar, P., Wainwright, M., Lafferty, J., et al. (2010). High-dimensional ising model selection using l_1-regularized logistic regression. *The Annals of Statistics*, 38(3) :1287–1319.

Reid, S. and Tibshirani, R. (2014). Regularization paths for conditional logistic regression : the clogitl1 package. *Journal of statistical software*, 58(12).

Riboli, E., Hunt, K., Slimani, N., Ferrari, P., Norat, T., Fahey, M., Charrondiere, U., Hemon, B., Casagrande, C., Vignat, J., Overvad, K., Tjønneland, A., Clavel-Chapelon, F., Thiébaut, A., Wahrendorf, J., Boeing, H., Trichopoulos, D., Trichopoulou, A., and et al. (2002). European prospective investigation into cancer and nutrition (epic) : study populations and data collection. *Public health nutrition*, 5(6b) :1113–1124.

Rothman, K. J., Greenland, S., and Lash, T. L. (2008). *Modern epidemiology, 3rd edition.* Philadelphia : Wolters Kluwer Health/Lippincott Williams & Wilkins.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2) :461–464.

Sennhenn-Reulen, H. and Kneib, T. (2016). Structured fusion lasso penalized multi-state models. *Statistics in medicine*, 35(25) :4637–4659.

Tao, Q., Huang, X., Wang, S., Xi, X., and Li, L. (2016). Multiple gaussian graphical estimation with jointly sparse penalty. *Signal Processing*, 128 :88–97.

Thacker, S. B. and Stroup, D. F. (1998). Public health surveillance and health services research. *Epidemiology and Health Services*, pages 61–82.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society : Series B (Methodological)*, 58(1) :267–288.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R. (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 74(2) :245–266.

van de Geer, S. and Bühlmann, P. (2013). ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41(2) :536–567.

van der Laan Mark J., C, P. E., and E., H. A. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1) :1–23.

Viallon, V., Banerjee, O., Jougla, E., Rey, G., and Coste, J. (2014). Empirical comparison study of approximate methods for structure selection in binary graphical models. *Biometrical Journal*, 56(2) :307–331.

Viallon, V., Lambert-Lacroix, S., Hoefling, H., and Picard, F. (2016). On the robustness of the generalized fused lasso to prior specifications. *Statistics and Computing*, 26(1) :285–301.

Wainwright, M., Lafferty, J., and Ravikumar, P. (2007). High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In *In Advances in neural information processing systems*, pages 1465–1472.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on*, 55(5) :2183–2202.

Wang, P., Chao, D., and Hsu, L. (2009). Learning networks from high dimensional binary data : An application to genomic instability data. *ArXiv e-prints arXiv :0908.3882*.

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., and Drazen, J. M. (2007). Statistics in medicine—reporting of subgroup analyses in clinical trials. *New England Journal of Medicine*, 357(21) :2189–2194.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6) :714–721.

Yang, E. and Ravikumar, P. (2011). On the use of variational inference for learning discrete graphical model. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1009–1016.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7 :2541–2563.

Zhou, N. and Zhu, J. (2010). Group Variable Selection via a Hierarchical Lasso and Its Oracle Property. *ArXiv e-prints arXiv :1006.2871*.