Intégration de données omiques pour la modélisation de l'impact de l'hétérogénéité inter-tumorale dans la survie de patients atteints de cancer

Sarah-Laure Rincourt

université PARIS-SACLAY

Intégration de données omiques pour la modélisation de l'impact de l'hétérogénéité inter-tumorale dans la survie de patients atteints de cancer

Omics integration to model the impact of inter-tumoral heterogeneity on the survival of patients with cancer

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 570, Santé Publique (EDSP)
Spécialité de doctorat : Santé publique - Biostatistiques
Graduate School : Santé publique, Référent : Faculté de Médecine

Thèse préparée dans l'unité de recherche CESP (Université Paris-Saclay, UVSQ, INSERM) sous la direction de Stefan MICHIELS, PhD-HDR, et sous l'encadrement de Damien DRUBAY, PhD.

Thèse soutenue à Paris-Saclay, le 29 septembre 2022, par

Sarah-Laure RINCOURT

Composition du jury

Pascale TUBERT-BITTER PhD, HDR, DR, Université Paris-Saclay	Présidente
Delphine MAUCORT-BOULCH PhD, HDR, PU-PH, Médecin de Santé Publique, Université Lyon 1	Rapporteur & Examinatrice
Thomas FILLERON PhD, HDR, Université Toulouse 1	Rapporteur & Examinateur
Macha Nikolski PhD, HDR, DR, Université Bordeaux	Examinatrice
Stefan MICHIELS PhD, HDR, DR, Université Paris-Saclay	Directeur de thèse

THESE DE DOCTORAT

NNT : 2022UPASR014

Remerciements

Me voilà à la fin de ces quatre ans de thèse. Que dire de cette thèse tant elle a été marquée par des hauts et des bas ! Mais je me tiens à la fin de cette page de ma vie en ne me remémorant que les bons souvenirs.

Mes remerciements vont à mon directeur Stefan Michiels et à mon encadrant Damien Drubay pour leur temps, leur dévouement et leur implication tout au long de cette thèse. Vous m'avez permis de découvrir le monde de la recherche, de développer mes compétences et d'aller toujours au-delà de moi-même.

Je souhaite remercier l'équipe Oncostat pour m'avoir accueilli et sans qui cette thèse n'aurait pas pu se faire. Je remercie aussi l'école doctorale pour m'avoir permis de recevoir une bourse doctorale et d'avoir permis à ce projet de naître.

Je tiens à remercier mon jury de thèse, notamment mes rapporteurs Delphine Maucort-Boulch et Thomas Filleron, et les membres du mon jury Pascale Tubert-Bitter et mon examinatrice Macha Nikolski qui ont pris de leur temps pour relire ma thèse et me donner un retour.

Ma pensée va aussi à ma famille qui m'a soutenue ces dernières années et ce malgré les pertes que nous avons subies. Merci à mes amis, Manu, Charlotte, Cloé, et tant d'autres vous m'avez toujours écoutée et vous m'avez aidée par nos discussions passionnées qui m'ont toujours remise sur le bon chemin !

Et finalement merci à mon compagnon Billy, sans qui ces années de thèse n'auraient pas pu se finir. Tu as cru en moi et m'a soutenu inconditionnellement malgré ta propre thèse. Tu as été ma force et ma motivation. Je n'ai pas les mots pour te dire que tu as été le plus beau piller de ma vie, merci.

Productions scientifiques

Articles

- Rincourt Sarah-Laure, Michiels Stefan, Drubay Damien. "Complex disease individual molecular characterization using infinite sparse graphical independent component analysis". In : *Cancer informatics*. 2022 Jul.
  DOI : 10.1177/11769351221105776.
- Rincourt Sarah-Laure, Michiels Stefan, Drubay Damien. "A non-parametric Bayesian joint model for latent individual molecular profils and survival in oncology". In : *Journal of Bioinformatics and Computational Biology*. 2022.
  DOI : 10.1142/S0219720022500226.

Communications orales en tant que présentatrice

- Rincourt Sarah-Laure, Michiels Stefan, Drubay Damien, Modélisation de la structure génomique tumorale du cancer du sein par analyse factorielle parcimonieuse non paramétrique. À : Journée GDR « Statistiques et santé »- SFB - groupe Biopharma de la SFDS, Paris, France, 2019.
- Rincourt Sarah-Laure, Michiels Stefan, Drubay Damien, Nonparametric Bayesian sparse infinite factor analysis with baseline profile to recover individual heterogeneity. À : $41^{\text{ème}}$ Conference of the International Society for Clinical Biostatistics (ISCB), en ligne, 2020.
- Rincourt Sarah-Laure, Michiels Stefan, Drubay Damien, Analyse factorielle non-paramétrique parcimonieuse et infinie avec profil de base pour la reconstitution

de l'hétérogénéité individuelle. À : Young Statisticians and Probabilists (YSP), en ligne, 2021.

– Rincourt Sarah-Laure, Michiels Stefan, Drubay Damien, Modèle bayésien conjoint pour l'analyse de profils latents moléculaires individuels et de survie en oncologie À : Epiclin, Paris, France, 2022.

Logiciels

Les simulations, l'optimisation des paramètres et la visualisation des données ont été réalisées à l'aide du logiciel R (version 3.6.0).

Les codes R et les données nécessaires correspondant aux analyses effectuées du Chapitre 3 basés sur le modèle d'analyse graphique avec un nombre infini de composantes indépendantes et parcimonieuses, sont disponibles sur :

— https://github.com/Oncostat/isgICA.

Les codes R et les données nécessaires correspondant aux analyses effectuées du Chapitre 4 basés sur le modèle conjoint seront disponibles sur :

— https://github.com/Oncostat/SisgICA.

Table des matières

Remerciements i
Productions scientifiques iii
Table des matières v
Liste des figures ix
Liste des tableaux x
Notations et Abréviations xi

Introduction générale **1**
Le cancer : une maladie complexe et hétérogène 1
Objectifs et plan de la thèse 3

1 Cadre applicatif : Hétérogénéité des patients atteints du cancer **5**
1.1 Le cancer 6
1.1.1 Principe de la cancérogénèse 6
1.1.2 Le cancer du sein 7
1.1.3 Essais cliniques en oncologie 7
1.1.4 Hétérogénéité des cancers - Apport de la médecine de précision 9
1.2 Méthodes de déconvolution (ou de réduction de dimension) parcimonieuse 11
1.3 Identifier des composantes pronostiques à la survie 12

2 Cadre méthodologique et inférentiel **15**
2.1 L'inférence Bayésienne 16
2.1.1 Principe 16
2.1.2 La définition des priors 17
2.1.3 Inférence Bayésienne non-paramétrique 19
2.2 Optimisation : l'inférence variationnelle 21
2.2.1 Difficultés d'inférence en grande dimension et en présence de labels échangeables des variables latentes 21
2.2.2 Principe 22
2.2.3 Algorithmes d'optimisation de l'ELBO 25
2.3 Modèles à variables latentes pour la réduction de dimension 28

2.3.1 Approche non-supervisée dans le cadre des modèles latents pour la réduction de dimension . . . 29
2.3.2 Comparaison des modèles de déconvolution quantitatives . . . 29

3 Caractérisation moléculaire individuelle à l'aide du modèle isgICA 41
3.1 Introduction . . . 43
3.2 Méthodes . . . 46
3.2.1 Identification du nombre de composantes présentes dans la population - Processus bêta-Bernoulli . . . 46
3.2.2 Modèle graphique pour une analyse en composantes indépendantes infinies et parcimonieuses . . . 50
3.2.3 Travaux connexes précédents . . . 52
3.2.4 Contrainte sur le modèle isgICA : apport du profil de base au modèle . . . 53
3.2.5 Mises à jour . . . 60
3.2.6 Optimisation des hyperparamètres . . . 62
3.2.7 Comparaison à un modèle de référence : l'ICA avec blanchissement 63
3.3 Résultats . . . 63
3.3.1 Données synthétiques . . . 63
3.3.2 Reconstruction des structures latentes . . . 65
3.4 Application : données sur le cancer du sein au stade précoce . . . 73
3.5 Discussion . . . 77
3.6 Conclusion . . . 79

4 Identification de mécanismes moléculaires individuelles pronostiques : un modèle conjoint 81
4.1 Introduction . . . 83
4.2 Travaux connexes précédents . . . 85
4.3 Méthode . . . 86
4.3.1 Modèle joint SisgICA . . . 86
4.3.2 Sous-modèle isgICA . . . 87
4.3.3 Sous-modèle paramétrique de survie : Régression paramétrique de Weibull sous l'hypothèse AFT . . . 88
4.3.4 Modèle complet Bayésien . . . 95
4.3.5 Équations variationnelles SisgICA . . . 97
4.3.6 Optimisation des hyperparamètres . . . 99
4.4 Résultats . . . 100
4.4.1 Données synthétiques . . . 100

4.4.2 Application : early breast cancer data 107
4.5 Discussion 111
4.6 Conclusion 113

Discussion générale **115**

Conclusion générale **121**

Bibliographie **133**

Annexes **134**

A Distribution conjuguée de la famille exponentielle **137**
A.1 Distributions 137
A.2 Exemple de relation conjuguée entre les distributions a priori, la vraisemblance et les distributions a posteriori 140

B Démonstration des équations du modèle isgICA **141**

C Article 1 **149**

D Article 2 **167**

Liste des figures

2.1 Représentation du *label-switching* entre deux variables latentes spécifiées par un paramètre avec en bleu le paramètre 1, en orange le paramètre 2, et par la ligne noire la valeur théorique des paramètres respectivement de -2 et de 3. 22
2.2 Représentation de l'ELBO défini en équation 2.8 comme la limite inférieure de l'évidence mais aussi de la log-vraisemblance. L'image est inspirée de C. Bishop (2006). 24
2.3 Factorisation matricielle de **X** la matrice des données 31
2.4 Adaptée de Lewicki et Sejnowski (2000) - Les composantes latentes sont définies par les axes en bleu, l'ACP suppose l'orthogonalité des composantes lorsque l'ICA suppose l'indépendance des composantes. . 36

3.1 Représentation des allocations de caractéristiques physiologiques de certains groupes d'espèces connues. 46
3.2 nterprétation de la structure de la matrice de parcimonie de l'analyse en composantes indépendantes sous forme d'hypergraphe 48
3.3 ICA (exemple non-bruitée) parcimonieuse pour dans le cas d'un modèle d'expression génique. 51
3.4 ICA parcimonieuse avec un profil de base Z_0 54
3.5 Représentation du prior *ridge* 57
3.6 ICA parcimonieuse avec un profil de base $\mathbf{Z_0}$ et la pseudo-parcimonie de $\mathbf{\Phi}$ 58
3.7 Représentation graphique du modèle isgICA. 59
3.8 Reconstruction du nombre de composantes latentes entre isgICA et l'ICA standard 67
3.9 Performance de la reconstruction de la matrice parcimonieuse des poids $Z \circ W$ des modèles isgICA, *ica*, et *fastICA* 68
3.10 Performance de la reconstruction de la pseudo-parcimonie de la matrice des sources $\mathbf{\Phi}$ des isgICA, ica et fastICA. 69

3.11 Reconstruction de la structure parcimonieuse des poids ($\mathbf{Z}$), de la matrice des poids parcimonieuse ($\mathbf{Z} \circ \mathbf{W}$) et de la matrice des sources ($\mathbf{\Phi}$) 71
3.12 Représentation graphique de la reconstruction de la structure de poids parcimonieux, de la structure de poids parcimonieux et des sources. . 72
3.13 Temps de calcul de l'ica, fastICA et isgICA 74
3.14 Matrice de l'hypergraphe de l'hétérogénéité individuelle (structure de dispersion des poids) extraite de l'ensemble de données sur le cancer du sein 75
3.15 Distribution de la valeur absolue des éléments de la matrice des sources associés aux *probes* des gènes 76

4.1 Transformations des fonctions utilisées dans l'analyse de survie ainsi que les symboles des fonctions associées. 89
4.2 Représentation graphique du modèle de Weibull sous l'hypothèse AFT. 95
4.3 Représentation graphique du modèle joint. 96
4.4 Critères de performance en fonction des différents taux de censure et du nombre de composantes du sous-modèle isgICA dans le modèle conjoint SisgICA. 102
4.5 Critères de performance du sous-modèle Weibull-AFT du modèle conjoint SisgICA 104
4.6 Représentation graphique de la reconstruction de la structure de poids parcimonieux, de la structure de poids parcimonieux et des sources. . 105
4.7 Représentation visuelle de la relation entre le biais et la couverture selon les scénarios avec taux de censure $\{0.3, 0.7\}$ et $K_{\text{th}} = \{10, 30\}$. 106
4.8 Représentation visuelle de la relation entre le biais et la sélection selon les scénarios avec taux de censure $\{0.3, 0.7\}$ et $K_{\text{th}} = \{10, 30\}$. 106
4.9 Matrice binaire parcimonieuse de l'hétérogénéité individuelle et des composantes de la matrice des sources extraites de l'ensemble de données sur le cancer du sein 107
4.10 Distribution de la valeur absolue des éléments de 6 des 15 composantes de la matrice des sources significativement associés aux données de survie 108
4.11 Distribution de la valeur absolue des éléments des 15 composantes de la matrice des sources 110

A.1 Extrait de Beal (2003), appendix A 138
A.2 Extrait de Beal (2003), appendix A (suite) 139

Liste des tableaux

1.1 Récapitulatif des phases d'un essai d'un médicament. DMT, dose maximale tolérée. 8
1.2 Définitions de la médecine stratifiée et de précision 9

2.1 Analyse en variables latentes pour la réduction de dimension pour les modèles à relation linéaire 30
2.2 Comparaison des priors imposés aux matrices en fonction du modèle voulu ; avec $\mathbf{\Psi_\Phi}$, $\mathbf{\Psi_W}$, et $\mathbf{\Psi_E}$ la matrice variance-covariance de la distribution de la matrice $\mathbf{\Phi}$, $\mathbf{W}$, et $\mathbf{E}$. 33

3.1 Scénarios des simulations. Soit N, P, and σ_E^2 le nombre d'individus, le nombre de gènes et la variance du bruit. 64
3.2 Nombre de composantes identifiées (médiane [2.5-97.5% intervalle percentiles]) en utilisant la méthode des valeurs propres pour l'ICA standard et pour l'isgICA. N, P et σ_E^2 sont respectivement le nombre d'individus, le nombre de gènes et la variance du bruit. 66

4.1 Critères de performance du sous-modèle isgICA dans le modèle conjoint SisgICA 103
4.2 Estimation des paramètres du sous-modèle de survie du modèle conjoint SisgICA : moyenne et intervalle de crédibilité [2.5% - 97.5%]. 109

A.1 Exemple de relation conjuguée entre les distributions a priori, la vraisemblance et les distributions a posteriori 140

Notations et Abréviations

Notations génériques

w	Scalaire w
w	Vecteur colonne w
w_i	i-ème élément du vecteur w
$\mathbf{W}$	Matrice $\mathbf{W}$
$w_{k,i}$	Élément à l'indice (k,i) de la matrice $\mathbf{W}$
$(.)^T$	Transposée
$\mathbf{\Phi W}$	Produit matriciel
$\mathbf{W} \circ \mathbf{Z}$	Produit de Hadamard
diag$(\mathbf{W})$	Matrice diagonale de $\mathbf{W}$
$\mathbf{1}_K$	Vecteur de taille k de 1

Probabilités

$x \sim p(\theta)$	x suit la loi $p(\theta)$
$p(x \mid \theta)$	Probabilité de x conditionnellement au paramètre θ
$\mathbb{E}[x]$	Espérance de x
$\mathbb{E}_q[x]$ ou $< x >$	Espérance de x sous la loi q

Fonctions

$\Gamma(.)$	Fonction digamma
$\mathbb{1}$	Fonction indicatrice

Ensembles

$\mathbb{R}$	Ensemble des nombres réels
$\mathbb{R}^K$	Ensemble des vecteurs réels de taille K
$[1, K]$	Intervalle fermé
$\mathbb{N}$	Ensemble des entiers positifs

Abréviations

AF	Analyse factorielle
AFT	Accelerated failure time
ACP	Analyse en composante principale
ACPP	Analyse probabiliste en composante principale
ARD	Détermination automatique de la pertinence
BBP	Processus bêta-Bernoulli
BO	Optimisation Bayésienne
CDF	Fonction de répartition
ELBO	Borne inférieure de l'évidence
HR	Hazard ratio
IBP	Processus du buffet indien
isgICA	Analyse avec un nombre infini de composantes indépendantes et parcimonieuses
ICA	Analyse en composante indépendante
KL	Kullback-Leibler
MCMC	Approche de MonteCarlo par chaîne de Markov
MVL	Modèle à variables latentes
PDF	Fonction de densité
PH	Risque proportionnel
SisgICA	Survie - Analyse avec un nombre infini de composantes indépendantes et parcimonieuses
VI	Inférence variationnelle

Introduction générale

Le cancer, une maladie complexe et hétérogène

Le coût décroissant de l'acquisition de données génomiques et la démocratisation de l'analyse de biomarqueurs « omiques » (génique, épigénomique, métabolomique, protéomique, radiomique, etc.) ont nourri l'espoir de développer une médecine de plus en plus personnalisée avec le développement de la médecine de précision Yates et al. (2018).

Le cancer est une maladie typiquement multifactorielle alliant des facteurs de risques internes (âge, sexe, histoire familiale, mutation génétique, contexte hormonal), et des facteurs de risques externes. L'utilisation de biomarqueurs moléculaires dans la prise en charge thérapeutique a permis un bénéfice important en termes de survie. Par exemple, dans le cadre du cancer du sein, la mise en place de la recherche systématique du statut HER2 (*Human Epidermal growth factor Receptor 2*, Slamon et al. (1987)) a permis aux femmes HER2 positives de bénéficier d'une thérapie ciblée anti-HER2. Malgré les moyens mis en place, l'implication thérapeutique de la recherche au niveau moléculaire est complexe et constitue une voie importante de recherche pour développer de nouveaux traitements personnalisés (Betensky, Louis et Cairncross (2002) ; Buyse et Michiels (2013)).

Pour des maladies complexes telles que le cancer, la compréhension approfondie des mécanismes moléculaires est l'un des principaux défis actuels pour la détection de biomarqueurs tumoraux permettant alors d'améliorer le pronostic des patients. La prise en compte de ces voies moléculaires pourra permettre aussi une meilleure caractérisation de la maladie.

L'identification des mécanismes oncologiques à partir de données omiques, telles

que l'expression des gènes, est difficile en raison des relations complexes qu'ils entretiennent dans diverses voies moléculaires impliquant des centaines ou des milliers d'acteurs (par exemple, des gènes) et du nombre relativement faible de patients dans les ensembles de données disponibles induisant des problèmes de stabilité des estimations statistiques dû au célèbre *fléau de la dimension*. On a alors un nombre de gènes (P), d'un ordre de grandeur bien plus important que le nombre d'individus (N) : $P >> N$. La séparation entre le signal et le bruit est en général très complexe dans les données de grande dimension et les calculs peuvent rapidement devenir très lourds. Les statistiques classiques ne sont pas adaptées là où des problèmes d'estimations sont courantes. On peut se tourner vers des méthodes permettant de limiter l'impact de la grande dimension des données et les relations complexes dans lesquels ils interviennent.

En raison du grand nombre de gènes, des modèles de variables latentes (Cunningham et Ghahramani (2015), MVL) pour la séparation/déconvolution en aveugle des composantes sont souvent utilisés pour identifier des structures de données géniques de grande dimension non-observées, qui fournissent un aperçu des mécanismes moléculaires.

Les méthodes parcimonieuses permettent d'identifier les variables actives parmi un grand nombre de candidats. Seul un nombre restreint de variables ont un rôle significatif, il faut alors des approches adaptées à la sélection de variables. Des méthodes parcimonieuses peuvent être appliquées pour interpréter les structures identifiées comme des mécanismes (ou voies) moléculaires, en sélectionnant un sous-ensemble de biomarqueurs associées à chaque composante. Plusieurs approches parcimonieuses ont été proposées, en particulier dans le cadre Bayésien, imposant par exemple un prior de parcimonie aux éléments de la matrice gènes-composantes (Rattray et al. (2009)).

Bien que cette approche soit adaptée pour donner un aperçu des mécanismes communs à tous les individus, en oncologie chaque tumeur peut résulter de mécanismes moléculaires différents selon les patients (hétérogénéité inter-individuelle), ce qui complique l'amélioration de la prise en charge thérapeutique des patients. Ces mécanismes moléculaires peuvent être impactés de façon unique chez chaque patient par des phénomènes épigénétiques complexes, ou par une erreur d'une des milliards de bases nucléotidiques constituant l'ADN... L'occurrence d'une mutation au niveau d'une zone codante d'un gène altère sa fonction et, par conséquent, tous les mécanismes moléculaires en

découlant.

L'identification de biomarqueurs moléculaires associés au pronostic d'un patient est essentielle en médecine personnalisée. Cependant ces algorithmes sont généralement dédiés à des tâches non-supervisées, et non à l'évaluation pronostique. Ainsi, la prise en compte de l'hétérogénéité inter-individuelle des mécanismes oncologiques dans les mécanismes pronostiques reste un domaine de recherche.

Pour pallier cette question, un grand nombre de travaux ont utilisé une procédure en deux étapes : (i) ajuster un MVL non-supervisé sur les données omiques (tel que l'analyse en composantes principales (*Principal Component Analysis*, PCA), l'analyse en composantes indépendantes (*Independent Component Analysis*, ICA) ou l'analyse factorielle (AF)) puis (ii) utiliser les poids individuels comme variables explicatives dans une régression de survie. Le travail principal de cette thèse a été le développement d'une approche basée sur un modèle commun entre la survie et le MVL afin d'améliorer l'inférence des composantes par rapport à la question clinique d'intérêt.

Objectifs et plan de la thèse

L'objectif de cette thèse a été de proposer une approche pour l'identification de structures de corrélation latentes pronostiques de différents biomarqueurs d'expression génique représentant de potentielles mécanismes moléculaires tout en prenant en compte l'hétérogénéité individuelle.

Le modèle développé est basé sur l'analyse des composantes latentes (analyse en composantes indépendantes, ICA). Le modèle a été étudié dans un cadre non-supervisé, puis dans un cadre supervisé à des fins pronostiques avec des données de survie. La modélisation supervisée de ces structures permet d'identifier l'hétérogénéité inter-individuelle moléculaire influençant la survie des patients, et de mieux comprendre les mécanismes de la maladie afin d'identifier de potentielles nouvelles cibles thérapeutiques.

L'approche Bayésienne non-paramétrique a été privilégiée pour faire face à la complexité de l'inférence de ce type de modèle.

Ce travail a été organisé en 5 chapitres.

Les chapitres 1 et 2 présentent l'état de l'art actuel avec les différents modèles, méthodes et inférences nécessaires à la compréhension de ces travaux de thèse.

Le chapitre 3 présente le cœur de ces travaux, où nous avons développé un modèle d'ICA bruité imposant deux parcimonies au modèle : une sur la matrice individus-composantes (pour modéliser l'hétérogénéité inter-individuelle) et une sur la matrice gènes-composantes (pour modéliser des composantes interprétables en terme de mécanismes moléculaires). Nous avons dénommé ce modèle isgICA pour *infinite sparse graphical independent component analysis* (analyse graphique avec un nombre infini de composantes indépendantes et parcimonieuses). À notre connaissance, il s'agit de la première approche imposant cette double parcimonie dans un modèle à dimension infinie et proposant une procédure d'optimisation afin d'étudier la reconstruction des structures latentes.

Le chapitre 4 présente notre extension de la proposition précédente, où nous avons développé un modèle pronostique conjoint à partir d'analyse de survie et du modèle précédemment étudié isgICA. Supposant que la présence de chacune de ces altérations moléculaires peut avoir un impact sur la survie du patient, nous avons considéré un modèle partageant l'espace latent binaire du processus bêta-Bernoulli entre l'isgICA et un modèle de Weibull à temps de vie accéléré (*Accelerated Failure Time*, AFT). Ce modèle permet d'intégrer l'information de la survie des patients à l'approche précédente pour définir des structures parcimonieuses expliquant l'expression génétique et la survie des patients.

Les notions théoriques de ces approches sont présentées dans les chapitres 3 et 4 et sont accompagnées d'une étude de simulation de ces structures pour illustrer la performance de cet algorithme et identifier en aveugle des signatures. Ces méthodes ont été appliquées dans un jeu de données réelles pour modéliser l'hétérogénéité de l'expression génique de tumeur de patientes atteintes d'un cancer du sein afin d'identifier des signatures d'expression génique pertinentes et connues du cancer du sein.

Le chapitre 5. Nous concluons sur nos travaux et présentons les prochaines perspectives envisagées.

Chapitre 1

Cadre applicatif : Hétérogénéité des patients atteints du cancer

Sommaire

1.1 Le cancer **6**

1.1.1 Principe de la cancérogénèse 6

1.1.2 Le cancer du sein 7

1.1.3 Essais cliniques en oncologie 7

1.1.4 Hétérogénéité des cancers - Apport de la médecine de précision 9

1.1.4.1 Impact de la médecine stratifiée ou de précision . . . 9

1.1.4.2 Biomarqueurs et données omiques : des gènes aux mécanismes moléculaires 10

1.2 Méthodes de déconvolution (ou de réduction de dimension) parcimonieuse **11**

1.3 Identifier des composantes pronostiques à la survie **12**

1.1 Le cancer

1.1.1 Principe de la cancérogénèse

L'Organisation Mondiale de la Santé (OMS) définit le cancer comme :

> le terme général appliqué à un grand groupe de maladies qui peuvent toucher n'importe quelle partie de l'organisme. L'une de ses caractéristiques est la prolifération rapide de cellules anormales qui peuvent essaimer dans d'autres organes, formant ce qu'on appelle des métastases.

Autrement dit, le cancer est provoqué par une instabilité du génome dû à une erreur aléatoire de réplication ou d'un agent cancérogène. On peut en citer trois classes : les agents chimiques (tel que l'amiante), physiques (tel que les rayonnements ultraviolets) ou biologiques (tel que le *Papillomavirus* pour le cancer du col de l'utérus). Sans le système de réparation cellulaire, les brins de l'ADN se détériorent et accumulent des détériorations pouvant entraîner l'initialisation du processus de cancérogénèse. En conséquence, les brins d'ADN endommagés modifient le comportement de la cellule et sa régulation. Ces cellules anormales vont se mettre à proliférer et à se disséminer, détournant le métabolisme énergétique cellulaire « normal », afin d'apporter les ressources nécessaires à leur prolifération et d'échapper à leur destruction par le système immunitaire.

En fonction de la propagation des cellules anormales, l'évolution avance de façon locale, régionale puis à distance (sous forme de métastases). Pour décrire l'histoire d'un cancer, il est courant d'utiliser une description en quatre stades basés sur le volume de la tumeur et son envahissement des tissus. Chaque type de cancer définit les stades en fonction de l'impact sur la santé des individus. Dès les premières réplications cellulaires (stade 1), la prolifération incontrôlée de ces cellules anormales aboutit à une tumeur maligne unique et de petite taille. Le stade 2 correspond à un envahissement local plus importante. La tumeur va se développer et commencer à envahir les tissus voisins (ou ganglions lymphatiques) en stade 3 pour finalement s'étendre plus largement dans l'organisme sous forme de métastases en stade 4.

1.1.2 Le cancer du sein

Chez les femmes, le premier cancer en termes de fréquence est le cancer du sein. Ce cancer se place en second, tous cancers et sexes confondus. Le rapport de Santé Publique France [1] de 2018 en France métropolitaine estime $382\ 000$ nouveaux cancers par an, dont $58\ 459$ nouveaux cas imputés au cancer du sein. Malgré la diminution du taux de mortalité de 1.6% par an entre 2010 et 2018, le taux d'incidence entre 1990 et 2018 a doublé. Santé publique France estime la survie nette à 5 ans de 88%, et à 10 ans de 76%, et estime à environ $12\ 000$ les décès annuels en France métropolitaine.

Le cancer est une maladie typiquement multifactorielle causée par des facteurs de risques internes, c'est-à-dire constitutifs des individus (âge, sexe, histoire familiale, mutation génétique, contexte hormonal), et les facteurs de risques externes, liés à l'environnement, aux modes et conditions de vie (comportements) sur lesquels il est parfois possible d'agir. Pour le cancer du sein les principaux facteurs de risque identifiés sont : l'âge, les antécédents personnels de pathologies mammaires, les antécédents familiaux de cancer du sein, dont les prédispositions génétiques et les antécédents de radiothérapie thoracique à haute dose. D'autres facteurs de risques sont aussi identifiés, tels que le surpoids (ou l'obésité) chez la femme ménopausée, la durée de l'exposition de l'organisme aux hormones, le tabagisme ou la consommation d'alcool. Selon le rapport 37% des cancers du sein seraient actuellement liés à des facteurs évitables, et 5 à 10% seraient liés à des mutations génétiques.

1.1.3 Essais cliniques en oncologie

Les essais cliniques ont pour but d'évaluation la tolérance/toxicité, la sécurité et l'efficacité d'un produit de santé (d'un médicament ou d'un dispositif de santé) sur des volontaires. En oncologie ces volontaires sont malades. Il est impératif de prouver la qualité du produit avant sa mise sur le marché. Cette démonstration passe par quatre phases, voir la table 1.1 pour plus de détail (Green, Benedetti et Benedetti (2012) ; Crowley et Hoering (2012)) :

1. https ://www.santepubliquefrance.fr/maladies-et-traumatismes/cancers/cancer-du-sein/documents/rapport-synthese/estimations-nationales-de-l-incidence-et-de-la-mortalite-par-cancer-en-france-metropolitaine-entre-1990-et-2018-etude-a-partir-des-registres-des

- Phase 1 : l'évaluation de la toxicité pour la détermination de la dose maximum tolérée (DMT) ;
- Phase 2 : l'évaluation de l'efficacité à court terme et l'évaluation de la balance toxicité - efficacité ;
- Phase 3 : l'évaluation de l'efficacité du traitement expérimental en le comparant à la procédure standard en vigueur au moment du départ de l'essai ;
- Phase 4 : l'évaluation à long terme après la mise sur le marché des effets indésirables du nouveau traitement sur la population ayant accès au traitement. non abordé

	Phase		
	Phase 1	Phase 2	Phase 3
Objectifs	DMT et profil de tolérance Pharmacologie Biomarqueur in vivo	Activité anti-tumorale ; Définition de la dose à utiliser	Comparaison avec le traitement ou la procédure standard pour prouver sa supériorité ou non-infériorité
Critère(s) de jugement (principal et secondaire(s))	Tolérance (toxicité) Pharmacocinétique-pharmacodynamique Activité (réponse, symptômes)	Activité (réponse, symptômes) : efficacité thérapeutique à court terme	Efficacité (survie, qualité de vie, ...) ; efficacité thérapeutique à long terme
Nombre de patients	< 50	50 et +	200 et +
Durée	en mois / années +	années +	années ++
Randomisation	Rarement	Souvent	Privilégié
Multicentrique	Parfois	Souvent	Privilégié

Table 1.1 – Récapitulatif des phases d'un essai d'un médicament. DMT, dose maximale tolérée.

Pour le développement de médicaments en oncologie, les essais cliniques évaluent si le produit de santé étudié peut servir à mesurer l'activité de la maladie ou à l'amélioration du bénéfice clinique dans divers contextes cliniques, comme la réduction de la tumeur ou la survie (générale, à la rechute, à la progression ...).

On définit un critère de jugement principal par rapport aux exigences attendues de l'essai clinique et doit permettre de conclure sur l'approbation de l'essai par les autorités de santé. Il doit être pertinent, en accord avec l'objectif de l'étude, bien défini, quantifiable, facile à mesurer et idéalement unique. Généralement, les critères de jugement

(qu'ils soient primaires ou secondaires) en essai clinique servent à mettre en évidence les différents objectifs qu'ils soient de tolérance, d'activité ou d'efficacité, comme le bénéfice d'un traitement expérimental par rapport à un traitement de référence.

1.1.4 Hétérogénéité des cancers - Apport de la médecine de précision

Dans le cadre du cancer du sein, la mise en place de la recherche systématique du statut HER2 (*Human Epidermal growth factor Receptor 2*, Slamon et al. (1987) ; Gennari et al. (2021) ; Burstein et al. (2021)) a permis aux femmes positives de bénéficier d'une thérapie ciblée anti-HER2. L'utilisation des biomarqueurs moléculaires (ER, PR, BRAC1 ou 2 ...) dans la prise en charge thérapeutique permet un bénéfice important Betensky, Louis et Cairncross (2002). Malgré les moyens mis en place, l'implication thérapeutique de la recherche génétique est complexe et constitue une voie importante de recherche pour développer de nouveaux traitements personnalisés Buyse et Michiels (2013).

1.1.4.1 Impact de la médecine stratifiée ou de précision

Il faut noter la différence entre la médecine de précision et la médecine stratifiée, voir table 1.2, qui sont souvent utilisées indifférenciées. Dans le cadre de la recherche de sous-groupes pour le développement de traitement, la médecine stratifiée cible le sous-groupe de patients les plus à même de répondre au traitement, leurs données génétiques étant utilisées à cette fin Yates et al. (2018).

Médecine stratifiée	Consiste à définir des sous-populations (groupes ou proportions de patients) en fonction du « sous-type » de maladie diagnostiquée chez un individu.
Médecine de précision	Le terme médecine de précision couvre les notions de stratification et de personnalisation,
	Repose sur un profil détaillé de l'individu, y compris la sous-population à laquelle le patient appartient,

Table 1.2 – Définitions de la médecine stratifiée et de précision

À chaque biomarqueur, il est possible de catégoriser les patients. La médecine stratifiée est fréquemment basée sur la définition de sous-populations (catégorisation) des

patients, afin d'étudier la question clinique d'intérêt. La catégorisation d'un biomarqueur continu engendre un biais sur l'estimation des effets. En grande dimension, les biomarqueurs se comptent par milliers. En supposant qu'il n'y aurait ne serait-ce qu'une vingtaine de biomarqueurs, il est impossible d'évaluer statistiquement la question d'intérêt (il n'y aura pas assez d'individus par strate). Sous cette forme catégorielle, elle rencontre ses limites dans la prise en compte de la spécificité de chaque individu. La médecine de précision est un cas extrême de la médecine stratifiée qui prend en compte le profil individuel en incluant des variables cliniques et moléculaires des patients.

Le cancer du sein est connu pour distinguer des sous-groupes de traitement et prédire le pronostic des patientes à partir de différents biomarqueurs génétiques (BRCA1 / BRCA2 ou HER2) ou hormonaux (ER et PR). Même si leur intérêt est établi, pouvoir améliorer le ciblage des patients reste encore au cœur des enjeux actuels.

1.1.4.2 Biomarqueurs et données omiques : des gènes aux mécanismes moléculaires

Un biomarqueur est défini par « une caractéristique définie qui est mesurée en tant qu'indicateur du déroulement normal des processus biologiques, des processus pathogènes ou des réponses à une exposition ou à une intervention » Silver Spring (MD) : Food and Drug Administration (US) et Bethesda (MD) : National Institutes of Health (US) (). Les biomarqueurs basés sur les données omiques permettent une compréhension des mécanismes moléculaires dans lesquels ils interviennent. La prise en compte de ces mécanismes moléculaires permettent une meilleure appréhension de la maladie ainsi que de la manière dont le traitement interagit avec l'organisme. Par ailleurs, les mutations provoquées par le cancer induisent des modifications du comportement des mécanismes moléculaires, qui sont dues à l'individualité des tumeurs, donc différentes entre les individus. Ainsi la prise en compte de l'hétérogénéité inter-individuelle devient un enjeu primordial afin d'améliorer la prise en charge thérapeutique des patients menant à la médecine de précision.

Le nombre de biomarqueurs (P) est d'un ordre de grandeur bien plus important que le nombre d'individus (N), le phénomène du *fléau de la dimension* induit des instabilités des estimations où les approches statistiques classiques ne sont pas adaptés. Plusieurs méthodes ont été développées ces dernières décennies pour tenter de prendre en compte la grande dimension des données, ainsi que leur relations complexes au sein

des mécanismes moléculaires.

1.2 Méthodes de déconvolution (ou de réduction de dimension) parcimonieuse

Les produits de l'expression des gènes interagissent entre eux pour construire les briques du vivant au sein de chaînes de réactions chimiques complexes appelées mécanismes moléculaires. L'étude de ces mécanismes permet une exploration de leur impact sur les mécanismes tumoraux. La compréhension approfondie de ces mécanismes moléculaires est l'un des défis actuels pour développer la médecine de précision où chaque parcours de patient est unique.

Pour limiter le fléau de la dimension, l'identification des structures de données non-observées résumant l'expression génique fournit un aperçu des mécanismes moléculaires. Dans ce but, des modèles de variables latentes (Cunningham et Ghahramani (2015), MVL) sont classiquement utilisés pour la séparation/déconvolution aveugle de ces composantes non-observées, incluant l'analyse en composantes principales (*Principal Component Analysis*, ACP), l'analyse en composantes indépendantes (*Independent Component Analysis*, ICA) ou l'analyse factorielle (AF).

Les modèles de réduction de dimension sont basés sur les MVL où le nombre de variables latentes est strictement inférieur aux nombres d'observations et de variables observées. Ces méthodes déterminent une représentation linéaire de faible dimension des données originales pour préserver certaines caractéristiques d'intérêt.

Les méthodes parcimonieuses permettent d'identifier les variables actives parmi un grand nombre de candidats. Supposant que seul un nombre restreint de variables ont un rôle dans l'explication ou la prédiction de la réponse d'un patient, il faut alors des approches adaptées à la sélection de variables.

Comme pour le modèle factoriel graphique proposé par Yoshida et West (2010), une méthode parcimonieuse peut être utilisée pour sélectionner un sous-ensemble de variables omiques associées à chaque composante afin d'interpréter la structure identifiée comme un mécanisme moléculaire. Par exemple, plusieurs approches parcimonieuses ont été proposé dans le cadre Bayésien, imposant une parcimonie à la matrice

gène-composantes en utilisant le prior *spike and slab* (West et al. (2003) ; D. Knowles et Ghahramani (2011)), le processus du buffet indien (T. L. Griffiths et Ghahramani (2005)), le prior de Laplace (Kabán (2007)), ou le prior horseshoe (Carvalho et al. (2008)). Bien que cette approche soit adaptée pour donner un aperçu des mécanismes communs à tous les individus, il est bien connu en oncologie que la tumeur peut résulter de mécanismes moléculaires différents selon les patients (hétérogénéité inter-individuelle), qui n'est pas pris en compte dans les approches actuelles de la littérature.

1.3 Identifier des composantes pronostiques à la survie

Les méthodes de déconvolution sont flexibles et comportent la capacité d'étudier l'impact des composantes sur la survie des patients tout en permettant de comprendre des mécanismes n'étant pas encore référencés et ainsi que d'étudier la différence des mécanismes moléculaires entre les individus.

Cependant, les algorithmes mentionnés dans la section précédente sont dédiés à des tâches non supervisées, mais n'ont pas été optimisés pour identifier des signatures moléculaires pronostiques. Un grand nombre de travaux ont utilisé une procédure en deux étapes (Peterson (2013) ; Witten et Tibshirani (2009) ; Baek, Ho et Ma (2020) ; Gal et al. (2020) ; Park et al. (2020)) :

1. Ajustement d'un MVL non supervisé sur les données omiques (tel que l'ACP, l'ICA, ...) ;
2. Utilisation des poids individuels comme variables explicatives dans une régression de survie.

En raison des processus d'inférence indépendants de ces deux étapes, l'information de survie n'est pas prise en compte dans les estimations des structures latentes du MVL. L'intégration de cette information permettrait d'identifier des structures associées au pronostic des patients.

Des approches de modélisation conjointe ont été proposées pour intégrer simultanément les données omiques et les données de survie pour l'inférence des variables latentes, comme la méthode des moindres carrés partiels adaptées aux données de survie (Bastien et al. (2014) ; Bertrand et al. (2014)). Dans le cadre Bayésien, différentes

versions du modèle conjoint analyse factorielle-survie sont également disponibles (M. Liu et al. (2019) ; Cai et Liang (2019) ; Pan et al. (2019) ; McCurdy, Molinaro et Pachter (2017)), ou en intégrant la réponse de survie comme covariable (mais sans tenir compte des données censurées) dans Bhattacharya et Dunson (2011). Cependant ces méthodes ne prennent pas en compte l'hétérogénéité inter-individuelle et la qualité de reconstruction des structures latentes pronostiques de ces modèles n'a pas été étudiée.

Chapitre 2

Cadre méthodologique et inférentiel

Sommaire

2.1 L'inférence Bayésienne 16
2.1.1 Principe 16
2.1.2 La définition des priors 17
2.1.3 Inférence Bayésienne non-paramétrique 19
2.1.3.1 Définition 19
2.1.3.2 Prior pour inférence en dimension infini 20
2.1.3.3 Limites de l'inférence non-paramétrique 20
2.2 Optimisation : l'inférence variationnelle 21
2.2.1 Difficultés d'inférence en grande dimension et en présence de labels échangeables des variables latentes 21
2.2.2 Principe 22
2.2.3 Algorithmes d'optimisation de l'ELBO 25
2.2.3.1 L'approche CAVI 25
2.2.3.2 Approche basée sur la descente de gradients 26
2.2.3.3 Limite de ces deux approches 28
2.3 Modèles à variables latentes pour la réduction de dimension 28
2.3.1 Approche non-supervisée dans le cadre des modèles latents pour la réduction de dimension 29
2.3.2 Comparaison des modèles de déconvolution quantitatives 29
2.3.2.1 Factorisation de matrice 30
2.3.2.2 Point de vue Bayésien des MVL classiques 33
2.3.2.3 L'identification du nombre de composantes 36

2.1 L'inférence Bayésienne

2.1.1 Principe

L'inférence Bayésienne est fondée sur le théorème de Bayes où l'on souhaite étudier l'évènement inconnu et non-observé A. À partir des connaissances d'un autre évènement B, nous pouvons définir le théorème de Bayes tel que :

$$p(A \mid B) = \frac{p(B \mid A)\ p(A)}{p(B)}, \tag{2.1}$$

avec $p(A \mid B)$ la probabilité de A conditionnellement à la réalisation de B. À l'inverse, $p(B \mid A)$ est la probabilité de B conditionnellement à la réalisation de A. Les probabilités $p(A)$ et $p(B)$ sont respectivement les probabilités marginales des évènements A et B.

Lorsqu'on étudie une quantité d'intérêt, tel que les paramètres du vecteur θ d'un modèle mathématique, nous pouvons reformuler le théorème de Bayes pour exprimer l'incertitude associée aux paramètres θ à partir de connaissances (de données ou d'hypothèses) a priori. Cette connaissance est définie sous forme de distribution en utilisant le théorème de Bayes :

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{p(x)}; \tag{2.2}$$

où $p(\theta)$ représente la connaissance a priori que nous avons des valeurs de θ nommé la *distribution a priori*. La probabilité $p(x)$ est la vraisemblance marginale nommée aussi l'*évidence* du modèle ou la *constante de normalisation*. La probabilité $p(x \mid \theta)$ est la *vraisemblance* du modèle, soit la probabilité d'observer les données x sachant l'ensemble des paramètres θ. À partir de ces quantités, nous pouvons actualiser nos connaissances en appliquant le théorème de Bayes afin de calculer la *probabilité a posteriori* $p(\theta \mid x)$.
Notre modèle s'écrira alors :

$$\begin{aligned} x_1 \quad ,..., \quad x_N \mid \theta &\sim p(x \mid \theta) \\ \theta &\sim p(\theta). \end{aligned} \tag{2.3}$$

L'inférence Bayésienne génère alors une probabilité $p(\theta \mid x)$ qui ajuste au mieux les paramètres aux données, basée sur les informations précédemment spécifiées (par la distribution a priori : en considérant un espace de valeurs possibles pour θ).

La modélisation Bayésienne permet la quantification de l'incertitude du paramètre θ autour de la prédiction d'une nouvelle observation (à travers la distribution prédictive a posteriori, *Posterior predictive distribution*). La distribution prédictive a priori est un ensemble de données générées à partir du posterior du modèle (la vraisemblance et les priors). Ayant obtenu les distributions a posteriori des paramètres après avoir pris en compte les données, les distributions a posteriori peuvent être utilisées pour générer des données futures à partir du modèle. Une fois les distributions a posteriori $p(\theta \mid x)$ estimées, les prédictions basées sur ces distributions sont définies, en intégrant les paramètres :

$$p(x_{pred} \mid x) = \int_{\theta} p(x_{pred}, \theta \mid x)\, d\theta = \int_{\theta} p(x_{pred} \mid \theta, x) p(\theta \mid x)\, d\theta. \tag{2.4}$$

En supposant que les observations passées et futures sont conditionnellement indépendantes étant donné θ, c'est-à-dire $p(x_{pred} \mid \theta, x) = p(x_{pred} \mid \theta)$, la distribution a posteriori prédictive peut être écrite comme :

$$p(x_{pred} \mid x) = \int_{\theta} p(x_{pred} \mid \theta) p(\theta \mid x)\, d\theta. \tag{2.5}$$

2.1.2 La définition des priors

Une des questions fondamentales en inférence Bayésienne est le choix des distributions a priori puisqu'elles vont représenter l'information connue qui sera intégrée à l'analyse pour « aider » l'inférence sur les paramètres inconnus θ Gelman et al. (2013). En fonction de la quantité ou du manque d'informations au préalable (solides et précisément quantifiées), un prior *informatif* ou *non-informatif* sera utilisé.

Comme indiqué par l'équation 2.2, la distribution a priori va avoir une influence sur le processus d'inférence, et donc sur la distribution a posteriori, la quantité d'intérêt en inférence Bayésienne. Aussi, lorsque nous n'avons pas de connaissances fiables au préalable, il est possible d'utiliser des priors dits « non-informatifs », c'est-à-dire qui ont une forme n'apportant pas d'informations supplémentaires que celle présentes dans

les observations. Du fait de leur impact négligeable, l'inférence n'est alors basée que sur la vraisemblance, résultant en des estimations a posteriori très proches de celles fréquentistes.

Cependant, notamment lorsque le nombre d'observations est limité (souvent le cas dans le cadre d'essais cliniques et pré-cliniques), il sera important de définir une distribution a priori, dite « informative », qui pourra intégrer la connaissance au préalable que nous avons sur les paramètres (ex : inférence sur un jeu de données précédent, savoir d'expert,...) afin de renforcer le processus d'inférence, pour limiter l'impact de la fluctuation d'échantillonnage, particulièrement dans le cadre de petits échantillons. Dans le cadre de grands échantillons, les 2 approches (informative et non-informative) sont asymptotiquement similaires (le poids du prior devient négligeable dans le calcul de la distribution a posteriori).

Les priors non-informatifs et informatifs ne sont pas formellement définis. Il est préférable de les considérer comme un continuum où certaines distributions préalables sont plus informatives que d'autres.

Une propriété importante en inférence Bayésienne est la capacité des distributions à pouvoir être *conjuguées*. Cette propriété permet de dériver une solution analytique des distributions a posteriori de modèles simples. Dans ce cas de figure, la distribution a posteriori sera de la même famille de distribution que la distribution a priori. Par exemple, le posterior du paramètre d'une loi de Bernoulli peut être calculé analytiquement en considérant une distribution Bêta comme distribution a priori. Leurs inférences sont alors bien connues et étudiées comme avec l'algorithme de Gibbs (S. Geman et D. Geman (1984)). Si les distributions a priori et la vraisemblance sont définies par des lois connues de la famille exponentielle (ex : la loi normale, gamma, Bernoulli..., voir l'annexe A pour des informations sur les distributions de la famille exponentielle et leur conjugaison) et sous certaines combinaisons (Fink (1997)), elles peuvent constituer alors des distributions a posteriori connues.

Pour des modèles plus complexes, l'inférence est réalisée à l'aide algorithmes, comme celui de Gibbs, qui sont basés sur des processus stochastiques, comme les chaines de Markov. Ces processus sont itératifs, générant séquentiellement des échan-

tillons des paramètres à partir de leur posterior analytique, permettant de parcourir l'espace du posterior lorsque le processus a atteint un état stationnaire. Pour de plus amples détails sur ces algorithmes, le lecteur pourra se référer à Gelman et al. (2013) pour les MCMC et pour les HMC.

2.1.3 Inférence Bayésienne non-paramétrique

2.1.3.1 Définition

A l'opposé du terme *paramétrique*, *non-paramétrique* n'indique pas qu'aucun paramètre n'est impliqué, mais au contraire que le modèle statistique est décrit par un nombre infini de paramètres (Phadia (2013)). Les méthodes *paramétriques* fixent la complexité du modèle a priori, alors que les méthodes *non-paramétriques* considèrent que le modèle sous-jacent possède un support sur un espace infini. Le nombre fini de paramètres de l'approche *paramétrique* peut être vu comme une contrainte de l'espace des modèles, au-delà de la structure choisie.

L'inférence Bayésienne non-paramétrique est une approche intéressante offrant une plus grande flexibilité de modélisation pour les données réelles qui sont généralement plus complexes que ce qui peut être décrit avec des modèles traditionnels. Le nombre restreint d'observations d'un échantillon ne permet d'estimer qu'une partie de la structure du modèle (à laquelle les observations contribuent). Le modèle augmente son niveau de complexité à mesure que de nouvelles observations deviennent disponibles.

Par exemple, un modèle de mélange de distributions normales est un modèle probabiliste qui suppose que les observations sont générées à partir d'un nombre fini m de distributions normales (pouvant représenter des groupes non-observés) dont nous cherchons à estimer les paramètres. Néanmoins, notre hypothèse a priori sur m pourrait être fausse, et la restriction de l'estimation sur cet espace pourrait résulter au choix d'un modèle sous-optimal. Il existe de nombreuses méthodes pour la sélection du nombre de *clusters*/groupes basés sur des critères d'information (AIC, BIC, DIC,...) ou sur des méthodes de stabilité (Sundqvist (2020)).

Nous proposons utiliser l'approche basée sur l'inférence Bayésienne non-paramétrique

(Blei et Jordan (2006) ; Blei, Kucukelbir et McAuliffe (2017)). Comme précisé précédemment, l'approche non-paramétrique suppose qu'il existe une infinité de paramètres. Nous supposerions qu'il existe une infinité de distributions normales composant le modèle de mélange. L'espace des modèles n'est ainsi restreint que par la forme du modèle (mélange de distributions normales).

Cette hypothèse suggère qu'un nombre infini de groupes existe, et qu'ils pourraient être observé avec un nombre d'observations infini. Néanmoins, seule une partie de ces groupes est présente dans un échantillon fini. L'objectif de l'inférence Bayésienne non-paramétrique dans ce contexte sera d'identifier les groupes présents dans notre échantillon. Par la suite, nous détaillerons plus en détail ces principes pour le modèle d'allocation qui est la base de notre développement.

2.1.3.2 Prior pour inférence en dimension infini

Le principe de base de la modélisation Bayésienne non-paramétrique est de définir une distribution de probabilité a priori sur un espace de dimension infinie, c'est-à-dire un processus stochastique. Un processus stochastique est défini comme une collection de variables aléatoires $X = \{X_t : t \in T\}$ définies sur un espace de probabilité commun, régit par un phénomène commun, prenant des valeurs dans un ensemble commun S, et indexé par un ensemble T. Les processus les plus communément utilisés comme prior en inférence Bayésienne sont les processus gaussiens, Dirichlet, bêta, Bernoulli, et gamma (Phadia (2013)).

2.1.3.3 Limites de l'inférence non-paramétrique

Les avantages des méthodes non-paramétriques sont à contrebalancer avec leurs inconvénients, notamment l'inférence est complexe (du fait du grand nombre de paramètres) et du risque accru de sur-apprentissage. Du fait de leur très grande souplesse, du bruit peut être intégré aux paramètres du modèle (qui ne lui sont pas dédiés), engendrant des biais d'estimation et donc des résultats peu reproductibles. D'autre part, l'inférence d'un très grand nombre de paramètres requière une méthode d'inférence efficiente. Pour cette raison, nous avons opté pour l'inférence variationnelle.

2.2 Optimisation : l'inférence variationnelle

2.2.1 Difficultés d'inférence en grande dimension et en présence de labels échangeables des variables latentes

L'inférence Bayésienne peut parfois être difficile à réaliser en fonction des configurations du modèle (complexité du modèle, dimensionnalité, ...). Dans les problèmes dûs à la grande dimension des données, le grand nombre de paramètres et la géométrie de la grande dimension nécessite des calculs lourds. Cela est d'autant plus vrai dans le cadre des MVL où les calculs deviennent souvent irréalisables avec l'augmentation de la dimension de l'espace des paramètres, et la présence de variables latentes, pouvant poser des problèmes d'identifiabilité.

L'inférence Bayésienne est classiquement réalisée à l'aide de l'approche de Monte Carlo par chaîne de Markov (MCMC) (Gilks et Wild (1992)). Cependant, dans le cadre de données de grande dimension, parcourir de façon stochastique l'espace des paramètres peut demander un très grand nombre d'itérations avant d'atteindre un état stationnaire. L'algorithme Monte Carlo Hamiltonien (HMC) a été proposé (Betancourt (2017)), utilisant les principes physiques de la dynamique Hamiltonienne afin d'orienter les propositions des déplacements des chaînes dans l'espace. Cette approche permet d'atteindre un état stationnaire en un nombre d'itérations beaucoup plus faible que l'approche par MCMC classique, ce qui contribue à son succès actuel.

Néanmoins, dans l'approche MCMC classique, la présence de variables latentes échangeables, hypothèse de base des MVL, suggère que la valeur de la vraisemblance (que l'on cherche à maximiser) est la même, peu importe l'ordre des variables latentes. Cela engendre le fait qu'une même chaîne peut échanger le label de plusieurs variables latentes : l'algorithme va explorer le posterior de plusieurs paramètres différents correspondant à différentes variables latentes au cours des itérations de l'algorithme. La figure 2.1 illustre ce phénomène où le label de chaque paramètre des variables latentes 1 et 2 est échangé plusieurs fois au cours du processus. De ce fait, la chaîne explore les posteriors des moyennes des deux variables latentes, conduisant à des posteriors mal formés si aucun traitement a posteriori n'est effectué. De plus, le *label-switching* peut ralentir l'inférence.

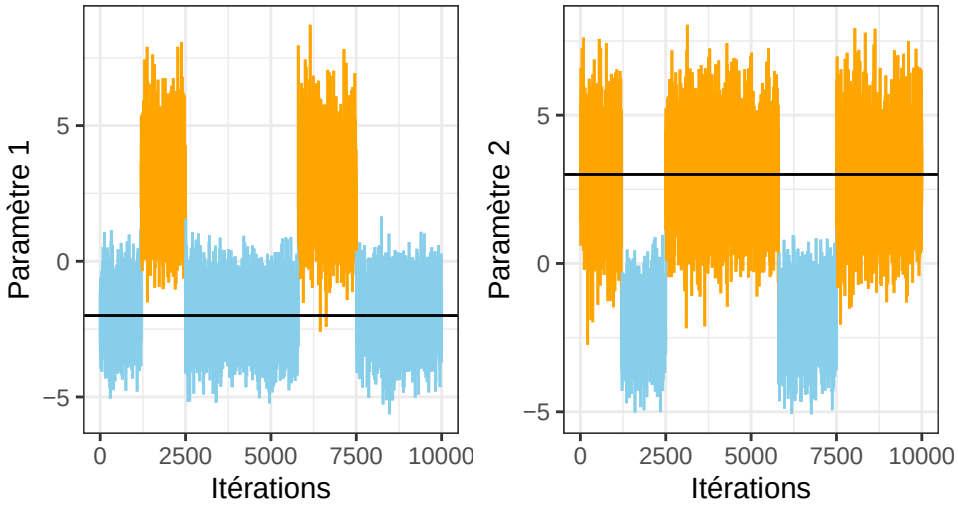


Figure 2.1 – Représentation du *label-switching* entre deux variables latentes spécifiées par un paramètre avec en bleu le paramètre 1, en orange le paramètre 2, et par la ligne noire la valeur théorique des paramètres respectivement de -2 et de 3.

Pour ces raisons algorithmiques, nous avons privilégié une toute autre approche, convergeant vers un optimum local. L'inférence variationnelle (IV) (Jordan, Ghahramani et al. (1999) ; Wainwright et Jordan (2008) ; Knoblauch, Jewson et Damoulas (2019)) est une approche d'inférence approximative, reformulant l'inférence Bayésienne stochastique en un problème d'optimisation, ce qui permet d'utiliser des algorithmes déterministes de descente/ascendance de gradients. Cette approche est beaucoup plus rapide que l'échantillonnage par MCMC, permettant de réaliser l'inférence pour des modèles comportant de nombreux paramètres sur de très grands jeux de données. En tant que stratégie alternative à l'échantillonnage par MCMC, l'inférence variationnelle est rapide en temps de calcul, relativement simple à adapter à de grandes données (Blei, Kucukelbir et McAuliffe (2017) ; Ranganath, Gerrish et Blei (2013)) en comparaison à la méthode MCMC.

2.2.2 Principe

L'idée principale de l'inférence variationnelle est d'approximer les densités conditionnelles des paramètres, difficiles à calculer.

L'inférence variationnelle va approximer la distribution cible a posteriori $p(\theta \mid x)$ par une famille de distribution paramétrique connue $\mathcal{L}$ (ex : la normale univariée) $q(\theta \mid \omega)$

avec respectivement x, θ et ω, les données observées, l'ensemble des paramètres étudiés et l'ensemble des paramètres variationnels. Elle optimise les paramètres variationnels associés, la méthode se base sur une divergence pré-spécifiée (classiquement, la minimisation de la divergence de Kullback-Leibler) entre la distribution approximée $q(\theta \mid \omega)$ et la distribution conditionnelle d'intérêt $p(\theta \mid x)$.

Ainsi, le problème d'inférence peut être réécrit comme un problème d'optimisation où l'objectif est d'identifier $q^*_\omega(\theta \mid \omega)$ tel que :

$$q^*_\omega(\theta \mid \omega) = \underset{q \in \mathcal{L}}{\operatorname{argmin}} \, \mathrm{KL}(q(\theta \mid \omega) \mid p(\theta|x)) \tag{2.6}$$

avec la divergence de KL définie telle que :

$$\begin{aligned}
\mathrm{KL}(q(\theta \mid \omega) \mid p(\theta|x)) &= -\int_\theta q(\theta \mid \omega) \ln \frac{p(\theta|x)}{q(\theta \mid \omega)} \\
&= -\mathbb{E}_{q(\theta|\omega)} \left[\ln \frac{p(\theta|x)}{q(\theta \mid \omega)} \right] \\
&= \mathbb{E}_{q(\theta|\omega)}[\ln q(\theta \mid \omega)] - \mathbb{E}_{q(\theta|\omega)}[\ln p(\theta|x)] \\
&= \mathbb{E}_{q(\theta|\omega)}[\ln q(\theta \mid \omega)] - \mathbb{E}_{q(\theta|\omega)}[\ln p(\theta, x)] + \ln p(x) \\
&= -\mathbb{E}_{q(\theta|\omega)} \left[\ln \frac{p(\theta, x)}{q(\theta \mid \omega)} \right] + \ln p(x) \\
&= -\mathrm{ELBO}(q(\theta \mid \omega)) + \ln p(x).
\end{aligned} \tag{2.7}$$

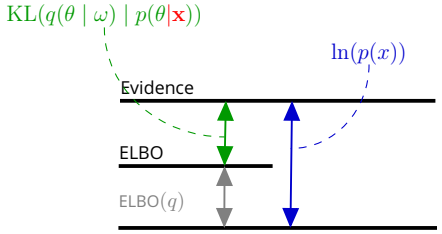
L'évidence $p(x)$ est constante par rapport à θ, permettant d'être retirée lors de l'actualisation du paramètre. Dans un but de clarification des notations, nous utiliserons par la suite la notation $\mathbb{E}_q$ à la place de $\mathbb{E}_{q(\theta|\omega)}$ pour dénommer l'espérance conditionnelle de $\theta \sim q(\theta \mid \omega)$.

En inversant les éléments de l'équation, la minimisation de la quantité $\mathrm{KL}(q(\theta \mid \omega) \mid p(\theta|x))$ est alors équivalente à la maximisation d'une nouvelle quantité : la borne inférieure de l'évidence $p(x)$, $ELBO(q(\theta \mid \omega))$, donnée par :

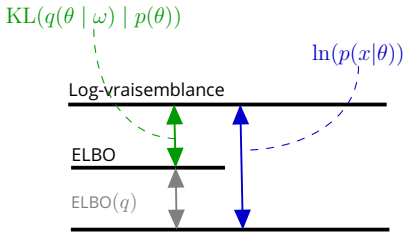
$$\begin{aligned}\mathrm{ELBO}(q(\theta \mid \omega)) &= \ln p(x) - \mathrm{KL}(q(\theta \mid \omega) \mid p(\theta|x)) \\ &= \mathbb{E}_q\left[\ln \frac{p(\theta, x)}{q(\theta)}\right] \\ &= \mathbb{E}_q[\ln p(\theta, x)] - \mathbb{E}_q[\ln q(\theta)] \\ &= \mathbb{E}_q[\ln\ p(x|\theta)] + \mathbb{E}_q[\ln p(\theta)] - \mathbb{E}_q[\ln q(\theta)] \\ &= \mathbb{E}_q[\ln\ p(x|\theta)] - \mathrm{KL}(q(\theta \mid \omega) \mid p(\theta)),\end{aligned} \tag{2.8}$$

avec $KL(q(\theta \mid \omega) \mid p(\theta)) = -\int_\theta q(\theta \mid \omega) \ln \frac{p(\theta)}{q(\theta|\omega)}$.

L'$\mathrm{ELBO}(q(\theta \mid \omega))$ est défini à partir de l'évidence $p(x)$. Cependant cette quantité n'est pas toujours disponible, on peut alors définir l'$\mathrm{ELBO}(q(\theta \mid \omega))$ à partir de la vraisemblance $p(x|\theta)$ (voir l'équation 2.8). Les divergences seront respectivement $\mathrm{KL}(q(\theta \mid \omega) \mid p(\theta|x))$ et $\mathrm{KL}(q(\theta \mid \omega) \mid p(\theta))$. On peut illustrer cette équivalence sous forme schématique en figure 2.2, avec en bleu sur la sous-figure 2.2a l'évidence et sur la sous-figure 2.2b la log-vraisemblance représentant la somme de la divergence de KL en vert et de l'ELBO en gris. L'ELBO correspond donc à la borne inférieure de l'évidence $p(x)$ mais aussi à la borne inférieure de la vraisemblance $p(x \mid \theta)$.



(a) Représentation de l'ELBO avec l'évidence



(b) Représentation de l'ELBO avec la log-vraisemblance

Figure 2.2 – Représentation de l'ELBO défini en équation 2.8 comme la limite inférieure de l'évidence mais aussi de la log-vraisemblance. L'image est inspirée de C. Bishop (2006).

La forme de l'$ELBO(q(\theta \mid \omega))$ basée sur la vraisemblance,

$$\mathrm{ELBO}(q(\theta \mid \omega)) = \mathbb{E}_q[\ln\ p(x|\theta)] - KL(q(\theta \mid \omega) \mid p(\theta)), \tag{2.9}$$

est maximisée pour l'optimisation des modèles. Cette formulation de l'ELBO permet de se rapprocher d'un problème familier en grande dimension, qui est celui d'une vraisem-

blance pénalisée par la divergence de KL entre l'approximation $q(\theta \mid \omega)$ et le prior $p(\theta))$ des paramètres. Le premier terme de cette équation est la vraisemblance attendue ; il encourage les densités à placer leur masse sur les configurations des paramètres qui expliquent le mieux les données observées. Le second terme est la divergence négative de Kullback-Leibler entre la distribution variationnelle $q(\theta \mid \omega)$ et la distribution a priori $p(\theta)$; il encourage les densités variationnelles à rester proches du prior. Ainsi, l'objectif variationnel reflète l'équilibre habituel de l'inférence Bayésienne entre la vraisemblance et la distribution a priori tout en prenant en compte les hypothèses des distributions.

2.2.3 Algorithmes d'optimisation de l'ELBO

Il existe deux approches principales pour maximiser l'ELBO : l'approche par ascension de coordonnées (CAVI, *Coordinate Ascent Variational Inference*) et l'approche basée sur le gradient.

2.2.3.1 L'approche CAVI

2.2.3.1.1 Principe Pour obtenir un maximum local pour l'ELBO$(q(\theta \mid \omega))$, l'algorithme CAVI (C. Bishop (2006)) optimise séquentiellement chaque distribution $q_1, \ldots, q_{j-1}, q_j, q_{j+1}, \ldots, q_J$ de la densité variationnelle, tout en maintenant les autres fixés.

Autrement dit, l'algorithme optimise séquentiellement les paramètres variationnels ω pour approximer les paramètres θ en maximisant ELBO$(q(\theta \mid \omega))$:

$$\text{ELBO}(q(\theta_j \mid \omega_j)) = \mathbb{E}_q[\ln p(\theta_j, x)] - \mathbb{E}_q[\ln q(\theta_j)], \tag{2.10}$$

ce qui donne par inversion des éléments :

$$\ln q(\theta_j \mid \omega_j) = \mathbb{E}_q[\ln q(\theta_j)] = \mathbb{E}_q[\ln p(\theta_j, x)] - \text{ELBO}(q(\theta_j \mid \omega_j)). \tag{2.11}$$

Pour l'estimation de $q(\theta_j)$, la quantité ELBO$(q(\theta_j \mid \omega_j))$ est constante. Ainsi, itérativement les paramètres variationnels de la distribution a posteriori du paramètre θ_j sont optimisés et sont proportionnels à l'exponentielle de l'espérance du logarithme de la distribution jointe :

$$q_j(\theta_j) \propto \exp\left\{\mathbb{E}_{-q}[\ln p(\theta, x)]\right\}. \tag{2.12}$$

avec $-q$ (la simplification de $-q(\theta_j \mid \omega_j)$) les distributions fixées de toutes les distributions à l'exception du paramètre j.

Parmi les limites de cette approche, il est à noter que l'espérance $\mathbb{E}_{-q}[\ln p(\theta_j, x)]$ peut être difficile à dériver analytiquement. De plus, l'algorithme CAVI est un algorithme itératif qui nécessite généralement de parcourir l'ensemble des données à chaque itération, ce qui peut s'avérer excessivement coûteux en termes de calcul pour les grands ensembles de données.

2.2.3.1.2 Cas particulier : la méthode du champ moyen (*Mean-field*) De nombreux algorithmes d'inférence se concentrent sur la famille variationnelle du champ moyen Blei, Kucukelbir et McAuliffe (2017). L'hypothèse principale de cette approche est que les différents paramètres de θ sont indépendants. Par conséquent, la distribution conjointe du vecteur θ peut être factorisée dans le produit des distributions marginales, régie par un ensemble de paramètres variationnels ω :

$$q(\theta \mid \omega) = \prod_{j=1}^{J} q_j(\theta_j \mid \omega_j) \tag{2.13}$$

Il est à noter que cette approximation reposant sur l'hypothèse d'indépendance peut conduire à des distributions a posteriori des paramètres dont la moyenne est non-biaisée, mais la variance est sous-estimée (Blei, Kucukelbir et McAuliffe (2017)). Il existe des méthodes plus élaborées, mais qui nécessite un temps de calcul plus important (Ranganath, Tran et Blei (2015)).

2.2.3.2 Approche basée sur la descente de gradients

2.2.3.2.1 Principe Le principe des approches basées sur le gradient est d'optimiser à partir du gradient de la fonction de l'ELBO les posteriors des distributions. L'algorithme optimise itérativement les paramètres à partir de la dérivée du premier ordre pour trouver un minimum local d'une fonction différentiable, ici l'ELBO :

$$\omega_{ite} = \omega_{ite-1} + \rho \hat{\nabla}_{\omega_{ite-1}} \text{ELBO}(q(\theta \mid \omega_{ite-1})) \tag{2.14}$$

avec ite l'indice de l'itération, ρ le taux d'apprentissage et $\hat{\nabla}_{\omega} \text{ELBO}(q(\theta \mid \omega))$ le gradient de l'ELBO(q($\theta \mid \omega$)).

2.2.3.2.2 La descente de gradient stochastique Inspirée de l'optimisation stochastique de gradient qui connait un grand succès dans le cadre des réseaux de neurones, Ranganath, Gerrish et Blei (2013) ont dérivé cette approche dans le cadre de l'inférence variationnelle. Cette approche permet de contourner la dérivation de gradients complexes (voir non-possibles) par une optimisation stochastique (Hoffman et al. (2013)). Ces méthodes permettent d'étudier des modèles non conjugués sans avoir à développer les équations.

Cette approche va estimer des gradients bruités (Robbins et Monro (1951)) à partir de l'échantillonnage des paramètres depuis leur distribution variationnelle, et conditionnellement aux hyperparamètres de l'itération actuelle de l'algorithme de descente de gradient. Pour chacun des paramètres, les échantillons vont être moyennés afin d'obtenir un gradient non-biaisé, sans nécessité de dériver analytiquement le gradient à partir des hyperparamètres du modèle. La mise à jour des paramètres au cours des itérations s'effectue de la façon suivante :

$$\omega_{ite} = \omega_{ite-1} + \rho \frac{1}{S} \sum_{s=1}^{S} \hat{\nabla}_{\omega_{ite-1}} \text{ELBO}(q(\theta_s \mid \omega_{ite-1})) \tag{2.15}$$

avec ite l'indice de l'itération, ρ le taux d'apprentissage, $\hat{\nabla}_{\omega} \text{ELBO}(q(\theta \mid \omega))$ le gradient de l'ELBO(q($\theta \mid \omega$)), S la taille de l'échantillonnage du gradient bruité de l'ELBO, et $\theta_s \sim q(\theta \mid \omega_{ite-1})$.

La descente de gradient stochastique conserve les propriétés de convergence du gradient vers un optimum local. Combinée à la descente de gradient de l'ELBO, on obtient alors un algorithme de descente de gradient stochastique appliquée à l'inférence variationnelle.

2.2.3.3 Limite de ces deux approches

L'approche CAVI et l'approche par descente de gradients sont deux approches basées sur des dérivées analytiques. L'approche CAVI impose de dériver analytiquement les équations de mise à jour de chaque paramètre variationnel du vecteur ω, et les approches basées sur la descente de gradients de l'ELBO nécessitent de dériver analytiquement le gradient en fonction de chacun des paramètres variationnels ω. Ces dérivations sont complexes, voir impossibles, dans le cadre d'un modèle comportant un nombre de paramètres élevés ou pour des distributions n'ayant pas de forme différentielle fermée (comme pour la distribution bêta). Par ailleurs ces méthodes requièrent aussi de traverser complètement le jeu de données et d'évaluer séquentiellement les paramètres ce qui est coûteux en termes de calcul pour les grands échantillons de données.

Néanmoins, il est toujours nécessaire de calculer les dérivées analytiques de l'ELBO, dérivées qui nécessitent la famille exponentielle. Cependant les distributions ne le sont pas toujours. Pour répondre à ce problème et dans le cas des méthodes basées sur le gradient, il existe des logiciels de différentiation automatique (*automatic differentiation*) comme autograd (Maclaurin, Duvenaud et Adams (2015)) ou tensorflow (Martín Abadi et al. (2015)) disposant d'un algorithme de calcul automatique du gradient.

2.3 Modèles à variables latentes pour la réduction de dimension

On peut différencier deux approches d'apprentissage : l'approche *supervisée* et l'approche *non-supervisée*.

Les modèles pour l'apprentissage supervisé visent à expliquer ou prédire une ou plusieurs réponses en fonction d'un ou plusieurs prédicteurs. Ils sont donc principalement utilisés pour tester des hypothèses d'association (exemple : expliquer l'occurrence d'une maladie en fonction d'expositions individuelles), ou pour développer des outils diagnostiques/pronostiques pour prédire une réponse à partir de prédicteurs (exemple : prédire la présence d'une tumeur à partir d'une image médicale). Pour ces objectifs d'explication/prédiction, la structure de ces modèles est contrainte par la forme $y = f(x)$, avec

y la(les) réponse(s) et x le(s) prédicteur(s).

L'approche non-supervisée a pour but de réaliser de la fouille de données (*data-mining*) à la recherche de nouvelles hypothèses pour l'application. Les modèles d'apprentissage non-supervisés sont en général très flexibles, avec peu de contraintes (par exemple par rapport à un modèle avec une structure linéaire) afin de pouvoir identifier des structures latentes. Ces méthodes sont à la source de nombreux outils de visualisation, dont les plus classiques font partie des 2 grandes familles des modèles à variables latentes (MVL, ex analyse en composantes principales) et du *clustering*.

2.3.1 Approche non-supervisée dans le cadre des modèles latents pour la réduction de dimension

Un modèle à variable(s) latente(s) (MVL) est un modèle statistique qui relie un ensemble de variables observables à un ensemble de variables latentes. Il suppose l'existence de variable(s) que l'on ne peut pas directement observer (dites « variable(s) latente(s) », « facteur(s) latent(s) » ou « composante(s) latente(s) ») où les réponses aux variables observables sont les résultats des réponses des individus à une ou des variable(s) latente(s) (Székely, Rizzo et Bakirov (2007)).

Les modèles de réduction de dimension sont basés sur les MVL où le nombre de variables latentes est strictement inférieur aux nombres d'observations et de variables observées. Les méthodes les plus utilisées (ACP, AF, ICA ...) projettent les données dans un espace à faible dimension des données originales à grande dimension (Cunningham et Ghahramani (2015)). Par conséquent, la réduction de la dimensionnalité peut être utilisée pour visualiser ou explorer la structure des données, débruiter ou compresser les données.

2.3.2 Comparaison des modèles de déconvolution quantitatives

Les différents types de modèles de réduction de dimension sont résumés en table 2.1) (Székely, Rizzo et Bakirov (2007)).

		Variable observée	
		Quantitative	Qualitative
Variable latente	Quantitative	Analyse Factorielle et ses variants (ACP, ICA)	Analyse des profils latents
	Qualitative	Analyse des traits latents	Analyse Factorielle des Correspondances

Table 2.1 – Analyse en variables latentes pour la réduction de dimension pour les modèles à relation linéaire

Nous allons maintenant aborder les techniques de réduction de la dimension les plus classiques qui seront la base de nos travaux à savoir l'AF, l'ACPP, l'ACP et l'ICA. Nous détaillerons leurs liens, leurs avantages et leurs limites, et nous évoquerons leur représentation Bayésienne à la compréhension du reste de notre développement.

2.3.2.1 Factorisation de matrice

Soit N le nombre d'individus, P le nombre de gènes, et K le nombre de composantes latentes (ou facteurs latents).

2.3.2.1.1 Le principe

La factorisation matricielle représente le processus de déconvolution (ou décomposition) de la matrice du jeu de données $\mathbf{X} \in \mathbb{R}^{P \times N}$ en un produit de deux matrices :

$$\mathbf{X} = \mathbf{\Phi}\mathbf{W} \tag{2.16}$$

avec $\mathbf{W} \in \mathbb{R}^{K \times N}$ la matrice des poids attribués entre les individus et les facteurs latents (*weight matrix*) et $\mathbf{\Phi} \in \mathbb{R}^{P \times K}$ la matrice des poids entre les facteurs et les gènes (*source matrix*).

Pour les lecteurs plus familiers avec les modèles de régression, nous pouvons remarquer que cette factorisation peut s'écrire sous la forme :

$$
\begin{aligned}
X_{1,i} &= \Phi_{1,1}W_{1,i} + \Phi_{1,2}W_{2,i} + ... + \Phi_{1,K}W_{K,i} \\
X_{2,i} &= \Phi_{2,1}W_{1,i} + \Phi_{2,2}W_{2,i} + ... + \Phi_{2,K}W_{K,i} \\
&... \\
X_{P,i} &= \Phi_{P,1}W_{1,i} + \Phi_{P,2}W_{2,i} + ... + \Phi_{P,K}W_{K,i}.
\end{aligned}
\tag{2.17}
$$

Avec cette écriture, nous pouvons voir que les éléments $\Phi_{p,k}$ sont les coefficients d'une régression multivariée où les réponses $\mathbf{X}$ sont expliquées par des variables explicatives non-observées $W_{k,i}$. Ces coefficients et ces variables latentes sont des paramètres du modèle que nous cherchons à estimer.

Suivant la convention pour la représentation de la modélisation par équations structurelles, la figure 2.3 illustre le principe de factorisation de matrices. Les ovales représentent les variables non-observées, les rectangles sont les variables observées et les flèches indiquent le lien entre les différentes variables observées et les facteurs latents. On peut le lire par exemple comme « les individus Ind_1, Ind_2, Ind_3, Ind_4, ..., Ind_N sont associés au premier facteur latent F_1 qui est associé lui-même aux expressions des biomarqueurs B_1, B_2, B_3, ... et B_P ». Sur cette figure, les flèches noires représentent les liens entre les individus et les facteurs qui peuvent être associés à $\mathbf{W}$ et les flèches grises représentent les liens entre les facteurs et les biomarqueurs associés à $\mathbf{\Phi}$.

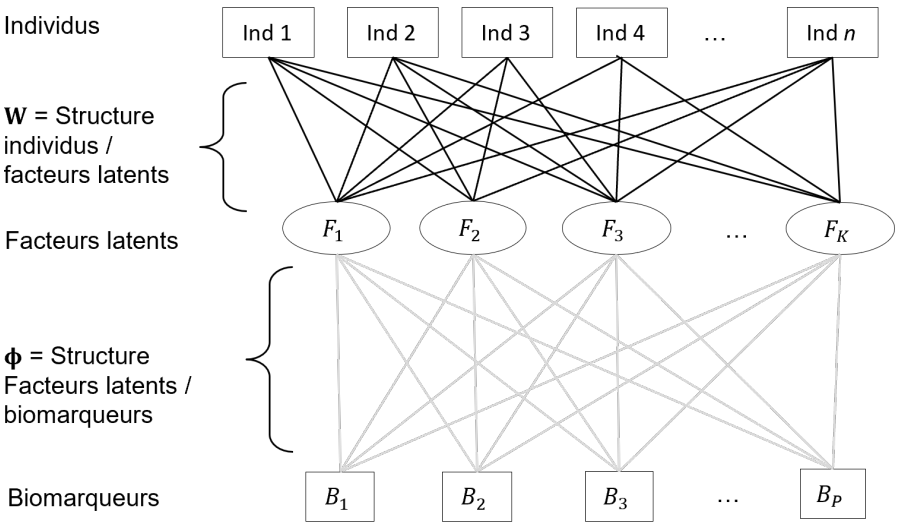


Figure 2.3 – Factorisation matricielle de **X** la matrice des données

2.3.2.1.2 Problèmes d'identifiabilité

Un modèle est dit identifiable s'il existe une solution asymptotique unique, c'est-à-dire, qu'il n'y a qu'une seule valeur des paramètres correspondant à un modèle optimal. Dans le cadre des MVL, une infinité de solutions peuvent correspondre à une même valeur optimale de fonction de coût (ex : maximum de vraisemblance). Les MVL sont donc non-identifiables du fait de plusieurs phénomènes, dont les principaux sont la rotation, la permutation, le signe et l'échelle. Nous pouvons en effet noter que le produit matriciel $\mathbf{W\Phi}$ peut s'écrire par $\mathbf{WRR^{-1}\Phi} = \mathbf{W'\Phi'}$, avec $\mathbf{R}$ une matrice définissant une rotation quelconque du modèle, avec $\mathbf{W} \neq \mathbf{W'}$ et $\mathbf{\Phi} \neq \mathbf{\Phi'}$, représentant différentes solutions optimales. Nous pouvons aussi définir le produit matriciel $\mathbf{W\Phi} = \mathbf{WPP^{-1}\Phi} = \mathbf{W'\Phi'}$ avec $\mathbf{P}$ la matrice de permutation, induisant l'échangeabilité des colonnes (si $\mathbf{P} \in \{0, 1\}$ avec la contrainte d'un seul « 1 » par ligne et par colonne), l'inversion des signes des composantes (si $\mathbf{P} = \text{diag}(a)$ avec $a \in \{-1, 1\}$ un vecteur de constante) ou d'échelle (si $\mathbf{P} = \text{diag}(a)$ avec $a \in \mathbb{R}^+$ un vecteur de constante).

2.3.2.1.3 Erreur de mesure

En reprennant l'equation 2.17, il est naturel de vouloir considérer l'existence d'un bruit statistique au-delà des composantes non-observées. Cette « composante de bruit » a pour but de capturer un signal qui ne résulterait pas d'un processus latent commun aux différents biomarqueurs (ex : voie biologique), mais qui pourrait être dû à une erreur de mesure aléatoire. Il est ainsi possible d'intégrer une erreur additive à la décomposition du jeu de données $\mathbf{X} \in \mathbb{R}^{P \times N}$ tel que :

$$\mathbf{X} = \mathbf{\Phi W} + \mathbf{E} \tag{2.18}$$

avec $\mathbf{W} \in \mathbb{R}^{K \times N}$ la matrice des poids attribués entre les individus et les facteurs latents (*weight matrix*), $\mathbf{\Phi} \in \mathbb{R}^{P \times K}$ la matrice des poids entre les facteurs et les gènes (*source matrix*) et $\mathbf{E} \in \mathbb{R}^{P \times N}$ la matrice des erreurs. Cette équation peut se réécrire sous la forme :

$$
\begin{aligned}
X_{1,i} &= \Phi_{1,1}W_{1,i} + \Phi_{1,2}W_{2,i} + ... + \Phi_{1,K}W_{K,i} + E_{1,i} \\
X_{2,i} &= \Phi_{2,1}W_{1,i} + \Phi_{2,2}W_{2,i} + ... + \Phi_{2,K}W_{K,i} + E_{2,i} \\
&... \\
X_{P,i} &= \Phi_{P,1}W_{1,i} + \Phi_{P,2}W_{2,i} + ... + \Phi_{P,K}W_{K,i} + E_{P,i}.
\end{aligned}
\tag{2.19}
$$

2.3.2.2 Point de vue Bayésien des MVL classiques

Au cours des dernières décennies, des procédures d'inférence Bayésiennes ont été développées pour les MVL les plus classiques. La définition probabiliste de ces modèles permet de les dériver à partir de la définition générale précédente du modèle de factorisation de matrices.

Chacun de ces modèles résulte du choix de considérer un terme d'erreur et des priors sur les éléments des différentes matrices (table 2.2). Cette définition permet d'étendre la gamme des MVL à d'autres modèles ayant d'autres propriétés, dépendant de celles des priors choisis.

Table 2.2 – Comparaison des priors imposés aux matrices en fonction du modèle voulu; avec $\mathbf{\Psi_\Phi}$, $\mathbf{\Psi_W}$, et $\mathbf{\Psi_E}$ la matrice variance-covariance de la distribution de la matrice $\mathbf{\Phi}$, $\mathbf{W}$, et $\mathbf{E}$.

Modèle	Prior W	Prior Phi	Prior E
AF (analyse factorielle)	$\mathcal{N}(0, \Psi_\mathbf{W})$	$\mathcal{N}(0, \Psi_\mathbf{\Phi})$	$\mathcal{N}(0, \Psi_\mathbf{E})$
ACP[a] (analyse en composante principale)	$\mathcal{N}(0, \Psi_\mathbf{W})$	$\mathcal{N}(0, \Psi_\mathbf{\Phi})$	-
ACPP[a] (analyse probabiliste en composante principale)	$\mathcal{N}(0, \Psi_\mathbf{W})$	$\mathcal{N}(0, \Psi_\mathbf{\Phi})$	$\mathcal{N}(0, \Psi_\mathbf{E})$ avec $\Psi_\mathbf{E}$ isotropique [c]
ICA[b](analyse en composante indépendante)	$\mathcal{N}(0, \Psi_\mathbf{W})$	Distribution non-normale	-
ICA-bruitée[b](analyse en composante indépendante bruitée)	$\mathcal{N}(0, \Psi_\mathbf{W})$	Distribution non-normale	$\mathcal{N}(0, \Psi_\mathbf{E})$ avec $\Psi_\mathbf{E}$ isotropique [c]

[a] Hypothèse d'orthogonalité des composantes.
[b] Hypothèse d'indépendence des composantes.
[c] Variance *isotropique lorsque* $\Psi_\mathbf{E} = \sigma_E^2 I$ (à opposer à la variance *anisotropique lorsque diag*$(\Psi_\mathbf{E}) = \sigma_E^2 = \{\sigma_{E_1}^2, \ldots, \sigma_{E_P}^2\}$). Tous les éléments non-diagonaux sont nuls pour la variance isotropique et anisotropique.

2.3.2.2.1 L'analyse factorielle

L'analyse factorielle (AF) est le modèle possédant la structure définie en équation 2.18. Elle est la plus connue et utilisée dans de nombreux domaines (Taherdoost, Sahibuddin et Jalaliyoon (2014)). Elle est définie par :

$$\mathbf{X} = \mathbf{\Phi W} + \mathbf{E} \tag{2.20}$$

avec $\mathbf{\Phi}$, $\mathbf{W}$, $\mathbf{E}$ suivent des lois a priori normales $\mathbf{\Phi} \sim \mathcal{N}(0, \Psi_{\mathbf{\Phi}})$, $\mathbf{W} \sim \mathcal{N}(0, \Psi_{\mathbf{W}})$, et $\mathbf{E} \sim \mathcal{N}(0, \Psi_{\mathbf{E}})$; avec $\mathbf{\Psi_{\Phi}}$, $\mathbf{\Psi_W}$, et $\mathbf{\Psi_E}$ la matrice variance-covariance de la distribution de la matrice $\mathbf{\Phi}$, $\mathbf{W}$, et $\mathbf{E}$..

Contrairement aux autres modèles qui sont décrits dans les paragraphes suivants (et précédemment résumés dans la table tab :chap2priorAF), l'inférence de l'AF doit être réalisée à partir d'une rotation choisie afin d'obtenir une interprétation simple. Cette rotation est réalisée a posteriori en fonction de l'interprétation que l'on souhaite. Parmi les rotations classiques, nous pouvons noter la rotation orthogonale nommée *varimax* (forme de la matrice $\mathbf{R}$), concentrant l'information dans un minimum de composantes, la rotation *quartimax* qui minimise le nombre de facteurs nécessaires (sans contrainte d'orthogonalité), la rotation oblique nommée *promax* qui permet aux facteurs d'être corrélés.

Contrairement à l'AF, les paramètres des autres modèles classiques décrits ci-dessous (ACP, ACPP, ICA et ICA-bruitée) sont optimisés selon une rotation spécifique, ce qui leur permet d'être identifiables que dans les limites de la mise à l'échelle, de l'inversion de signe et de la permutation des colonnes (Sokol, Maathuis et Falkeborg (2014)).

2.3.2.2.2 L'analyse en composante principale (ACP)

L'ACP définit différentes composantes pour que chacune implique une direction qui maximise la variance des données projetées. Cette hypothèse permet de fixer la rotation, afin que les composantes de la matrice des sources soient orthogonales, c'est à dire que $\mathbf{\Phi \Phi^T} = \mathbb{I}$. Il est à noter qu'il faut différencier l'orthogonalité de la rotation comme dans *varimax* et l'orthogonalité de la matrice des sources comme dans l'ACP. Le modèle est défini tel que :

$$\mathbf{X} = \mathbf{\Phi W} \tag{2.21}$$

avec $\mathbf{\Phi}$, $\mathbf{W}$ suivent des lois a priori normales $\mathbf{\Phi} \sim \mathcal{N}(0, \Psi_{\mathbf{\Phi}})$, $\mathbf{W} \sim \mathcal{N}(0, \Psi_{\mathbf{W}})$.

2.3.2.2.3 L'analyse probabiliste en composante principale (ACPP)

L'ACPP est une extension de l'ACP introduisant la modélisation d'une erreur aléatoire isotropique. Elle a été introduite pour la première fois par Lawley (1953). Elle est définie telle que :

$$\mathbf{X} = \mathbf{\Phi W} + \mathbf{E} \tag{2.22}$$

avec $\mathbf{\Phi}$, $\mathbf{W}$, $\mathbf{E}$ qui suivent des lois a priori normales $\mathbf{\Phi} \sim \mathcal{N}(0, \Psi_{\mathbf{\Phi}})$, $\mathbf{W} \sim \mathcal{N}(0, \Psi_{\mathbf{W}})$, et $\mathbf{E} \sim \mathcal{N}(0, \Psi_{\mathbf{E}})$ avec $\Psi = \tau_E I$.

Tipping et C. M. Bishop (1999) ont montré en se basant sur les résultats de Theobald (1975) que ce modèle est apparenté à l'ACP à travers leur vraisemblance. Plus précisément, si $\mathbf{A}$ est la matrice $p \times k$ des vecteurs propres principaux ordonnés de $\mathbf{X^T X}$ et si $\mathbf{\Lambda}$ est la matrice diagonale $k \times k$ avec les valeurs propres correspondantes, nous avons :

$$\widehat{\mathbf{\Phi}}_{ML} = \widehat{\mathbf{A}}(\widehat{\mathbf{\Lambda}} - \widehat{\sigma}^2 I_d)^{1/2}\widehat{\mathbf{R}}, \tag{2.23}$$

où $\widehat{\mathbf{R}}$ est une matrice orthogonale arbitraire, $\widehat{\mathbf{\Phi}}_{ML}$ est l'estimateur du maximum de vraisemblance (ML) de $\mathbf{\Phi}$.

2.3.2.2.4 L'analyse en composantes indépendantes (ICA)

L'ICA est un autre cas particulier de l'analyse factorielle et de l'ACP, où la rotation est définie pour caractériser des composantes qui sont indépendantes et non-corrélées. Ainsi la matrice de rotation sera définie pour respecter cette contrainte (figure 2.4). Le modèle est défini tel que :

$$\mathbf{X} = \mathbf{\Phi W} \tag{2.24}$$

avec $\mathbf{W}$ qui suit une loi a priori normale $\mathbf{W} \sim \mathcal{N}(0, \Psi_{\mathbf{W}})$.

$\mathbf{\Phi}$ suit une **loi non-normale** dont le support est sur $\mathbb{R}$ avec une distribution symétrique afin que les composantes soient indépendantes. Il est à noter qu'une ICA est identifiable si toutes les composantes indépendantes de Φ sont non-normales, à l'exception d'une composante.

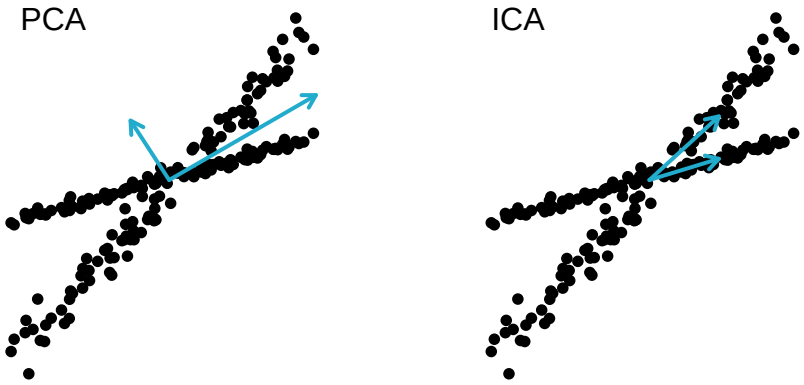


Figure 2.4 – Adaptée de Lewicki et Sejnowski (2000) - Les composantes latentes sont définies par les axes en bleu, l'ACP suppose l'orthogonalité des composantes lorsque l'ICA suppose l'indépendance des composantes.

2.3.2.2.5 L'analyse en composantes indépendantes bruitée (ICA - bruitée)

L'ICA-bruitée est une extension de l'ICA introduisant la modélisation d'une erreur aléatoire isotropique. Elle est définie tel que :

$$\mathbf{X} = \mathbf{\Phi W} + \mathbf{E} \tag{2.25}$$

avec $\mathbf{W} \sim \mathcal{N}(0, \Psi_{\mathbf{W}})$, $\mathbf{E} \sim \mathcal{N}(0, \Psi_{\mathbf{E}})$ avec $\Psi = \tau_E I$, et $\mathbf{\Phi}$ qui suit une **loi non-normale** dont le support est sur $\mathbb{R}$ avec une distribution symétrique.

2.3.2.3 L'identification du nombre de composantes

2.3.2.3.1 Méthodes classiques

L'utilisation des méthodes de déconvolution requiert un paramètre important : le nombre de composantes définissant la taille du sous-espace où seront projetées les données. Il tente de préserver les parties où le plus de variations existent dans les données. La contrainte d'un nombre de composantes trop faibles pourrait concentrer l'information et ne pas refléter la structure latente des données. A l'opposé, augmenter

le nombre de composantes pour augmenter la flexibilité du modèle peut entraîner une sur-décomposition de la variance des données, y-compris celle d'une erreur de mesure aléatoire qui serait intégrée dans les composantes (= sur-apprentissage). La détermination statistique du nombre suffisant de composantes est encore aujourd'hui un champ constant de recherche.

Fréquemment utilisées pour l'ACP, différentes méthodes classiques existent pour la détermination du nombre de composantes notamment basées sur les valeurs propres :

— les composantes sélectionnées ayant une valeur propre supérieure à 1, nommé la règle de Kaiser (1960)) : chaque composante doit expliquer la variance d'au moins une variable ;

— les composantes sélectionnées décrivant au moins x % de la variance ;

— les composantes sont sélectionnées jusqu'à ce que l'apport de la nouvelle composante ait peu d'impact sur l'explication de la variance, nommé la règle du coude ou de Cattell (1966).

Le vecteur des valeurs propres λ d'une matrice de variance-covariance $\mathbf{\Sigma}$ est défini tel que :

$$\Sigma v = \lambda v \tag{2.26}$$

avec v le vecteur propre de Σ. Le vecteur des valeurs propres λ correspond à la transformation d'échelle. Cette quantité est disponible pour les différents modèles de factorisation. Les modèles de factorisation de matrices, pour estimer les éléments des matrices, doivent considérer la matrice de variance-covariance comme non-singulière, c'est à dire que toutes les valeurs propres sont strictement supérieures à 0.

Il existe d'autres procédures de sélection implémentées dans des librairies comme la librairie *psych* (Revelle (2022)) avec sa fonction d'analyse en parallèle (Horn (1965)) pour l'analyse factorielle. Cependant les méthodes de cette librairie sont spécialisées pour certains modèles de déconvolution et peuvent ne pas convenir pour des données non-inversibles (déterminant de la matrice du jeu de données égal à 0).

Comme toute quantité estimée à partir des données, il existe une incertitude autour de ces estimateurs qui n'est pas prise en compte par ces méthodes. Ces méthodes ont tendance à sur-estimer ou sous-estimer le nombre de composantes du modèle ayant

pour conséquence à conduire à des résultats erronés et ininterprétables (Preacher et MacCallum (2003)). Une comparaison de ces méthodes avec des données génomiques est présentée par Cangelosi et Goriely (2007). Des procédures de ré-échantillonnages ont été proposées pour essayer de quantifier et prendre en compte cette incertitude au travers de la stabilité des composantes. La procédure *icasso* a été implémentée sous plusieurs langages de programmation dont notamment la bibliothèque *icasso* de Matlab (Himberg et Hyvärinen (2003)), et la bibliothèque R *MineICA* (Biton (2021)). Les bibliothèques implémentant l'ICA, dont *FastICA* (l'implémentation sur laquelle est fondée *icasso*) reposent sur des initialisations aléatoires des paramètres, qui peuvent conduire à des estimations des composantes différentes, du fait de la présence de maxima locaux. Une façon d'atténuer ce problème comme dans la procédure *icasso* est d'exécuter fastICA plusieurs fois, puis de regrouper les estimations à partir de clusters et d'utiliser comme estimations finales les centrotypes des groupes (Biton et al. (2014) ; Kairov et al. (2017)).

2.3.2.3.2 Limites de ces approches

Le rôle principal des méthodes de fouille de données est d'identifier des structures afin de générer de nouvelles hypothèses. Ces structures doivent donc être interprétables, ce qui, en fonction du contexte, influence le choix de la rotation. Dans les cas précédents, le nombre de paramètres est fixé en amont de l'inférence et nécessite d'optimiser les paramètres de plusieurs modèles en considérant un nombre de variables latentes différent afin de faire de la sélection de modèle. Par exemple, *Varimax* est souvent choisie pour l'AF pour concentrer l'information dans peu de composantes et ainsi faciliter l'interprétation. Néanmoins, un nombre faible de composantes n'est pas toujours désirable, notamment lorsque nous essayons de modéliser des processus biologiques dont nous ignorons le nombre, mais qui est certainement grand.

À l'inverse des méthodes précédentes où le nombre de composantes est réduit afin de faciliter l'interprétation, l'approche Bayésienne non-paramétrique se place dans un cadre où le nombre de composantes est infini, mais seul un sous-ensemble est présent dans notre échantillon. Nous supposerons donc que le nombre de composantes n'est pas fixe dans la population et peu grandir avec le nombre d'observations (Sethuraman (1994) ; Ishwaran et James (2001) ; T. L. Griffiths et Ghahramani (2005)). Cette

hypothèse est raisonnable pour la présence d'anomalies moléculaires qui peuvent être rares, et donc avoir une probabilité très faible d'être observée dans un petit échantillon, mais qui grandit avec le nombre d'observations. Cette approche sera détaillée dans le chapitre suivant.

Chapitre 3

Caractérisation moléculaire individuelle à l'aide de l'analyse graphique avec un nombre infini de composantes indépendantes et parcimonieuses

Article accepté :
RINCOURT, Sarah-Laure, MICHIELS, Stefan, DRUBAY Damien (2022). "Complex disease individual molecular characterization using infinite sparse graphical independent component analysis". In : *Cancer informatics*. DOI : 10.1177/11769351221105776.

Chapitre adapté de l'article

La base de données sur l'expression génétique du cancer du sein est disponible dans la librairie R biospear. Des exemples simulés et le code du modèle isgICA sont disponibles à l'URL suivante https ://github.com/Oncostat/isgICA.

Du fait de l'hétérogénéité tumorale, il est nécessaire de développer une médecine de précision, afin d'apporter un soin personnalisé pour une meilleure efficacité. Pour caractériser cette hétérogénéité, nous avons proposé un modèle pour étudier l'hétérogénéité inter-individuelle moléculaire, en supposant que l'expression tumorale résulte d'un mélange d'un sous-ensemble de signatures indépendantes. Nous avons déconvolué les données omiques à l'aide d'une analyse non-paramétrique Bayésienne en composantes indépendantes avec une double structure de parcimonie pour les matrices des sources et des poids, correspondant respectivement aux associations gènes-composantes et individus-composantes.

Sommaire

3.1 Introduction **43**

3.2 Méthodes **46**

3.2.1 Identification du nombre de composantes présentes dans la population - Processus bêta-Bernoulli 46

3.2.1.1 Modèle d'allocation 46

3.2.1.2 Modèles d'allocation et hypergraphes 47

3.2.1.3 Processus bêta-Bernoulli 48

3.2.2 Modèle graphique pour une analyse en composantes indépendantes infinies et parcimonieuses 50

3.2.3 Travaux connexes précédents 52

3.2.4 Contrainte sur le modèle isgICA : apport du profil de base au modèle 53

3.2.5 Mises à jour 60

3.2.6 Optimisation des hyperparamètres 62

3.2.7 Comparaison à un modèle de référence : l'ICA avec blanchissement 63

3.3 Résultats **63**

3.3.1 Données synthétiques 63

3.3.1.1 Scénarios 63

3.3.1.2 Critères de reconstruction 64

3.3.2 Reconstruction des structures latentes 65

3.4 Application : données sur le cancer du sein au stade précoce **73**

3.5 Discussion **77**

3.6 Conclusion **79**

3.1 Introduction

Une compréhension approfondie des mécanismes moléculaires pour des maladies complexes, comme le cancer, est l'un des principaux défis actuels pour développer la médecine de précision. L'identification de ces mécanismes à partir de données omiques, telles que l'expression des gènes, est difficile en raison des relations complexes dans diverses voies moléculaires impliquant des centaines ou des milliers d'acteurs (par exemple, des gènes) et du nombre relativement faible de patients dans les bases de données disponibles.

Pour limiter le fléau de la dimension, l'identification des structures de données omiques de haute dimension non-observées, qui fournissent un aperçu des mécanismes moléculaires, est souvent effectuée en utilisant des modèles à variables latentes (Cunningham et Ghahramani (2015), MVL) pour la séparation/déconvolution en aveugle des sources, y compris l'analyse en composantes principales (PCA), l'analyse en composantes indépendantes (ICA) ou l'analyse factorielle (AF). Pour identifier les composantes moléculaires indépendantes, nous avons basé notre travail sur le modèle d'ICA, et nous utilisons ci-dessous la terminologie correspondante (voir section 2.3.2.2), c'est-à-dire que les matrices des sources et des poids correspondent aux paramètres représentant respectivement l'association entre les composantes avec les gènes et avec les observations.

Pour interpréter les structures identifiées comme des mécanismes (ou des voies) moléculaires, des méthodes parcimonieuses [1] peuvent être utilisées pour sélectionner un sous-ensemble de variables omiques associées à chaque composante, de façon similaire au modèle factoriel par graphe proposé par Yoshida et West (2010). Chacune de ces structures latentes peut représenter différents mécanismes moléculaires, tels que des voies biologiques, associés à seulement l'expression d'un sous-ensemble de gènes. Plusieurs approches parcimonieuses ont été proposées, en particulier dans le cadre Bayésien, imposant la parcimonie à la matrice des sources en utilisant le prior *spike and slab* (West et al. (2003) ; D. Knowles et Ghahramani (2011)), le processus du buffet indien (T. L. Griffiths et Ghahramani (2005)), le prior de Laplace (Kabán

1. Méthodes utilisant un nombre restreint de paramètres non-nuls par rapport au nombre de variables incluses. Les paramètres nuls sont pénalisés (ou régularisés) et représentés par des zéros dans un vecteur (ou une matrice) binarisé(e).

(2007)), ou le prior horseshoe (Carvalho et al. (2008)).

Bien que cette approche soit adaptée pour donner un aperçu des mécanismes communs à tous les individus, il est connu en oncologie que la tumeur peut résulter de mécanismes moléculaires différents (ou d'altérations / états différents du même mécanisme) selon les patients, ce qui complique le développement de thérapies efficaces pour l'ensemble de la population. Une façon naturelle de considérer cette hétérogénéité inter-individuelle avec le MVL est d'imposer une seconde structure de parcimonie à la matrice des poids, associant les individus aux composantes.

Comme le nombre de mécanismes moléculaires impliqués est inconnu (mais probablement important), et que le nombre de leurs altérations possibles peut croître rapidement avec le nombre d'individus, il est préférable de ne pas fixer la dimension du modèle mais de déduire le nombre de composantes présentes dans la population étudiée. Ces deux perspectives de modélisation peuvent être considérées simultanément en utilisant le processus bêta-Bernoulli (BBP, Hjort (1990)) comme prior de l'hypergraphe de la matrice des poids (D. Knowles et Ghahramani (2011) ; Paisley et Carin (2009)).

Cette approche considère une distribution a priori sur l'espace du modèle de dimension infinie n'attribuant presque sûrement qu'un nombre fini de 1 dans la matrice de l'hypergraphe (donc un nombre fini de composantes parcimonieuses) dans un échantillon fini (Hjort et al. (2010)). Cela permet de considérer qu'il existe un nombre infini d'altérations moléculaires, mais que seul un sous-ensemble est présent dans notre échantillon fini. Cette approche a la propriété attrayante de permettre à la complexité du modèle de croître avec le nombre d'observations sous la régularisation des hyperparamètres du BBP (Paisley et Carin (2009) ; Hjort et al. (2010) ; Gershman et Blei (2012)).

Alors que les travaux mentionnés précédemment se sont concentrés sur le codage parcimonieux de la matrice des poids ou de la matrice des sources, nous proposons d'imposer la parcimonie sur la matrice des poids et sur la matrice des sources. À notre connaissance, il s'agit de la première étude imposant cette double parcimonie dans un modèle à dimension infinie et proposant une procédure d'optimisation pour améliorer la reconstruction des structures latentes sous-jacentes. Cependant, l'interprétation des composantes résultantes reste complexe et hasardeuse car elles peuvent représen-

ter différents mécanismes/voies moléculaires ou leurs différentes altérations chez les patients. Au lieu de caractériser précisément ces mécanismes moléculaires, nous proposons d'identifier des « altérations » s'écartant d'un profil moléculaire de base. Notre choix est motivé par l'hypothèse que la progression différentielle de la maladie ou de la résistance aux médicaments résulte d'un mélange de multiples altérations moléculaires, qui pourraient être différentes entre les patients. Nous avons imposé cette contrainte de base en faisant en sorte que la première composante de la matrice des poids soit un vecteur de 1, c'est-à-dire que tous les individus soient associés à cette composante, représentant le profil moléculaire moyen de la population (que nous avons appelé le profil moléculaire de base). Nous considérons également une composante de bruit (comme l'ICA bruitée Hyvärinen (1999)) pour capturer les caractéristiques spécifiques de chaque échantillon, incluant une potentielle erreur de mesure.

Dans ce travail, nous avons d'abord évalué la capacité de notre modèle isgICA (*infinite sparse graphical independent component analysis*) à reconstruire les structures de parcimonie des matrices des poids et des sources par une étude de simulation. Nous avons ensuite comparé ses performances d'identification du nombre de composantes et pour la reconstruction des matrices aux algorithmes de référence. Enfin, nous avons appliqué notre méthode pour modéliser l'hétérogénéité de l'expression génique dans un ensemble de données d'expression génique de tumeurs de patientes atteintes d'un cancer du sein incluses dans des essais cliniques de chimiothérapie à base d'anthracycline et nous avons illustré la pertinence de cet algorithme pour identifier en aveugle des signatures d'expression génique pertinentes et connues du cancer du sein.

3.2 Méthodes

3.2.1 Identification du nombre de composantes présentes dans la population - Processus bêta-Bernoulli

3.2.1.1 Modèle d'allocation

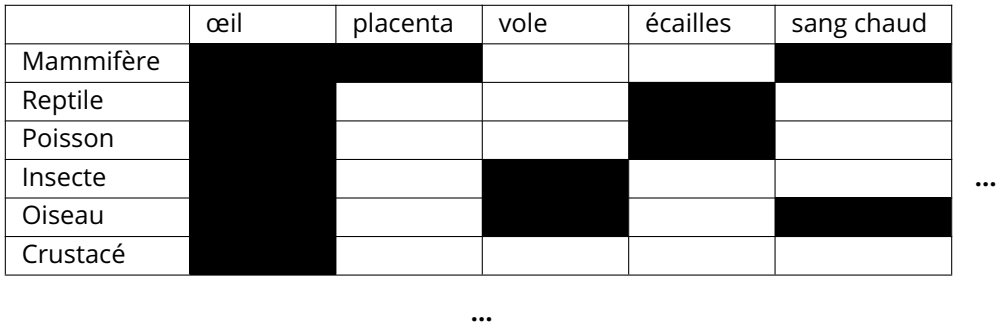


Figure 3.1 – Représentation des allocations de caractéristiques physiologiques de certains groupes d'espèces connues.
Chaque groupe d'espèces possède ou non certaines caractéristiques, on peut créer un vecteur binaire (oui (1, case noire) / non (0, case blanche)) définissant notre groupe d'espèces. Dans cet exemple, nous souhaitons décrire plusieurs groupes d'espèces par leurs caractéristiques.
Il existe des caractéristiques qui peuvent être communes à tous les groupes d'espèces comme ici les yeux, partagées par certains comme les écailles, ou unique comme le placenta pour les mammifères. Ici cinq caractéristiques nous ont permis de caractériser nos groupes d'espèces, mais il serait possible d'en rajouter, ce nombre n'est donc pas fixé.

Le clustering est la méthode la plus connue d'allocation de groupes latents à des observations. Nous pouvons illustrer cette allocation par un exemple simple par la figure 3.1. Chaque observation (une ligne de la figure) possède ou non certaines caractéristiques que l'on peut résumer sous forme d'un vecteur binaire (oui (1, case noire) / non (0, case blanche)). Les méthodes d'allocations, comme présenter en figure 3.1, permettent pour chaque observation d'avoir plusieurs caractéristiques. Le clustering est une approche d'allocation où chaque observation ne possède qu'une seule caractéristique nommé « *cluster* ». Par exemple on pourra résumer en un cluster les caractéristiques « oeil = 1 ; Placenta = 1 ; vole = 0 ; écaille = 0 ; Sang chaud = 1 » qui décriraient les mammifères ou les caractéristiques « oeil = 1 ; Placenta = 0 ; vole = 0 ; écaille = 1 ; Sang chaud = 0 » qui décriraient les reptiles et les poissons. Chacun des

clusters est composé de plusieurs variables correspondant aux caractéristiques, chacune associée à un paramètre qu'il faut estimer (ex : vecteur des moyennes d'une gaussienne multivariée représentant un des clusters dans un modèle de mélange gaussien). Certains de ces paramètres étant communs à différentes espèces, il est assez inefficient de devoir estimer le même paramètre plusieurs fois pour différents clusters. De plus, dans cet exemple, le nombre de clusters grandit très vite, car résulte du nombre de combinaison de caractéristiques possibles, et serait infini si les paramètres avaient un support continu (ex mélange de gaussiennes multivariées)

Le modèle d'allocation permet de réduire le nombre de paramètres à estimer en allouant chaque caractéristique à chaque espèce au lieu des combinaisons de caractéristiques. Dans l'approche que nous présenterons dans les prochains chapitres, chaque ligne correspondra à un patient, et chaque colonne à une variable latente. Cette représentation nous permettra d'allouer différents mécanismes biologiques (variables latentes) à différents patients pour représenter l'hétérogénéité inter-individuelle. N'ayant pas d'a priori sur le nombre de variables latentes, nous utiliserons l'inférence Bayésienne non-paramétrique (section 2.1.3), réalisant l'inférence pour un matrice binaire d'allocation comportant un nombre infini de colonnes.

3.2.1.2 Modèles d'allocation et hypergraphes

Pour interpréter les structures identifiées comme des mécanismes (ou des voies) moléculaires, des méthodes parcimonieuses peuvent être utilisées pour sélectionner un sous-ensemble de variables omiques associées à chaque composante, de façon similaire au modèle factoriel par graphe proposé par Yoshida et West (2010).

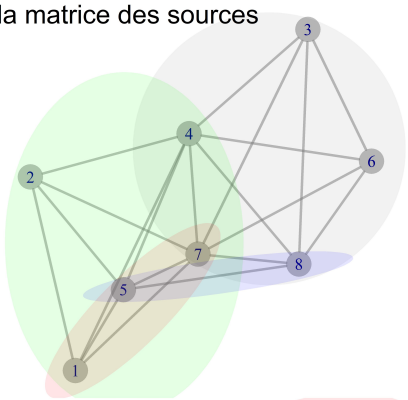
Comme l'illustre la sous-figure 3.2.a, la structure de parcimonie de la matrice des sources peut être considérée comme une matrice d'hypergraphes. L'approche par hypergraphes est une généralisation des méthodes de graphes prenant en compte les interactions d'ordre supérieur des nœuds (par exemple, les gènes) pour modéliser des relations complexes (Feng et al. (2021)), qui sont représentées par différents sous-ensembles de nœuds (potentiellement superposés) associés à différents hypergraphes.

Cette seconde structure de parcimonie peut également être considérée comme un hypergraphe dans lequel, cette fois, chaque hyperedge représente le sous-ensemble d'individus présentant les mécanismes moléculaires correspondants (voir la sous-figure

3.2.b).

a. Hypergraphe de la matrice des sources

	C1	C2	C3	C4
Gène 1	0	1	1	0
Gène 2	0	0	1	0
Gène 3	1	0	0	0
Gène 4	1	0	1	0
Gène 5	0	1	1	1
Gène 6	1	0	0	0
Gène 7	1	1	1	0
Gène 8	1	0	0	1



b. Hypergraphe de la matrice des poids

	C1	C2	C3	C4
Individu 1	0	1	0	1
Individu 2	0	1	0	1
Individu 3	0	1	1	1
Individu 4	1	1	1	0
Individu 5	0	0	1	1
Individu 6	1	0	1	0
Individu 7	1	0	1	0

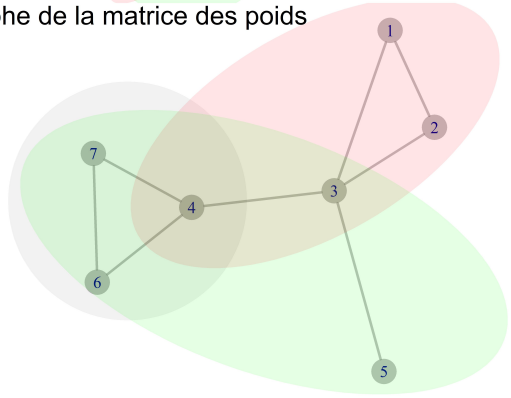


Figure 3.2 – Interprétation de la structure de la matrice de parcimonie de l'analyse en composantes indépendantes sous forme d'hypergraphe. La parcimonie de la matrice des sources (sous-figure a) représente K composantes moléculaires (*hyperedge*) associées à différentes combinaisons de gènes (nœud) qui peuvent représenter différents mécanismes moléculaires ou leurs altérations. La matrice des poids (sous-figure b) définit différents profils par des combinaisons individuelles de ces composantes moléculaires pour caractériser l'hétérogénéité de la maladie du patient.

3.2.1.3 Processus bêta-Bernoulli

Comme discuté précédemment, il est nécessaire de définir un processus stochastique qui sera adapté pour la génération de cette matrice binaire avec un nombre infini de colonne. Le processus de Bernoulli est un choix adapté pour la génération d'un vecteur binaire infini d'allocation des variables latentes à un individu. La distribution bêta étant conjuguée à la distribution de Bernoulli, le processus bêta peut être utilisé comme prior des paramètres des processus de Bernoulli afin d'apporter de l'information, notamment en terme de régularisation (cf paragraphe ci-dessous). Ainsi, le processus bêta-Bernoulli

va permettre de générer pour chaque individu un vecteur binaire infini d'allocation, et permettra donc d'obtenir une matrice similaire à la figure 3.1 avec un nombre infini de colonne.

En effet, il est désirable de définir un prior qui ne se limitera pas à un nombre fini le nombre de composantes allouées aux observations d'un échantillon fini, ce que permet presque sûrement le processus beta-Bernoulli. Cette propriété est importante, dans le sens où la diversité des mécanismes biologiques tumoraux est supposé très grands, mais pas infini. De plus, plusieurs résultats empiriques dans différents domaines suggèrent que le nombre de composantes et leur parcimonie ont un comportement « loi puissance » (*power law*) de différents types (Mitzenmacher (2003) ; M. Newman (2005) ; Clauset, Shalizi et M. E. J. Newman (2009)). Le premier type de comportement *power law* correspond au fait que le nombre de composantes allouées à au moins une observation augmente avec le nombre d'observations disponible selon la loi : $K \sim cN^a$ avec $a \in (0, 1)$ et $c > 0$ les paramètres de cette loi puissance (Broderick, Jordan et Pitman (2012)). De plus, un second type de comportement *power law* a également été décrit, correspondant au fait que nombre j d'allocations à chaque composante suit une distribution puissance telle que :

$$\frac{\Gamma(k-a)}{\Gamma(k+1=} \sim k^{-1-a}, \quad k \to \infty \tag{3.1}$$

avec Γ la fonction gamma. Le processus beta bernouilli permet d'apporter cette information a priori tout en restant flexible à l'écart à cette hypothèse (comme le montrera notre étude de simulation ci-dessous).

Ainsi, les composantes possédant une forte probabilité pourraient être allouées à tous les individus de notre échantillon, alors que celles avec une faible probabilité pourraient ne pas être allouée. Plus la taille d'échantillon grandit, plus la probabilité que ces dernières soient allouées à une des observations grandit, ce qui permet au modèle de se complexifier avec l'apport de nouvelles données.

Historiquement le processus bêta-Bernoulli a été proposé par Ghahramani et T. Griffiths (2006) avec le processus du buffet indien (IBP). Plus tard, Thibaux et Jordan (2007) ont fait le lien entre l'IBP et le processus bêta à 1 paramètre, permettant sa généralisation avec la représentation en deux paramètres Phadia (2013), en trois para-

mètres Armagan, Dunson et Clyde (2011). Il est aussi possible de définir le processus bêta-Bernoulli à partir du *stick-breaking* en deux paramètres (Paisley et Carin (2009) ; Paisley et Jordan (2016)) ou en trois paramètres (Broderick, Jordan et Pitman (2012)) ou d'un spike and slab basé sur la distribution de dirac (Mitchell et Beauchamp (1988)).

3.2.2 Analyse graphique en composantes indépendantes infinies et parcimonieuses (*Infinite sparse graphical independent component analysis*, isgICA)

Soit N le nombre d'individus, P le nombre de gènes et K le nombre de composantes latentes. L'analyse en composantes indépendantes bruitée vise à décomposer une matrice de données $\mathbf{X} \in \mathbb{R}^{P \times N}$ en le produit de deux matrices plus le bruit résiduel comme suit :

$$\mathbf{X} = \mathbf{\Phi}\mathbf{W} + \mathbf{E}, \tag{3.2}$$

où $\mathbf{W} \in \mathbb{R}^{K \times N}$ représente la matrice des poids, $\mathbf{\Phi} \in \mathbb{R}^{P \times K}$ représente la matrice des sources, et $\mathbf{E} \in \mathbb{R}^{P \times N}$ représente la matrice de bruit normal.

Il est d'usage d'imposer des priors non-normaux à l'ICA afin de favoriser l'indépendance entre les composantes, l'exemple classique étant un mélange de distribution normale. L'utilisation de priors de régularisation (*shrinkage*) est également commun, permettant d'introduire de la parcimonie pour n'associer qu'un sous-ensemble de chaque variable à chaque composante.

Une approche alternative serait d'introduire de la parcimonie en considérant une matrice binaire $\mathbf{\Theta}$, représentant la structure de parcimonie de $\mathbf{\Phi}$, c'est-à-dire :

$$\mathbf{X} = (\mathbf{\Theta} \circ \mathbf{\Phi})\mathbf{W} + \mathbf{E}, \tag{3.3}$$

avec $\circ$ représentant le produit par élément. Cette équation correspond à la formulation du modèle graphique factoriel parcimonieux de Yoshida et West (2010), qui a inspiré le nom de notre approche présenté dans la section précédente (3.2.1.2).

Nous avons utilisé cette approche pour imposer la parcimonie sur $\mathbf{W}$, permettant l'attribution d'un sous-ensemble des K composantes à chaque individu. En considérant

la matrice binaire parcimonieuse $\mathbf{Z} \in \mathbb{1}^{K \times N}$, le modèle est défini comme :

$$\mathbf{X} = \mathbf{\Phi}(\mathbf{W} \circ \mathbf{Z}) + \mathbf{E}. \tag{3.4}$$

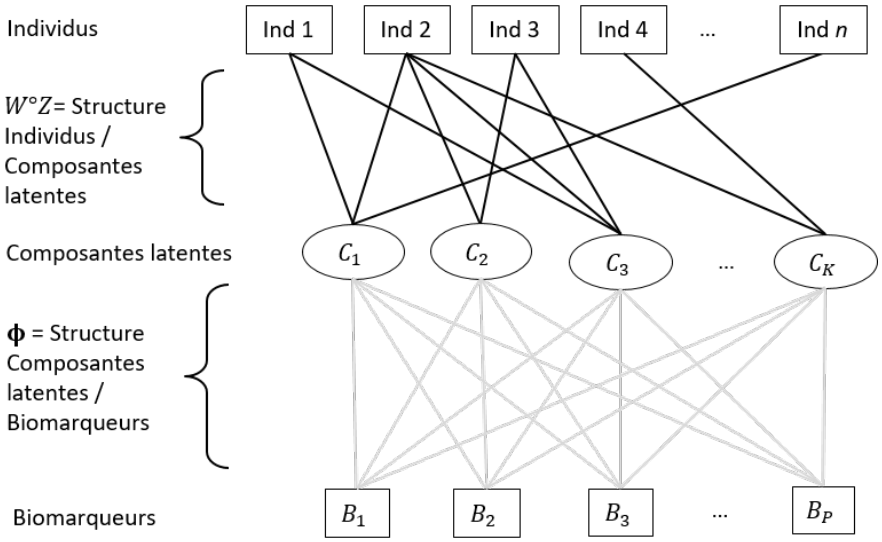


Figure 3.3 – ICA (exemple non-bruitée) parcimonieuse pour dans le cas d'un modèle d'expression génique. Les individus sont liés aux composantes latentes par la présence de « 1 » dans la matrice $\mathbf{Z}$ (représenté par la présence de liens noirs). Les composantes latentes sont quant à elles toutes associées à tous les biomarqueurs.

La figure 3.3 permet de visualiser le principe du modèle isgICA avec une parcimonie de la matrice $\mathbf{Z}$. Les rectangles et les ovales représentent respectivement les variables observées (les individus Ind et biomarqueurs B) et les variables non-observées (latentes ; les composantes, C). Cette forme du modèle essaye d'optimiser le trade-off entre l'ajustement aux données et la simplicité du modèle : seul un petit nombre de liens sera suffisant pour expliquer les données (liens représentés par des « 1 » dans la matrice de parcimonie $\mathbf{Z}$ avec :

$$\mathbf{Z} = \begin{array}{c} \\ \mathsf{Ind}_1 \\ \mathsf{Ind}_2 \\ \mathsf{Ind}_3 \\ \mathsf{Ind}_4 \\ \dots \\ \mathsf{Ind}_n \end{array} \begin{array}{c} \begin{array}{ccccc} \mathsf{C}_1 & \mathsf{C}_2 & \mathsf{C}_3 & \dots & \mathsf{C}_K \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & \dots & 0 \\ 1 & 1 & 1 & \dots & 1 \\ 0 & 1 & 1 & \dots & 0 \\ 0 & 0 & 0 & \dots & 1 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 0 & \dots & 0 \end{pmatrix} \end{array}$$

et par des traits noirs dans la figure 3.3). On parle alors de représentation parcimonieuse (*sparse representation*). La matrice $\mathbf{Z}$ représente l'allocation des composantes aux individus avec des composantes qui peuvent être communes à tous les individus, partagées, ou uniques. Chaque individu est composé d'un vecteur binaire d'allocation qui représente son profil moléculaire. Au niveau de l'interprétation biologique, la matrice $\mathbf{Z}$ permet de modéliser et de décomposer les profils de biomarqueurs (présents dans la matrice des sources) des individus en la somme de différents sous-profils non-observés.

Comme le nombre de potentielles altérations moléculaires non-observées est inconnu mais probablement élevé, nous considérons une ICA non-paramétrique avec un nombre infini de composantes (c'est-à-dire $K = \infty$). Le processus bêta-Bernoulli (BBP) est un prior non-paramétrique approprié pour la matrice binaire ($\mathbf{Z}$) avec un nombre infini de lignes (ou de colonnes), fournissant un nombre fini de lignes non nulles presque sûrement dans le cas d'un échantillon fini. La nature non-paramétrique de cette approche permet à la complexité du modèle de croître avec les données (c'est-à-dire que K augmente avec N), ce qui est une propriété intéressante pour déduire le nombre de composantes moléculaires présents dans la population étudiée, qui devrait augmenter avec le nombre d'individus.

3.2.3 Travaux connexes précédents

Ce modèle est inspiré du travail fourni par B. Chen, M. Chen et al. (2010) en formulant notre modèle à partir d'ICA et non d'analyse factorielle. Ce modèle a été aussi repris dans les dictionnaires comme dans Zhou et al. (2012).

Plusieurs méthodes ont été développées pour induire de la parcimonie au sein de

méthodes de réduction de dimension à partir d'un processus bêta-Bernoulli :

- Basées sur la parcimonie de la matrice des sources $\mathbf{\Phi}$ de l'analyse factorielle induite par le prior *spike-and-slab* (Rattray et al. (2009) ; Sharp et Rattray (2010) ; Foo et Shim (2021) ; W. Wang et Stephens (2021) etc.) ;
- Basées sur la parcimonie de la matrice $\mathbf{\Phi}$ de l'analyse factorielle induit par la version du processus bêta-Bernoulli à deux paramètres (Buettner et al. (2017) ; Argelaguet et al. (2018)) ;
- Basées sur la parcimonie de la matrice des poids $\mathbf{W}$ du modèle BPFA (*Beta Process Factor Analysis* induit par la version du processus bêta-Bernoulli à deux paramètres (Paisley et Carin (2009)) et son variant à trois paramètres en *stick-breaking* (Broderick, Jordan et Pitman (2012))) ;
- Basées sur la parcimonie de la matrice des sources $\mathbf{\Phi}$ du modèle NSFA (*non-Parametric Sparse Factor Analysis*) induit par le prior IBP D. Knowles et Ghahramani (2011) ; une alternative est présentée dans l'article de Yang et Koeppl (2018) où la parcimonie est sur la matrice des poids $\mathbf{W}$ induit par le prior IBP ;
- Une extension avec les modèles en tensor 3D (individus/gènes/tissus) (Hore et al. (2016)) où le processus à deux paramètres est défini sur les tissus.

Toutes ces versions du modèle se sont concentrées sur l'optimisation de différents critères d'ajustement comme le RMSE (*root-mean-square error*) ou l'ELBO ; cependant ils ne se sont pas intéressés à l'interprétabilité de leur modèle ainsi qu'à la capacité de leur modèle à reconstruire les structures de parcimonie des matrices des poids $\mathbf{W}$ et des sources $\mathbf{\Phi}$.

3.2.4 Contrainte sur le modèle isgICA : apport du profil de base au modèle

Nous définissons un profil de base, $\mathbf{Z_0} = [\mathbf{1}, ..., \mathbf{1}]_\mathbf{N}$ comme une composante associée à tous les individus, telle que la matrice $\mathbf{Z} = [\mathbf{Z_0}, \mathbf{Z^*}]$ avec $\mathbf{Z^*}$ la résultante de la réalisation du processus bêta-Bernoulli.

En raison de leur interprétation définie précédemment, nous dénommerons ces composantes binaires parcimonieuses composant $\mathbf{Z^*}$ par « altérations » (du profil de base) dans la suite du manuscrit. Les dimensions des matrices isgICA de profil de base correspondantes sont $\mathbf{Z} \in \mathbb{1}^{(K+1)\times N}$, $\mathbf{W} \in \mathbb{R}^{(K+1)\times N}$, $\mathbf{\Phi} \in \mathbb{R}^{P\times(K+1)}$, et $\mathbf{E} \in \mathbb{R}^{P\times N}$.

La figure 3.4 illustre un exemple de matrice $\mathbf{Z}$ parcimonieuse du modèle isgICA, incluant le profil de base $\mathbf{Z_0}$ (nommé dans la figure C_0) avec :

$$\mathbf{Z} = \begin{array}{c} \\ \text{Ind}_1 \\ \text{Ind}_2 \\ \text{Ind}_3 \\ \text{Ind}_4 \\ \dots \\ \text{Ind}_n \end{array} \begin{array}{c} \begin{array}{cccccc} C_0 & C_1 & C_2 & C_3 & \dots & C_K \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 1 & \dots & 0 \\ 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & 0 & 1 & 1 & \dots & 0 \\ 1 & 0 & 0 & 0 & \dots & 1 \\ 1 & \dots & \dots & \dots & \dots & \dots \\ 1 & 1 & 0 & 0 & \dots & 0 \end{pmatrix} \end{array} .$$

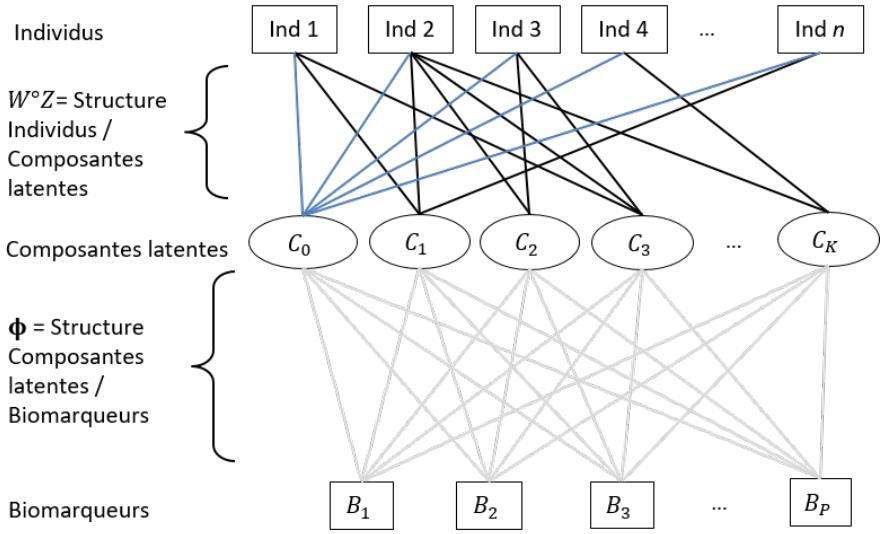


Figure 3.4 – ICA parcimonieuse avec un profil de base Z_0 (liens bleus) et un ensemble d'altérations $\mathbf{Z}^*$ (liens noirs) dans le cade de la modélisation de l'expression génique. L'allocation d'une composante latente à un individu est représentée par la présence d'un « 1 » dans la matrice $\mathbf{Z}$. Dans cet exemple, les composantes latentes sont toutes associées à tous les biomarqueurs.

Cette représentation parcimonieuse est basée sur la figure 3.3. La composante C_0 représente le profil de base et les composantes C_1 à C_K représentent les composantes altérées, soit les sous-profils. Les individus sont la somme du profil de base et des différents sous-profils non-observés.

Nous avons considéré des priors conjugués pour les éléments des matrices du modèle, ce qui permet de calculer analytiquement les posteriors variationnels par l'approche par ascension de coordonnées (section 2.2.3.1). La représentation graphique du

modèle proposé est illustrée à la figure 3.7 et le modèle complet est formulé comme suit :

$$
\begin{aligned}
&\mathbf{X} = \mathbf{\Phi}(\mathbf{W} \circ \mathbf{Z}) + \mathbf{E} \\
&X_i \sim \mathcal{N}(\mathbf{\Phi}(\mathbf{W} \circ \mathbf{Z}), \tau_E^{-1}\mathbb{I}_P) \\
&\Phi_{j,k} \sim \mathcal{N}(0, \tau_{\phi_{j,k}}^{-1}) \\
&W_{k,i} \sim \mathcal{N}(0, \tau_W^{-1}\mathbb{I}_K) \\
&\tau_{\phi_{j,k}} \sim \mathcal{G}(c_{j,k}, d_{j,k}) \\
&\tau_W \sim \mathcal{G}(e, f) \\
&Z_{k^*,i} \sim \mathcal{B}e(\pi_k) \\
&\pi_{k^*} \sim \mathcal{B}(\frac{a}{K}, \frac{b(K-1)}{K}) \\
&Z_{0,} = [1, .., 1]_N
\end{aligned}
\tag{3.5}
$$

où i = 1, ..., N ; j = 1, ..., P ; k = 0, ..., K ; and k*= 1,...,K. Tous les éléments des matrices $\mathbf{W}$ et $\mathbf{\Phi}$ sont des paramètres à estimer qui nécessitent de définir un prior. Les priors spécifiés, appelés détermination automatique de la pertinence (*automatic relevance determination*, ARD), appliquent une régularisation sur les éléments des matrices $\mathbf{W}$ et $\mathbf{\Phi}$. Cette régularisation est induite par la forme de super-gaussienne de la densité de ce prior qui est centrée sur 0. Comme montré dans la section , l'inférence Bayésienne une vraisemblance pénalisée par la divergence de KL, les estimations élevées des éléments seront régularisés vers 0 par ce prior. Cette forme de super-gaussienne possède également des queues dites lourdes, c'est-à-dire qui ne sont pas bordées exponentiellement, assurant une densité de probabilité plus élevée qu'une distribution gaussienne pour les valeurs les plus élevées, permettant aux estimations des éléments les plus élevées d'échapper à cette régularisation (illustration figure 3.6). Du fait de la nature continue de la fonction de densité du posterior des éléments, leur estimation ne peut être égale à 0. Dans ce cadre, nous utilisons le terme de pseudo-parcimonie pour distinguer cette structure de parcimonie plus stricte imposée à $\mathbf{W}$ par $\mathbf{Z}$. Dans le cadre de l'ICA, considèrer que des hyperparamètres du prior gamma du prior ARD (i.e. prior du vecteur des variances) égaux à 1, la combinaison définit un prior d'une distribution

super-gaussienne sur les éléments de $\mathbf{\Phi}$, favorisant la parcimonie des sources et donc leur indépendance. Ce prior est une généralisation du *ridge* Bayésien, qui considère un paramètre de variance identique pour tous les éléments de la même matrice (voir l'encadré 3.1).

Remarque 3.1 : Prior ARD et *ridge*

On peut définir un modèle gaussien multivarié pour expliquer des données $\mathbf{X}$ tel que :

$$\begin{aligned} p(\mathbf{X} \mid w, \tau) &= \mathcal{N}(\mathbf{X} \mid w, \tau^{-1}) \\ \tau &= \mathcal{G}(\tau \mid a_\tau, b_\tau) \end{aligned} \tag{3.6}$$

où w est le vecteur des paramètres de moyenne et τ est la précision de la distribution normale. Les paramètres w et τ en inférence Bayésienne sont considérés comme des variables où les distributions qui les définissent varient en fonction des objectifs. En grande dimension, il est classique de régulariser les paramètres du modèle afin de limiter le phénomène de sur-apprentissage. Un exemple classique est la régularisation l_2 du prior *ridge* qui est défini par

$$\begin{aligned} p(w \mid \lambda) &= \mathcal{N}(w \mid 0, \lambda^{-1}\mathbb{I}_P)\lambda \\ &= \mathcal{G}(\lambda \mid a_\lambda, b_\lambda), \end{aligned} \tag{3.7}$$

avec λ le paramètre de précision des paramètres suivant une distributions gamma (prior conjugué du paramètre de précision dans une distribution normale). On nomme aussi ce prior la distribution Normal-Gamma. On peut représenter la régularisation *ridge* comme en figure 3.5. Néanmoins, la contrainte d'une précision identique λ pour tous les paramètres implique d'attribuer la même régularisation à tous les paramètres. Afin de relâcher cette contrainte, il est possible d'utiliser le prior ARD qui est de la forme :

$$p(w \mid \lambda) = \mathcal{N}(w \mid 0, \Lambda^{-1}), \tag{3.8}$$

avec Λ la matrice des paramètres de précision définie par $\text{diag}(\Lambda) = \lambda = \{\lambda_1, \cdots, \lambda_P\}$ et

$$\lambda_p = \mathcal{G}(\lambda_p \mid a_{\lambda_p}, b_{\lambda_p}). \tag{3.9}$$

Ainsi, contrairement au prior *ridge*, chaque élément w_p de w possède sa propre précision λ_p.

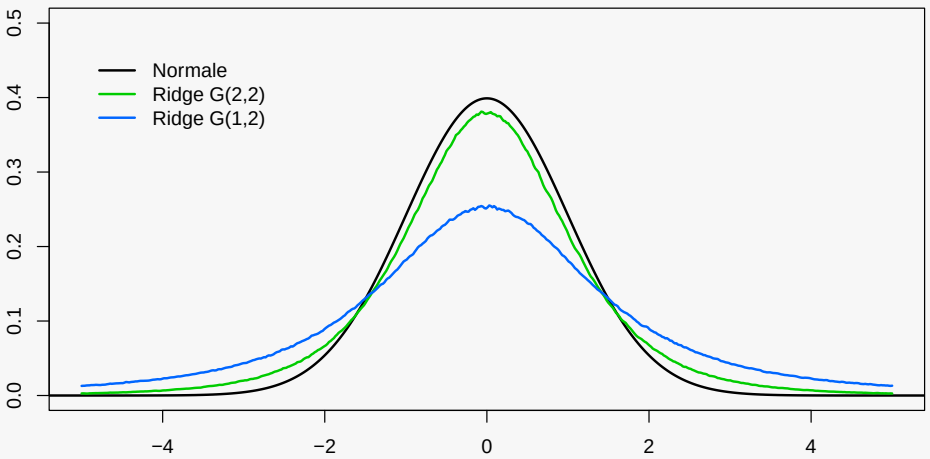


Figure 3.5 – Représentation du prior *ridge* par la fonction de densité en fonction des différentes régularisations imposées à la distribution gamma (G(forme,taux)) sur le paramètre τ.

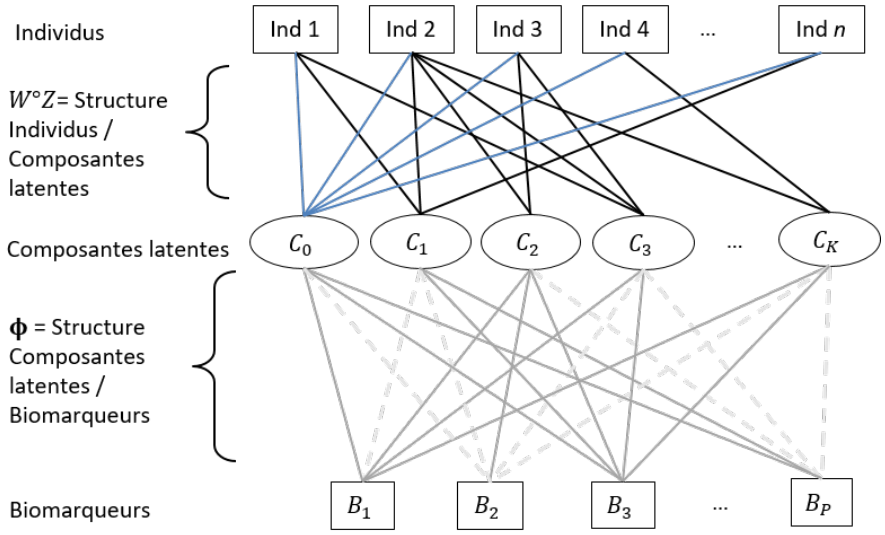


Figure 3.6 – ICA parcimonieuse avec un profil de base $\mathbf{Z_0}$ (liens bleus) et un ensemble d'altérations $\mathbf{Z}^*$ (liens noirs) dans le cade de la modélisation de l'expression génique. L'allocation d'une composante latente à un individu est représentée par la présence d'un « 1 » dans la matrice $\mathbf{Z}$. Dans cet exemple, les composantes latentes de la matrice des sources $\mathbf{\Phi}$ sont toutes associées à tous les biomarqueurs avec une régularisation appliquée aux éléments qui est représentée en traits pointillés où les éléments sont proches de 0 mais ne peuvent être égaux à 0 (pseudo-parcimonie).

Les méthodes de réduction de dimension ayant pour objectif de décomposer la variance des données peuvent intégrer, à certaine composante, le bruit aléatoire (ex : erreur de mesure) si celui-ci n'est pas (ou mal) pris en compte (Bouveyron, Latouche et Mattei (2020)). L'estimation des paramètres de précision du bruit joue donc un rôle essentiel pour éviter une sur-décomposition de la variance des données en un nombre trop important de composantes, et/ou une inflation des estimations des éléments des matrices, diminuant la parcimonie.

En effet, nos résultats empiriques issus d'une étude de simulation ont confirmé cette affirmation, mettant en évidence que le modèle s'adapte trop aux données, décomposant une partie de la variance du bruit en composantes supplémentaires non pertinentes (résultats non montrés). Pour limiter ce problème, nous avons normalisé les gènes de départ et fixé les précisions du bruit τ_E à 1, c'est-à-dire que la variance des expressions génétiques sont centrées et réduites. Cette contrainte régularise indirectement les paramètres des matrices des poids et des sources, ce qui permet d'identifier uniquement les signaux forts, mais peut conduire à des faux négatifs dans l'identification des relations

de l'hypergraphe (voir la section des résultats).

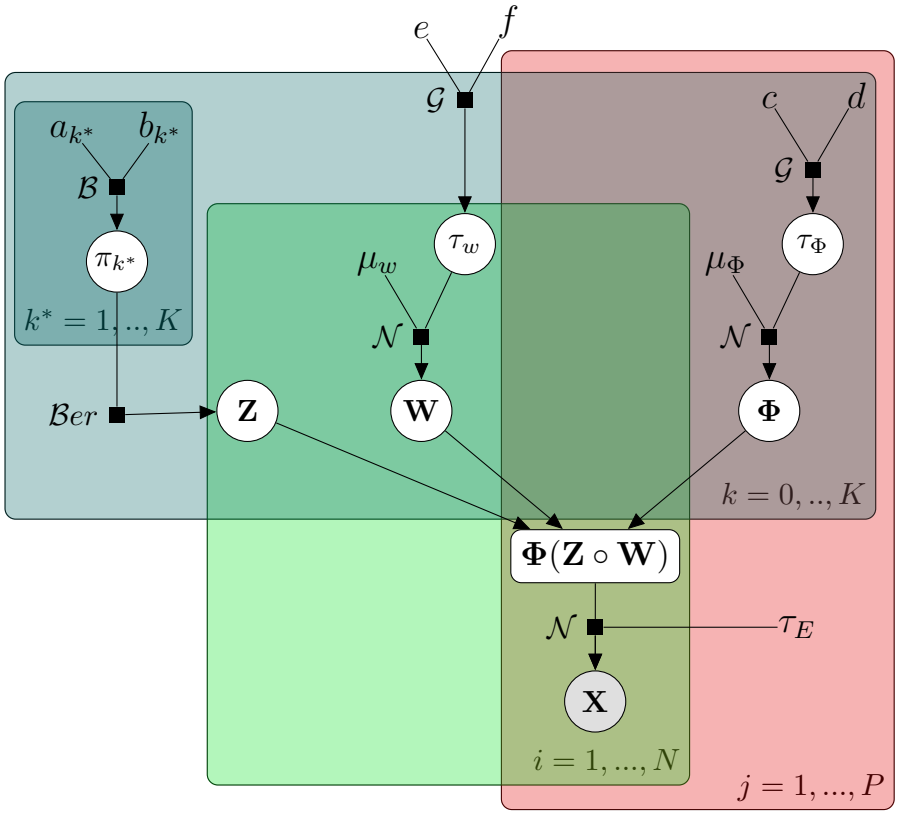


Figure 3.7 – Représentation graphique du modèle isgICA. Les variables observées sont désignées par des nœuds grisés, tandis que les variables non-observées sont représentées par des nœuds blancs et les hyperparamètres sont affichés sans nœuds. Abréviations : $\mathcal{N}$, distribution normale ; $\mathcal{B}$, distribution bêta ; $\mathcal{B}er$, distribution Bernoulli ; $\mathcal{G}$, distribution gamma.

L'identifiabilité du modèle isgICA

L'identifiabilité du modèle isgICA repose sur la non-normalité des composantes. La parcimonie stricte imposée par la matrice $\mathbf{Z}$ rend la matrice $\mathbf{Z} \circ \mathbf{W}$ non-normale et empêche le déconvolution d'effectuer une rotation, le modèle est donc invariant par rotation. La non-normalité est aussi renforcée par la pseudo-parcimonie imposée par le prior ARD sur la matrice $\mathbf{\Phi}$.

Notre algorithme basé sur l'ICA reste sujet aux problèmes de mise à l'échelle, de l'identifiabilité des signes et de l'échangeabilité des composantes. Cependant, nous pouvons étudier la contribution des gènes (et des individus) à chaque composante, qui est facilement obtenue en classant les valeurs des composantes

de la matrice des sources $\mathbf{\Phi}$ (et de la matrice des poids $\mathbf{W}$).

3.2.5 Mises à jour

Comme le calcul de la distribution a posteriori à l'aide de la méthode MCMC est connu pour sa lenteur lorsque le nombre de paramètres est élevé, nous avons utilisé l'inférence Bayésienne variationnelle (VI) sous l'hypothèse du champ moyen afin d'approximer la véritable distribution a posteriori (Beal (2003) ; Blei, Kucukelbir et McAuliffe (2017)).

Inspirés par B. Chen, M. Chen et al. (2010), nous avons utilisé l'algorithme CAVI pour minimiser la limite inférieure de l'évidence avec une approximation du champ moyen. Nous avons dérivé la limite inférieure de l'évidence variationnelle (ELBO) et le développement des équations variationnelles des paramètres est présenté dans l'annexe B. Les équations de mise à jour des paramètres variationnels sont décrites ci-dessous, avec $i = \{1, ..., N\}, j = \{1, ..., P\}, k = \{0, ..., K_{max}\}$, $k^*= \{1, ..., K_{max}\}$ et le symbole $< \bullet >$ définissant l'espérance du paramètre.

Mise à jour variationnelle du paramètre $<z^*_{k^*,i}>$:

$$\begin{aligned} q(z^*_{k^*,i}|-) &= Bernoulli(z^*_{k^*,i}; \pi_{k^*}) \\ &= \frac{q(z_{k^*,i} = 1|-)}{q(z_{k^*,i} = 1|-) + q(z_{k^*,i} = 0|-)}. \end{aligned} \tag{3.10}$$

On définit alors :

$$\begin{aligned} q(z^*_{k^*,i} = 1|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z_{-k^*,i}}, \boldsymbol{W}, \boldsymbol{\Phi}, \tau_E) \\ \propto \exp(&< \ln(\pi_{k^*}) > \quad - \frac{1}{2}(<\Phi_{k^*}>^T diag(\tau_E) <\Phi_{k^*}><w_{k^*,i}>^2 \\ &- 2<\Phi_{k^*}>^T diag(\tau_E)\mathbf{x}_i^{-k^*} <w_{k^*,i}>)) \\ &+ \delta_i(<\epsilon_i> - \log <\sigma_Y>) - exp(<\epsilon_i>)) \\ \text{with } \mathbf{x}_{j,i}^{-k^*} &= x_{j,i} - \sum_{l=0, l\neq k^*}^{K} <\Phi_{j,l}><z_{l,i}><w_{l,i}> \end{aligned} \tag{3.11}$$

et

$$\begin{aligned} q(z_{k^*,i} = 0 | \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z_{-k^*,i}}, \boldsymbol{W}, \boldsymbol{\Phi}, \tau_E) \\ \propto \exp(< \ln(1 - \pi_{k^*}) >). \end{aligned} \tag{3.12}$$

Mise à jour variationnelle du paramètre $<\pi_{k^*}>$:

$$\begin{aligned} q(\pi_{k^*}|-) &= \text{Beta}(\pi_{k^*,i}; a'_{k^*}, b'_{k^*}) \\ <a'_{k^*}> &= \sum_{i=1}^{N} <z_{k^*,i}> + \frac{a}{K} \\ <b'_{k^*}> &= N + \frac{b(K-1)}{K} - \sum_{i=1}^{N} <z_{k^*,i}>; \end{aligned} \tag{3.13}$$

Mise à jour variationnelle du paramètre $<\Phi_{j,k}>$:

$$\begin{aligned} q(\Phi_{j,k}|-) &= \mathcal{N}(\Phi_{j,k}; \mu_{\Phi_{j,k}}, \tau^{-1}_{\Phi_{j,k}}) \\ <\tau_{\Phi_{j,k}}> &= \sum_{i=1}^{N} \tau_{E_j} <w_{k,i}>^2 <z_{k,i}>^2 + <\tau_{\Phi_{j,k}}> \\ <\mu_{\Phi_{j,k}}> &= <\tau_{\Phi_{j,k}}>^{-1} \Big(\sum_{i=1}^{N} \tau_{E_j} <w_{k,i}><z_{k,i}> \mathbf{X}^{-k}_{j,i} \Big); \end{aligned} \tag{3.14}$$

Mise à jour variationnelle du paramètre $<W_i>$:

$$\begin{aligned} q(W_i|-) &= \mathcal{N}(W_i; \mu_{W_i}, \tau^{-1}_{W_i}) \\ <\tau_{W_i}> &= (<\Phi^T> \circ \widetilde{Z}_i) diag(\tau_E) (<\Phi> \circ \widetilde{Z}_i^T) + <\tau_W> I_K \\ <\mu_{W_i}> &= <\tau^{-1}_{W_i}> (<\Phi> \circ \widetilde{Z}_i) diag(\tau_E) x_i \\ \text{avec } \widetilde{Z}_i &:= [<z_i>, ..., <z_i>] \text{ , et } <z_i> \text{ le vecteur répété K fois ;} \end{aligned} \tag{3.15}$$

Mise à jour variationnelle du paramètre $<\tau_{\Phi_{j,k}}>$:

$$
\begin{aligned}
q(\tau_{\Phi_{j,k}}|-) &= \text{Gamma}(\tau_{\Phi_{j,k}}; c_{j,k}, d_{j,k}) \\
<c_{j,k}> &= c_0 + \frac{1}{2} \\
<d_{j,k}> &= d_0 + \frac{1}{2}<\Phi_{j,k}>^2;
\end{aligned}
\tag{3.16}
$$

Mise à jour variationnelle du paramètre $<\tau_W>$:

$$
\begin{aligned}
p(\tau_W|-) &= \text{Gamma}(\tau_W; e, f) \\
<e> &= e_0 + \frac{NK}{2} \\
<f> &= f_0 + \frac{1}{2}\sum_{i=1}^{N} <w_i^T><w_i>.
\end{aligned}
\tag{3.17}
$$

3.2.6 Optimisation des hyperparamètres

Nous avons utilisé le processus bêta tronqué pour l'inférence, avec un nombre maximal de composantes noté K_{max} (Paisley et Jordan (2016)). Pour toutes les simulations et analyses de données, nous avons considéré les valeurs d'hyperparamètres a priori : $K_{max} = 100, a = 1, c = d = 1, e = f = 10^{-6}$.

La distribution bêta n'ayant pas de loi conjuguée avec une forme fermée permettant de dériver simplement les mises à jour de ses paramètres dans l'algorithme CAVI, nous avons ajusté l'hyperparamètre b du BBP par optimisation Bayésienne à l'aide du module R ParBayesianOptimization (Samuel Wilson (2020)) en nous basant sur six évaluations d'initialisation et 24 époques (pour un total de 30 évaluations).

Le support de b étant ouvert, l'utilisation de l'optimisation Bayésienne nécessite de définir une borne supérieure de l'intervalle dans lequel chercher une solution optimale. Cette borne étant subjective, il est possible que la valeur optimale ne soit pas dans l'intervalle défini. Afin de s'affranchir de cela, nous nous sommes inspirés de la paramétrisation proposée par Ferrari et Cribari-Neto (2004) pour la régression bêta. En fixant $a = 1$, le paramètre de la moyenne de cette paramétrisation correspond à $\mu = 1/(1 + b)$, qui a un support sur (0 ;1) ($\mu = 0$ correspondant à $b = +\infty$ et $\mu = 1$ à $b = 0$)), sur lequel nous pouvons rechercher la valeur optimale sans restreindre le

support de b.

3.2.7 Comparaison à un modèle de référence : l'ICA avec blanchissement

Nous avons évalué la capacité de notre méthode à reconstruire les structures de parcimonie des matrices à partir de données simulées, en comparaison avec des algorithmes de référence. Nous avons utilisé le modèle ICA *whitening* (avec blanchissement) standard mis en œuvre dans les bibliothèques R *ica* (Helwig (2018)) et *fastICA* (Marchini, Heaton et Ripley (2019)). En raison de problèmes de calcul pour notre scénario le plus large (temps et/ou mémoire), nous avons présélectionné les composantes dont la valeur propre est supérieure à un (voir section 2.3.2.3). Le critère basé sur la valeur propre spécifie que seules les composantes dont la valeur propre est supérieure à 1 doivent être conservées puisque chaque composante doit expliquer la variance d'au moins une variable.

3.3 Résultats

3.3.1 Données synthétiques

3.3.1.1 Scénarios

Nous avons simulé des scénarios de données synthétiques à partir de l'équation (3.5), selon différents canevas (voir table 3.1) avec $N = \{100, 500\}$ individus, $P = \{500, 1000, 5000\}$ gènes, et $K = \{10, 30\}$ composantes. La structure de Z a été générée aléatoirement pour contenir environ 35% de « 1 ». Nous avons considéré quatre paramètres de bruit ($\sigma_E^2 = \{$ 0.5, 1, 1.5, 2 $\}$) pour évaluer l'impact du rapport signal/bruit.

Pour tous les scénarios, les éléments de la matrice des sources $\mathbf{\Phi}$ ont été générés pour chaque composante à partir de distributions normales avec une variance égale à un, et différentes moyennes (de -3 à 3) pour évaluer si le modèle peut identifier des composantes avec des caractéristiques spécifiques, telles que des valeurs majoritairement positives ou négatives, ou les deux (pour des moyennes proches de 0). Des blocs aléatoires ont été générés pour attribuer une structure de parcimonie à cette matrice

présentée par la figure 3.12. Les éléments de la matrice des poids W ont été tirés d'une distribution normale standard (moyenne = 0, variance = 1). Nous avons simulé 10 jeux de données pour chaque scénario.

Table 3.1 – Scénarios des simulations. Soit N, P, and σ_E^2 le nombre d'individus, le nombre de gènes et la variance du bruit.

	N=100	**N=500**
	10 composantes simulées	
P = 500	$\sigma_E^2 = \{0.5,\ 1,\ 1.5,\ 2\}$	$\sigma_E^2 = \{0.5,\ 1,\ 1.5,\ 2\}$
P = 1000	$\sigma_E^2 = \{0.5,\ 1,\ 1.5,\ 2\}$	$\sigma_E^2 = \{0.5,\ 1,\ 1.5,\ 2\}$
P = 5000	$\sigma_E^2 = \{0.5,\ 1,\ 1.5,\ 2\}$	$\sigma_E^2 = \{0.5,\ 1,\ 1.5,\ 2\}$
	30 composantes simulées	
P = 500	$\sigma_E^2 = \{0.5,\ 1,\ 1.5,\ 2\}$	$\sigma_E^2 = \{0.5,\ 1,\ 1.5,\ 2\}$
P = 1000	$\sigma_E^2 = \{0.5,\ 1,\ 1.5,\ 2\}$	$\sigma_E^2 = \{0.5,\ 1,\ 1.5,\ 2\}$
P = 5000	$\sigma_E^2 = \{0.5,\ 1,\ 1.5,\ 2\}$	$\sigma_E^2 = \{0.5,\ 1,\ 1.5,\ 2\}$

3.3.1.2 Critères de reconstruction

3.3.1.2.1 Corrélation absolue moyenne des composantes : étude de la pseudo-parcimonie

Comme l'ICA standard, notre modèle est identifiable sous réserve d'une mise à l'échelle, d'une inversion de signe et d'une permutation de colonnes (Sokol, Maathuis et Falkeborg (2014)). Pour évaluer la reconstruction du modèle, nous avons aligné les composantes estimées sur les composantes simulées en utilisant comme distance les coefficients de corrélation absolue de Pearson de chaque paire de la colonne de la matrice source simulée $\mathbf{\Phi}$ et des colonnes estimées non nulles. Nous avons utilisé l'algorithme hongrois (Kuhn (1955)) pour un réordonnancement efficace à partir de cette distance, en utilisant la fonction HungarianSolver de la bibliothèque RcppHungarian (Silverman (2022)).

Sachant que l'ordre des valeurs des éléments est conservé pour chaque composante de $\widehat{\Phi_{.,j}}$, nous pouvons estimer la corrélation absolue de Pearson entre la composante $\Phi_{.,i}$ et la composante reconstruite $\widehat{\Phi_{.,j}}$. Ce calcul est fait pour chaque paire de la colonne de la matrice source simulée $\mathbf{\Phi}$ avec les colonnes estimées non nulles de $\widehat{\mathbf{\Phi}}$. Nous supposons alors que la corrélation la plus importante correspond à la reconstruction de

la composante simulée avec la contrainte que chaque composante ne peut être alignée qu'une seule fois et nous ne conserverons que cette valeur.

Pour la composante i de la matrice simulée des sources $\mathbf{\Phi}$ et la composante j de la matrice estimée des sources $\widehat{\mathbf{\Phi}}$, cette distance est estimée par :

$$\left| \frac{\mathrm{cov}(\Phi_{\cdot,i}, \widehat{\Phi_{\cdot,j}})}{\sigma_{\Phi_{\cdot,i}} \sigma_{\widehat{\Phi_{\cdot,j}}}} \right|, \tag{3.18}$$

qui est invariant au changement d'échelle et à la réversion de signe.

Après ordonnancement, la moyenne des coefficients de corrélation absolue de Pearson est alors définie par la moyenne de ces corrélations :

$$\frac{1}{K} \sum_{k=1}^{K} \left| \frac{\mathrm{cov}(\Phi_{\cdot,k}, \widehat{\Phi_{\cdot,k}})}{\sigma_{\Phi_{\cdot,k}} \sigma_{\widehat{\Phi_{\cdot,k}}}} \right|, \tag{3.19}$$

La moyenne des coefficients de corrélation absolue de Pearson entre les matrices simulées $\mathbf{\Phi}$ et $\mathbf{Z} \circ \mathbf{W}$ et les matrices estimées $\widehat{\mathbf{\Phi}}$ et $\widehat{\mathbf{Z} \circ \mathbf{W}}$ ordonnés en colonne a été présentée pour évaluer la reconstruction de la matrice source et de la matrice poids respectivement. Pour la suite nous prendrons exemple avec la matrice $\mathbf{\Phi}$.

3.3.1.2.2 La précision (*Accuracy*) : étude de la parcimonie

Nous avons utilisé la précision (*Accuracy*) comme second critère pour étudier la reconstruction d'une matrice binaire. En supposant que la matrice binaire $\mathbf{Z}$ estimée est ordonnée de la même façon que la matrice $\mathbf{Z}$ simulée, la reconstruction de la structure de parcimonie a été évaluée à l'aide du critère de précision, défini par : $\frac{\text{Vrais uns} + \text{Vrais zéros}}{N \times K} \in [0, 1]$.

3.3.2 Reconstruction des structures latentes

Nous avons d'abord évalué la capacité des algorithmes à identifier le nombre de composantes (figure 3.8 et table 3.2). L'isgICA a retrouvé le nombre exact de composantes latentes dans la majorité des simulations, mais a sous-estimé le nombre dans les scénarios de plus faible dimension (P=500), en particulier lorsque le nombre d'observations était faible par rapport au nombre de composantes (N=100, K=30). Ce comportement était légèrement plus visible lorsque la variance du bruit était accrue.

Le nombre de composantes sélectionnées par l'ICA débruitée et la méthode des valeurs propres augmente rapidement avec la dimension (rapport P/N), jusqu'au nombre maximal de composantes permis par cette approche (minimum entre N-1 et P-1). Il augmentait aussi légèrement lorsque la variance du bruit augmentait dans notre algorithme.

Table 3.2 – Nombre de composantes identifiées (médiane [2.5-97.5% intervalle percentiles]) en utilisant la méthode des valeurs propres pour l'ICA standard et pour l'isgICA. N, P et σ_E^2 sont respectivement le nombre d'individus, le nombre de gènes et la variance du bruit.

Modèle	σ_E^2	N=100 P=500	N=100 P=1000	N=100 P=5000	N=500 P=500	N=500 P=1000	N=500 P=5000
		10 composantes simulées					
Méthode des valeurs propres	0.5	35 [19, 48]	66 [50, 79]	99 [99, 99]	28 [14, 43]	48 [29, 68]	412 [320, 457]
	1.0	54 [33, 66]	92 [82, 99]	99 [99, 99]	42 [20, 62]	82 [54, 106]	496 [470, 499]
	1.5	66 [46, 77]	99 [96, 99]	99 [99, 99]	56 [26, 76]	114 [79, 140]	499 [499, 499]
	2.0	74 [56, 84]	99 [99, 99]	99 [99, 99]	67 [33, 88]	139 [102, 167]	499 [499, 499]
		30 composantes simulées					
	0.5	30 [27, 30]	31 [30, 34]	99 [99, 99]	30 [29, 32]	32 [30, 33]	48 [38, 95]
	1.0	31 [28, 32]	38 [32, 49]	99 [99, 99]	30 [29, 34]	32 [30, 38]	196 [167, 243]
	1.5	32 [30, 34]	56 [43, 70]	99 [99, 99]	31 [30, 35]	33 [31, 42]	338 [313, 371]
	2.0	36 [33, 39]	72 [60, 85]	99 [99, 99]	32 [30, 35]	38 [32, 48]	424 [404, 447]
		10 composantes simulées					
Méthode isgICA	0.5	10 [8, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]
	1.0	10 [8, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]
	1.5	10 [7, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]
	2.0	9 [6, 10]	10 [9, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]
		30 composantes simulées					
	0.5	11 [8, 14]	16 [13, 17]	20 [19, 22]	25 [20, 26]	29 [27, 30]	30 [29, 30]
	1.0	11 [9, 13]	16 [13, 16]	20 [19, 22]	22 [19, 24]	28 [27, 30]	30 [29, 30]
	1.5	11 [9, 13]	15 [13, 16]	20 [19, 22]	22 [18, 24]	28 [27, 30]	30 [29, 30]
	2.0	11 [9, 13]	15 [13, 16]	20 [18, 22]	20 [16, 23]	28 [27, 30]	30 [29, 30]

En raison de cette sur-décomposition par la méthode des valeurs propres dans la majorité des scénarios, le calcul des critères de reconstruction (la moyenne des corrélations absolues des composantes) pour *ica* et *fastICA* n'a pas été possible. Dans ce cas, nous avons sélectionné les 10 composantes estimées (ou 30, selon le scénario) pour lesquelles les composantes dans la matrice des sources étaient les plus corrélées aux composantes simulées. Nous avons également effectué une analyse de sensibilité de

type oracle, en fixant a priori le nombre de composantes aux composantes simulées pour ces deux approches, afin de pouvoir comparer leur reconstruction à la reconstruction aveugle d'isgICA (voir les résultats en figures 3.9 et 3.10).

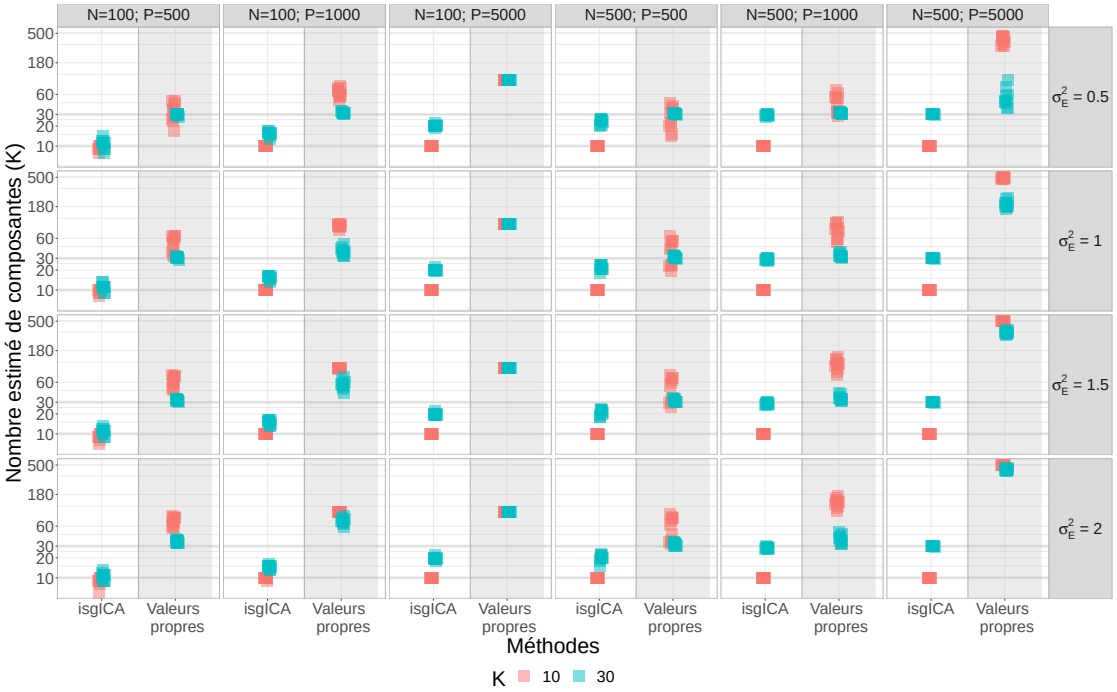


Figure 3.8 – Reconstruction du nombre de composantes latentes entre isgICA et l'ICA standard en utilisant le critère des valeurs propres avec 10 (rouge) ou 30 (bleu) composantes simulées (K) . Chaque ligne correspond aux scénarios avec différentes variances de bruit (σ_E^2) et chaque colonne correspond à différentes dimensions (N individus et P gènes), pour 10 simulations dans chaque scénario.

isgICA a surpassé *ica* pour la reconstruction de la matrice des sources ($\mathbf{\Phi}$) lorsque le nombre de composantes a été estimé par la méthode des valeurs propres (figure 3.10). Cette différence était moins nette pour *fastICA* (corrélation avec la matrice des sources simulée égale à 0.818 [0.191, 0.934] (médiane [2.5-97.5% percentiles]) pour l'isgICA, 0.767 [0.513, 0.982] pour la *fastICA*, et 0.189 [0.028, 0.834] pour l'*ica*), notamment en raison de la performance inférieure de l'isgICA dans les scénarios avec les dimensions les plus faibles (N=100/P=500, et N=500/P=500). La performance de toutes les méthodes a diminué avec l'augmentation de la variance du bruit, mais l'isgICA a été moins affectée pour les scénarios à haute dimension. De plus, isgICA présentait des performances équivalentes à celles des modèles oracle *fastICA* et *ica*, à l'exception des scénarios où isgICA a été moins performant, c'est-à-dire dans le cas de

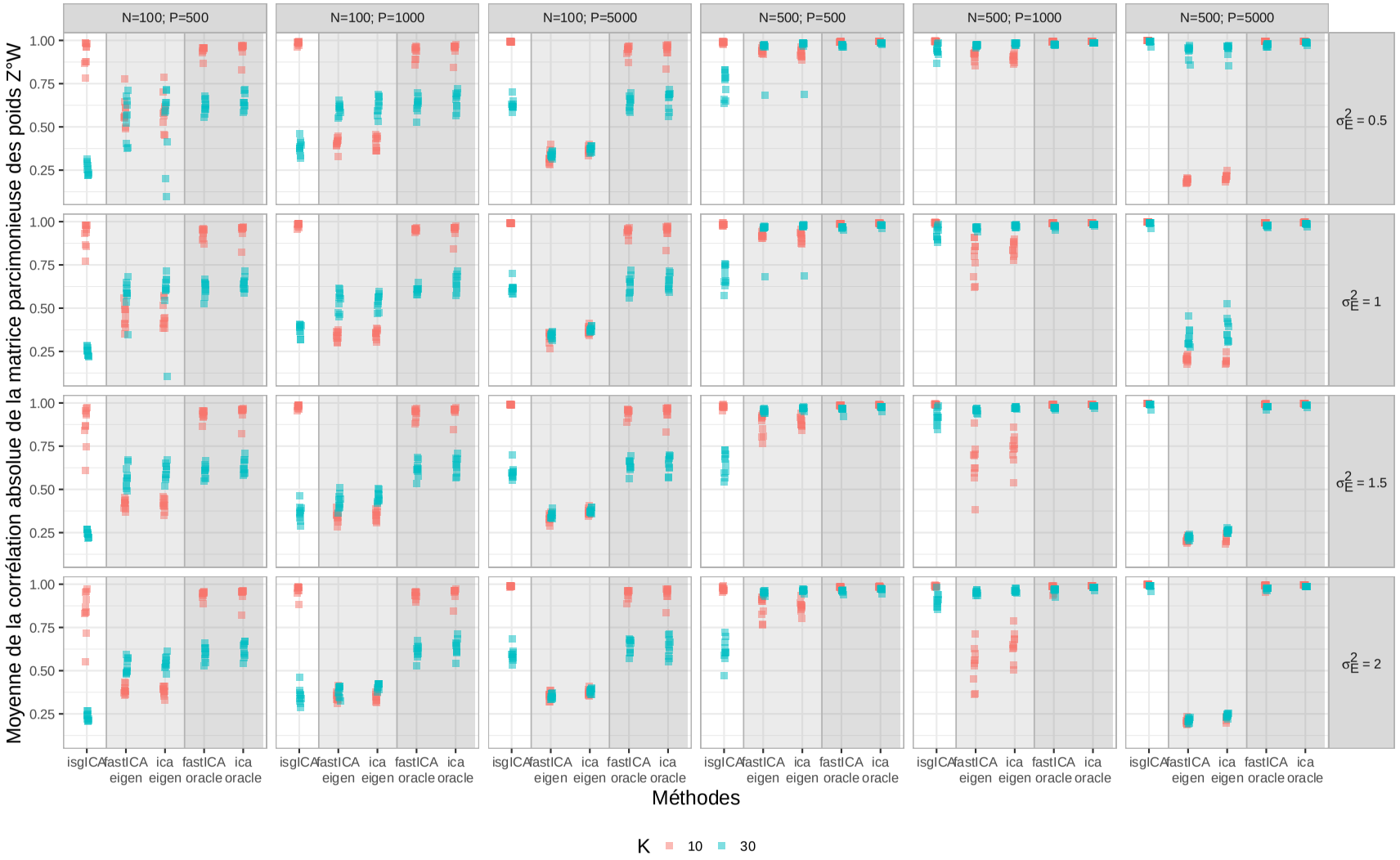


Figure 3.9 – Performance de la reconstruction de la matrice parcimonieuse des poids $Z \circ W$ des modèles isgICA, *ica*, et *fastICA*, pour les différents scénarios avec 10 (rouge) ou 30 (bleu) composantes (K) simulées, pour différentes variances du bruit (σ^2_E) en lignes, et différentes dimensions (N individus et P gènes) en colonnes, pour 10 simulations dans chaque scénario.

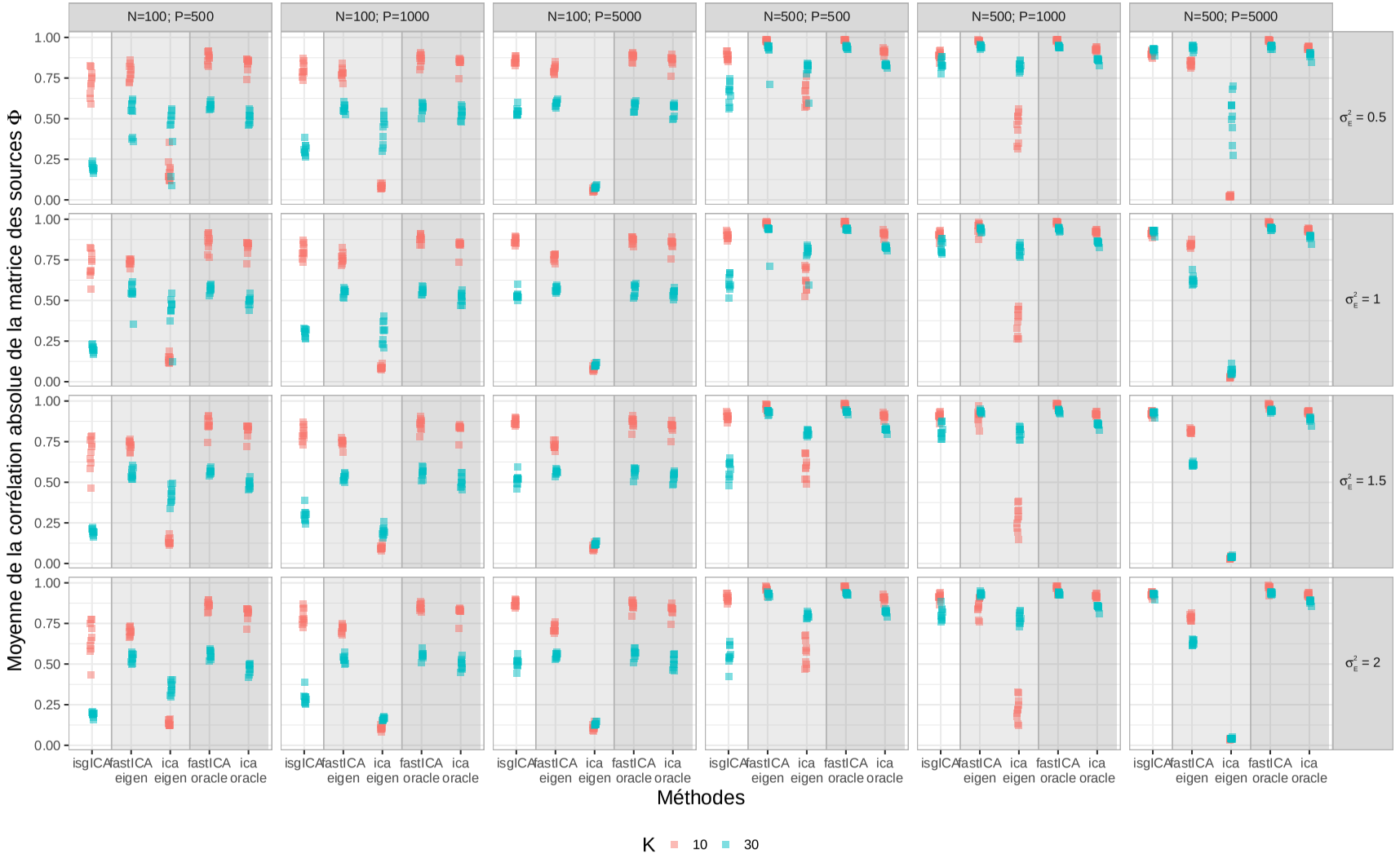


Figure 3.10 – Performance de la reconstruction de la pseudo-parcimonie de la matrice des sources $\mathbf{\Phi}$ des isgICA, ica et fastICA, pour les différents scénarios avec 10 (rouge) ou 30 (bleu) composantes simulées (K), pour différentes variances du bruit (σ^2_E) en lignes, et différentes dimensions (N individus et P gènes) en colonnes, pour 10 simulations dans chaque scénario.

la faible dimension.

Pour tous les scénarios, l'isgICA a surpassé les autres méthodes pour la reconstruction des matrices parcimonieuses des poids $\mathbf{Z} \circ \mathbf{W}$ (figure 3.9) (corrélations absolues moyennes de Pearson égales à 0.968 [0.225, 0. 999] (médiane [2.5-97.5% percentiles]) pour isgICA, 0.453 [0.194, 0.972] pour *fastICA*, 0.515 [0.199, 0.983] pour *ica*), sauf pour les scénarios avec N=100/500, P=500 et K=30. Ce résultat s'explique par la surdécomposition à l'aide de la méthode des valeurs propres. En considérant les modèles oracle *fastICA* et *ica*, les performances de reconstruction étaient similaires pour toutes les méthodes.

Les résultats d'isgICA sont résumés dans la figure 3.11. La capacité d'isgICA à reconstruire la structure de parcimonie de la matrice des poids ($\mathbf{Z}$) était précise dans la majorité des scénarios (précision > 0.8), mais diminuait lorsque le nombre de paramètres augmentait (c'est-à-dire en augmentant P ou K) pour atteindre les 35% simulés de « 1 » dans les composantes non zéros, ce qui correspond à la précision des matrices $\mathbf{Z}$ (toutes d'uns). Cependant, la bonne reconstruction de la matrice des poids parcimonieuse ($\mathbf{Z} \circ \mathbf{W}$) dans les scénarios de haute dimension indique que cette diminution de la précision pour le caractère parcimonieux strict est contrebalancée par le caractère pseudo-parcimonieux induit par le prior ARD sur $\mathbf{W}$.

Pour illustrer la capacité de reconstruction du modèle isgICA, la figure 3.12 illustre la reconstruction de la matrice parcimonieuse des poids ($\mathbf{Z} \circ \mathbf{W}$) et de la matrice des sources ($\mathbf{\Phi}$) pour N=500, P=5000, K=10 et σ_E^2=0.5. Dans cet exemple, notre approche a pu identifier le nombre exact de composantes non-nulles et reconstruire la précision de la matrice des poids parcimonieux ($\mathbf{Z}$) avec une précision de 0.982, la matrice parcimonieuse des poids ($\mathbf{Z} \circ \mathbf{W}$) avec une corrélation absolue moyenne de 0.999 et la matrice des sources ($\mathbf{\Phi}$) avec une corrélation absolue moyenne de 0.873. Comme l'illustre la troisième ligne de la figure 3.12, l'algorithme basé sur l'ICA souffre des problèmes d'identifiabilité typiques de l'ICA pour la matrice des sources : le signe des éléments de certaines composantes peut être inversé par rapport à la matrice simulée, et ils peuvent présenter des valeurs plus élevées (problèmes d'identifiabilité de l'échelle et du signe). Cependant, le rang des valeurs simulées avec les valeurs estimées est fortement corrélé, ce qui permet d'interpréter les valeurs les plus élevées de la matrice des sources comme les gènes contribuant le plus aux composantes.

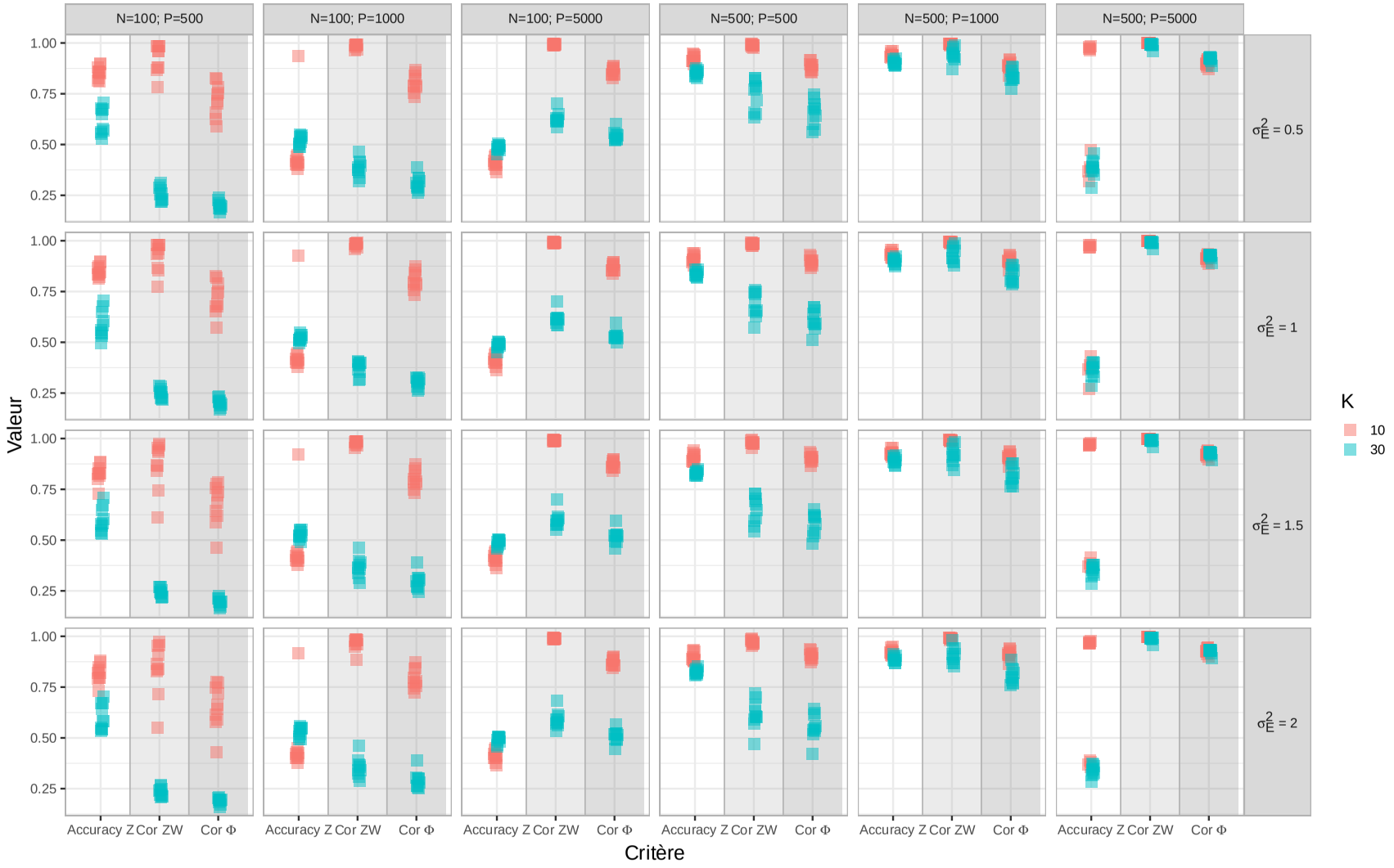


Figure 3.11 – Reconstruction de la structure parcimonieuse des poids ($\mathbf{Z}$), de la matrice des poids parcimonieuse ($\mathbf{Z} \circ \mathbf{W}$) et de la matrice des sources ($\mathbf{\Phi}$) en fonction de différentes variances du bruit en lignes avec 10 (rouge) et 30 (bleu) composantes simulées (K) et différentes dimensions (N individus et P gènes) en colonnes : la précision de la reconstruction de Z (premier critère), la corrélation absolue moyenne de $\mathbf{Z} \circ \mathbf{W}$ (deuxième critère) et la corrélation absolue moyenne de $\mathbf{\Phi}$ (troisième critère), pour 10 simulations dans chaque scénario.

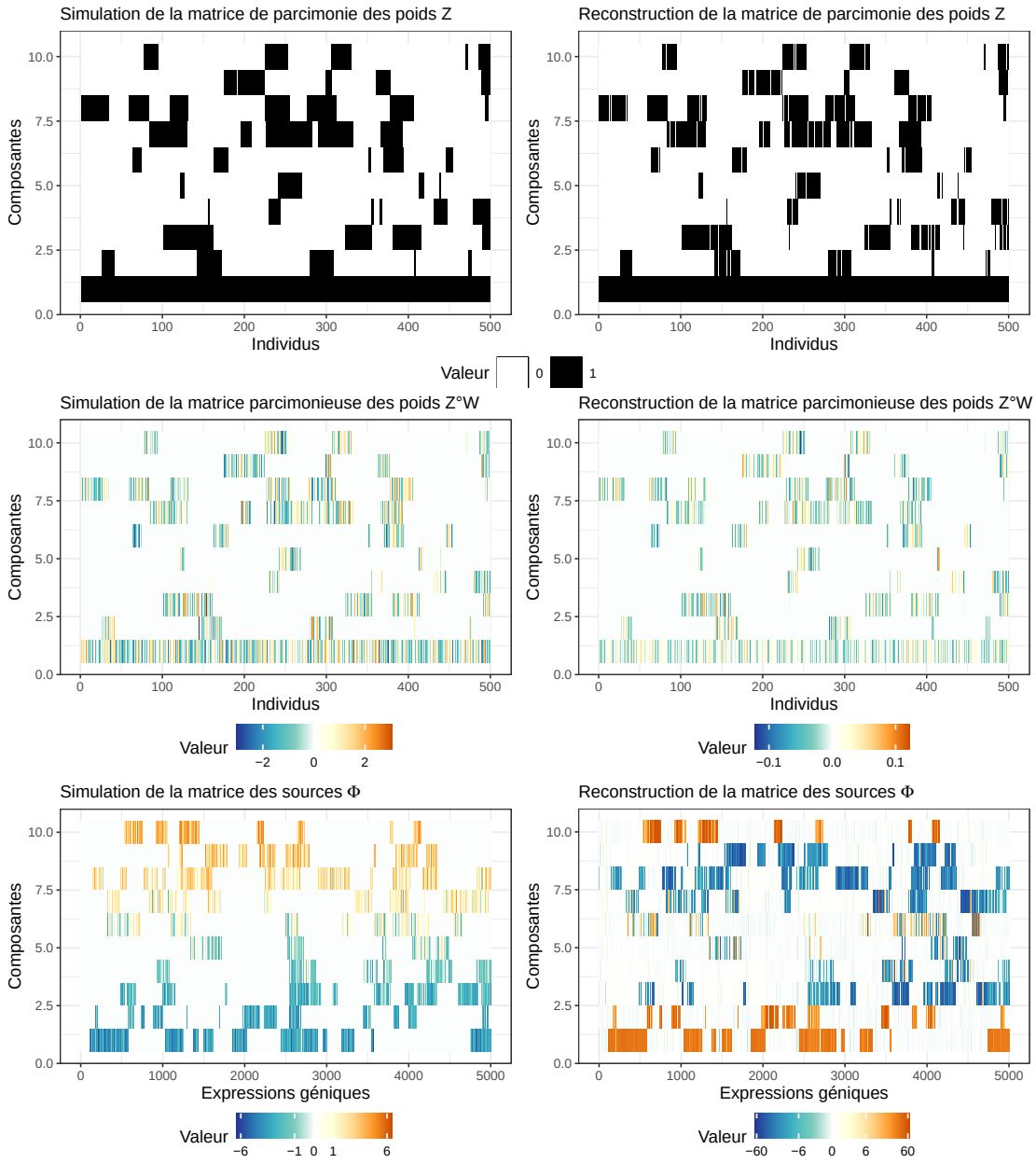


Figure 3.12 – Représentation graphique de la reconstruction de la structure de poids parcimonieux ($\mathbf{Z}$, ligne supérieure, précision de 0.982), de la structure de poids parcimonieux ($\mathbf{W} \circ \mathbf{Z}$, ligne du milieu, corrélation de 0.999), et des sources ($\mathbf{\Phi}$, ligne inférieure, corrélation de 0.873) à partir d'une structure diagonale et avec N=500, P=5000, K=10 et $\sigma_E^2) = 0.5$.

On peut noter que la méthode *ica* n'a pas convergé pour 2.9% (14/480) des simulations, concernant le scénario à haute dimension (N=500, P=5000, K=10) en raison du grand nombre de composantes identifiées par la méthode des valeurs propres. Par conséquent, le résultat de cette méthode est légèrement trop optimiste et doit être interprété en conséquence (y compris dans le contexte d'une application réelle). Le gain de précision pour la reconstruction matricielle (en aveugle) et le taux de convergence d'isgICA par rapport aux autres méthodes ont été obtenus au prix d'un temps de calcul plus important par rapport à la méthode des valeurs propres (de quelques minutes à plusieurs heures), qui augmente rapidement avec la dimension (figure 3.13).

3.4 Application : données sur le cancer du sein au stade précoce

Nous avons appliqué notre méthode à des données d'expression génique disponibles publiquement, obtenues à partir de biopsies tumorales chez 614 patientes atteintes d'un cancer du sein au stade précoce et incluses dans des essais cliniques de chimiothérapie à base d'anthracycline (Desmedt et al. (2011) ; Hatzis et al. (2011)), disponibles dans le package R de biospear Ternès, Rotolo et Michiels (2018). Les données d'expression de 22 277 *probes* (puce Affymetrix) ont été prétraitées par l'algorithme *Frozen robust multiarray analysis* (McCall, Bolstad et Irizarry (2010)) et par une normalisation multiplateforme (Shabalin et al. (2008)). Les *probes* ont été filtrées si l'écart interquartile ≤ 1. Les 1 689 *probes* restantes ont été normalisées puis filtrées avec le package jetset (Li et al. (2011)) pour conserver une seule *probe* par gène, ce qui a donné un jeu de données final comprenant l'expression de 1 063 gènes. De façon similaire à Belhechmi et al. (2020), nous avons associé les *probes* à trois signatures moléculaires étant connues pour être associées au pronostic du cancer du sein au stade précoce (Système immunitaire, Prolifération et invasion du stroma Ignatiadis et al. (2012)), mais également à une signature non-associée afin d'avoir un contrôle négatif (signature d'activation du SRC Ignatiadis et al. (2012)). Toutes les autres *probes* ont été classées dans la catégorie « Autres ».

La figure 3.14 présente la matrice de l'hypergraphe de l'hétérogénéité individuelle.

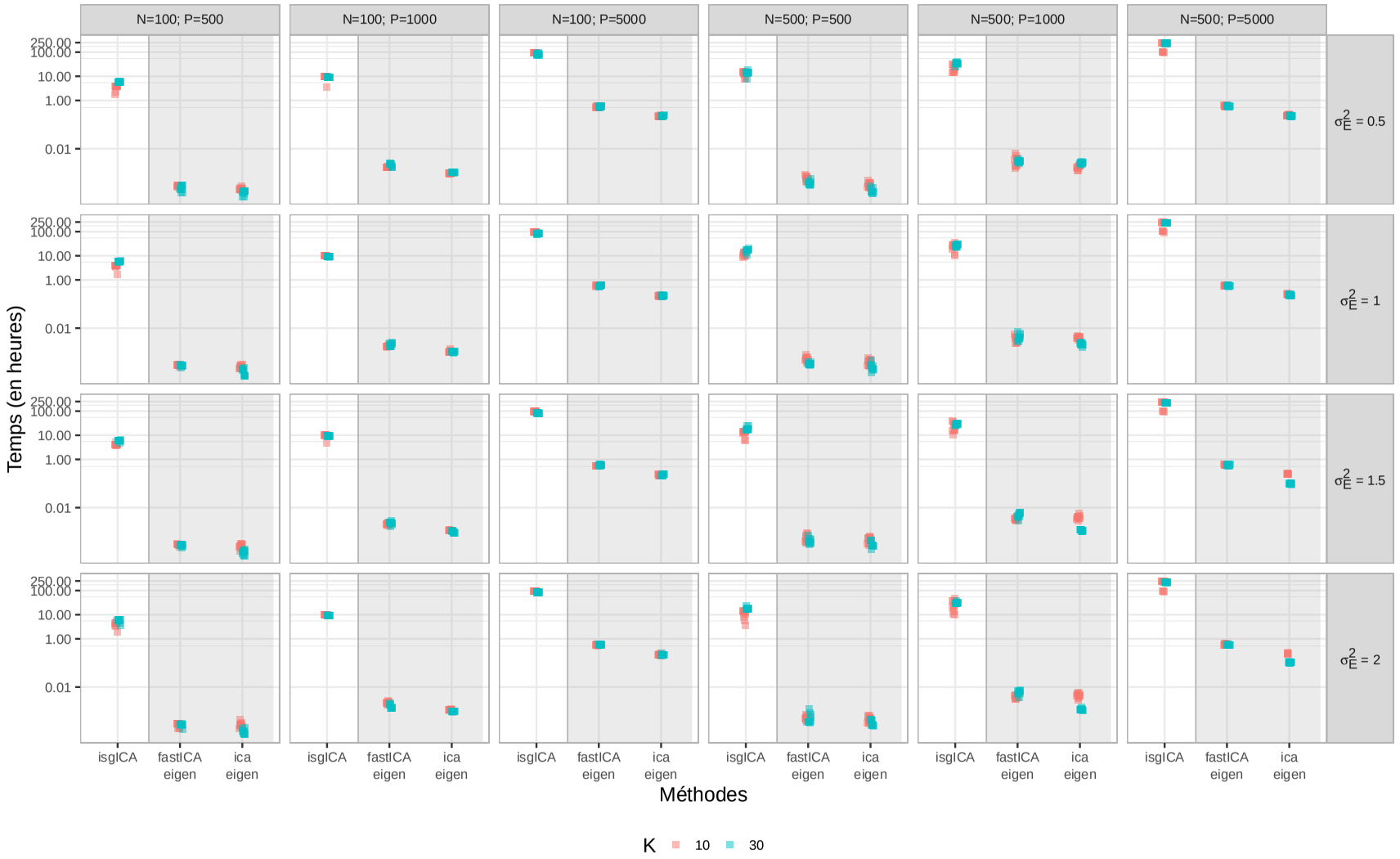


Figure 3.13 – Temps de calcul de l'ica, fastICA et isgICA pour les différents scénarios avec 10 (rouge) ou 30 (bleu) composantes simulées (K), différentes variances du bruit (σ^2_E)) en lignes, et différentes dimensions (N individus et P gènes) en colonnes, pour 10 simulations dans chaque scénario.

Le modèle a identifié 22 composantes non nulles (y compris le profil de base).

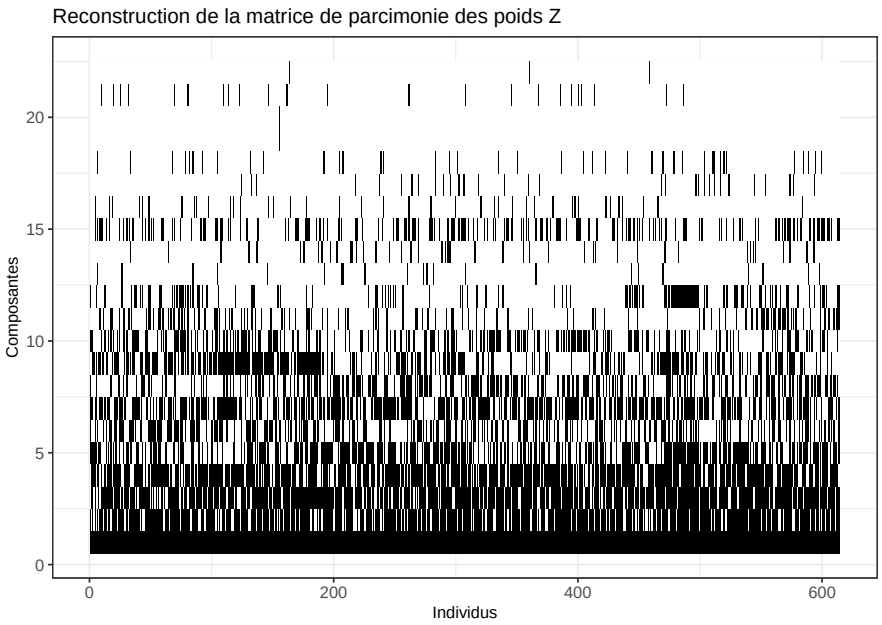


Figure 3.14 – Matrice de l'hypergraphe de l'hétérogénéité individuelle (structure de dispersion des poids) extraite de l'ensemble de données sur le cancer du sein à l'aide du modèle d'analyse graphique infiniment dispersée en composantes indépendantes avec profil de base. Le modèle a identifié 22 composantes (y compris le profil de base).

Pour étudier la pertinence moléculaire de ces résultats, nous avons contourné les problèmes d'inversion de signe et de changement d'échelle en classant les valeurs absolues des éléments de chaque composante de la matrice des sources pour identifier les gènes qui contribuent le plus à chaque altération moléculaire identifiée. La figure 3.15 montre la distribution des valeurs absolues des éléments de la matrice des sources de chaque composante en fonction des différentes signatures du cancer du sein. Les signatures basées sur la prolifération, sur le système immunitaire et sur le stroma semblent respectivement être liées aux composantes 2/5, 4/7/8/15 et 3. La signature SRC, qui a été choisie comme signature de « contrôle négatif » dans Belhechmi et al. (Belhechmi et al. (2020)), ne semblait pas être associée à une composante moléculaire particulière.

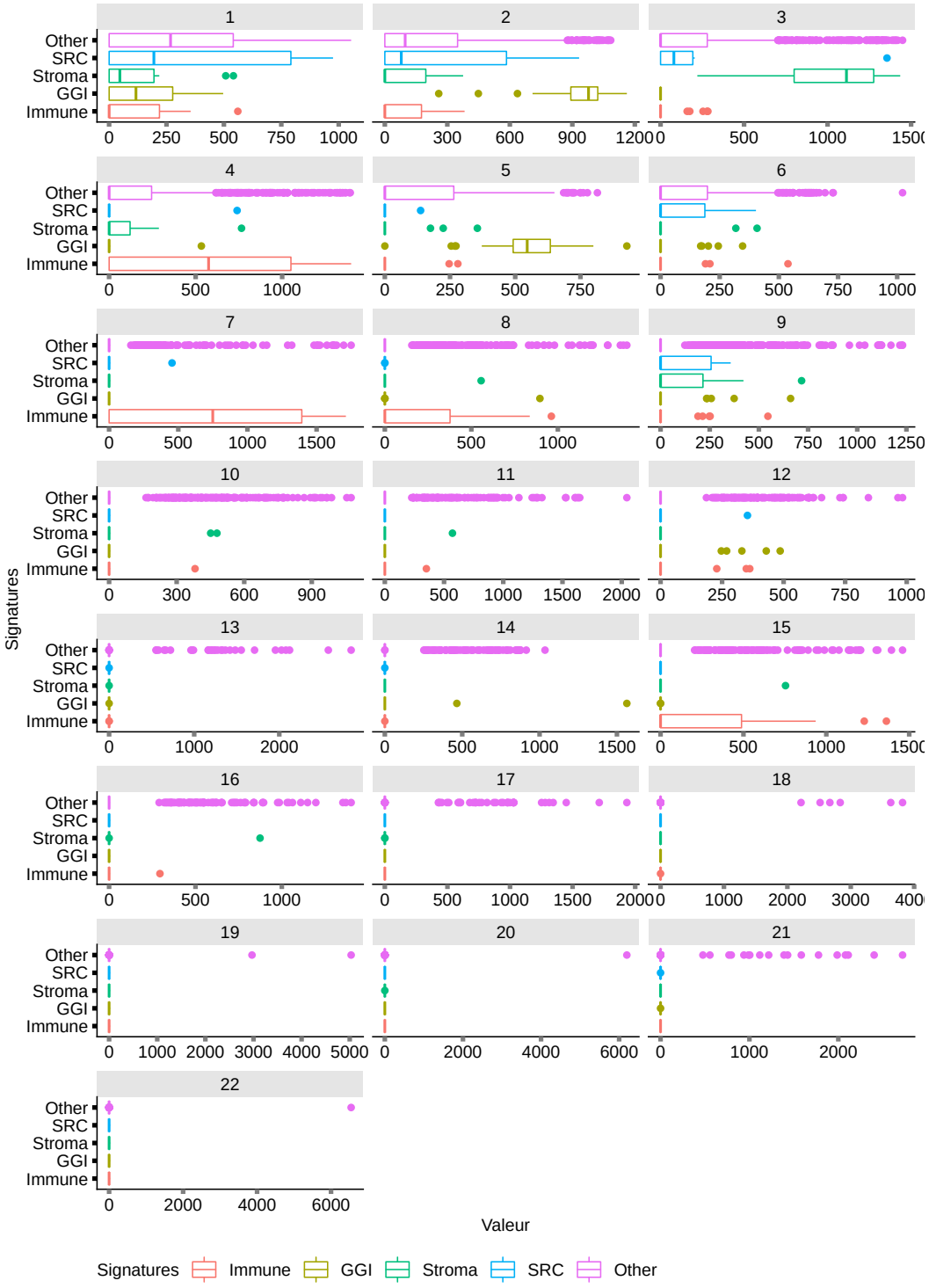


Figure 3.15 – Distribution de la valeur absolue des éléments de la matrice des sources associés aux *probes* des gènes associés à trois signatures moléculaires ayant un effet pronostique connu dans le cancer du sein au stade précoce (GGI, Immune2, et Stroma2) et une sans (signature d'activation SRC). Tous les autres gènes ont été classés dans la catégorie « Autres ». Les signatures liées à la prolifération, à l'immunité et au stroma semblaient être liées respectivement aux composantes 2/5, 4/7/8/15 et 3. La signature SRC n'était pas liée à une composante spécifique.

3.5 Discussion

Dans cet article, nous avons proposé une nouvelle approche pour caractériser les mécanismes moléculaires hétérogènes entre les patients. À notre connaissance, il s'agit de la première approche qui suppose que le profil moléculaire de chaque patient est un mélange de différentes composantes moléculaires, qui peuvent être partagées avec les autres patients. Nous avons modélisé ces composantes comme des altérations d'un profil moléculaire de base partagé par tous les individus, représentant les mécanismes communs à tous les patients, tandis que le paramètre d'erreur capture le bruit de fond moléculaire individuel. En supposant que chaque composante moléculaire représente les altérations d'une voie moléculaire ou d'un groupe de voies connexes, cette approche peut nous aider à comprendre les mécanismes moléculaires et à identifier des cibles potentielles pour le développement de médicaments.

Nous avons illustré ce concept à l'aide d'un ensemble de données sur l'expression génique d'échantillons de tumeurs du cancer du sein provenant de patientes incluses dans des essais cliniques de chimiothérapie à base d'anthracycline. La correspondance des profils moléculaires identifiés avec les voies moléculaires connues jouant un rôle pronostique dans le cancer du sein au stade précoce (prolifération, système immunitaire et voie du stroma) suggère que notre approche peut aider à caractériser le contexte moléculaire de sous-populations particulières.

Dans l'étude de simulation, notre méthode a été capable d'identifier en aveugle le nombre de composantes générées, alors que l'approche classique basée sur les valeurs propres tendait à inclure un plus grand nombre de composantes. De plus, la reconstruction des structures latentes par isgICA était similaire à celle des méthodes classiques lorsque l'on fixe a priori le nombre de composantes au nombre simulé, sous réserve des problèmes d'identifiabilité connus de l'ICA classique de la mise à l'échelle et de la réversion de signe, qui sont des problèmes courants d'identifiabilité. De plus, isgICA reconstruisait la matrice des poids et des sources avec une meilleure précision que les méthodes classiques dans les scénarios avec un faible effectif. Lorsque la performance était similaire, la parcimonie stricte imposée à la matrice des poids ($\mathbf{Z} \circ \mathbf{W}$) possède l'avantage de l'interprétation de Z comme la matrice de l'hypergraphe représentant l'hétérogénéité individuelle. Notre algorithme était également moins sensible à l'aug-

mentation de la dimension et de la variance du bruit.

Cependant ses performances inférieures dans les scénarios de dimension les plus faibles suggèrent que la régularisation imposé par la variance du bruit est probablement trop importante. En raison de la sous-estimation bien connue de la variance du bruit dans les modèles parcimonieux de grande dimension (Moran, Rockova et George (2018)), nous avons décidé de la fixer à 1 (c'est-à-dire la variance des données $\mathbf{X}$ après normalisation et standardisation) pour éviter la sur-décomposition de la variance qui entraîne un excès de composantes (résultat non-montré). Ce choix a un impact sur les estimations car la variance du bruit est théoriquement inférieure à 1 si la variance de $\mathbf{X}$ est expliquée par certaines composantes. Cela induit un rapport signal/bruit trop faible entraînant des moyennes a posteriori des éléments des composantes proches de zéro (Moran, Rockova et George (2018)) (du fait d'une sur-régularisation), et donc une sous-estimation du nombre de composantes. L'augmentation de la dimension entraîne une augmentation de la moyenne a posteriori des paramètres qui échappent à la régularisation *spike* (comme le reflète le maximum a posteriori de $\mathbf{Z}$, correspondant à une matrice de 1). Cependant, le modèle isgICA présente de bonnes performances dans ces contextes, car ces paramètres ont été régularisés par le motif *slab* (prior ARD), résultant en une pseudo-parcimonie observée comme pour les approches classiques (*ica* et *fastICA*). Dans des recherches ultérieures, l'utilisation de priors indépendants pourrait être explorée pour atténuer la sous-estimation de la variance du bruit (Moran, Rockova et George (2018)). Mais cette approche basée sur des priors non conjugués nécessite des algorithmes complexes pour l'inférence, ce qui pourrait augmenter le temps de calcul pour un ensemble de données de grande dimension. Comme l'illustre notre benchmark, les performances de notre algorithme ont été obtenues au prix d'un temps de calcul important qui pourrait constituer une limitation pratique. Une première étape consistera à réimplémenter l'algorithme actuel avec des calculs sur GPU pour l'adapter à de grands ensembles de données.

Nous avons montré que cette approche est capable d'identifier en aveugle les composantes qui dévient d'un profil de base. Les recherches futures se concentreront sur l'amélioration de leur identifiabilité et de leur interprétabilité, y compris par l'intégration d'informations externes supplémentaires. Nous nous attendons à ce que les poids de certaines composantes reconstruisent certaines des variables individuelles observées

non prises en compte par le modèle. Il pourrait être possible d'étendre ce modèle, en fixant les éléments de certaines composantes de poids aux valeurs des variables individuelles observées, qui peuvent être pertinentes pour expliquer les expressions géniques (par exemple, le sexe, l'âge, le stade de la tumeur). Cette extension pourrait être utilisée pour réaliser une analyse différentielle (ex : impact d'un traitement). De plus, alors que les données de séquençage en masse résultent d'un mélange de plusieurs éléments (non seulement des cellules tumorales, mais aussi de cellules provenant de tissus sains (Petralia et al. (2018)) ou du microenvironnement tumoral (B. Chen, Khodadoust et al. (2018))), d'autres sources de données telles que des profils moléculaires de référence pourraient être utilisées pour améliorer l'identifiabilité et l'interprétabilité des composantes. Une autre façon intéressante d'intégrer des informations externes est de considérer les caractéristiques du patient comme une variable explicative de la structure de parcimonie de la matrice des sources pour modéliser différents états du graphe, comme suggéré par Z. Wang et al. (2022). La modélisation conjointe de notre modèle isgICA et d'un résultat clinique, tel que la survie du patient, peut présenter un intérêt particulier pour la médecine de précision, favorisant l'identification de profils moléculaires indépendants plus spécifiques au pronostic du patient. Cette extension permettra d'estimer l'interaction composante/traitement dans le modèle de survie afin de mettre en évidence les voies liées à la réponse au traitement dans le contexte de la médecine de précision.

Enfin, comme le propose la vaste littérature sur les méthodes de déconvolution des données omiques, notre approche peut également être étendue à d'autres données omiques non-normales, telles que les données de comptage (RNAseq brut, protéomique) ou binaires (mutations) en utilisant différentes fonctions de liaison en s'inspirant des méthodes parcimonieuses proposées par Ray et al. (2014) et Argelaguet et al. (2018).

3.6 Conclusion

Nous avons développé un modèle isgICA muni d'un profil de base pour caractériser en aveugle l'hétérogénéité individuelle à partir d'un profil de base dans un cadre de grande dimension. Cette approche illustre un nouveau concept pour l'identification de profils moléculaires individuels qui pourraient être la clé pour comprendre les diffé-

rents mécanismes de la maladie et identifier des cibles potentielles pour développer de nouveaux traitements.

Chapitre 4

Identification de mécanismes moléculaires individuels pronostiques

Modèle Bayésien conjoint pour l'analyse de profils moléculaires individuels et de survie en oncologie

Article en révision :
RINCOURT, Sarah-Laure, MICHIELS, Stefan, DRUBAY Damien. "A non-parametric Bayesian joint model for latent individual molecular profiles and survival in oncology". In : *Journal of Bioinformatics and Computational Biology*. 2022. DOI : 10.1142/S0219720022500226.

Chapitre adapté de l'article

Le développement de signatures moléculaires pronostiques tenant compte de l'hétérogénéité inter-patients est un défi majeur pour la médecine de précision. En se basant sur l'algorithme isgICA présenté dans le chapitre précédent, nous avons proposé un modèle pronostique basé sur la survie de patients considérant la présence/absence des altérations moléculaires comme facteurs de risque. Pour cela, nous avons développé un modèle permettant de modéliser conjointement des données de survie et la structure latente sous-jacente de données omiques, ces deux parties partageant la structure d'hétérogénéité individuelle latente $\mathbf{Z}$.

Sommaire

4.1 Introduction **83**

4.2 Travaux connexes précédents **85**

4.3 Méthode **86**

4.3.1 Modèle joint SisgICA 86

4.3.2 Sous-modèle isgICA 87

4.3.3 Sous-modèle paramétrique de survie : Régression paramétrique de Weibull sous l'hypothèse AFT 88

4.3.3.1 Introduction à la survie 88

4.3.3.2 Notions 88

4.3.3.3 L'utilisation du modèle de Weibull avec l'hypothèse AFT 93

4.3.4 Modèle complet Bayésien 95

4.3.5 Équations variationnelles SisgICA 97

4.3.6 Optimisation des hyperparamètres 99

4.4 Résultats **100**

4.4.1 Données synthétiques 100

4.4.1.1 Scénarios des simulations 100

4.4.1.2 Critères de reconstruction 101

4.4.1.3 Reconstruction de la structure latente et sélection des paramètres de survie 102

4.4.2 Application : early breast cancer data 107

4.5 Discussion **111**

4.6 Conclusion **113**

4.1 Introduction

L'identification de biomarqueurs moléculaires associés au pronostic d'un patient est essentielle en médecine personnalisée. Cette médecine de précision est aujourd'hui basée sur des biomarqueurs ayant un impact pronostique majeur chez des patients atteints de cancer. Dans le cadre du cancer du sein, par exemple, la mise en place de la recherche systématique du statut HER2 (*Human Epidermal growth factor Receptor 2*, Slamon et al. (1987)) a permis d'améliorer le statut de santé des patients. Néanmoins, les traitements ciblant ces biomarqueurs n'induisent pas toujours une réponse chez les patients porteurs d' anomalie, suggérant qu'il existe d'autres altérations qui n'ont pas encore été détectées. La détection des biomarqueurs est rendue difficile par le grand nombre de patients qu'il faut pour obtenir une puissance statistique suffisante. Différentes hypothèses peuvent être faites sur les raisons de ces difficultés, comme leur rareté, ou leur signal faible, et/ou un impact synergiques/antagonistes d'ensemble de biomarqueurs au sein de voies moléculaires complexes.

Les structures non-observées dans les données omiques sont généralement identifiées à l'aide de modèles de variables latentes (Székely, Rizzo et Bakirov (2007) ; Cunningham et Ghahramani (2015), MVL). Comme pour le modèle factoriel graphique (Yoshida et West (2010)), l'interprétabilité des composantes peut être renforcée par l'introduction de la parcimonie, pour limiter la contribution des biomarqueurs à un sous-ensemble de composantes. Les composantes identifiées peuvent ainsi être interprétées comme des voies ou des signatures moléculaires (D. A. Knowles et Ghahramani (2007) ; Paisley et Carin (2009) ; Vidaurre et al. (2013)).

En oncologie, différentes tumeurs d'un même organe peuvent résulter de divers mécanismes moléculaires chez différents patients, ce qui complique le développement de traitements stratifiés ou ciblés. Nous avons proposé un nouveau MVL à double parcimonie (Rincourt, Michiels et Drubay (in-revision 2022), chapitre précédent) afin de représenter une hétérogénéité inter-individuelle dans le processus latent impliqué dans la structure des données omiques observées. Nous avons considéré une matrice de poids parcimonieuse pour identifier les signatures moléculaires potentielles, en supposant que ces mécanismes biologiques sont des altérations moléculaires différentes qui ne sont pas partagées par tous les individus. Cette parcimonie est induite par un prior non-

normal (l'ICA Bayésienne standard), favorisant l'indépendance des composantes et limitant les problèmes d'identifiabilité par rotation. Inspirés par B. Chen, M. Chen et al. (2010), nous avons utilisé le processus bêta-Bernoulli (Hjort (1990)) pour identifier simultanément le nombre de composantes et la parcimonie à partir d'un modèle à composantes infinies (en considérant qu'il existe une infinité de composantes mais que seul un sous-ensemble est présent chez les individus observés). Nous avons introduit une composante qui est partagée par tous les individus, qui représente le bruit moléculaire de la population. Cela nous a permis d'interpréter les composantes parcimonieuses comme des altérations moléculaires partagées par plusieurs individus. Nous avons ensuite ajouté un terme de bruit pour capturer les spécificités moléculaires individuelles.

Cet algorithme, et ceux mentionnés précédemment, sont dédiés à des tâches non-supervisées et ne sont pas destinés à identifier des signatures moléculaires pronostiques. Un grand nombre de travaux ont utilisé une procédure en deux étapes (Peterson (2013) ; Witten et Tibshirani (2009) ; Baek, Ho et Ma (2020) ; Gal et al. (2020) ; Park et al. (2020)), comme suit : i) ajustement d'un MVL non-supervisé sur les données omiques (par exemple, PCA, ICA, etc.) ; et ii) utilisation des poids individuels comme variables explicatives dans une régression de survie. Réalisant l'inférence de ces deux modèles indépendamment, l'information de survie n'est pas prise en compte dans les estimations des paramètres du MVL et inversement. Tout comme les approches multi-omiques visent à intégrer l'information de différentes données moléculaires, nous proposons d'intégrer l'information de la survie des patients au MVL. Des approches de modélisation conjointe ont été proposées pour intégrer simultanément les données omiques et les données de survie pour l'inférence des variables latentes, comme le modèle PLS-survie (Bastien et al. (2014)) et implémenté dans le package *plsRcox* (Bertrand et al. (2014)). Plusieurs versions du modèle conjoint analyse factorielle-survie sont également disponibles dans le cadre Bayésien (M. Liu et al. (2019) ; Cai et Liang (2019) ; Pan et al. (2019) ; McCurdy, Molinaro et Pachter (2017)). Dans Bhattacharya et Dunson (2011), les auteurs intègrent la réponse de survie comme une covariable (sans tenir compte des données censurées).

Pour tenir compte de l'hétérogénéité interpatient, nous proposons un modèle conjoint d'analyse de survie et de l'ICA graphique parcimonieuse à nombre infini de composantes (isgICA, voir chapitre précédent), appelé SisgICA. En partant du principe que

les composantes moléculaires proposées peuvent avoir un impact sur la survie du patient, nous avons envisagé un modèle qui partage l'espace latent binaire du processus bêta-Bernoulli (entre l'isgICA et un modèle de Weibull à temps de défaillance accéléré (AFT) (c'est-à-dire que la structure parcimonieuse de la matrice de poids est utilisée comme un ensemble de variables explicatives du sous-modèle de survie). Ce modèle utilise simultanément les informations des deux sources (c'est-à-dire la survie et l'expression génique) pour définir des structures parcimonieuses qui nous permettraient de déconvoluer l'expression génique et d'expliquer la survie.

La performance du modèle SisgICA a été évaluée sur des ensembles de données synthétiques et sur un jeu de données d'expression génique de patientes atteintes d'un cancer du sein qui ont été incluses dans des essais cliniques de chimiothérapie à base d'anthracycline.

Ce chapitre est organisé comme suit : la section 2 présente les sous-modèles du modèle conjoint, et la section 3 les performances du modèle dans les ensembles de données synthétiques et réelles. Nous discutons ensuite nos résultats et concluons dans les sections 4 et 5, respectivement.

4.2 Travaux connexes précédents

Un certain nombre de méthodes statistiques sont disponibles pour étudier l'identification des caractéristiques associées à la survie. À cette fin, différentes méthodes supervisées pour étudier l'hétérogénéité individuelle sont disponibles : un modèle de survie-LDA (Dawson et Kendziorski (2012)), un modèle de survie par classification (Kohjima, Matsubayashi et Toda (2018) ; Chapfuwa et al. (2020)), ou un modèle de survie-VAE (Beaulac, Rosenthal et Hodgson (2018)). Ces méthodes étudient l'hétérogénéité individuelle par classification, mais n'aborde pas l'hétérogénéité moléculaire individuelle.

Bhattacharya et Dunson (2011) ont proposé un modèle Bayésien non-paramétrique avec un nombre infini de facteurs parcimonieux pour prédire les temps de survie en présence de mécanismes moléculaires. La déconvolution impose une pseudo-parcimonie à la matrice gènes-composantes (sans imposer de (pseudo)parcimonie à la matrice individus-composantes). Le nombre de facteurs latents est estimé à partir d'un pro-

cessus gamma multiplicatif. De plus, l'approche proposée par ces auteurs modélise la survie des patients selon un modèle paramétrique (log-normal), mais sans prendre en compte la censure, inhérente aux données réelles. On peut citer également que cette étude ne fournit pas d'évaluation de la reconstruction de la structure latente et des paramètres de survie de la régression.

Plusieurs versions du modèle conjoint analyse factorielle (non parcimonieuse) - données de survie (modèle de Cox) sont également disponibles dans le cadre Bayésien paramétrique (M. Liu et al. (2019) ; Cai et Liang (2019) ; Pan et al. (2019) ; McCurdy, Molinaro et Pachter (2017)). Le nombre de composantes est estimé à travers différentes méthodes : par DIC (M. Liu et al. (2019)), par cross-validation (McCurdy, Molinaro et Pachter (2017)) ou par facteur de Bayes (Pan et al. (2019)). Ces modèles intègrent aussi différentes données génétiques (McCurdy, Molinaro et Pachter (2017)) ou cliniques (Cai et Liang (2019) ; Pan et al. (2019)), ou bien le modèle intègre de données sous forme longitudinale (M. Liu et al. (2019)).

4.3 Méthode

4.3.1 Modèle joint SisgICA

Considérant que l'indépendance de l'expression génique X et le temps de survie des patients T est conditionnelle à l'allocation des patients Z (cf section 4.3.3.3), nous pouvons écrire :

$$p(Z, X, \Delta, T) = p(Z \mid \pi)\ p(X \mid Z)\ p(T \mid \Delta, Z) \tag{4.1}$$

où $p(Z \mid \pi)$ est le prior de l'allocation des patients, $p(X \mid Z)$ est le sous-modèle de déconvolution de l'expression génique, et $p(T \mid \Delta, Z)$ est le sous-modèle de survie.

En conséquence, l'ELBO du modèle joint est défini telle que :

$$
\begin{aligned}
ELBO(q) &= \mathbb{E}_q[\log(p(X,T|\theta)] - KL(q(\theta)|p(\theta)) \\
&= \underbrace{\mathbb{E}_q[\log(p(X|\theta_X)] + \mathbb{E}_q[\log(p(T|\theta_T)]}_{\text{X indépendant de T conditionnellement à Z}} - KL(q(\theta_X)|p(\theta_X)) - KL(q(\theta_T)|p(\theta_T)) \\
&= \underbrace{\mathbb{E}_q[\log(p(X|\theta_X)] - KL(q(\theta_X)|p(\theta_X))}_{\text{isgICA}} + \underbrace{\mathbb{E}_q[\log(p(T|\theta_T)] - KL(q(\theta_T)|p(\theta_T))}_{\text{Survie}}.
\end{aligned}
\tag{4.2}
$$

4.3.2 Sous-modèle isgICA

Le sous-modèle isgICA étant décrit en détail dans le chapitre précédent (voir 3), ce paragraphe est dédié aux rappels des notions essentielles à sa compréhension. Soit N le nombre d'individus, P le nombre de gènes et K le nombre de composantes latentes. L'analyse en composantes indépendantes bruitées vise à décomposer une matrice de données $\mathbf{X} \in \mathbb{R}^{P\times N}$ en le produit de deux matrices plus le bruit résiduel comme suit :

$$
\mathbf{X} = \mathbf{\Phi}(\mathbf{W} \circ \mathbf{Z}) + \mathbf{E}, \tag{4.3}
$$

où $\mathbf{W} \in \mathbb{R}^{K\times N}$ représente la matrice de poids, $\mathbf{Z} \in \mathbb{1}^{K\times N}$ la matrice de parcimonie de $\mathbf{W}$, $\mathbf{\Phi} \in \mathbb{R}^{P\times K}$ représente la matrice des sources non-observées, et $\mathbf{E} \in \mathbb{R}^{P\times N}$ représente la matrice de bruit normal additif. Voir la section 3.2.2 pour plus de détails.

Comme précédemment, nous appliquons une régularisation de type ARD sur les estimations des éléments de $\mathbf{W}$ et de $\mathbf{\Phi}$ avec la même paramétrisation des hyper-paramètres (avec $c = d = 1, e = f = 10^{-6}$, voir la section 4.3.6).

L'estimation consistante de la précision du bruit (τ_E) est essentielle pour déterminer le nombre optimal de composantes latentes (voir la section 3.2.4) (Bouveyron, Latouche et Mattei (2020)). Pour atténuer ce problème, nous avons normalisé les variables d'entrée et fixé la précision du bruit à $\tau_E = 1$.

4.3.3 Sous-modèle paramétrique de survie : Régression paramétrique de Weibull sous l'hypothèse AFT

4.3.3.1 Introduction à la survie

La modélisation du délai de survenue du cancer permet d'étudier différentes situations comme le délai avant la survenue du décès (survie générale) ou d'une rechute à un traitement (survie sans progression). C'est un critère de jugement dans le domaine de la cancérologie qui est très utilisé car il permet d'étudier un bénéfice clinique direct pour un patient (l'amélioration de la survie). La survie est un critère facile à mesurer, objectif et non-ambigu.

L'analyse de survie étudie la distribution de probabilité d'un événement dans le temps. La présence de la censure - la non-observation de l'événement d'intérêt après une période de suivi - est la principale caractéristique des données de survie. On suppose que les patients qui sont censurés ont les mêmes perspectives de survie que ceux qui continuent à être suivis, c'est-à-dire que la censure n'est pas informative.

4.3.3.2 Notions

Également appelé le temps de survie, le temps jusqu'à occurrence de l'évènement d'intérêt $\widetilde{t_i}$ chez le patient i n'est pas observé pour tous les patients, du fait des perdus de vue ou d'un suivi trop court dans une étude. Nous observons le temps jusqu'à la dernière information de suivi du patient t_i, et le statut du patient δ_i, à ce moment donné ($\delta_i = 0$ si censuré avec un temps jusqu'à censure c_i, $\delta_i = 1$ si l'événement d'intérêt est observé pour le i-ième patient).

Dans le cadre d'une analyse de survie, on dispose d'un échantillon de N individus composé du temps de suivi $t_1, \ldots, t_N$ et d'indicatrice d'évènements $\delta_1, \ldots, \delta_N$ avec la relation :

$$t_i = \min(\widetilde{t_i}, c_i) \text{ et } \delta_i = \begin{cases} 1, & \text{si } \widetilde{t_i} \leq c_i, \\ 0, & \text{sinon.} \end{cases} \tag{4.4}$$

La probabilité qu'un événement se produise à un moment T avant un certain mo-

ment t peut s'écrire comme suit :

$$Pr(T \leq t) = \int_0^t f(s)ds = F(t), \tag{4.5}$$

où les fonctions f(t) et F(t) sont respectivement la densité de probabilité (PDF) et la fonction de probabilité cumulée (CDF). Ces différentes fonctions représentant l'analyse de durée sont représentées en figure 4.1 avec le symbole associé de la fonction et les transformations associées.

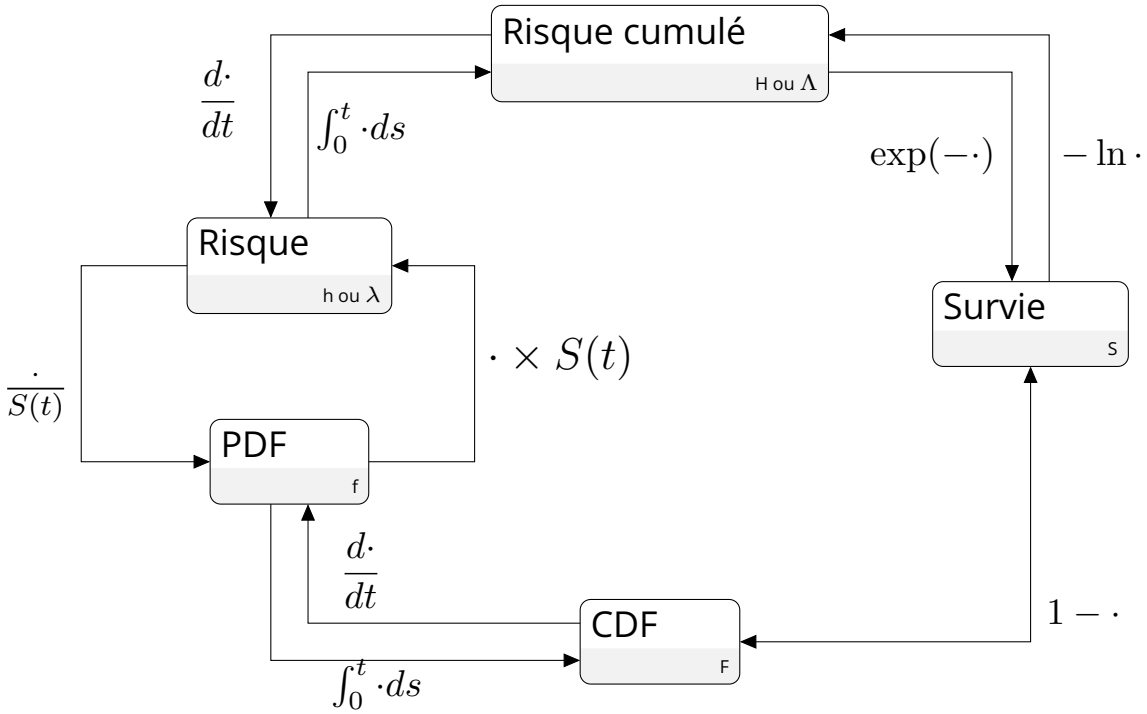


Figure 4.1 – Transformations des fonctions utilisées dans l'analyse de survie ainsi que les symboles des fonctions associées.

Dans le cas contraire, la fonction de survie S(t) est la probabilité que l'instant T de l'évènement ait lieu après un certain temps t, soit :

$$Pr(T > t) = S(t) = 1 - F(t). \tag{4.6}$$

La fonction de risque est définie par le logarithme de la fonction de survie :

$$\Lambda(t) = -\ln(S(t)), \quad \Lambda(t) = \int_0^t \lambda(s)ds, \tag{4.7}$$

avec λ le risque instantané, défini par la probabilité que l'événement se produise dans un petit intervalle de temps après s sachant qu'il n'a pas eu lieu jusqu'à s.

Le temps jusqu'à l'évènement $\widetilde{t}$ n'étant pas toujours observé, il faut prendre en compte la censure dans les estimations des paramètres. On définit alors la fonction de vraisemblance censurée des données de survie comme :

$$\begin{aligned} p(t \mid \theta) &= \prod_{i=1}^{N} h(t)^{\delta_i} S_i(t) \\ &= \prod_{i=1}^{N} f_i(t)^{\delta_i} S_i(t)^{1-\delta_i}. \end{aligned} \tag{4.8}$$

Nous pouvons estimer l'ELBO (voir section 3.2.4) du modèle de survie par :

$$\begin{aligned} ELBO(q(\theta \mid \omega)) &= \ln p(t) - KL(q(\theta \mid \omega)|p(\theta \mid t)) \\ &= \mathbb{E}_{q(\theta|\omega)}[\log(p(t|\theta)] - KL(q(\theta \mid \omega)|p(\theta)). \end{aligned} \tag{4.9}$$

4.3.3.2.1 Modélisation de la fonction de survie : modèle paramétrique, semi-paramétrique ou non-paramétrique Dans l'analyse de survie, nous avons trois possibilités pour modéliser la fonction de survie : non-paramétrique, semi-paramétrique et paramétrique.

La méthode non-paramétrique la plus utilisée pour modéliser la fonction de survie est l'estimateur de Kaplan-Meier.

La méthode semi-paramétrique la plus utilisée est la régression de Cox. Elle décompose la fonction du hasard (risque instantané) en une fonction du risque de base non-paramétrique afin de garder la flexibilité de sa modélisation, et un risque relatif pour modéliser l'impact des covariables sur ce risque de base par une fonction paramétrique, permettant de donner une évaluation quantitative de cet impact (voir la section 4.3.3.2.2 sur l'interprétation des *hazards ratios*) HR. Le modèle est estimé à partir de la vraisemblance partielle qui permet de traiter le risque de base comme un paramètre

de nuisance (d'où le terme de vraisemblance *partielle*). L'intérêt de cette approche est que le risque de base n'est pas nécessaire à estimer pour estimer les HR.

Malgré que la vraisemblance partielle du modèle de Cox permette de s'affranchir de l'estimation du risque de base, la modélisation de ce risque de base est utile pour calculer facilement les statistiques correspondantes aux distributions (grâce aux formules analytiques), de prédire les temps de survie ou d'extrapoler les résultats en dehors des limites de l'échantillon.

La troisième possibilité est de modéliser la fonction de survie à partir d'une distribution paramétrique. Par exemple, on peut supposer une distribution exponentielle, une distribution de Weibull, une distribution de Gompertz... Cette approche modélise la fonction du risque instantané à partir d'une distribution connue. Ainsi le risque de base est paramétrique, comme l'est le risque relatif pour la modélisation de l'impact des covariables.

4.3.3.2.2 Hypothèse sur la fonction de risque Les hypothèses fréquemment utilisées pour modéliser la durée de survie sont : la fonction de risque avec l'hypothèse des risques proportionnels (PH), ou l'hypothèse AFT (*Accelerated Failure Time*). Un descriptif complet entre l'hypothèse PH et AFT dans les modèles paramétriques est disponible dans la thèse de Qi (2009) :

— les modèles PH décrivent un effet multiplicatif des covariables par rapport au **risque** ;

— et les modèles AFT décrivent l'effet multiplicatif des covariables par rapport au **temps de survie**.

L'hypothèse des risques proportionnels est impliquée dans le modèle semi-paramétrique de Cox, ou certains modèles paramétriques dont les plus classiques sont le modèle Weibull-PH ou le modèle Gompertz-PH. Lorsque l'on définit les risques proportionnels par une distribution paramétrique quelconque, on définit λ le paramètre d'échelle, et γ le paramètre de forme. L'hypothèse suppose que l'effet d'une covariable est de multiplier le risque par une certaine constante :

$$\begin{aligned} h(t|Z_i) &= \lambda h_0(t)^{\gamma-1}\theta \\ \text{avec } \theta &= \exp(+Z^T\beta) \end{aligned} \tag{4.10}$$

avec $\beta \in \mathbb{R}^P$ le vecteur de paramètres de la régression de dimension P, $\mathbf{Z} \in \mathbb{R}^{N,P}$ la matrice de covariables, et $h_0(t)$ le risque de base.

En comparaison l'hypothèse AFT impliquée dans les modèles paramétriques (tels que dans la Weibull-AFT, la log-logistique, la log-normale ou la gamma) suppose que l'effet d'une covariable est d'accélérer ou de ralentir le cours de la vie d'une maladie par une certaine constante :

$$\begin{aligned} h(t|Z_i) &= \lambda h_0(t)^{\gamma-1}\theta^{\gamma} \\ \text{avec } \theta &= \exp(-Z^T\beta). \end{aligned} \tag{4.11}$$

Hazard Ratio. Indépendamment de l'hypothèse du modèle, le vecteur estimé θ des paramètres de la régression exprime le logarithme des risques instantanés (ou logarithme du *Hazard Ratio*), noté $\ln(HR)$:

$$HR = \frac{h(t|Z_i)}{h(t|Z_{-i})} = \frac{\exp(\pm Z_i^T\beta)}{\exp(\pm Z_{-i}^T\beta)} = \exp(\pm\beta(Z_i - Z_{-i})). \tag{4.12}$$

avec i et $-i$, les indicatrices pour les groupes disjoints et complémentaires de la variable étudiée (ex : homme et femme). Dans le cadre de variable binaire on aura alors $Z_i = 1$ et $Z_{-i} = 0$, ce qui donne :

$$HR = h(t|Z) = \exp(\pm\beta). \tag{4.13}$$

Le *hazard ratio* (HR) est la quantité qui est le plus souvent d'intérêt. Son interprétation va cependant dépendre de l'hypothèse utilisée dans le modèle :

— Avec l'hypothèse de PH, si $\beta > 0$ (soit le HR > 1) le risque **augmente**, et si

$\beta < 0$ (soit le HR < 1) le risque **diminue**.

— Avec l'hypothèse d'AFT, si $\beta > 0$ (soit le HR > 1) le risque **diminue**, et si $\beta < 0$ (soit le HR < 1) le risque **augmente**. **Attention :** en fonction de l'écriture implémentée dans les logiciels, β peut prendre la forme inversée avec les paramètres positifs (en écrivant $\theta = \frac{1}{\exp(Z^T\beta)}$) et obtenir une interprétation similaire à l'hypothèse PH.

4.3.3.3 L'utilisation du modèle de Weibull avec l'hypothèse AFT

Nous avons supposé que la présence d'une altération k chez un patient i ($Z_{k,i} = 1$) pourrait être associée à la survie de ce patient. Notre modèle fait donc l'hypothèse que l'expression génique X et le temps de survie T sont indépendants, conditionnellement à l'allocation des altérations Z. Nous avons choisi d'utiliser le modèle paramétrique basé sur la distribution de Weibull sous l'hypothèse des temps de défaillance accéléré (AFT), qui correspond à un modèle *log-linéaire* du logarithme du temps de survie Y de la forme (E. Liu et Lim (2018)) :

$$\begin{aligned} T &\sim \text{Weibull}(Z^T\beta, \frac{1}{\sigma_Y}), \\ \ln T = Y &= Z^T\beta + \sigma_Y\epsilon \sim \text{Gumbel}(Z^T\beta, \sigma_Y), \\ \epsilon &= \frac{log(T) - Z^T\beta}{\sigma_Y} \sim \text{Gumbel}(\ln(Z^T\beta), \frac{1}{\sigma_Y}), \end{aligned} \tag{4.14}$$

où $\beta = \{\beta_0 \dots \beta_K\} \in \mathbb{R}^{(K+1)\times 1}$ sont les paramètres de régression, σ_Y est le paramètre d'échelle, et $\epsilon \in \mathbb{R}^{N\times 1}$ est le vecteur d'erreur. Comme l'allocation du profil de base est égale à un pour toutes les observations ($Z_0 = 1$), le paramètre β_0 associé au profil de base Z_0 est confondu avec le paramètre associé à l'intercept du modèle de survie. Les altérations du profil moléculaire de base correspondent aux covariables explicatives du modèle. Nous avons considéré comme priors la distribution normal multivariée pour les paramètres de régression et la distribution de Wishart pour la matrice de précision.

Comme la censure est la principale caractéristique des données de survie, on observe le temps jusqu'à la dernière information de suivi du patient T, et le statut du patient δ au dernier temps de suivi. La fonction de risque $h_\epsilon(Y)$, la fonction de densité $f_\epsilon(Y)$,

la fonction de survie $S_\epsilon(Y)$, la vraisemblance $L(\beta, \sigma_Y; Y)$ et la log-vraisemblance $\ln L(\beta, \sigma_Y; Y)$ du modèle Weibull-AFT sont respectivement :

$$\begin{aligned}
h_\epsilon(Y) &= \exp(\epsilon), \\
f_\epsilon(Y) &= \exp(\epsilon)\exp(-\exp(\epsilon)), \\
S_\epsilon(Y) &= \exp(-\exp(\epsilon)), \\
p(Y \mid \beta, \sigma_Y) &= \prod_{i=1}^{N} f_i(y_i)^{\delta_i} S_i(y_i)^{1-\delta_i}, \\
&= \prod_{i=1}^{N} \left[\frac{1}{\sigma_Y}\exp(\epsilon)\right]^{\delta_i} \exp(-\exp(\epsilon)), \\
\ln p(Y \mid \beta, \sigma_Y) &= \sum_{i=1}^{N} \delta_i(\epsilon_i - \ln \sigma_Y) - \exp(\epsilon_i).
\end{aligned} \tag{4.15}$$

En inférence Bayésienne, la représentation Bayésienne graphique du modèle est visible en figure 4.2 et le modèle complet est défini tel que (Ibrahim, M.-H. Chen et Sinha (2001)) :

$$\begin{aligned}
&\ln T = Y = b^T Z + \sigma_Y \epsilon \\
&Y \sim \text{Gumbel}(Z^T b, \sigma_Y) \\
&b \sim \mathcal{MN}(\mu_b, (\kappa_0 \Lambda_b)^{-1}) \\
&\Lambda_b \sim \mathcal{W}is(\nu_{\Lambda_b}, V_{\Lambda_b}) \\
&\sigma_Y \sim \mathcal{LN}(m_{\sigma_Y}, s^2_{\sigma_Y}).
\end{aligned} \tag{4.16}$$

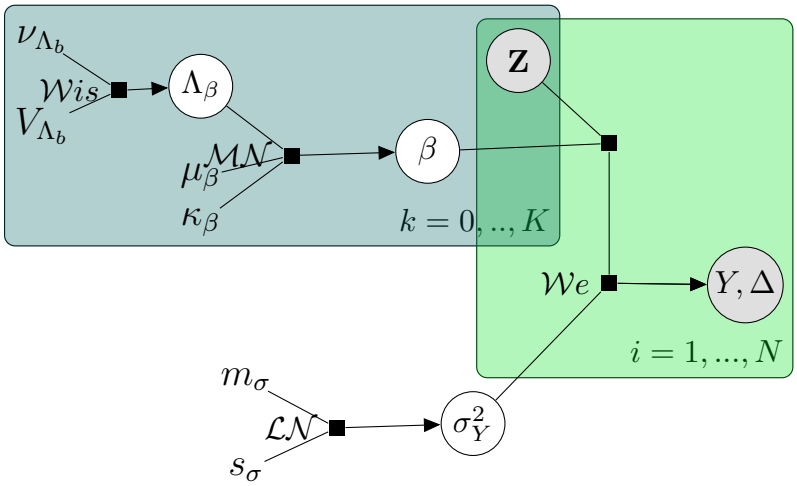


Figure 4.2 – Représentation graphique du modèle de Weibull sous l'hypothèse AFT. Les variables observées sont désignées par des nœuds grisés, tandis que les variables non-observées sont représentées par des nœuds blancs et les hyperparamètres sont affichés sans nœud. Abréviations : $\mathcal{N}$, distribution normale ; $\mathcal{LN}$, distribution log-normale ; $\mathcal{MN}$, distribution multinormale ; $\mathcal{W}e$, distribution de Weibull ; $\mathcal{W}is$, distribution de Wishart.

4.3.4 Modèle complet Bayésien

Le modèle conjoint SisgICA complet est défini par l'équation (4.17) tel que :

$$
\begin{aligned}
&X = \Phi(W \circ Z) + E \\
&X_i \sim \mathcal{N}(\Phi(W \circ Z), \tau_E^{-1}\mathbb{I}_P) \\
&\Phi_{j,k} \sim \mathcal{N}(0, \tau_{\phi_{j,k}}^{-1}) \\
&W_{k,\cdot} \sim \mathcal{N}(0, \tau_W^{-1}\mathbb{I}_K) \\
&\tau_{\phi_{j,k}} \sim \mathcal{G}(c_{j,k}, d_{j,k}) \\
&\tau_W \sim \mathcal{G}(e, f) \\
\\
&\log T = Y = Z^T\beta + \sigma_Y\epsilon \\
&Y \sim \text{Gumbel}(Z^T\beta, \sigma_Y) \\
&\beta \sim \mathcal{MN}(0, (\kappa_\beta\Lambda_\beta)^{-1}) \\
&\Lambda_\beta \sim \mathcal{W}is(\nu_{\Lambda_\beta}, V_{\Lambda_\beta}) \\
&\sigma_Y \sim \mathcal{LN}(0, s^2_{\sigma_Y})
\end{aligned}
\tag{4.17}
$$

$$
\begin{aligned}
Z_{k^*,i} &\sim \mathcal{B}e(\pi_k) \\
\pi_{k^*} &\sim \mathcal{B}(\frac{a}{K}, \frac{b(K-1)}{K}) \\
Z_{0,} &= [1,..,1]_N,
\end{aligned}
$$

où i = 1, ..., N ; j = 1, ..., P ; k = 0, ..., K ; et k*= 1,...,K., avec τ le symbole des précisions des différentes distributions des matrices, {g,h,c,d,e,f} les différents hyperparamètres associés aux priors des précisions définies par des distributions gamma, $\{a, b\}$ les deux paramètres du processus bêta-Bernoulli(BBP), $\{V, \nu\}$ les hyperparamètres associées à de la distribution Wishart, $\{\kappa, \mu_\beta, \Lambda\}$ les hyperparamètres associées à de la distribution normale multivariée.

Le modèle est représenté graphiquement en figure 4.3 avec les abréviations des distributions.

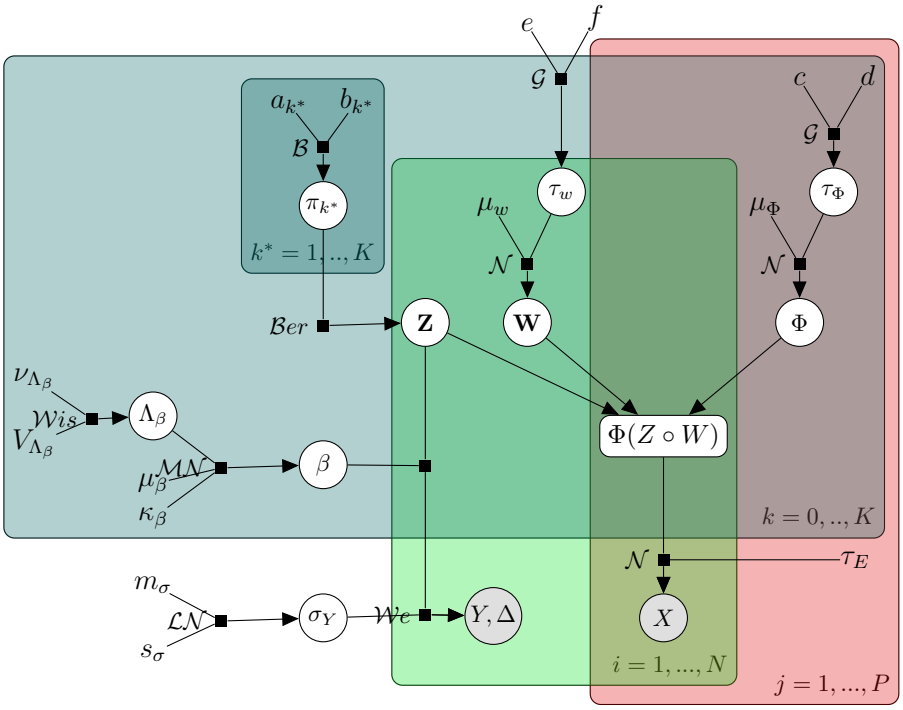


Figure 4.3 – Représentation graphique du modèle joint. Les variables observées sont désignées par des nœuds ombragés, tandis que les variables non-observées sont représentées par des nœuds blancs et que les hyperparamètres ne sont pas entourés. Abréviations : $\mathcal{B}$, distribution bêta ; $\mathcal{B}er$, distribution de Bernoulli ; $\mathcal{G}$, distribution gamma ; $\mathcal{N}$, distribution normale ; $\mathcal{MN}$, distribution multinormale ; $\mathcal{W}e$, distribution de Weibull ; $\mathcal{W}is$, distribution de Wishart.

4.3.5 Équations variationnelles SisgICA

Inspiré par B. Chen, M. Chen et al. (2010), nous avons utilisé l'algorithme CAVI pour minimiser la limite inférieure de l'évidence avec une approximation du champ moyen. S'inspirant du développement du chapitre précédent (présenté dans l'annexe B) nous avons dérivé, les équations de mise à jour des paramètres variationnels, qui sont décrites ci-dessous, avec $i = \{1, ..., N\}, j = \{1, ..., P\}, k = \{0, ..., K_{max}\}$, $k^*= \{1, ..., K_{max}\}$ et le symbole $< \bullet >$ définissant l'espérance du paramètre.

Mise à jour variationnelle du paramètre $<z^*_{k^*,i}>$:

$$
\begin{aligned}
q(z^*_{k^*,i}|-) &= Bernoulli(z^*_{k^*,i}; \pi_{k^*}) \\
&= \frac{q(z_{k^*,i} = 1|-)}{q(z_{k^*,i} = 1|-) + q(z_{k^*,i} = 0|-)}
\end{aligned}
$$

avec

$$
\begin{aligned}
q(z^*_{k^*,i} = 1|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z_{-k^*,i}}, \boldsymbol{W}, \boldsymbol{\Phi}, \tau_E) \\
&\propto \exp(<\ln(\pi_{k^*})> \quad - \frac{1}{2}(<\Phi_{k^*}>^T diag(\tau_E) <\Phi_{k^*}><w_{k^*,i}>^2 \\
&\quad - 2<\Phi_{k^*}>^T diag(\tau_E) \mathbf{X}_i^{-k^*} <w_{k^*,i}>)) \\
&\quad + \delta_i(<\epsilon_i> - \log <\sigma_Y>) - exp(<\epsilon_i>)) \\
&\text{avec } \mathbf{X}_{j,i}^{-k^*} = x_{j,i} - \sum_{l=0, l\neq k^*}^{K} <\Phi_{j,l}><z_{l,i}><w_{l,i}>
\end{aligned}
\tag{4.18}
$$

et

$$
\begin{aligned}
q(z_{k^*,i} = 0|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z_{-k^*,i}}, \boldsymbol{W}, \boldsymbol{\Phi}, \tau_E) \\
&\propto \exp(<\ln(1 - \pi_{k^*})>)
\end{aligned}
$$

Mise à jour variationnelle du paramètre $<\pi_{k^*}>$:

$$
\begin{aligned}
q(\pi_{k^*}|-) &= Beta(\pi_{k^*,i}; a'_{k^*}, b'_{k^*}) \\
&<a'_{k^*}> = \sum_{i=1}^{N} <z_{k^*,i}> + \frac{a}{K_{max}} \\
&<b'_{k^*}> = N + \frac{b(K_{max}-1)}{K_{max}} - \sum_{i=1}^{N} <z_{k^*,i}>;
\end{aligned}
\tag{4.19}
$$

Mise à jour variationnelle du paramètre $<\Phi_{j,k}>$:

$$
\begin{aligned}
q(\Phi_{j,k}|-) &= \mathcal{N}(\Phi_{j,k}; \mu_{\Phi_{j,k}}, \tau^{-1}_{\Phi_{j,k}}) \\
&<\tau_{\Phi_{j,k}}> = \sum_{i=1}^{N} \tau_{E_j} <w_{k,i}>^2 <z_{k,i}>^2 + <\tau_{\Phi_{j,k}}> \\
&<\mu_{\Phi_{j,k}}> = <\tau_{\Phi_{j,k}}>^{-1} \Big(\sum_{i=1}^{N} \tau_{E_j} <w_{k,i}><z_{k,i}> \mathbf{X}^{-k}_{j,i} \Big);
\end{aligned}
\tag{4.20}
$$

Mise à jour variationnelle du paramètre $<W_i>$:

$$
\begin{aligned}
q(W_i|-) &= \mathcal{N}(W_i; \mu_{W_i}, \tau^{-1}_{W_i}) \\
&<\tau_{W_i}> = (\tau_E)(<\Phi> \circ \widetilde{Z}_i^T) + <\tau_W> I_K \\
&<\mu_{W_i}> = <\tau^{-1}_{W_i}>(<\Phi> \circ \widetilde{Z}_i) diag(\tau_E) x_i \\
&\text{avec } \widetilde{Z}_i := [<z_i>, ..., <z_i>] \ , \ <z_i> \text{vecteur répété K fois ;}
\end{aligned}
\tag{4.21}
$$

Mise à jour variationnelle du paramètre $<\tau_{\Phi_{j,k}}>$:

$$
\begin{aligned}
&q(\tau_{\Phi_{j,k}}|-) = Gamma(\tau_{\Phi_{j,k}}; c_{j,k}, d_{j,k}) \\
&<c_{j,k}> = c_0 + \frac{1}{2} \\
&<d_{j,k}> = d_0 + \frac{1}{2} <\Phi_{j,k}>^2;
\end{aligned}
\tag{4.22}
$$

Mise à jour variationnelle du paramètre $<\tau_W>$:

$$
\begin{aligned}
p(\tau_W|-) &= Gamma(\tau_W; e, f) \\
<e> &= e_0 + \frac{NK_{max}}{2} \\
<f> &= f_0 + \frac{1}{2}\sum_{i=1}^{N} <w_i^T><w_i>.
\end{aligned}
\tag{4.23}
$$

Il n'existe pas de forme définie analytique pour approximer les distribution a posteriori des paramètres du sous-modèle de survie. Nous avons utilisé l'inférence variationnelle « boîte noire » (*black-box variational inference*, Ranganath, Gerrish et Blei (2013) ; Knoblauch, Jewson et Damoulas (2019)) que nous avons implémentée à l'aide de Tensorflow (Martín Abadi et al. (2015)).

4.3.6 Optimisation des hyperparamètres

Dans un souci de temps de calcul raisonnables, nous avons utilisé le processus bêta tronqué pour l'inférence, avec une troncature définie par un nombre maximal de composantes noté K_{max} (Paisley et Jordan (2016)). Pour toutes les simulations et l'analyse des données, nous avons considéré les valeurs des hyperparamètres a priori du sous-modèle isgICA comme étant $K_{max} = 100, a = 1, c = d = 1, e = f = 10^{-6}$. Nous avons choisi les valeurs des hyperparamètres des distributions a priori du sous-modèle de survie à $\nu_0 = K; V_0 = 0.1 \times \mathbb{I}_K$ pour la distribution de Wishart et la matrice de variance-covariance S_0 à $S_0 = 0.1 \times \mathbb{I}_K$ pour la distribution normale multivariée. La variance s_{σ_0} de la distribution lognormale est choisi à 1.

En raison de l'absence d'une forme analytique simple de la conjugaison entre la loi a priori de la distribution bêta et la distribution bêta pour les moments, nous avons ajusté l'hyperparamètre b du BBP par optimisation Bayésienne (B0) à l'aide du module R ParBayesianOptimization Samuel Wilson (2020) en nous basant sur six évaluations d'initialisation et neuf époques (pour un total de 15 évaluations). En considérant $a = 1$, nous avons reparamétré b comme $\mu = 1/(1 + b)$ (selon Ferrari et Cribari-Neto (2004)) qui ont pour support l'intervalle [0 ; 1] ($\mu = 0$ correspondant à $b = +\infty$ et $\mu = 1$ à $b = 0$), pour éviter de restreindre le support de $b \in [0; +\infty]$ avec une valeur maximale a priori pour la gamme des points de l'évaluation par BO.

4.4 Résultats

Les simulations, l'optimisation des paramètres et la visualisation des données ont été réalisées à l'aide du logiciel R. Les codes et les données R sont disponibles sur https://github.com/Oncostat/SisgICA.

4.4.1 Données synthétiques

4.4.1.1 Scénarios des simulations

Nous avons simulé des jeux de données synthétiques à partir de l'équation (3.5), avec N=600 sujets et P=1200 variables, pour se rapprocher de notre cas d'application sur le cancer du sein décrite ci-dessous. Nous avons envisagé quatre scénarios en utilisant deux taux de censure (30% et 70%) pour 10 et 30 composantes. Nous avons simulé 100 bases de données pour chaque scénario.

Pour tous les scénarios, les éléments de la matrice des sources Φ ont été générés pour chaque composante à partir de distributions normales avec une variance égale à un, et différentes moyennes (de -3 à 3) pour évaluer si le modèle est capable d'identifier des composantes avec des intensités différentes. Des blocs aléatoires ont été générés pour attribuer une structure de sparsité à cette matrice qui peut être facilement visualisée (voir la figure 4.6). Les éléments de la matrice de poids W ont été simulés d'une distribution normale standard (moyenne = 0, variance = 1). La structure binaire parcimonieuse Z a été générée aléatoirement pour contenir environ 35% de « 1 ». Nous avons considéré que le paramètre de bruit était égal à $\sigma_E^2 = 1$.

Nous avons de la matrice d'allocation des altérations $\mathbf{Z}$ comme variables explicatives de la survie des patients permettant d'interpréter les paramètres de la régression en évitant les problèmes de signe et d'échelle. De plus, comme cette matrice ne souffre pas des problèmes d'identifiabilité connus de l'ICA, notamment l'inversion de signe, cette paraméterisation permet d'obtenir des paramètres dont le signe pourra être interprété.

Nous ne considérons que six paramètres non-nuls dans le vecteur β associé à la composante latente dans le sous-modèle de Weibull de l'équation 4.14, avec des valeurs arbitraires fixées à $[\,-1,\ -0.5,\ -0.1,\ 0.1,\ 0.5,\ 1]$, ce qui permet de considérer des

signatures ayant un impact pronostique faible (0.1) à élevé (1).

Pour générer le temps de suivi t (à partir de l'équation 4.4) défini par le minimum entre le temps avant l'occurrence de l'évènement d'intérêt $\widetilde{t}$ et le temps avant censure c, nous avons simulé deux distributions de Weibull correspondant respectivement au temps avant évènement $\widetilde{t}$ et au temps avant censure c. Le paramètre de forme de la distribution de Weibull des temps avant l'évènement $\widetilde{t}$ et des temps censurés c a été fixée à 2 et le paramètre d'échelle de la distribution des temps censurés a été adaptée pour assurer le taux de censure. Le statut δ est défini par un « 1 » lorsque le temps avant l'évènement $\widetilde{t}$ est inférieur au temps avant censure c et respectivement par « 0 » lorsque le temps avant l'évènement $\widetilde{t}$ est supérieur au temps avant censure c et respectivement.

4.4.1.2 Critères de reconstruction

Notre modèle est identifiable, comme étant l'ICA standard, sous réserve de la mise à l'échelle, de l'inversion de signe et de la permutation des colonnes (Sokol, Maathuis et Falkeborg (2014)). Nous avons utilisé la moyenne des coefficients de corrélation absolue de Pearson (section 3.3.1.2.1) pour ordonner les colonnes et pour quantifier la reconstruction de la matrice des sources $\mathbf{\Phi}$. Selon l'ordre des composantes défini précédemment, la reconstruction de la structure de parcimonie de la matrice de poids (à l'exception du profil de base, c'est-à-dire Z^*) a été évaluée avec le critère de précision (section 3.3.1.2.2).

Nous avons utilisé trois critères pour évaluer la reconstruction des paramètres du sous-modèle de survie : la probabilité de sélection, la probabilité de la couverture et la moyenne du biais.

Nous avons considéré d'utiliser l'intervalle de crédibilité pour déterminer les variables impactant la réponse. Nous définissons l'intervalle de crédibilité bilatéral à 95% à partir des quantiles 2.5 et 97.5% :

$$IC_{95\%} := m \left[m - t_{\frac{\alpha}{2}} * s^2; m + t_{\frac{\alpha}{2}} * s^2 \right] \tag{4.24}$$

avec la moyenne m, la variance s^2 et $t_{\frac{\alpha}{2}}$ le seuil du quantile $\frac{\alpha}{2}$ de la loi normale centrée-réduite.

La sélection des paramètres a été définie par la proportion de fois où l'intervalle de crédibilité à 95% n'inclut pas 0. Ce critère est important lorsque le nombre de variables est élevé et que le nombre de faux positifs est considéré comme plus important que celui des faux négatifs.

La couverture à 95% a été définie comme la proportion de fois où la valeur théorique θ du paramètre est comprise dans l'intervalle de crédibilité à 95 % $IC_{95\%}$.

Le troisième critère est le biais qui a été défini par la différence entre l'estimation $\widehat{\theta}$ et la valeur théorique du paramètre θ :

$$\mathbb{E}\left[\widehat{\theta} - \theta\right]. \tag{4.25}$$

4.4.1.3 Reconstruction de la structure latente et sélection des paramètres de survie

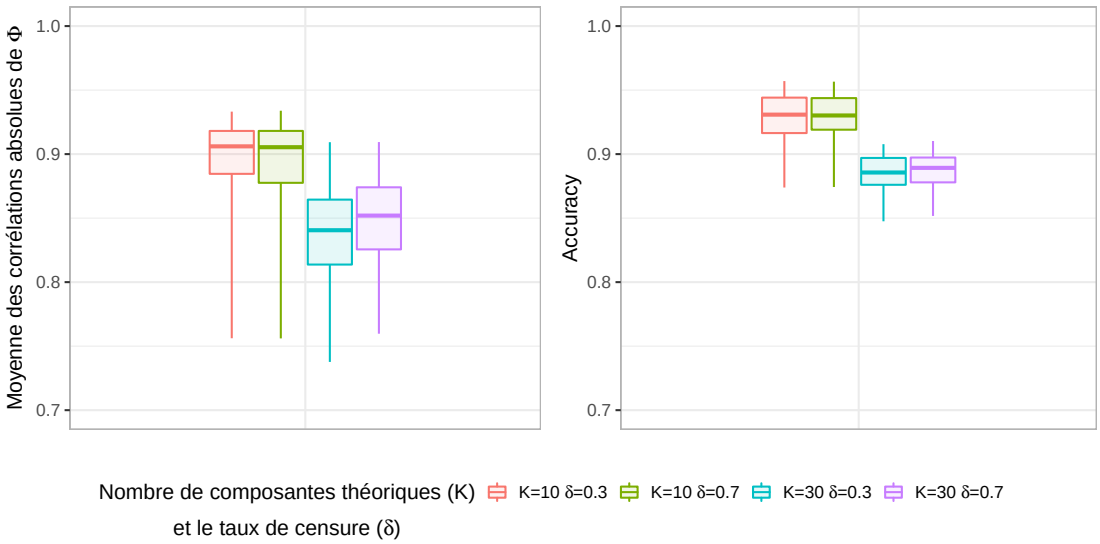


Figure 4.4 – Critères de performance en fonction des différents taux de censure et du nombre de composantes du sous-modèle isgICA dans le modèle conjoint SisgICA, où les boxplots sont définis par la médiane, les quartiles et l'intervalle de percentiles [2.5% - 97.5%] : la reconstruction de la moyenne de corrélation absolue de Φ (colonne de gauche) et de la précision (colonne de droite).

La robustesse de l'algorithme de reconstruction des matrices d'intérêt est visualisée

K_{th}	taux de censure$_{th}$	K	Précision	$\overline{\lvert \text{Cor}_\Phi \rvert}$
10	0.3	10 [10, 10]	0.931 [0.874, 0.957]	0.906 [0.756, 0.933]
10	0.7	10 [10, 10]	0.930 [0.874, 0.957]	0.905 [0.756, 0.934]
10	all	10 [10, 10]	0.930 [0.873, 0.957]	0.906 [0.755, 0.934]
30	0.3	29 [26.475, 30]	0.886 [0.848, 0.908]	0.841 [0.738, 0.909]
30	0.7	29 [27.475, 30]	0.889 [0.852, 0.910]	0.852 [0.760, 0.909]
30	all	29 [26.975, 30]	0.887 [0.849, 0.909]	0.849 [0.743, 0.910]

Table 4.1 – Critères de performance en fonction des différents taux de censure et du nombre de composantes du sous-modèle isgICA dans le modèle conjoint SisgICA avec la médiane et l'intervalle de percentiles [2.5% - 97.5%] : la reconstruction du nombre de composantes (troisième colonne), de la précision (quatrième colonne) et de la moyenne des corrélations absolues de Φ (cinquième colonne).

dans la figure 4.4, et résumée dans le tableau 4.1. Dans tous les scénarios où les données simulés inclut moins de 10 composantes, l'algorithme a reconstruit le nombre exact de composantes latentes et a fourni une bonne reconstruction de la matrice Z (précision $>$ 0.930 [0.873, 0.957] (médiane avec l'intervalle percentile [2.5 - 97.5 %])) et de la matrice Φ (corrélation absolue moyenne $>$ 0.906 [0.755, 0.934]) quel que soit le taux de censure. Quant à la simulation pour 30 composantes latentes, l'algorithme a fourni une bonne reconstruction du nombre de composantes latentes 29 [27-30] et de la reconstruction des matrices Z (0.887 [0.849, 0.909]) et Φ (0.849 [0.743, 0.910]), quel que soit le taux de censure.

Nous avons évalué la probabilité de sélection (figure 4.5, première ligne), de la couverture (figure 4.5, deuxième ligne), et la moyenne du biais des paramètres de survie (figure 4.5, troisième ligne). Les paramètres non-nuls ont été sélectionnés avec une probabilité proche de 1 pour les effets pronostiques les plus forts ($|\beta| = 1$), et diminuant avec l'intensité de l'effet. La probabilité de sélection des paramètres des effets pronostiques plus faibles ($|\beta| = 0.1$) était proche de celle des paramètres nuls ($\approx 25\%$). Ce faible taux de sélection pour les paramètres nuls peut s'expliquer par un faible biais et une probabilité de couverture élevée. Les résultats des figures 4.7 et 4.8 ont montré la régularisation induite par les priors des paramètres du modèle de survie avec l'augmentation du nombre de composantes ou du taux de censure. On observe que les paramètres non-nuls étaient sous-estimés, et que le biais augmentait en fonction de l'ampleur de l'effet pronostique et s'accompagnait d'une couverture et

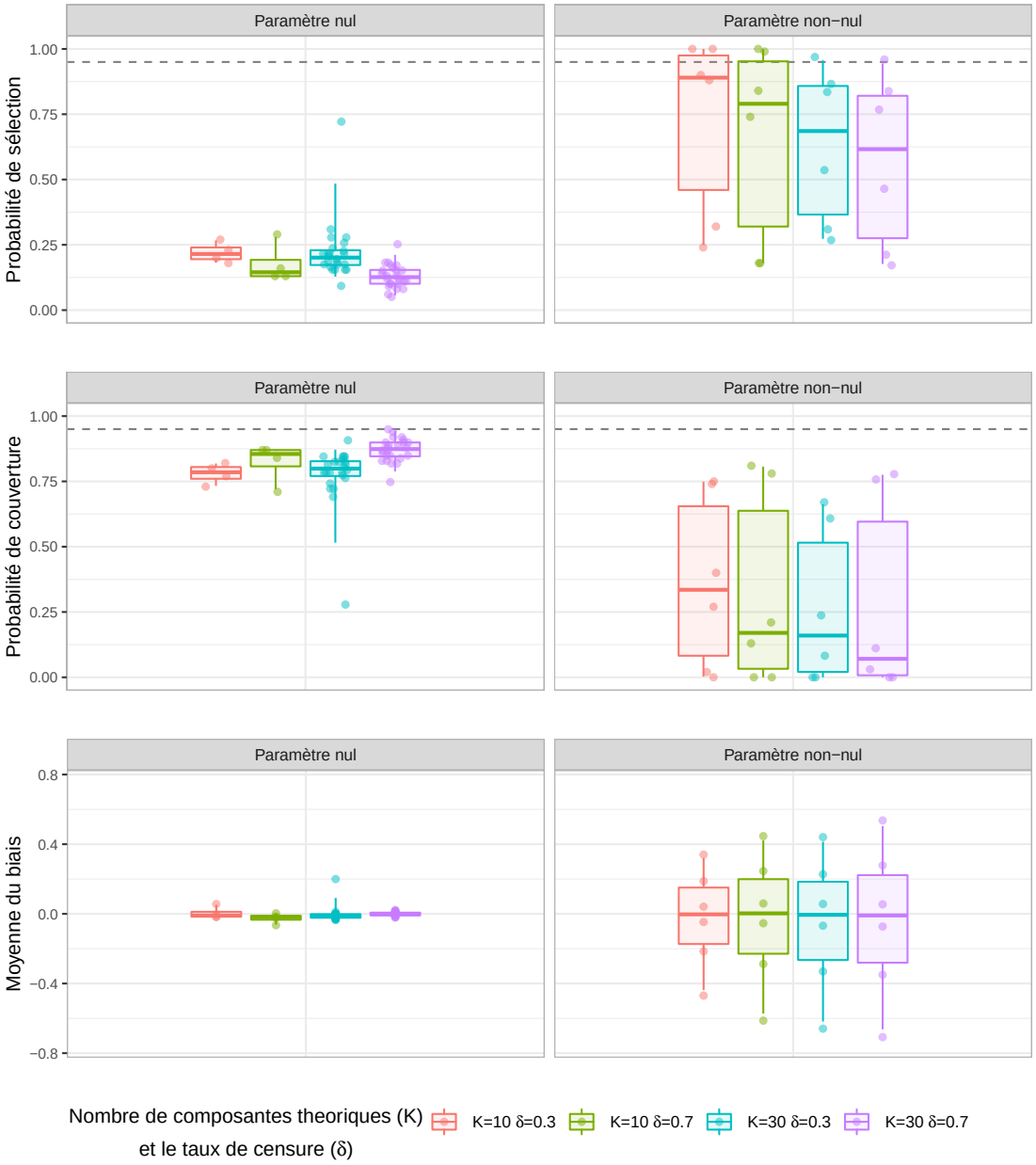


Figure 4.5 – Critères de performance en fonction des différents taux de censure et du nombre de composantes du sous-modèle Weibull-AFT du modèle conjoint SisgICA, où les boxplots sont définis par la médiane, les quartiles et l'intervalle percentile [2.5% - 97.5%] : la sélection (ligne supérieure), la couverture (ligne centrale) et le biais (ligne inférieure).

d'une sélection décroissantes.

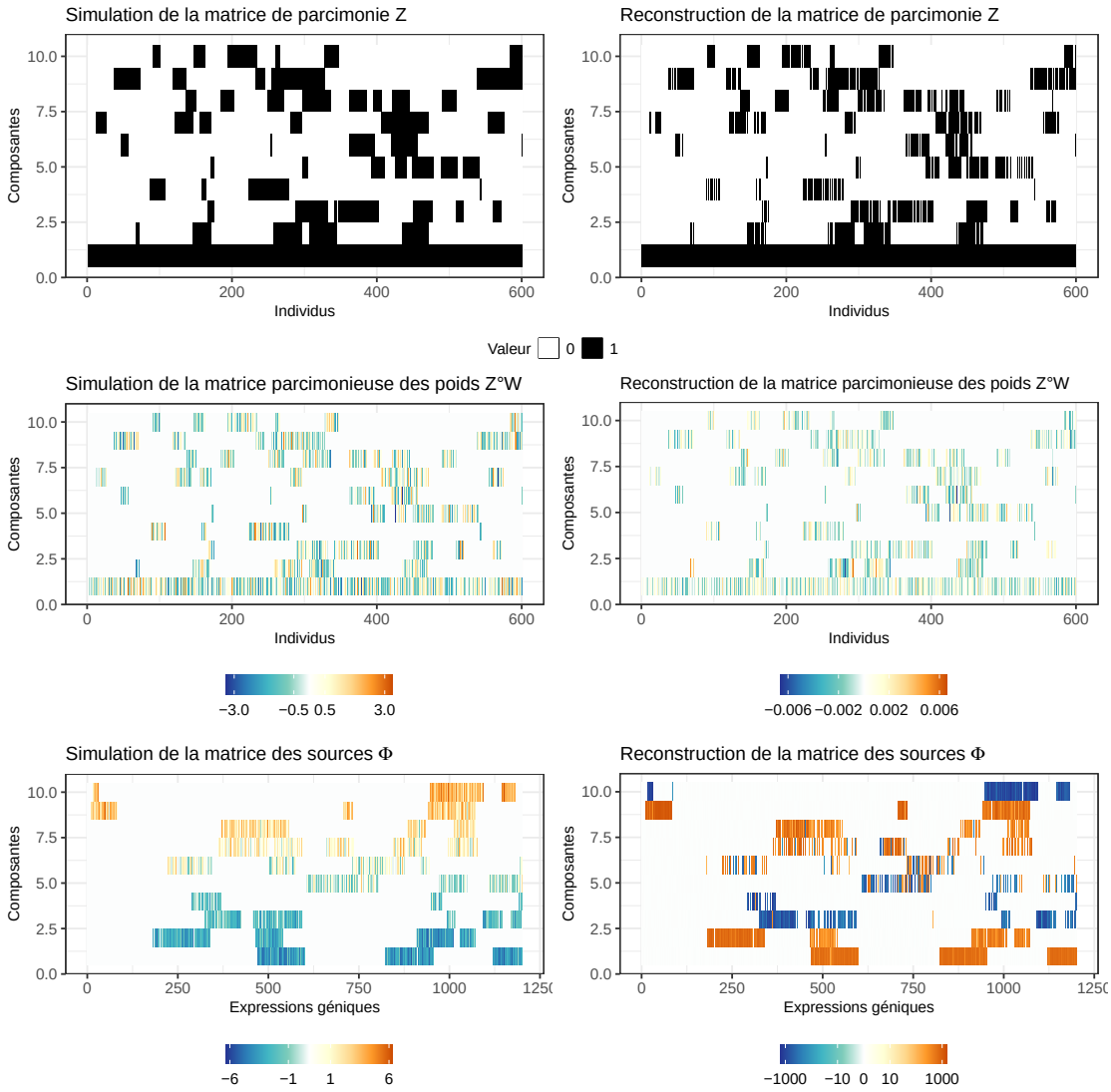


Figure 4.6 – Représentation graphique de la reconstruction de la structure de poids parcimonieux ($\mathbf{Z}$, ligne supérieure, précision de 0.962), de la structure de poids parcimonieux ($\mathbf{W} \circ \mathbf{Z}$, ligne du milieu, corrélation de 0.997), et des sources ($\mathbf{\Phi}$, ligne inférieure, corrélation de 0.900) à partir du scénario $\delta = 0.7$ et $K_{\text{th}} = 10$.

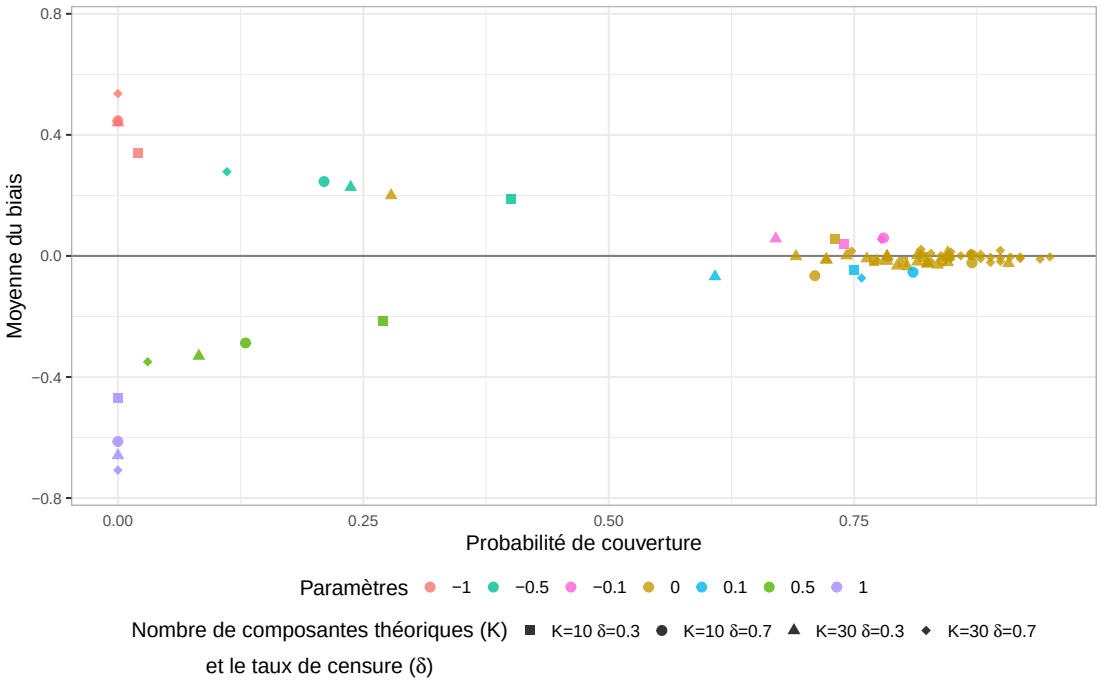


Figure 4.7 – Représentation visuelle de la relation entre le biais et la couverture selon les scénarios avec taux de censure $\{0.3, 0.7\}$ et $K_{\text{th}} = \{10, 30\}$.

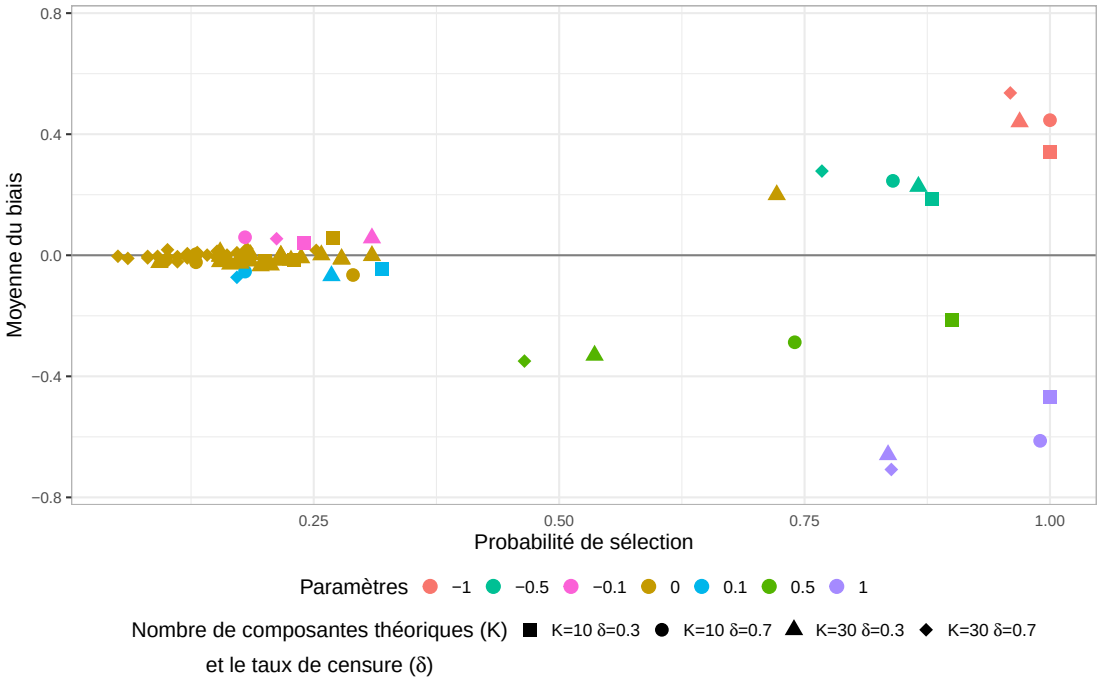


Figure 4.8 – Représentation visuelle de la relation entre le biais et la sélection selon les scénarios avec taux de censure $\{0.3, 0.7\}$ et $K_{\text{th}} = \{10, 30\}$.

4.4.2 Application : early breast cancer data

Nous avons appliqué notre méthode à des données d'expression génique publiquement disponibles, obtenues à partir de biopsies de tumeurs chez 614 patientes atteintes d'un cancer du sein et incluses dans des essais cliniques de chimiothérapie à base d'anthracycline (Desmedt et al. (2011) ; Hatzis et al. (2011)), disponibles dans le package R *biospear* (Ternès, Rotolo et Michiels (2018)). Les 1 689 *probes* normalisées ont été filtrées avec le package *jetset* (Li et al. (2011)) pour ne retenir qu'une seule *probe* par gène. Le jeu de données final comprend l'expression de 1063 gènes. De façon similaire à dans Rincourt, Michiels et Drubay (in-revision 2022), nous avons associé les *probes* à trois signatures moléculaires ayant un effet pronostique dans le cancer du sein précoce (Système immunitaire, Prolifération et invasion du stroma Ignatiadis et al. (2012)) et une sans (signature d'activation du SRC Ignatiadis et al. (2012)), toutes les autres *probes* ont été classées dans la catégorie « Autre ».

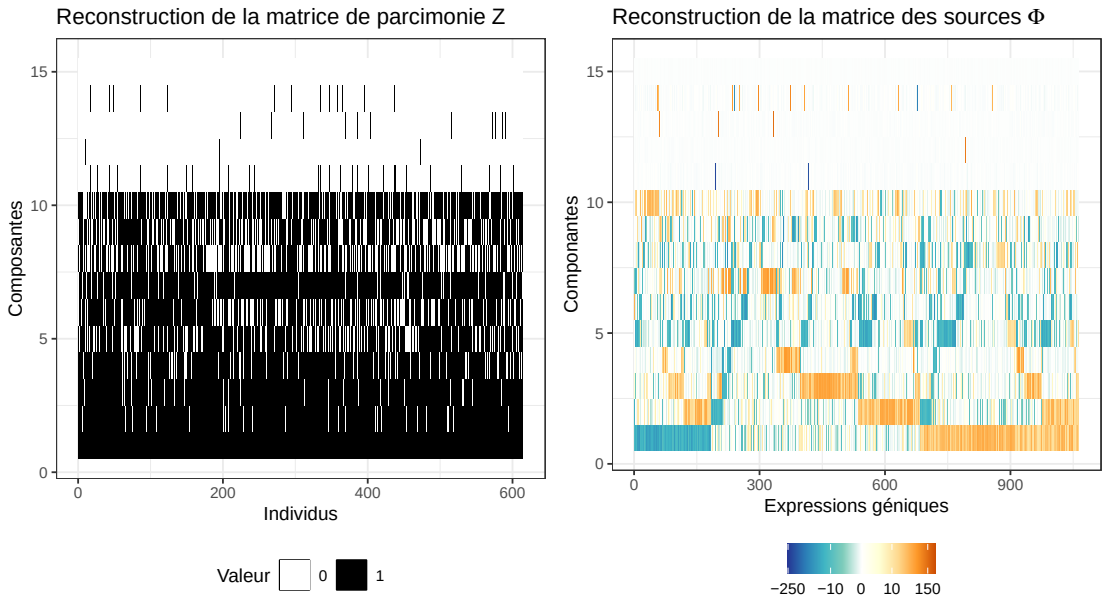


Figure 4.9 – Matrice binaire parcimonieuse de l'hétérogénéité individuelle ($\mathbf{Z}$,structure parcimonieuse de poids, colonne de gauche) et des composantes de la matrice des sources (colonne de droite) extraites de l'ensemble de données sur le cancer du sein à l'aide du modèle conjoint SisgICA. Le modèle a identifié 15 composantes (y compris le profil de base) avec six paramètres non-nuls associés de manière significative au sous-modèle de survie (le profil de base et les composantes d'altération 1, 2, 3, 4 et 6.

La figure 4.9 présente la matrice binaire parcimonieuse $\mathbf{Z}$ de l'hétérogénéité indi-

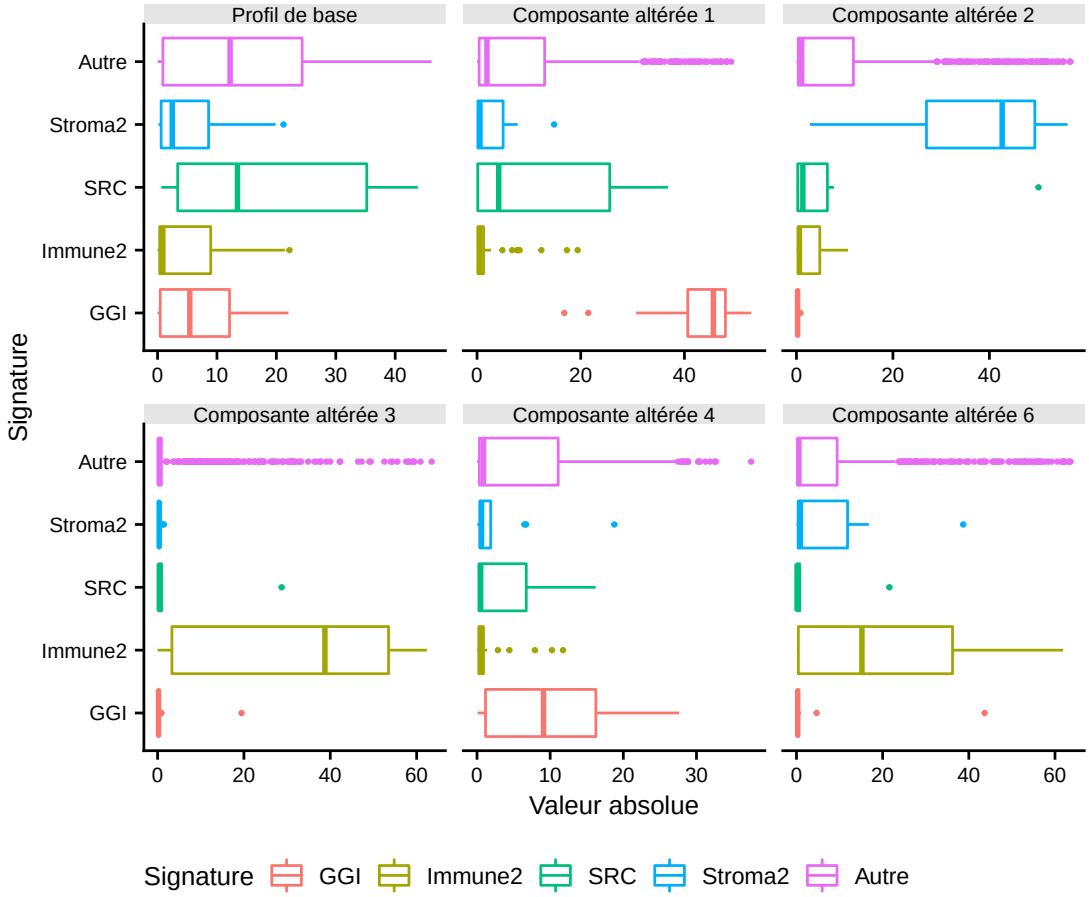


Figure 4.10 – Distribution de la valeur absolue des éléments de 6 des 15 composantes de la matrice des sources significativement associés aux données de survie, et associés aux composantes de gènes appartenant à trois signatures moléculaires ayant un effet pronostique dans le cancer du sein précoce (GGI, Immune2, et Stroma2) et une sans (signature d'activation SRC). Tous les autres gènes ont été classés dans la catégorie « Autre ».

viduelle et la parcimonie de la matrice des sources Φ. Le modèle a identifié 15 composantes non-vides (incluant le profil de base) avec six paramètres non-nuls résumés dans le tableau 4.2.

Pour étudier la pertinence biologique potentielle des composantes moléculaires, la figure 4.10 montre la distribution des valeurs absolues (pour pallier au problème d'identifiabilité et du signe) des éléments de la matrice des sources de chaque composante en fonction des différentes signatures du cancer du sein sélectionnées par le sous-modèle de survie (pour la figure complète voir la figure 4.11). La signature basée sur la prolifération, la signature du système immunitaire et la signature liée au stroma semblent être fortement liées respectivement aux composantes d'altération 1, 3 et 2. La signature

Composante	Paramètre estimé	Hazard ratio
Profile de base	0.605 [0.428, 0.782]	
Composante altérée 1	0.372 [0.187, 0.557]	1.451 [1.206, 1.745]
Composante altérée 2	0.412 [0.231, 0.593]	1.509 [1.259, 1.809]
Composante altérée 3	0.327 [0.143, 0.510]	1.386 [1.154, 1.665]
Composante altérée 4	0.355 [0.166, 0.544]	1.426 [1.181, 1.723]
Composante altérée 6	0.256 [0.067, 0.445]	1.292 [1.069, 1.560]

Table 4.2 – Estimation des paramètres du sous-modèle de survie du modèle conjoint SisgICA : moyenne et intervalle de crédibilité [2.5% - 97.5%].

SRC, choisie comme « contrôle négatif » dans Belhechmi et al. (2020), ne semblait pas être directement liée à une composante moléculaire en particulier.

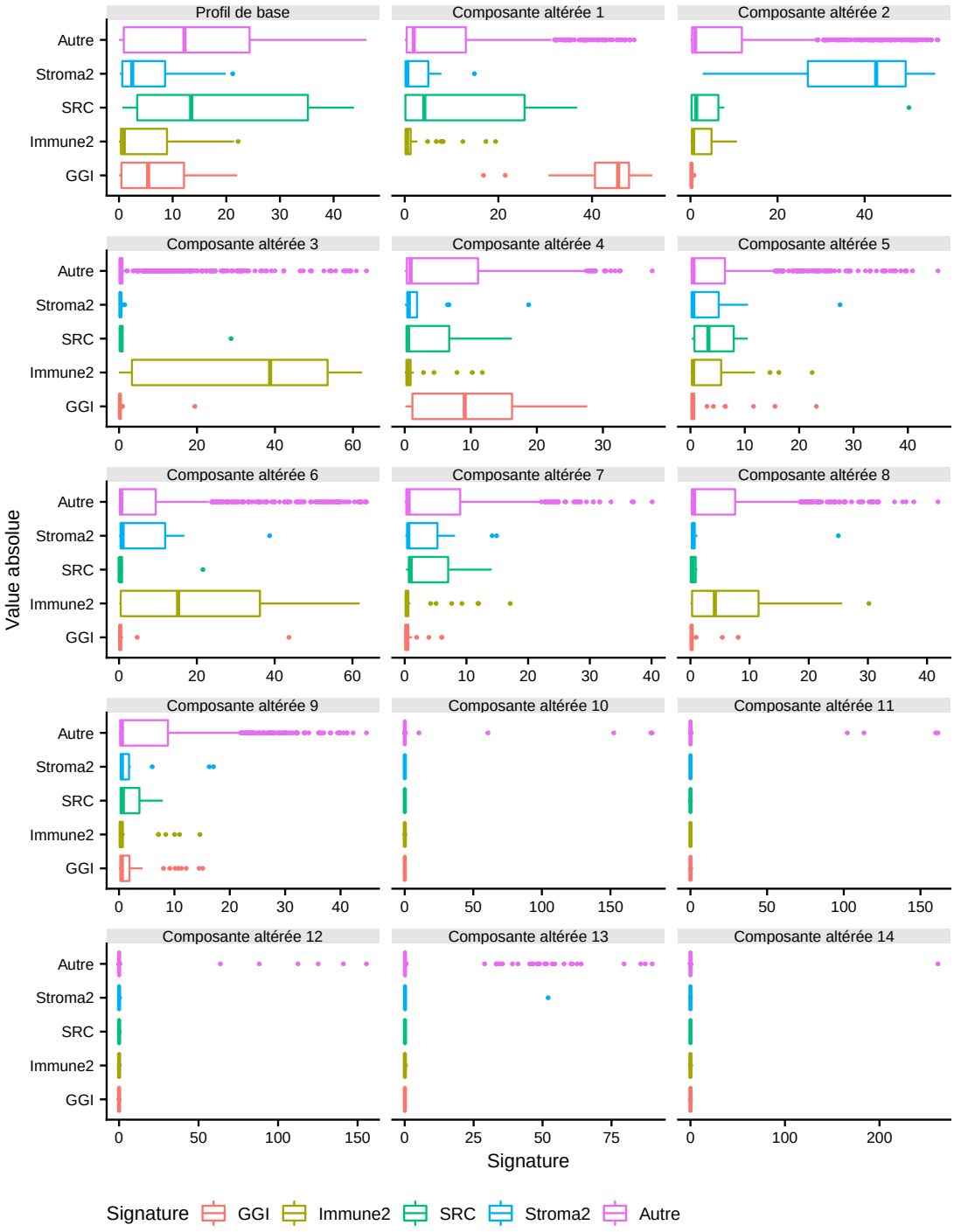


Figure 4.11 – Distribution de la valeur absolue des éléments des 15 composantes de la matrice des sources, et associés aux *probes* des gènes associés à trois signatures moléculaires ayant un effet pronostique connu dans le cancer du sein précoce (GGI, Immune2, et Stroma2) et une sans (signature d'activation SRC). Tous les autres gènes ont été classés dans la catégorie « Autres ». Six des 15 composantes sont significativement associées au sous-modèle de survie (le profil de base et les composantes d'altération 1, 2, 3, 4 et 6).

4.5 Discussion

Dans cet article, nous avons proposé SisgICA, un modèle combinant la survie du patient et de son expression génétique tumorale, afin d'identifier un sous-ensemble de signatures moléculaires pouvant être associées au pronostic du patient. Nous avons supposé que chaque tumeur est un mélange d'un sous-ensemble spécifique d'un nombre limité de signatures moléculaires, ce qui permet de modéliser les altérations moléculaires individuelles d'un profil de population de base. Il s'agit, à notre connaissance, de la première approche abordant cette question qui pourrait fournir une nouvelle vision pour comprendre l'hétérogénéité inter-patients.

Sur la base d'un a priori non-paramétrique, notre algorithme est capable d'identifier le nombre de composantes présents dans l'échantillon observé à partir d'un nombre hypothétiquement infini dans nos simulations. En considérant une double contrainte de parcimonie pour une ICA bruitée, nous avons pu ré-identifié avec précision la structure d'hétérogénéité individuelle (précision > 85 %) dans les ensembles de données synthétiques. Alors que le profil de base semble capturer le bruit moléculaire présent dans la population, le terme de bruit capture la spécificité individuelle, et les composantes parcimonieuses identifiées capturent les signatures moléculaires partagées.

En raison des problèmes d'identifiabilité bien connus de l'ICA (mise à l'échelle et inversion de signe), la matrice des sources n'est pas directement interprétable ; nous avons donc utilisé les valeurs absolues des éléments sources pour l'interprétation des composantes. L'identification des gènes qui contribuent le plus à chaque composante permet de décrire les mécanismes moléculaires potentiels qui peuvent être liés à chacune de ces composantes. L'utilisation de statistiques basées sur le rang (pour ignorer le problème d'échelle) et sur la valeur absolue des éléments de la matrice des sources (pour atténuer le problème d'inversion de signe) a montré que notre algorithme était capable de reconstruire la contribution des gènes aux composantes (corrélation des composantes estimées avec les composantes simulées > 75 %). De plus, la régularisation appliquée à cette matrice (à l'aide d'un prior ARD) a permis d'obtenir une structure de pseudo-parcimonie proche de la structure simulée, ce qui aide à préciser l'interprétation des composantes.

En utilisant la structure de l'allocation des composantes aux individus ($\mathbf{Z}$) comme

variables explicatives du sous-modèle de survie, nous avons montré que notre algorithme était capable d'identifier les composantes pronostiques ayant des effets élevés à modérés. En raison de la régularisation de la distribution à priori que nous avons considérée, le taux de faux positifs est limitée à une valeur raisonnable ($< 25\%$), avec les composantes ayant un faible impact pronostique qui n'ont pas été sélectionnées. Le taux de censure élevé a légèrement augmenté cette régularisation, augmentant le taux de vrais positifs et le taux de faux positifs, mais il n'a pas d'impact significatif sur la reconstruction des matrices du sous-modèle isgICA. L'augmentation du nombre de composantes a eu un rôle similaire sur la régularisation.

Nous avons illustré cette approche à l'aide de données d'expression génique mesurées dans des essais cliniques incluant des patientes atteintes d'un cancer du sein traitées par une chimiothérapie à base d'anthracycline. La méthode a permis d'identifier des composantes particulières, associées à des voies moléculaires avec une valeur pronostique dans le cancer du sein précoce (prolifération, système immunitaire et voie stromale Ignatiadis et al. (2012)). Les composantes non sélectionnées n'étaient pas liées aux voies pronostiques moléculaires que nous avons considéré, ce qui supporte notre hypothèse selon laquelle l'algorithme peut identifier des mécanismes tumoraux.

La prochaine perspective majeure de ce travail sera d'inclure les facteurs clinico-pathologiques standard dans le modèle (par exemple, l'âge, le stade du cancer, ...), et les traitements pour identifier les signatures moléculaires associées à un bénéfice de survie associé au traitement. Une option pourrait être d'utiliser des méthodes d'inférence causale pour fournir un modèle plus interprétable et limiter les problèmes d'identifiabilité. Enfin, comme le propose la (vaste) littérature sur les méthodes de déconvolution pour les données omiques, notre approche pourrait également être adaptée à d'autres données omiques non-normales, telles que les données de comptage (ARNseq brut, protéomique) ou binaires (mutations), en utilisant différentes fonctions de liaison. Par extension, SisgICA pourrait être étendue à l'analyse multi-blocs de données multi-omiques, considérant une structure latente commune des processus générant les différents types de données biologiques ou par une structure latente hiérarchique (Ray et al. (2014)).

4.6 Conclusion

Dans cet article, nous avons proposé un modèle conjoint SisgICA basé sur l'expression génique et la survie des patients, afin d'identifier les composantes liées aux différents mécanismes moléculaires impliqués dans les mécanismes du cancer et le pronostic des patients. Considérant que chaque tumeur ne résulte pas que d'une seule signature moléculaire, mais d'un mélange de plusieurs signatures spécifiques au patient, notre approche peut contribuer à la compréhension des mécanismes moléculaires.

Discussion générale

Du fait de l'hétérogénéité tumorale, il est nécessaire de développer une médecine de précision, afin d'apporter un soin personnalisé pour une meilleure efficacité. L'objectif de cette thèse a été de développer une méthode pour identifier l'hétérogénéité inter-individuelle des mécanismes oncologiques. Dans un premier temps, nous rappellerons les résultats de la thèse, puis nous aborderons les limites et les perspectives de ces travaux.

L'identification des mécanismes moléculaires individuels

L'identification des mécanismes oncologiques à partir de données omiques, telles que l'expression des gènes, est difficile en raison des relations complexes qu'ils entretiennent dans diverses voies moléculaires impliquant des milliers de gènes, ainsi que du nombre relativement faible de patients. Dans ce travail de thèse, nous avons développé deux modèles Bayésiens pour identifier des structures non-observées de l'hétérogénéité individuelle de l'expression génique dans les mécanismes tumoraux. Ces travaux sont pionniers dans un contexte où l'hétérogénéité inter-individuelle est peu étudiée malgré l'importance des développements thérapeutiques ciblés en oncologie. Ces méthodes se basent sur des modèles de variables latentes pour la déconvolution en aveugle des composantes afin d'identifier les structures de données géniques de grande dimension non-observées, qui fournissent un aperçu des mécanismes moléculaires.

Afin de prendre en compte la diversité des mécanismes oncologiques entre les individus, le premier travail a été de proposer un modèle de factorisation de matrice d'analyse en composantes indépendantes (ICA) en supposant que l'expression tumorale de chaque patient résulte d'un mélange d'un sous-ensemble de signatures moléculaires

indépendantes. Les données d'expression génique ont été déconvoluées à partir d'une analyse avec un nombre infini de composantes indépendantes avec une double structure de parcimonie pour les matrices des sources et des poids, correspondant respectivement aux associations gènes-composantes et individus-composantes. Cette structure de parcimonie individus-composantes est un modèle d'allocation basée sur le processus bêta-Bernoulli, qui permet de supposer qu'il existe un nombre infini de composantes dans la population, dont un sous-ensemble fini est présent dans notre échantillon de patients. Nous avons montré que cette approche est capable d'identifier en aveugle les composantes dans un échantillon fini qui représentent des déviations de l'expression d'un profil de base commun à tous les patients.

La compréhension des mécanismes pronostiques tenant compte de l'hétérogénéité inter-patients est un défi majeur pour la médecine de précision. Des algorithmes dédiés à des tâches non-supervisées existent mais ils ne sont pas destinés à identifier des signatures moléculaires pronostiques. Nous avons donc, dans un deuxième travail, proposé une extension de notre approche vers une modélisation conjointe de l'hétérogénéité inter-individuelle moléculaire et de la survie de patients. Ce travail a permis d'intégrer l'information de la survie à la déconvolution de l'expression génique.

Limites et perspectives

Cette approche a permis de développer une nouvelle approche de la modélisation de l'hétérogénéité tumorale inter-patients. Nous avons mis l'accent sur la reconstruction de la parcimonie et de la pseudo-parcimonie des matrices afin de faciliter l'interprétation des composantes par rapport à leur allocation aux patients et aux gènes. Néanmoins, en raison des problèmes d'identifiabilité bien connus de l'ICA (mise à l'échelle et inversion de signe), la matrice des sources n'est pas directement interprétable ; nous avons utilisé les valeurs absolues des éléments sources pour l'interprétation des composantes.

En supposant que la présence d'une altération chez un patient pourrait être associée à la survie de ce patient, l'utilisation de la matrice binaire correspondante aux associations individus-composantes nous a permis d'étudier les mécanismes oncologiques individuels pronostiques. La matrice des poids fournit des quantités continues de chaque composante associée aux individus, mais souffre des mêmes problèmes d'iden-

tifiabilité que la matrice des sources (mise à l'échelle et inversion de signe). Pour cette raison, la matrice des poids ne peut pas être interprétée directement, et l'inversion de signe pourrait résulter en des résultats inverses pour une même composante (effet protecteur/délétère). Nous avons donc choisi d'utiliser la structure de parcimonie de l'allocation composantes/individus, non-sujette à ces problèmes, comme structure latente partagée du modèle joint (c'est à dire comme variable explicative du modèle de survie).

D'un point de vue de l'identifiabilité, nous avons vu que la modélisation du bruit avait un impact très important sur l'identification du nombre de composantes non-nulles (allouées à au moins 1 individu). L'utilisation des estimateurs du maximum de vraisemblance (ou l'ELBO qui en est dérivée) favorise l'adéquation du modèle aux données. En ce sens, nous avons pu observer que notre modèle est très sensible au sur-apprentissage du fait de sa complexité, ce qui nous a obligé à appliquer une contrainte sur la valeur du paramètre de variance du bruit. En effet, comme connu en régression plus classique, le sur-apprentissage résulte en une intégration du bruit dans les paramètres du modèle pour expliquer la réponse, résultant en des paramètres plus élevé que théoriquement (justifiant leur régularisation vers 0). Pour notre modèle, nous avons observé que le paramètre de variance du bruit tendait vers 0, l'information aléatoire étant intégrée dans un nombre excessif de composantes (souvent le seuil de troncature choisi, c'est à dire le nombre maximum autorisé par l'algorithme), qui étaient très peu parcimonieuses. Cela indique que l'estimation des paramètres nuls du modèle d'allocation échappent à la régularisation imposée par le processus bêta-Bernoulli. Bien que nous avons pu montrer que notre modèle garde une bonne performance avec la contrainte que nous avons imposée, cela a pour effet d'imposer une régularisation un peu stricte dans le cadre de la dimension faible à modérée. Une des prochaines extensions majeures de ce modèle sera d'en définir une forme générative plus simple (ex : *variational autoencoder*) afin de pouvoir en générer une prédiction plus facilement et pouvoir ajuster les hyperparamètres du BBP (source de la régularisation) à l'aide d'une fonction de coût de mesure d'erreur d'extrapolation sur un jeu de données de validation, mesurée par exemple par validation croisée (au lieu du maximum de vraisemblance qui est une mesure d'adéquation aux données actuelles).

Néanmoins, les deux modèles de cette thèse possèdent de bonnes performances qui sont obtenues au prix d'un temps de calcul important qui pourrait constituer une limitation pratique, notamment si nous souhaitons réaliser des étapes de validation croisée, nécessitant de répéter les calculs un grand nombre de fois. Une étape de ré-implémentation de l'algorithme sera nécessaire avec le calcul GPU pour l'adapter à de grands ensembles de données et à la ultra-haute dimension. Les codes des algorithmes sont disponibles en ligne mais ne permettent pas d'utiliser ces méthodes aussi facilement qu'une bibliothèque dans un logiciel, il serait pratique d'implémenter la nouvelle version en GPU dans une bibliothèque R ou python.

En comparant la performance des deux méthodes développées au cours de cette thèse, on observe que les résultats sont similaires. L'apport de la survie correspond à l'ajout d'une variable parmi P+1 variables. De ce fait, plus le nombre de gènes augmente, plus la survie exerce une influence faible sur la définition des composantes. Ces résultats ouvrent une question plus large autour de la très populaire intégration de données hétérogènes : les méthodes d'analyse de données *multi-omiques* sont généralement des analyses multi-blocs, avec 1 bloc par type de données omiques (expression, nombre de copie, immunologie, protéosome, lipidosome,...). Ces blocs ayant chacun des dimensions différentes, leur poids dans l'analyse sera différent, les plus petits blocs auront une plus faible influence pour la définition des structures latentes. Certains de ces petits blocs représentent des mécanismes biologiques communs ou biologiquement plus pertinents (p.ex : les quelques dizaines de populations immunitaires vs expression de plusieurs dizaines de milliers de gènes), et une extension de ces travaux de thèse à l'intégration de données multi-omiques devra en tenir compte. En effet, à l'instar des approches multi-omiques mentionnées précédemment, notre approche pourrait être adaptée à d'autres données omiques non-normales, telles que les données de comptage (ARNseq brut, protéomique) ou binaires (mutations), en utilisant différentes fonctions de lien. Par extension, SisgICA pourrait être étendue à l'analyse multi-blocs de données multi-omiques, en définissant une structure latente commune des processus générant les différents types de données biologiques ou par une structure latente hiérarchique (Ray et al. (2014)).

La seconde perspective majeure de ce travail sera d'intégrer de l'information a priori pour faciliter l'identifiabilité du modèle. Dans un premier temps, le prior de la structure de l'hypergraphe pourrait intégrer l'information acquise depuis des décennies sur de nombreuses voies biologiques, identifiées in vitro ou in vivo. De plus, les facteurs clinico-pathologiques standards (par exemple, l'âge, le stade du cancer, ...), et les traitements pourraient être intégrés au modèle pour identifier les signatures moléculaires associées à un bénéfice de survie, et associées au traitement. Une option pourrait être d'utiliser des méthodes d'inférence causale pour fournir un modèle plus interprétable et limiter les problèmes d'identifiabilité.

Conclusion générale

Pour des maladies complexes telles que le cancer, la médecine de précision vise à allouer à chaque patient le traitement qui lui apporterait le meilleur bénéfice, en fonction de son profil moléculaire. Le coût décroissant de l'acquisition de données génomiques et la démocratisation de l'analyse d'autres acteurs « omiques » ont nourri l'espoir de développer une médecine de plus en plus personnalisée. Bien que les biomarqueurs les plus associés à la réponse des patients aient été identifiés, le développement de thérapies ciblées ne permet pas de soigner l'ensemble des patients, du fait de l'interaction avec de multiples autres acteurs au sein de mécanismes biologiques complexes. De plus, de nombreux biomarqueurs ont été identifiés, mais très peu ont été validés, ce qui peut suggérer qu'il existe des biomarqueurs ayant un faible impact sur la réponse du patient mais qui agissent en synergie ou antagonisme les uns avec les autres. L'objectif de ce travail de thèse a été de proposer deux nouveaux outils pour identifier ces structures de biomarqueurs oncologiques au sein de potentielles voies moléculaires. Les perspectives de cette thèse sont multiples : autant méthodologique, afin d'améliorer le contexte clinique, que technique, afin de le rendre plus fonctionnel, et clinique, afin d'orienter la compréhension et la recherche de nouveaux biomarqueurs pour aider au développement de nouvelles thérapies de précision.

Bibliographie

Argelaguet, R. et al. (2018). « Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets ». In : *Molecular Systems Biology* 14.6, e8124. doi : https://doi.org/10.15252/msb.20178124. eprint : https://www.embopress.org/doi/pdf/10.15252/msb.20178124.

Armagan, A., D. B. Dunson et M. Clyde (2011). *Generalized Beta Mixtures of Gaussians*. doi : 10.48550/ARXIV.1107.4976.

Baek, S., Y.-Y. Ho et Y. Ma (jan. 2020). « Using sufficient direction factor model to analyze latent activities associated with breast cancer survival ». In : *Biometrics* 76.4, p. 1340-1350. doi : 10.1111/biom.13208.

Bastien, P. et al. (oct. 2014). « Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data ». In : *Bioinformatics* 31.3, p. 397-404. issn : 1367-4803. doi : 10.1093/bioinformatics/btu660.

Beal, M. J. (mai 2003). « Variational algorithms for approximate bayesian inference ». Thèse de doct. Gatsby Computational Neuroscience Unit, University College London, p. 1-281. isbn : 9789090257501.

Beaulac, C., J. S. Rosenthal et D. Hodgson (2018). *A Deep Latent-Variable Model Application to Select Treatment Intensity in Survival Analysis*. arXiv : 1811.12323 [stat.ML].

Belhechmi, S. et al. (juill. 2020). « Accounting for grouped predictor variables or pathways in high-dimensional penalized Cox regression models ». In : *BMC Bioinformatics 2020 21 :1* 21.1, p. 1-20. issn : 1471-2105. doi : 10.1186/S12859-020-03618-Y.

Bertrand, F. et al. (2014). « plsRcox, Cox-models in a high dimensional setting in R ». In : *Proceedings of User2014 !*, p. 177.

Betancourt, M. (2017). *A Conceptual Introduction to Hamiltonian Monte Carlo*. doi : 10.48550/ARXIV.1701.02434.

Betensky, R. A., D. N. Louis et J. G. Cairncross (mai 2002). « Influence of unrecognized molecular heterogeneity on randomized clinical trials ». In : *Journal of Clinical Oncology* 20.10, p. 2495-2499. doi : 10.1200/jco.2002.06.140.

Bhattacharya, A. et D. B. Dunson (2011). « Sparse Bayesian infinite factor models ». In : *Biometrika*. issn : 00063444. doi : 10.1093/biomet/asr013.

Bishop, C. (août 2006). *Pattern Recognition and Machine Learning*. en. 1re éd. Information Science and Statistics. New York, NY : Springer.

Biton, A. (2021). *MineICA : Analysis of an ICA decomposition obtained on genomics data*. R package version 1.34.0.

Biton, A. et al. (2014). « Independent Component Analysis Uncovers the Landscape of the Bladder Tumor Transcriptome and Reveals Insights into Luminal and Basal Subtypes ». In : *Cell Reports* 9.4, p. 1235-1245. issn : 2211-1247. doi : https://doi.org/10.1016/j.celrep.2014.10.035.

Blei, D. M., A. Kucukelbir et J. D. McAuliffe (2017). « Variational inference : a review for statisticians ». In : *Journal of the American Statistical Association* 112.518, p. 859-877. issn : 1537274X. doi : 10.1080/01621459.2017.1285773.

Blei, D. M. et M. I. Jordan (2006). « Variational inference for Dirichlet process mixtures ». In : *Bayesian Analysis* 1.1, p. 121-143. doi : 10.1214/06-BA104.

Bouveyron, C., P. Latouche et P. A. Mattei (2020). « Exact dimensionality selection for Bayesian PCA ». In : *Scandinavian Journal of Statistics* 47.1, p. 196-211. issn : 14679469. doi : 10.1111/sjos.12424.

Broderick, T., M. I. Jordan et J. Pitman (2012). « Beta Processes, Stick-Breaking and Power Laws ». In : *Bayesian Analysis* 7.2, p. 439-476. doi : 10.1214/12-BA715.

Buettner, F. et al. (nov. 2017). « f-scLVM : scalable and versatile factor analysis for single-cell RNA-seq ». In : *Genome Biology* 18.1. doi : 10.1186/s13059-017-1334-8.

Burstein, H. J. et al. (2021). « Customizing local and systemic therapies for women with early breast cancer : the St. Gallen International Consensus Guidelines for treatment of early breast cancer 2021 ». In : *Annals of Oncology* 32.10, p. 1216-1235. issn : 0923-7534. doi : https://doi.org/10.1016/j.annonc.2021.06.023.

Buyse, M. et S. Michiels (mai 2013). « Omics-based clinical trial designs ». In : *Current Opinion in Oncology* 25.3, p. 289-295. doi : 10.1097/cco.0b013e32835ff2fe.

Cai, J. et C. Liang (juill. 2019). « Bayesian analysis of semi-parametric Cox models with latent variables ». en. In : *Stat. Methods Med. Res.* 28.7, p. 2150-2164.

Cangelosi, R. et A. Goriely (2007). « Component retention in principal component analysis with application to cDNA microarray data ». In : *Biology Direct* 2.1, p. 2. doi : 10.1186/1745-6150-2-2.

Carvalho, C. M. et al. (2008). « High-dimensional sparse factor modeling : Applications in gene expression genomics ». In : *Journal of the American Statistical Association* 103.484, p. 1438-1456. issn : 01621459. doi : 10.1198/016214508000000869. arXiv : NIHMS150003.

Cattell, R. B. (1966). « The Scree Test For The Number Of Factors ». In : *Multivariate Behavioral Research* 1.2. PMID : 26828106, p. 245-276. doi : 10.1207/s15327906mbr0102_10. eprint : https://doi.org/10.1207/s15327906mbr0102_10.

Chapfuwa, P. et al. (fév. 2020). « Survival cluster analysis ». In : *ACM CHIL 2020 - Proceedings of the 2020 ACM Conference on Health, Inference, and Learning*, p. 60-68. doi : 10.1145/3368555.3384465.

Chen, B., M. S. Khodadoust et al. (2018). « Profiling tumor infiltrating immune cells with CIBERSORT ». In : *Methods in molecular biology (Clifton, N.J.)* 1711, p. 243-259. issn : 1940-6029. doi : 10.1007/978-1-4939-7493-1_12.

Chen, B., M. Chen et al. (2010). « Bayesian inference of the number of factors in gene-expression analysis : application to human virus challenge studies ». In : *BMC Bioinformatics* 11.1, p. 552. issn : 14712105. doi : 10.1186/1471-2105-11-552.

Clauset, A., C. R. Shalizi et M. E. J. Newman (2009). « Power-Law Distributions in Empirical Data ». In : *SIAM Review* 51.4, p. 661-703. issn : 00361445, 10957200.

Crowley, J. et A. Hoering, éd. (mars 2012). *Handbook of Statistics in Clinical Oncology*. Chapman et Hall/CRC. doi : 10.1201/b11800.

Cunningham, J. P. et Z. Ghahramani (2015). « Linear Dimensionality Reduction : Survey, Insights, and Generalizations ». In : *Journal of Machine Learning Research* 16.89, p. 2859-2900.

Dawson, J. A. et C. Kendziorski (2012). *Survival-supervised latent Dirichlet allocation models for genomic analysis of time-to-event outcomes*. arXiv : 1202.5999 [stat.ME].

Desmedt, C. et al. (avr. 2011). « Multifactorial approach to predicting resistance to anthracyclines ». In : *Journal of Clinical Oncology* 29.12, p. 1578-1586. issn : 0732-183X. doi : 10.1200/JCO.2010.31.2231.

Feng, S. et al. (déc. 2021). « Hypergraph models of biological networks to identify genes critical to pathogenic viral response ». In : *BMC Bioinformatics* 22 (1), p. 1-21. issn : 14712105. doi : 10.1186/S12859-021-04197-2/FIGURES/5.

Ferrari, S. et F. Cribari-Neto (2004). « Beta regression for modelling rates and proportions ». In : *Journal of Applied Statistics* 31.7, p. 799-815. doi : 10.1080/0266476042000214501.

Fink, D. (1997). « A Compendium of Conjugate Priors ». In.

Foo, Y. S. et H. Shim (2021). *A Comparison of Bayesian Inference Techniques for Sparse Factor Analysis*. doi : 10.48550/ARXIV.2112.11719.

Gal, J. et al. (2020). « Comparison of unsupervised machine-learning methods to identify metabolomic signatures in patients with localized breast cancer ». In : *Computational and Structural Biotechnology Journal* 18, p. 1509-1524. doi : 10.1016/j.csbj.2020.05.021.

Gelman, A. et al. (nov. 2013). *Bayesian Data Analysis*. Chapman et Hall/CRC. doi : 10.1201/b16018.

Geman, S. et D. Geman (nov. 1984). « Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.6, p. 721-741. doi : 10.1109/tpami.1984.4767596.

Gennari, A. et al. (déc. 2021). « ESMO Clinical practice guideline for the diagnosis, staging and treatment of patients with metastatic breast cancer ». In : *Annals of Oncology* 32.12, p. 1475-1495. doi : 10.1016/j.annonc.2021.09.019.

Gershman, S. J. et D. M. Blei (2012). *A tutorial on Bayesian nonparametric models*. doi : 10.1016/j.jmp.2011.08.004.

Ghahramani, Z. et T. Griffiths (2006). « Infinite latent feature models and the Indian buffet process ». In : *Advances in Neural Information Processing Systems*. Sous la dir. d'Y. Weiss, B. Schölkopf et J. Platt. T. 18. MIT Press.

Gilks, W. R. et P. Wild (1992). « Adaptive rejection sampling for gibbs sampling ». In : *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 41.2, p. 337-348. issn : 00359254, 14679876.

Green, S., J. Benedetti et J. Benedetti (mai 2012). *Clinical Trials in Oncology*. CRC Press. doi : 10.1201/b12125.

Griffiths, T. L. et Z. Ghahramani (2005). « Infinite latent feature models and the Indian buffet process ». In : *Proceedings of the 18th International Conference on Neural Information Processing Systems*. NIPS-05. Cambridge, MA, USA : MIT Press, p. 475-482.

Hatzis, C. et al. (mai 2011). « A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer ». In : *JAMA* 305.18, p. 1873. issn : 0098-7484. doi : 10.1001/jama.2011.593.

Helwig, N. E. (2018). *ica : Independent Component Analysis*. R package version 1.0-2.

Himberg, J. et A. Hyvärinen (2003). « Icasso : software for investigating the reliability of ICA estimates by clustering and visualization ». In : *2003 IEEE XIII Workshop on Neural Networks for Signal Processing (IEEE Cat. No.03TH8718)*, p. 259-268.

Hjort, N. L. (juill. 1990). « Nonparametric Bayes estimators based on beta processes in models for life history data ». In : *The Annals of Statistics* 18.3, p. 1259-1294. issn : 0090-5364. doi : 10.1214/aos/1176347749.

Hjort, N. L. et al. (2010). *Bayesian nonparametrics*. Sous la dir. de N. L. Hjort et al. Cambridge : Cambridge University Press, p. 1-299. isbn : 9780511802478. doi : 10.1017/CBO9780511802478.

Hoffman, M. D. et al. (2013). « Stochastic Variational Inference ». In : *Journal of Machine Learning Research* 14.4, p. 1303-1347.

Hore, V. et al. (août 2016). « Tensor decomposition for multiple-tissue gene expression experiments ». In : *Nature Genetics* 48.9, p. 1094-1100. doi : 10.1038/ng.3624.

Horn, J. L. (juin 1965). « A rationale and test for the number of factors in factor analysis ». In : *Psychometrika* 30.2, p. 179-185. doi : 10.1007/bf02289447.

Hyvärinen, A. (1999). « Gaussian moments for noisy independent component analysis - IEEE Signal Processing Letters ». In : *IEEE Signal Processing Letters* 6.6, p. 145-147.

Ibrahim, J. G., M.-H. Chen et D. Sinha (2001). *Bayesian Survival Analysis*. Springer New York. doi : 10.1007/978-1-4757-3447-8.

Ignatiadis, M. et al. (juin 2012). « Gene modules and response to neoadjuvant chemotherapy in breast cancer subtypes : a pooled analysis ». In : *Journal of Clinical Oncology* 30.16, p. 1996-2004. issn : 0732183X. doi : 10.1200/JCO.2011.39.5624.

Ishwaran, H. et L. F. James (2001). « Gibbs sampling methods for stick-breaking priors ». In : *Journal of the American Statistical Association* 96.453, p. 161-173. doi : 10.1198/016214501750332758. eprint : https://doi.org/10.1198/016214501750332758.

Jordan, M. I., Z. Ghahramani et et al. (1999). « An introduction to variational methods for graphical models ». In : *MACHINE LEARNING*. MIT Press, p. 183-233.

Kabán, A. (2007). « On Bayesian classification with Laplace priors ». In : *Pattern Recognition Letters* 28.10, p. 1271-1282. issn : 01678655. doi : 10.1016/j.patrec.2007.02.010.

Kairov, U. et al. (sept. 2017). « Determining the optimal number of independent components for reproducible transcriptomic data analysis ». In : *BMC Genomics* 18.1. doi : 10.1186/s12864-017-4112-9.

Kaiser, H. F. (1960). « The Application of Electronic Computers to Factor Analysis ». In : *Educational and Psychological Measurement* 20.1, p. 141-151. doi : 10.1177/001316446002000116. eprint : https://doi.org/10.1177/001316446002000116.

Knoblauch, J., J. Jewson et T. Damoulas (2019). *Generalized variational inference : three arguments for deriving new posteriors*. arXiv : 1904.02063 [stat.ML].

Knowles, D. et Z. Ghahramani (2011). « Nonparametric Bayesian sparse factor ». In : *Annals of Applied Statistics* 5.2B, p. 1534-1552. issn : 1941-7330. doi : 10.1214/10-AOAS435.

Knowles, D. A. et Z. Ghahramani (2007). « Infinite sparse factor analysis and infinite independent components analysis. » In : *ICA 2007*. T. 4666. Lecture Notes in Computer Science. Springer, p. 381-388. isbn : 978-3-540-74493-1.

Kohjima, M., T. Matsubayashi et H. Toda (2018). « Variational Bayes for mixture models with censored data ». In : *ECMLPKDD*, p. 605-620. issn : 16113349. doi : 10.1007/978-3-030-10928-8_36.

Kuhn, H. W. (1955). « The Hungarian method for the assignment problem ». In : *Naval Research Logistics Quarterly* 2 (1-2), p. 83-97. doi : https://doi.org/10.1002/nav.3800020109.

Lewicki, M. S. et T. J. Sejnowski (fév. 2000). « Learning Overcomplete Representations ». In : *Neural Computation* 12.2, p. 337-365. doi : 10.1162/089976600300015826.

Li, Q. et al. (déc. 2011). « Jetset : selecting the optimal microarray probe set to represent a gene ». In : *BMC Bioinformatics* 12.1, p. 474. issn : 14712105. doi : 10.1186/1471-2105-12-474.

Liu, E. et K. Lim (juill. 2018). « Using the Weibull accelerated failure time regression model to predict time to health events ». In : *bioRxiv*. doi : 10.1101/362186.

Liu, M. et al. (sept. 2019). « Joint models for time-to-event data and longitudinal biomarkers of high dimension ». In : *Statistics in Biosciences* 11.3, p. 614-629. doi : 10.1007/s12561-019-09256-0.

Maclaurin, D., D. Duvenaud et R. Adams (juill. 2015). « Gradient-based Hyperparameter Optimization through Reversible Learning ». In : *Proceedings of the 32nd International Conference on Machine Learning*. Sous la dir. de F. Bach et D. Blei. T. 37. Proceedings of Machine Learning Research. Lille, France : PMLR, p. 2113-2122.

Marchini, J. L., C. Heaton et B. D. Ripley (2019). *fastICA : FastICA algorithms to perform ICA and projection pursuit*. R package version 1.2-2.

Martín Abadi et al. (2015). *TensorFlow : Large-scale machine learning on heterogeneous systems*. Software available from tensorflow.org.

McCall, M. N., B. M. Bolstad et R. A. Irizarry (avr. 2010). « Frozen robust multiarray analysis (fRMA) ». In : *Biostatistics* 11.2, p. 242-253. issn : 14654644. doi : 10.1093/biostatistics/kxp059.

McCurdy, S. R., A. Molinaro et L. Pachter (2017). *A latent variable model for survival time prediction with censoring and diverse covariates*. arXiv : 1706.06995 [stat.AP].

Mitchell, T. J. et J. J. Beauchamp (déc. 1988). « Bayesian Variable Selection in Linear Regression ». In : *Journal of the American Statistical Association* 83.404, p. 1023-1032. doi : 10.1080/01621459.1988.10478694.

Mitzenmacher, M. (2003). « A Brief History of Generative Models for Power Law and Lognormal Distributions ». In : *Internet Mathematics* 1.2, p. 226-251. doi : im/1089229510.

Moran, G. E., V. Rockova et E. I. George (2018). « Variance prior forms for high-dimensional Bayesian variable selection ». In : doi : 10.48550/ARXIV.1801.03019.

Newman, M. (2005). « Power laws, Pareto distributions and Zipf's law ». In : *Contemporary Physics* 46.5, p. 323-351. doi : 10.1080/00107510500052444. eprint : https://doi.org/10.1080/00107510500052444.

Paisley, J. et L. Carin (2009). « Nonparametric factor analysis with beta process priors ». In : *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*. isbn : 9781605585161. doi : 10.1145/1553374.1553474.

Paisley, J. et M. I. Jordan (2016). *A constructive definition of the beta process*. arXiv : 1604.00685 [math.ST].

Pan, D. et al. (2019). « Bayesian proportional hazards model with latent variables ». In : *Statistical Methods in Medical Research* 28.4. PMID : 29226787, p. 986-1002. doi : 10.1177/0962280217740608.

Park, M. et al. (nov. 2020). « Integrative analysis of multi-omics data based on blockwise sparse principal components ». In : *International Journal of Molecular Sciences* 21.21, p. 8202. doi : 10.3390/ijms21218202.

Peterson, D. R. (2013). « Constructing multivariate prognostic gene signatures with censored survival data ». In : *Methods in Molecular Biology*. Springer New York, p. 85-101. doi : 10.1007/978-1-60327-337-4_6.

Petralia, F. et al. (juill. 2018). « A new method for constructing tumor specific gene co-expression networks based on samples with tumor purity heterogeneity ». In : *Bioinformatics (Oxford, England)* 34 (13), p. i528-i536. issn : 1367-4811. doi : 10.1093/BIOINFORMATICS/BTY280.

Phadia, E. G. (2013). *Prior processes and their applications*. Springer Berlin Heidelberg. doi : 10.1007/978-3-642-39280-1.

Preacher, K. J. et R. C. MacCallum (fév. 2003). « Repairing Tom Swift's electric factor analysis machine ». In : *Understanding Statistics* 2.1, p. 13-43. doi : 10.1207/s15328031us0201_02.

Qi, J. (2009). « Comparison of proportional hazards and accelerated failure time models ». Thèse de doct., p. 1-89.

Ranganath, R., S. Gerrish et D. M. Blei (2013). *Black Box Variational Inference*. arXiv : 1401.0118 [stat.ML].

Ranganath, R., D. Tran et D. M. Blei (2015). *Hierarchical Variational Models*. doi : 10.48550/ARXIV.1511.02386.

Rattray, M. et al. (déc. 2009). « Inference algorithms and learning theory for Bayesian sparse factor analysis ». In : *Journal of Physics : Conference Series* 197, p. 012002. doi : 10.1088/1742-6596/197/1/012002.

Ray, P. et al. (jan. 2014). « Bayesian joint analysis of heterogeneous genomics data ». In : *Bioinformatics* 30.10, p. 1370-1376. issn : 1367-4803. doi : 10.1093/bioinformatics/btu064.

Revelle, W. (2022). *psych : Procedures for psychological, psychometric, and personality research*. R package version 2.2.3. Northwestern University. Evanston, Illinois.

Rincourt, S.-L., S. Michiels et D. Drubay (in-revision 2022). « Complex disease individual molecular characterization using infinite sparse graphical independent component analysis ». In : *Cancer informatics*.

Robbins, H. et S. Monro (1951). « A Stochastic Approximation Method ». In : *The Annals of Mathematical Statistics* 22.3, p. 400-407. doi : 10.1214/aoms/1177729586.

Samuel Wilson (2020). « ParBayesianOptimization : parallel Bayesian optimization of hyperparameters ». In : *https ://cran.r-project.org/package=parbayesianoptimization*.

Sethuraman, J. (1994). « A constructive definition of Dirichlet priors ». In : *Statistica Sinica* 4.2, p. 639-650. issn : 10170405, 19968507.

Shabalin, A. A. et al. (mai 2008). « Merging two gene-expression studies via cross-platform normalization ». In : *Bioinformatics* 24.9, p. 1154-1160. issn : 13674803. doi : 10.1093/bioinformatics/btn083.

Sharp, K. et M. Rattray (mai 2010). « Dense Message Passing for Sparse Principal Component Analysis ». In : *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Sous la dir. d'Y. W. Teh et M. Titterington. T. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy : PMLR, p. 725-732.

Silver Spring (MD) : Food and Drug Administration (US) et Bethesda (MD) : National Institutes of Health (US) (s. d.). *BEST (Biomarkers, EndpointS, and other Tools) Resource - NCBI Bookshelf*.

Silverman, J. (2022). *RcppHungarian : Solves Minimum Cost Bipartite Matching Problems*. R package version 0.2.

Slamon, D. J. et al. (jan. 1987). « Human Breast Cancer : Correlation of Relapse and Survival with Amplification of the HER-2/ ineu/i Oncogene ». In : *Science* 235.4785, p. 177-182. doi : 10.1126/science.3798106.

Sokol, A., M. H. Maathuis et B. Falkeborg (2014). « Quantifying identifiability in independent component analysis ». In : *Electronic Journal of Statistics* 8, p. 1438-1459. issn : 1935-7524. doi : 10.1214/14-EJS932.

Sundqvist, M. (déc. 2020). « Stability and selection of the number of groups in unsupervised clustering : application to the classification of triple negative breast cancers ». Theses. Université Paris-Saclay.

Székely, G. J., M. L. Rizzo et N. K. Bakirov (déc. 2007). « Measuring and testing dependence by correlation of distances ». In : *Annals of Statistics* 35.6, p. 2769-2794. issn : 00905364. doi : 10.1214/009053607000000505.

Taherdoost, H., S. Sahibuddin et N. Jalaliyoon (2014). « Exploratory Factor Analysis ; Concepts and Theory ». In : *Advances in Applied and Pure Mathematics*. Sous la dir. de J. Balicki. T. 27. Mathematics and Computers in Science and Engineering Series. Proceedings of the 2nd International Conference on Mathematical, Computational and Statistical Sciences (MCSS '14)Proceedings of the 7th International Conference on Finite Differences, Finite Elements, Finite Volumes, Boundary Elements (F-and-B '14)Gdansk, PolandMay 15-17, 2014. WSEAS, p. 375-382.

Ternès, N., F. Rotolo et S. Michiels (jan. 2018). « biospear : an R package for biomarker selection in penalized Cox regression ». In : *Bioinformatics (Oxford, England)* 34.1, p. 112-113. doi : 10.1093/bioinformatics/btx560.

Thibaux, R. et M. I. Jordan (2007). « Hierarchical beta processes and the Indian buffet process ». In : *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS)*. issn : 15324435. doi : 10.1.1.121.455.

Tipping, M. E. et C. M. Bishop (août 1999). « Probabilistic Principal Component Analysis ». In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 61.3, p. 611-622. doi : 10.1111/1467-9868.00196.

Vidaurre, D. et al. (déc. 2013). « Bayesian sparse partial least squares ». In : *Neural Computation* 25.12, p. 3318-3339. doi : 10.1162/neco_a_00524.

Wainwright, M. J. et M. I. Jordan (2008). « Graphical Models, Exponential Families, and Variational Inference ». In : *Foundations and Trends® in Machine Learning* 1.1–2, p. 1-305. issn : 1935-8237. doi : 10.1561/2200000001.

Wang, W. et M. Stephens (2021). « Empirical Bayes Matrix Factorization ». In : *Journal of Machine Learning Research* 22.120, p. 1-40.

Wang, Z. et al. (2022). « Bayesian edge regression in undirected graphical models to characterize interpatient heterogeneity in cancer ». In : *Journal of the American Statistical Association* 0 (0), p. 1-14. doi : 10.1080/01621459.2021.2000866.

West, M. et al. (2003). « Bayesian factor regression models in the "large p, small n" paradigm ». In : *Bayesian Statistics* 7. September. Oxford University Press, p. 723-732.

Witten, D. M. et R. Tibshirani (août 2009). « Survival analysis with high-dimensional covariates ». In : *Statistical Methods in Medical Research* 19.1, p. 29-51. doi : 10.1177/0962280209105024.

Yang, S. et H. Koeppl (2018). *Collapsed Variational Inference for Nonparametric Bayesian Group Factor Analysis*. doi : 10.48550/ARXIV.1809.03566.

Yates, L. et al. (jan. 2018). « The European Society for Medical Oncology (ESMO) Precision Medicine Glossary ». In : *Annals of Oncology* 29.1, p. 30-35. doi : 10.1093/annonc/mdx707.

Yoshida, R. et M. West (2010). « Bayesian learning in sparse graphical factor models via variational mean-field annealing ». In : *Journal of Machine Learning Research* 11, p. 1771-1798. issn : 15324435.

Zhou, M. et al. (jan. 2012). « Nonparametric Bayesian Dictionary Learning for Analysis of Noisy and Incomplete Images ». In : *IEEE Transactions on Image Processing* 21.1, p. 130-144. doi : 10.1109/tip.2011.2160072.

Annexes

Annexe A

Distribution conjuguée de la famille exponentielle

A.1 Distributions

Les deux tableaux suivants viennent de Beal (2003) et présentent des informations sur une série de distributions de la famille exponentielle, qui incluent les entropies, les divergences de KL, et les moments couramment employés. Lorsqu'ils sont utilisés, les symboles tilde (par exemple $\tilde{\theta}$), désignent les paramètres d'une autre distribution de même forme. Par conséquent, $\mathrm{KL}(\tilde{\theta} \mid \theta)$ est un raccourci pour la divergence de KL entre la distribution avec le paramètre $\tilde{\theta}$ et la distribution avec le paramètre θ (en faisant la moyenne par rapport à la première distribution qui est spécifiée). Le reste de la notation est défini par distribution dans les tableaux.

Distribution	Notation & Parameters	Density function	Moments, entropy, KL-divergence, etc.
Exponential Family	$\boldsymbol{\theta} \sim ExpFam(\eta, \boldsymbol{\nu})$ number η and value $\boldsymbol{\nu}$ of pseudo-observations	$p(\boldsymbol{\theta} \mid \eta, \boldsymbol{\nu}) = \frac{1}{Z_{\eta\boldsymbol{\nu}}} g(\boldsymbol{\theta})^{\eta} e^{\boldsymbol{\phi}(\boldsymbol{\theta})^{\top}\boldsymbol{\nu}}$	$H_{\boldsymbol{\theta}} = \ln Z_{\eta\boldsymbol{\nu}} - \eta\langle \ln g(\boldsymbol{\theta})\rangle - \boldsymbol{\nu}^{\top}\langle \boldsymbol{\phi}(\boldsymbol{\theta})\rangle$
Uniform	$\theta \sim U(a, b)$ boundaries a,b with $b > a$	$p(\theta \mid a, b) = \frac{1}{b-a}, \theta \in [a, b]$	$H_{\theta} = \ln(b-a)$ $\langle\theta\rangle = \frac{a+b}{2}, \langle\theta^2\rangle - \langle\theta\rangle^2 = \frac{(b-a)^2}{12}$
Laplace	$\theta \sim Laplace(\mu, \lambda)$ μ mean λ decay scale	$p(\theta \mid \mu, \lambda) = \frac{1}{2\lambda} e^{-\frac{\lvert\theta-\mu\rvert}{\lambda}}$ $\lambda > 0$	$H_{\theta} = 1 + \ln(2\lambda)$
Multivariate normal (Gaussian)	$\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \Sigma)$ $\boldsymbol{\mu}$ mean vector Σ covariance	$p(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \Sigma) = (2\pi)^{-d/2} \lvert\Sigma\rvert^{-1/2} e^{-\frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})(\boldsymbol{\theta}-\boldsymbol{\mu})^{\top}\right]}$	$H_{\boldsymbol{\theta}} = \frac{d}{2}(\ln 2\pi e) + \frac{1}{2}\ln\lvert\Sigma\rvert$ $\mathrm{KL}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma} \mid\mid \boldsymbol{\mu}, \Sigma) = -\frac{1}{2}\left(\ln\left\lvert\tilde{\Sigma}\Sigma^{-1}\right\rvert\right.$ $\left. +\mathrm{tr}\left[I - \left[\tilde{\Sigma} + (\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu})(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu})^{\top}\right]\Sigma^{-1}\right]\ln e\right)$ $\langle\boldsymbol{\theta}\rangle = \boldsymbol{\mu}$ $\langle\boldsymbol{\theta}\boldsymbol{\theta}^{\top}\rangle = \Sigma$ $K_{\boldsymbol{\theta}} = \frac{\langle\boldsymbol{\theta}^4\rangle}{\langle\boldsymbol{\theta}^2\rangle^2} - 3 = 0$ (relative kurtosis)
Gamma	$\tau \sim Ga(\alpha, \beta)$ shape $\alpha > 0$ inv. scale $\beta > 0$	$p(\tau \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau}$	$H_{\tau} = \ln\Gamma(\alpha) - \ln\beta + (1-\alpha)\psi(\alpha) + \alpha$ $\langle\tau^n\rangle = \frac{\Gamma(\alpha+n)}{\beta^n\Gamma(\alpha)}$ $\langle(\ln\tau)^n\rangle = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\frac{\partial^n}{\partial\alpha^n}\left(\frac{\Gamma(\alpha)}{\beta^{\alpha}}\right)$ $\langle\tau\rangle = \alpha/\beta$ $\langle\tau^2\rangle - \langle\tau\rangle^2 = \alpha/\beta^2$ $\langle\ln\tau\rangle = \psi(\alpha) - \ln\beta$ $\mathrm{KL}(\tilde{\alpha}, \tilde{\beta} \mid\mid \alpha, \beta) = \tilde{\alpha}\ln\tilde{\beta} - \alpha\ln\beta - \ln\frac{\Gamma(\tilde{\alpha})}{\Gamma(\alpha)}$ $+(\tilde{\alpha} - \alpha)(\psi(\tilde{\alpha}) - \ln\tilde{\beta}) - \tilde{\alpha}(1 - \frac{\beta}{\tilde{\beta}})$

Figure A.1 – Extrait de Beal (2003), appendix A

Distribution	Notation & Parameters	Density function	Moments, entropy, KL-divergence, etc.
Wishart	$W \sim Wishart_\nu(S)$ deg. of freedom ν precision matrix S	$p(W \mid \nu, S) = \frac{1}{Z_{\nu S}} \lvert W\rvert^{(\nu-k-1)/2} e^{-\frac{1}{2}\mathrm{tr}\left[S^{-1}W\right]}$ $Z_{\nu S} = 2^{\nu k/2}\pi^{k(k-1)/4} \lvert S\rvert^{\nu/2} \prod_{i=1}^{k} \Gamma\left(\frac{\nu+1-i}{2}\right)$	$H_W = \ln Z_{\nu S} - \frac{\nu-k-1}{2}\langle \ln \lvert W\rvert\rangle + \frac{1}{2}\nu k$ $\langle W\rangle = \nu S$ $\langle \ln \lvert W\rvert\rangle = \sum_{i=1}^{k} \psi\left(\frac{\nu+1-i}{2}\right) + k\ln 2 + \ln \lvert S\rvert$ $\mathrm{KL}(\tilde{\nu}, \tilde{S}\|\nu, S) = \ln \frac{Z_{\nu S}}{Z_{\tilde{\nu}\tilde{S}}}$ $+\frac{\tilde{\nu}-\nu}{2}\langle \ln \lvert W\rvert\rangle_{\tilde{Q}} + \frac{1}{2}\tilde{\nu}\,\mathrm{tr}\left[S^{-1}\tilde{S} - \mathrm{I}\right]$
Inverse-Wishart	$W \sim Inv\text{–}Wishart_\nu(S^{-1})$ deg. of freedom ν covariance matrix S	$p(W \mid \nu, S^{-1}) = \frac{1}{Z} \lvert W\rvert^{-(\nu+k+1)/2} e^{-\frac{1}{2}\mathrm{tr}\left[SW^{-1}\right]}$ $Z = 2^{\nu k/2}\pi^{k(k-1)/4} \prod_{i=1}^{k} \Gamma\left(\frac{\nu+1-i}{2}\right) \times \lvert S\rvert^{-\nu/2}$	$\langle W\rangle = (\nu-k-1)^{-1}S$
Student-t (1)	$\theta \sim t_\nu(\mu, \sigma^2)$ deg. of freedom $\nu > 0$ mean μ; scale $\sigma > 0$	$p(\theta \mid \nu, \mu, \sigma^2) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}\sigma} \left(1 + \frac{1}{\nu}\left(\frac{\theta-\mu}{\sigma}\right)^2\right)^{-(\nu+1)/2}$	$\langle\theta\rangle = \mu$, for $\nu > 1$ $\langle\theta^2\rangle - \langle\theta\rangle^2 = \frac{\nu}{\nu-2}\sigma^2$, for $\nu > 2$
Student-t (2)	$\theta \sim t(\mu, \alpha, \beta)$ shape $\alpha > 0$; mean μ scale^2 $\beta > 0$	$p(\theta \mid \mu, \alpha, \beta) = \frac{\Gamma(\alpha+1/2)}{\Gamma(\alpha)\sqrt{2\pi\beta}} \left(1 + \frac{(\theta-\mu)^2}{2\beta}\right)^{-(\alpha+1/2)}$	$H_\theta = \left[\psi(\alpha+\frac{1}{2}) - \psi(\alpha)\right](\alpha+\frac{1}{2})$ $+\ln\sqrt{2\beta}B(\frac{1}{2}, \alpha)$ $K_\theta = \frac{3}{\alpha-2}$ (relative to Gaussian) equiv. $\alpha \to \frac{\nu}{2}; \beta \to \frac{\nu}{2}\sigma^2$
Multivariate Student-t	$\boldsymbol{\theta} \sim t_\nu(\boldsymbol{\mu}, \Sigma)$ deg. of freedom $\nu > 0$ mean $\boldsymbol{\mu}$; scale^2 matrix Σ	$p(\boldsymbol{\theta} \mid \nu, \boldsymbol{\mu}, \Sigma) = \frac{1}{Z}\left(1 + \frac{1}{\nu}\mathrm{tr}\left[\Sigma^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})(\boldsymbol{\theta}-\boldsymbol{\mu})^\top\right]\right)^{-(\nu+d)/2}$ $Z = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)(\nu\pi)^{d/2}} \lvert\Sigma\rvert^{-1/2}$	$\langle\boldsymbol{\theta}\rangle = \boldsymbol{\mu}$, for $\nu > 1$ $\langle\boldsymbol{\theta}\boldsymbol{\theta}^\top\rangle - \langle\boldsymbol{\theta}\rangle\langle\boldsymbol{\theta}\rangle^\top = \frac{\nu}{\nu-2}\Sigma$, for $\nu > 2$
Beta	$\theta \sim Beta(\alpha, \beta)$ prior sample sizes $\alpha > 0, \beta > 0$	$p(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$ $\theta \in [0, 1]$	See Dirichlet with $k = 2$
Dirichlet	$\boldsymbol{\pi} \sim Dir(\boldsymbol{\alpha})$ prior sample sizes $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_k\}$ $\alpha_j > 0; \alpha_0 = \sum_{j=1}^{k} \alpha_j$	$p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)} \pi_1^{\alpha_1-1} \cdots \pi_k^{\alpha_k-1}$ $\pi_1, \ldots, \pi_k \geq 0; \sum_{j=1}^{k} \pi_j = 1$	$\langle\boldsymbol{\pi}\rangle = \boldsymbol{\alpha}/\alpha_0$ $\langle\boldsymbol{\pi}\boldsymbol{\pi}^\top\rangle - \langle\boldsymbol{\pi}\rangle\langle\boldsymbol{\pi}\rangle^\top = \frac{\alpha_0\mathrm{diag}(\boldsymbol{\alpha}) - \boldsymbol{\alpha}\boldsymbol{\alpha}^\top}{\alpha_0^2(\alpha_0+1)}$ $\langle\ln\pi_j\rangle = \psi(\alpha_j) - \psi(\alpha_0)$ $\mathrm{KL}(\tilde{\boldsymbol{\alpha}}\|\boldsymbol{\alpha}) = \ln\frac{\Gamma(\tilde{\alpha}_0)}{\Gamma(\alpha_0)} - \sum_{j=1}^{k}\Big[\ln\frac{\Gamma(\tilde{\alpha}_j)}{\Gamma(\alpha_j)}$ $-(\tilde{\alpha}_j - \alpha_j)\left(\psi(\tilde{\alpha}_j) - \psi(\tilde{\alpha}_0)\right)\Big]$

Figure A.2 – Extrait de Beal (2003), appendix A (suite)

A.2 Exemple de relation conjuguée entre les distributions a priori, la vraisemblance et les distributions a posteriori

Table A.1 – Exemple de relation conjuguée entre les distributions a priori, la vraisemblance et les distributions a posteriori

Vraisemblance	Paramètre du modèle estimé	Prior conjugué	Hyper-paramètres du prior	Posterior	Hyperparamètres du posterior	Interprétation des hyper-paramètres
Bernoulli	p (probabilité)	Beta	$\alpha,\ \beta \in \mathbb{R}$	Beta	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + n - \sum_{i=1}^{n} x_i$	α succès, β échecs
Normale avec la précision connue τ	μ (moyenne)	Normale	$\mu_0,\ \tau_0^{-1}$	Normale	$\frac{\tau_0\mu_0 + \tau\sum_{i=1}^{n} x_i}{\tau_0 + n\tau},\ (\tau_0 + n\tau)^{-1}$	la moyenne a été estimée à partir des observations avec une précision τ_0 et avec une moyenne d'échantillon μ_0
Normale avec la moyenne connue μ	τ (précision)	Gamma	$\alpha,\ \beta$	Gamma	$\alpha + \frac{n}{2},\ \beta + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}$	La précision a été estimée à partir de 2α observations avec une variance d'échantillon β/α (c'est-à-dire avec une somme des écarts au carré 2β, où les écarts sont par rapport à une moyenne connue μ).
Log-normale	Comme pour la distribution normale après avoir appliqué le logarithme aux données pour estimer les posterieurs des hyperparamètres.					
Gamma avec le paramètre de forme connu α	β (taux)	Gamma	$\alpha_0,\ \beta_0$	Gamma	$\alpha_0 + n\alpha,\ \beta_0 + \sum_{i=1}^{n} x_i$	α_0/α observations avec la somme β_0
Gamma avec le paramètre de taux connu β	α (forme)	$\propto \frac{a^{\alpha-1}\beta^{\alpha c}}{\Gamma(\alpha)^b}$	$a,\ b,\ c$	sans-nom	$a\prod_{i=1}^{n} x_i,\ b + n,\ c + n$	b ou c observations (b pour estimer α, c pour estimer β) avec le produit a

Annexe B

Démonstration des équations du modèle isgICA

B.1 Notions - Rappels

La vraisemblance est définie telle que :

$$\begin{aligned} p(x|\boldsymbol{Z},\boldsymbol{W},\boldsymbol{\Phi},\Psi,-) &= \frac{\Psi^{\frac{-1}{2}}}{(2\pi)^{\frac{N}{2}}} exp(-\Psi\frac{||X-\Phi(W\circ Z)||_2^2}{2}) \\ &= \frac{\Psi^{\frac{-1}{2}}}{(2\pi)^{\frac{N}{2}}} exp(-\frac{1}{2}(X-\Phi(W\circ Z))^T\Psi(X-\Phi(W\circ Z))) \end{aligned} \tag{B.1}$$

Pour rappel nous avons choisi $\Psi = \tau_E = 1$. Pour la suite nous avons : la matrice $\widetilde{Z}_i$:

$$\widetilde{Z}_i := [z_i, ..., z_i] \text{ vecteur } z_i \text{ répété P fois} \tag{B.2}$$

et la matrice $\mathbf{X}_{j,i}^{-k}$:

$$\mathbf{X}_{j,i}^{-k} = x_{j,i} - \sum_{l=1,l\neq k}^{K} \Phi_{j,l} z_{l,i} w_{l,i}. \tag{B.3}$$

Tous les paramètres variationnels ($<\pi_{k^*}>$, $<z^*_{k^*,i}>$, $<\Phi_{j,k}>$, $<W_i>$, $<\tau_{\Phi_{j,k}}>$ et $<\tau_W>$) sont définis sans le symbole de l'espérance variationnelle $<\bullet>$.

B.2 Démonstration de π_k

Nous utilisons la loi bêta, comme prior conjugué de la loi Bernoulli de la matrice $\mathbf{Z}$, avec un prior beta(a_0, b_0) alors le posterior est de la forme beta($a_0 + \sum_{i=1}^{n}$ succès, $b_0 + n - \sum_{i=1}^{n}$ succès). Le posterior variationnel s'écrit :

$$\begin{aligned} q(\pi_k \mid -) &= Beta(\pi_{k,i} \mid a, b) \\ a &= \sum_{i=1}^{n} z_{k,i} + \frac{a_0}{K} \\ b &= n + \frac{b_0(K-1)}{K} - \sum_{i=1}^{n} z_{k,i} \end{aligned} \tag{B.4}$$

Espérance de $\ln(\pi_k)$: $\mathbb{E}[\ln(\pi_k)] = \psi(a) - \psi(a+b+N)$.
Espérance de $\ln(1-\pi_k)$: $\mathbb{E}[\ln(1-\pi_k)] = \psi(b) - \psi(a+b+N)$.

B.3 Démonstration de $z_{k,i}$

Le prior de $\mathbf{Z}$ est définie par une loi de Bernoulli :

$$q(z \mid \pi_k) = P(Z = z_{k,i}) = \pi_k^{z_{k,i}}(1-\pi_k^{1-z_{k,i}}), \tag{B.5}$$

et le posterior variationnel s'ecrit :

$$q(z_{k,i} \mid X_i, \boldsymbol{Z_{-k,i}}, \boldsymbol{W}, \boldsymbol{\Phi}, \Psi) \tag{B.6}$$

$$\begin{aligned} &= \prod_{j=1}^{p} q(z_{k,i}|x_{j,i}, W_i, \boldsymbol{\Phi}, \boldsymbol{\Psi}, z_{-k,i}, \pi_k) \\ &= \prod_{j=1}^{p} \frac{q(z_{k,i}, x_{j,i}|W_i, \boldsymbol{\Phi}, \boldsymbol{\Psi}, z_{-k,i}, \pi_k)}{p(x_{j,i}|W_i, \boldsymbol{\Phi}, \boldsymbol{\Psi}, z_{-k,i}, \pi_k)} \\ &= \prod_{j=1}^{p} \frac{p(x_{j,i} \mid W_i, \boldsymbol{\Phi}, \boldsymbol{\Psi}, z_{-k,i}, z_{k,i}, \pi_k) q(z_{k,i}|\pi_k)}{p(x_{j,i} \mid W_i, \boldsymbol{\Phi}, \boldsymbol{\Psi}, z_{-k,i}, \pi_k)} \\ &\propto \prod_{j=1}^{p} p(x_{j,i} \mid W_i, \boldsymbol{\Phi}, \boldsymbol{\Psi}, z_{-k,i}, z_{k,i}, \pi_k) q(z_{k,i}|\pi_k) \\ \propto &\frac{\Psi^{\frac{-1}{2}}}{(2\pi)^{\frac{N}{2}}} \exp(-\frac{1}{2}(X_i - \Phi(W_i \circ Z_i))^T \Psi (X_i - \Phi(W_i \circ Z_i))) \exp(\mathbb{E}[\ln(\pi_k)]) \\ \propto &\exp(-\frac{1}{2}(X_i - \Phi(W_i \circ Z_i))^T \Psi (X_i - \Phi(W_i \circ Z_i))) \exp(\mathbb{E}[\ln(\pi_k)]) \end{aligned} \tag{B.7}$$

$$\propto \exp(-\frac{1}{2}(X_i - \sum_{l=1}^{K} \Phi_l w_{l,i} z_{l,i}))^T \Psi (X_i - \sum_{l=1}^{K} \Phi_l w_{l,i} z_{l,i}))) \exp(\mathbb{E}[\ln(\pi_k)])$$

$$\propto \exp(-\frac{1}{2}(X_i - \left(\Phi_k w_{k,i} z_{k,i} + \sum_{l=1, l\neq k}^{K} \Phi_l w_{l,i} z_{l,i}\right))^T \Psi (X_i - \left(\Phi_k w_{k,i} z_{k,i} + \sum_{l=1, l\neq k}^{K} \Phi_l w_{l,i} z_{l,i}\right)))) \exp(\mathbb{E}[\ln(\pi_k)])$$

$$\propto \exp(-\frac{1}{2}(X_i - \Phi_k w_{k,i} z_{k,i} - \sum_{l=1, l\neq k}^{K} \Phi_l w_{l,i} z_{l,i}))^T \Psi (X_i - \Phi_k w_{k,i} z_{k,i} - \sum_{l=1, l\neq k}^{K} \Phi_l w_{l,i} z_{l,i}))) \exp(\mathbb{E}[\ln(\pi_k)])$$

$$\begin{aligned}
\propto \exp(-\frac{1}{2}(&X_i^T \mathbf{\Psi} X_i + (\Phi_l w_{k,i} z_{k,i})^T \mathbf{\Psi} (\Phi_l w_{k,i} z_{k,i}) + (\sum_{l=1, l\neq k}^{K} \Phi_l w_{l,i} z_{l,i})^T \mathbf{\Psi} (\sum_{l=1, l\neq k}^{K} \Phi_l w_{l,i} z_{l,i}) \\
&- 2(\sum_{l=1, l\neq k}^{K} \Phi_l w_{l,i} z_{l,i})^T \mathbf{\Psi} (X_i) \sum_{l=1, l\neq k}^{K} \Phi_l w_{l,i} z_{l,i} - 2(\Phi_l w_{k,i} z_{k,i})^T \mathbf{\Psi} (X_i) \\
&+ 2(\Phi_l w_{k,i} z_{k,i})^T \mathbf{\Psi} (\sum_{l=1, l\neq k}^{K} \Phi_l w_{l,i} z_{l,i}))) \exp(\mathbb{E}[\ln(\pi_k)])
\end{aligned}$$

$$\begin{aligned}
\propto \exp(-\frac{1}{2}(&X_i^T \mathbf{\Psi} X_i + (\Phi_l w_{k,i} z_{k,i})^T \mathbf{\Psi} (\Phi_l w_{k,i} z_{k,i}) + (\sum_{l=1, l\neq k}^{K} \Phi_l w_{l,i} z_{l,i})^T \mathbf{\Psi} (\sum_{l=1, l\neq k}^{K} \Phi_l w_{l,i} z_{l,i}) \\
&- 2(\sum_{l=1, l\neq k}^{K} \Phi_l w_{l,i} z_{l,i})^T \mathbf{\Psi} (X_i) \sum_{l=1, l\neq k}^{K} \Phi_l w_{l,i} z_{l,i} \\
&- 2(\Phi_l w_{k,i} z_{k,i})^T \mathbf{\Psi} (X_i - \sum_{l=1, l\neq k}^{K} \Phi_l w_{l,i} z_{l,i}))) \exp(\mathbb{E}[\ln(\pi_k)])
\end{aligned}$$

$$\begin{aligned}
\propto \exp(-\frac{1}{2}(&(\Phi_l w_{k,i} z_{k,i})^T \mathbf{\Psi} (\Phi_l w_{k,i} z_{k,i}) \\
&- 2(\Phi_l w_{k,i} z_{k,i})^T \mathbf{\Psi} (X_i - \sum_{l=1, l\neq k}^{K} \Phi_l w_{l,i} z_{l,i}) \\
&+ (X_i - \sum_{l=1, l\neq k}^{K} \Phi_l w_{l,i} z_{l,i})^T \mathbf{\Psi} (X_i - \sum_{l=1, l\neq k}^{K} \Phi_l w_{l,i} z_{l,i}))) \exp(\mathbb{E}[\ln(\pi_k)])
\end{aligned}$$

$$\begin{aligned}
\propto \exp(-\frac{1}{2}(&(\Phi_l w_{k,i} z_{k,i})^T \mathbf{\Psi} (\Phi_l w_{k,i} z_{k,i}) \\
&- 2(\Phi_l w_{k,i} z_{k,i})^T \mathbf{\Psi} (X_i - \sum_{l=1, l\neq k}^{K} \Phi_l w_{l,i} z_{l,i}) \\
&+ (\mathbf{X}_{j,i}^{-k})^T \mathbf{\Psi} \mathbf{X}_{j,i}^{-k})) \exp(\mathbb{E}[\ln(\pi_k)])
\end{aligned}$$

$(\mathbf{X}_{j,i}^{-k})^T \mathbf{\Psi} \mathbf{X}_{j,i}^{-k}$ ne dépend pas de $z_{k,i}$, donc il peut être retirer de l'équation :

$$\propto exp(-\frac{1}{2}((\Phi_l w_{k,i} z_{k,i})^T \mathbf{\Psi}(\Phi_l w_{k,i} z_{k,i}) - 2(\Phi_l w_{k,i} z_{k,i})^T \mathbf{\Psi}(X_i - \sum_{l=1, l\neq k}^{K} \Phi_l w_{l,i} z_{l,i}))) \exp(\mathbb{E}[\ln(\pi_k)]). \tag{B.8}$$

La mise à jour variationnelle du paramètre $<z^*_{k^*,i}>$ s'écrit :

$$\begin{aligned} q(z^*_{k^*,i}|-) &= Bernoulli(z^*_{k^*,i}; \pi_{k^*}) \\ &= \frac{q(z_{k^*,i}=1|-)}{q(z_{k^*,i}=1|-) + q(z_{k^*,i}=0|-)}. \end{aligned} \tag{B.9}$$

avec $z_{k,i}$ ==1 :

$$\begin{aligned} q(z_{k,i}=1|\boldsymbol{X}, \boldsymbol{Z_{-k,i}}, \boldsymbol{W}, \mathbf{\Phi}, \tau_E) &\propto \exp(\mathbb{E}[\ln(\pi_k)] \ -\frac{1}{2}(\Phi_k^T \mathbf{\Psi} \Phi_k w_{k,i}^2 - 2\Phi_k^T \mathbf{\Psi} \mathbf{X}_i^{-k} w_{k,i})) \\ &\propto \exp((\psi(a) - \psi(a+b+N)) \ -\frac{1}{2}(\Phi_k^T \mathbf{\Psi} \Phi_k w_{k,i}^2 - 2\Phi_k^T \mathbf{\Psi} \mathbf{X}_i^{-k} w_{k,i})) \end{aligned} \tag{B.10}$$

et $z_{k,i}$ ==0

$$\begin{aligned} q(z_{k,i}=0|\boldsymbol{X}, \boldsymbol{Z_{-k,i}}, \boldsymbol{W}, \mathbf{\Phi}, \tau_E) &\propto \exp(\mathbb{E}[\ln(1-\pi_k)]) \\ &\propto \exp((\psi(b) - \psi(a+b+N)). \end{aligned} \tag{B.11}$$

B.4 Démonstration de $\Phi_{j,k}$

$$q(\Phi_{j,k}|\boldsymbol{X}, \boldsymbol{\Phi_{-j,k}}, \boldsymbol{W}, \boldsymbol{Z}, \Psi, -) \tag{B.12}$$

$$\begin{aligned}
&= \mathcal{N}(\Phi_{j,k}; \mu_{\Phi_{j,k}}, \tau^{-1}_{\Phi_{j,k}}) \\
&= \prod_{i=1}^{n} q(\Phi_{j,k}|x_{j,i}, \boldsymbol{\Phi_{-j,k}}, z_{k,i}, w_{k,i}, \Psi, \mu_{\Phi_{j,k}}, \tau^{-1}_{\Phi_{j,k}}) \\
&= \prod_{i=1}^{n} \frac{q(\Phi_{j,k}, x_{j,i}|\boldsymbol{\Phi_{-j,k}}, z_{k,i}, w_{k,i}, \Psi, \mu_{\Phi_{j,k}}, \tau^{-1}_{\Phi_{j,k}})}{p(x_{j,i}|\boldsymbol{\Phi_{-j,k}}, z_{k,i}, w_{k,i}, \Psi, \mu_{\Phi_{j,k}}, \tau^{-1}_{\Phi_{j,k}})} \\
&= \prod_{i=1}^{n} \frac{p(x_{j,i}|\boldsymbol{\Phi_{-j,k}}, \Phi_{j,k}, z_{k,i}, w_{k,i}, \Psi, \mu_{\Phi_{j,k}}, \tau^{-1}_{\Phi_{j,k}}) q(\Phi_{j,k}|\mu_{\Phi_{j,k}}, \tau^{-1}_{\Phi_{j,k}})}{p(x_{j,i}|\boldsymbol{\Phi_{-j,k}}, z_{k,i}, w_{k,i}, \Psi, \mu_{\Phi_{j,k}}, \tau^{-1}_{\Phi_{j,k}})} \\
&\propto \prod_{i=1}^{n} p(x_{j,i}|\boldsymbol{\Phi_{-j,k}}, \Phi_{j,k}, z_{k,i}, w_{k,i}, \Psi, \mu_{\Phi_{j,k}}, \tau^{-1}_{\Phi_{j,k}}) q(\Phi_{j,k}|\mu_{\Phi_{j,k}}, \tau^{-1}_{\Phi_{j,k}})
\end{aligned} \tag{B.13}$$

$$\begin{aligned}
&\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\tau_{E_j}} exp(-\frac{\tau_{E_j}}{2}(x_{j,i} - \sum_{l=1}^{K} \Phi_{j,l} w_{l,i} z_{l,i})^2) \frac{1}{\sqrt{2\pi}\tau_{\Phi_{j,k}}} exp(-\frac{\tau_{\Phi_{j,k}}}{2}(\Phi_{j,k} - \mu_{\Phi_{j,k}})^2) \\
&\propto exp(\sum_{i=1}^{n} -\frac{\tau_{E_j}}{2}(x_{j,i} - \sum_{l=1}^{K} \Phi_{j,l} w_{l,i} z_{l,i})^2) exp(-\frac{\tau_{\Phi_{j,k}}}{2}(\Phi_{j,k} - \mu_{\Phi_{j,k}})^2) \\
&\propto exp(-\frac{1}{2}(\tau_{E_j} \sum_{i=1}^{n} (x_{j,i} - \sum_{l=1}^{K} \Phi_{j,l} w_{l,i} z_{l,i})^2 + \tau_{\Phi_{j,k}}(\Phi_{j,k} - \mu_{\Phi_{j,k}})^2)) \\
&\propto exp(-\frac{1}{2}(\tau_{E_j} \sum_{i=1}^{n} (x_{j,i} - \sum_{l=1,l\neq k}^{K} \Phi_{j,l} w_{l,i} z_{l,i} - \Phi_{j,k} w_{k,i} z_{k,i})^2 + \tau_{\Phi_{j,k}}(\Phi_{j,k} - \mu_{\Phi_{j,k}})^2))
\end{aligned}$$

$$\begin{aligned}
\propto \; exp(& -\frac{1}{2}(\tau_{E_j} \sum_{i=1}^{n} (x_{j,i}^2 + \Phi_{j,k}^2 w_{k,i}^2 z_{k,i}^2 + (\sum_{l=1,l\neq k}^{K} \Phi_{j,l} w_{l,i} z_{l,i})^2 - 2x_{j,i}\Phi_{j,k} w_{k,i} z_{k,i} \\
& - 2x_{j,i}(\sum_{l=1,l\neq k}^{K} \Phi_{j,l} w_{l,i} z_{l,i}) + 2\Phi_{j,k} w_{k,i} z_{k,i} (\sum_{l=1,l\neq k}^{K} \Phi_{j,l} w_{l,i} z_{l,i})) \\
& + \tau_{\Phi_{j,k}}(\Phi_{j,k}^2 + \mu_{\Phi_{j,k}}^2 - 2\Phi_{j,k}\mu_{\Phi_{j,k}}))) \\
\propto \; exp(& -\frac{1}{2}(\Phi_{j,k}^2 (\tau_{E_j} \sum_{i=1}^{n} w_{k,i}^2 z_{k,i}^2 + \tau_{\Phi_{j,k}}) \\
& - 2\Phi_{j,k}(\tau_{E_j} \sum_{i=1}^{n} (x_{j,i} w_{k,i} z_{k,i} - w_{k,i} z_{k,i} (\sum_{l=1,l\neq k}^{K} \Phi_{j,l} w_{l,i} z_{l,i})) + \mu_{\Phi_{j,k}} \tau_{\Phi_{j,k}})
\end{aligned} \tag{B.14}$$

$$
\begin{aligned}
& \tau_{E_j} \sum_{i=1}^{n} (x_{j,i}^2 + (\sum_{l=1,l\neq k}^{K} \Phi_{j,l} w_{l,i} z_{l,i})^2 - 2x_{j,i}(\sum_{l=1,l\neq k}^{K} \Phi_{j,l} w_{l,i} z_{l,i})) + \mu_{\Phi_{j,k}}^2 \tau_{\Phi_{j,k}})) \\
\propto \; & exp(-\frac{1}{2}(\Phi_{j,k}^2 (\tau_{E_j} \sum_{i=1}^{n} w_{k,i}^2 z_{k,i}^2 + \tau_{\Phi_{j,k}}) \\
& - 2\Phi_{j,k}(\tau_{E_j} \sum_{i=1}^{n} (w_{k,i} z_{k,i} (x_{j,i} - (\sum_{l=1,l\neq k}^{K} \Phi_{j,l} w_{l,i} z_{l,i}))) + \mu_{\Phi_{j,k}} \tau_{\Phi_{j,k}}) \\
& \tau_{E_j} \sum_{i=1}^{n} (x_{j,i} - (\sum_{l=1,l\neq k}^{K} \Phi_{j,l} w_{l,i} z_{l,i}))^2 + \mu_{\Phi_{j,k}}^2 \tau_{\Phi_{j,k}})) \\
\propto \; & exp(-\frac{1}{2}(\Phi_{j,k}^2 (\tau_{E_j} \sum_{i=1}^{n} w_{k,i}^2 z_{k,i}^2 + \tau_{\Phi_{j,k}}) \\
& - 2\Phi_{j,k}(\tau_{E_j} \sum_{i=1}^{n} (w_{k,i} z_{k,i} \mathbf{X}_{j,i}^{-k}) + \mu_{\Phi_{j,k}} \tau_{\Phi_{j,k}}) \\
& \tau_{E_j} \sum_{i=1}^{n} (\mathbf{X}_{j,i}^{-k})^2 + \mu_{\Phi_{j,k}}^2 \tau_{\Phi_{j,k}}))
\end{aligned}
\tag{B.15}
$$

Comme le prior de la moyenne est nulle, on a au final (Rappel $\mathbf{X}_{j,i}^{-k} = x_{j,i} - \sum_{l=1,l\neq k}^{K} \Phi_{j,l} z_{l,i} w_{l,i}$) :

$$
\begin{aligned}
\tau_{\Phi_{j,k}} &= \tau_{E_j} \sum_{i=1}^{n} w_{k,i}^2 z_{k,i}^2 + \tau_{\Phi_{j,k}} \\
\mu_{\Phi_{j,k}} &= \tau_{\Phi_{j,k}}^{-1} (\sum_{i=1}^{n} \tau_{E_j} w_{k,i} z_{k,i} \mathbf{X}_{j,i}^{-k})
\end{aligned}
\tag{B.16}
$$

B.5 Démonstration de w_i

0n définit pour la suite :

(1) $\mathbf{\Phi}(W_i \circ Z_i) \Leftrightarrow (\mathbf{\Phi} \circ \widetilde{Z}_i^T)W_i$
(2) $(\mathbf{\Phi}(W_i \circ Z_i))^T \Leftrightarrow W_i^T(\mathbf{\Phi}^T \circ \widetilde{Z}_i)$.

Note : $\boldsymbol{\mu_W}$ est un vecteur de taille K de valeur μ_W.

$$q(W_i|-) = \mathcal{N}(W_i; \mu_W, \tau_W^{-1}) \tag{B.17}$$

$$\begin{aligned}
&= \prod_{j=1}^{P} q(W_i|x_{j,i}, \mathbf{\Phi_{-j}}, \boldsymbol{W_{-i}}, \mathbf{\Psi}, Z_i, \mu_W, \tau_W^{-1}) \\
&= \prod_{j=1}^{P} \frac{q(W_i, x_{j,i}|\mathbf{\Phi_{-j}}, \boldsymbol{W_{-i}}, \mathbf{\Psi}, Z_i, \mu_W, \tau_W^{-1})}{p(x_{j,i}|\mathbf{\Phi_{-j}}, \boldsymbol{W_{-i}}, \mathbf{\Psi}, Z_i, \mu_W, \tau_W^{-1})} \\
&= \prod_{j=1}^{P} \frac{p(x_{j,i}|\mathbf{\Phi_{-j}}, \boldsymbol{W}, \mathbf{\Psi}, Z_i)q(W_i|\mu_W, \tau_W^{-1})}{p(x_{j,i}|\mathbf{\Phi_{-j}}, \boldsymbol{W_{-i}}, \mathbf{\Psi}, Z_i, \mu_W, \tau_W^{-1})} \\
&\propto \prod_{j=1}^{P} p(x_{j,i}|\mathbf{\Phi_{-j}}, \boldsymbol{W}, \mathbf{\Psi}, Z_i)q(W_i|\mu_W, \tau_W^{-1}) \\
&\propto \frac{\Psi}{(2\pi)^{\frac{P}{2}}} exp(-\frac{1}{2}(X_i - \mathbf{\Phi}(W_i \circ Z_i))^T \mathbf{\Psi}(X_i - \mathbf{\Phi}(W_i \circ Z_i))) \frac{\tau_W}{(2\pi)^{\frac{K}{2}}} exp(-\frac{1}{2}(W_i - \boldsymbol{\mu_W})^T \\
&(\tau_W I_K)(W_i - \boldsymbol{\mu_W})) \\
&\propto exp(-\frac{1}{2}((X_i - \mathbf{\Phi}(W_i \circ Z_i))^T \mathbf{\Psi}(X_i - \mathbf{\Phi}(W_i \circ Z_i)) + (W_i - \boldsymbol{\mu_W})^T(\tau_W I_K)(W_i - \boldsymbol{\mu_W}))) \\
\propto exp(-\frac{1}{2}(&X_i^T \mathbf{\Psi} X_i - 2(\mathbf{\Phi}(W_i \circ Z_i))^T \mathbf{\Psi} X_i + (\mathbf{\Phi}(W_i \circ Z_i))^T \mathbf{\Psi}(\mathbf{\Phi}(W_i \circ Z_i)) \\
&+ W_i^T(\tau_W I_K)W_i + (\tau_W^2 I_K \boldsymbol{\mu}_{\boldsymbol{W}}^2) - 2\tau_W W_i^T I_K \boldsymbol{\mu_W})) \\
\propto exp(-\frac{1}{2}(&X_i^T \mathbf{\Psi} X_i - 2W_i^T(\mathbf{\Phi}^T \circ \widetilde{Z}_i)\mathbf{\Psi} X_i + W_i^T(\mathbf{\Phi}^T \circ \widetilde{Z}_i)\mathbf{\Psi}(\mathbf{\Phi} \circ \widetilde{Z}_i^T)W_i \\
&+ W_i^T(\tau_W I_K)W_i + (\tau_W^2 I_K \boldsymbol{\mu}_{\boldsymbol{W}}^2) - 2\tau_W W_i^T I_K \boldsymbol{\mu_W})) \\
\propto exp(-\frac{1}{2}(&W_i^T((\mathbf{\Phi}^T \circ \widetilde{Z}_i)\mathbf{\Psi}(\mathbf{\Phi} \circ \widetilde{Z}_i^T) + \tau_W I_K)W_i \\
&- 2W_i^T((\mathbf{\Phi}^T \circ \widetilde{Z}_i)\mathbf{\Psi} X_i + \tau_W I_K \boldsymbol{\mu_W}) \\
&X_i^T \mathbf{\Psi} X_i + (\tau_W^2 I_K \boldsymbol{\mu}_{\boldsymbol{W}}^2)))
\end{aligned} \tag{B.18}$$

Comme le prior de la moyenne est nulle, on a au final :

$$
\begin{aligned}
q(W_i|-) &= \mathcal{N}(W_i; \mu_{W_i}, \tau_{W_i}^{-1}) \\
\tau_{W_i} &= (\Phi^T \circ \widetilde{Z}_i)\mathbf{\Psi}(\Phi \circ \widetilde{Z}_i^T) + \tau_W I_K \\
\mu_{W_i} &= \tau_{W_i}^{-1}(\Phi \circ \widetilde{Z}_i)\mathbf{\Psi} X_i
\end{aligned}
\tag{B.19}
$$

B.6 Démonstration de $\tau_{\Phi_{j,k}}$

Utilisation de la loi gamma (prior pour la précision conjuguée à la loi normale) avec un prior gamma(a,b) alors le posterior est de la forme gamma($a + \frac{n}{2}$, $b + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}$)

$$
\begin{aligned}
q(\tau_{\Phi_{j,k}}|-) &= Gamma(\tau_{\Phi_{j,k}}; c_{j,k}, d_{j,k}) \\
c_{j,k} &= c_0 + \frac{1}{2} \\
d_{j,k} &= d_0 + \frac{1}{2}\Phi_{j,k}^2
\end{aligned}
\tag{B.20}
$$

B.7 Démonstration de τ_W

$$
\begin{aligned}
q(\tau_W|-) &= Gamma(\tau_W; e, f) \\
e &= e_0 + \frac{nK}{2} \\
f &= f_0 + \frac{1}{2}\sum_{i=1}^{n} s_i^T s_i
\end{aligned}
\tag{B.21}
$$

Annexe C

Article 1

Complex Disease Individual Molecular Characterization Using Infinite Sparse Graphical Independent Component Analysis

Cancer Informatics
Volume 21: 1–16

DOI: 10.1177/11769351221105776
SAGE

Sarah-Laure Rincourt[1], Stefan Michiels[1,2] and Damien Drubay[1,2]

[1]Oncostat U1018, Inserm, University Paris-Saclay, Labelled Ligue Contre le Cancer, Villejuif, France. [2]Department of Biostatistics and Epidemiology, Gustave Roussy, University Paris-Saclay, Villejuif, France.

ABSTRACT: Identifying individual mechanisms involved in complex diseases, such as cancer, is essential for precision medicine. Their characterization is particularly challenging due to the unknown relationships of high-dimensional omics data and their inter-patient heterogeneity. We propose to model individual gene expression as a combination of unobserved molecular mechanisms (molecular components) that may differ between the individuals. Considering a baseline molecular profile common to all individuals, these molecular components may represent molecular pathways differing from the population background. We defined an infinite sparse graphical independent component analysis (isgICA) to identify these molecular components. This model relies on double sparseness: the source matrix sparseness defines the subset of genes involved in each molecular component, whereas the weight matrix sparseness identifies the subset of molecular components associated with each patient. As the number of molecular components is unknown but likely high, we simultaneously inferred it and the weight matrix sparseness using the beta-Bernoulli process (BBP). We simulated data from a double sparse ICA with 10/30 components with specific sparseness structures for 100/500 individuals and 500/1000/5000 genes with different noise variance levels to evaluate the reconstruction of the latent structures by our model. For all simulations, the isgICA was able to reconstruct with higher accuracy than 2 state-of-the-art methods (*ica* and *fastICA*) the number of components, the weight and source matrix sparsenesses (correlation simulated/estimated $>.8$). Applying our model to the expression of 1063 genes of 614 breast cancer patients, the isgICA identified 22 components. According to the source matrix, 7 of these 22 components seemed to be specifically related to 3 known molecular pathways with a prognostic effect in early breast cancer (immune system, proliferation, and stroma invasion). This proposed algorithm provides an insight into individual molecular heterogeneity to better understand complex disease mechanisms.

KEYWORDS: Nonparametric Bayesian model, independent component analysis, individual heterogeneity, gene expression, molecular mechanisms

RECEIVED: November 9, 2021. **ACCEPTED:** May 22, 2022.

TYPE: Original Research

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Damien Drubay, Service de Biostatistique et d'Epidémiologie, Gustave Roussy, Oncostat U1018, Inserm, University Paris-Saclay, Labeled Ligue Contre le Cancer, Institut Gustave Roussy-B2M, 114 rue Edouard Vaillant, Villejuif cedex, 94807, France. Email: damien.drubay@gustaveroussy.fr

Introduction

A comprehensive understanding of the molecular mechanisms of complex diseases, such as cancer, is one of the main current challenges to develop precision medicine. Identifying these mechanisms from omics data, such as gene expression, is challenging due to the complex relationships in various molecular pathways involving hundreds or thousands of actors (eg, genes) and the relatively small number of patients in the available data sets.

To limit the curse of dimensionality, the identification of non-observed high dimensional omics data structures, which provide an insight into the molecular mechanisms, is often performed using latent variable models[1] (LVM) for blind source separation/deconvolution, including principal component analysis (PCA), independent component analysis (ICA), or factor analysis (FA). To identify independent molecular components, we based our work on the ICA model, and we use below the corresponding terminology, that is, the source and the weight matrices corresponding to the parameters representing the association of the components with the genes and the observations, respectively.

To interpret the identified structures as molecular mechanisms or pathways, sparse methods may be used to select a subset of the omics variables associated with each component, similarly to the graphical factor model proposed by Yoshida and West.[2] As illustrated by Figure 1, the sparseness structure of the source matrix may be considered as a hypergraph matrix. The hypergraph approach is a generalization of the graph methods considering higher-order interactions of the nodes (eg, genes) to model complex relationships,[3] which are represented by different (potentially overlapped) subsets of nodes associated to different hyperedges. Each of these latent structures may represent different molecular mechanisms such as pathways, associated to only a subset of the gene expressions. Several sparse approaches have been proposed, especially in the Bayesian framework, imposing sparseness on the source matrix using the spike and slab prior,[4,5] Indian Buffet Process,[6] Laplace prior,[7] or the horseshoe prior.[8]

	K1	K2	K3	K4
Gene 1	0	1	1	0
Gene 2	0	0	1	0
Gene 3	1	0	0	0
Gene 4	1	0	1	0
Gene 5	0	1	1	1
Gene 6	1	0	0	0
Gene 7	1	1	1	0
Gene 8	1	0	0	1

	K1	K2	K3	K4
Individual 1	0	1	0	1
Individual 2	0	1	0	1
Individual 3	0	1	1	1
Individual 4	1	1	1	0
Individual 5	0	0	1	1
Individual 6	1	0	1	0
Individual 7	1	0	1	0

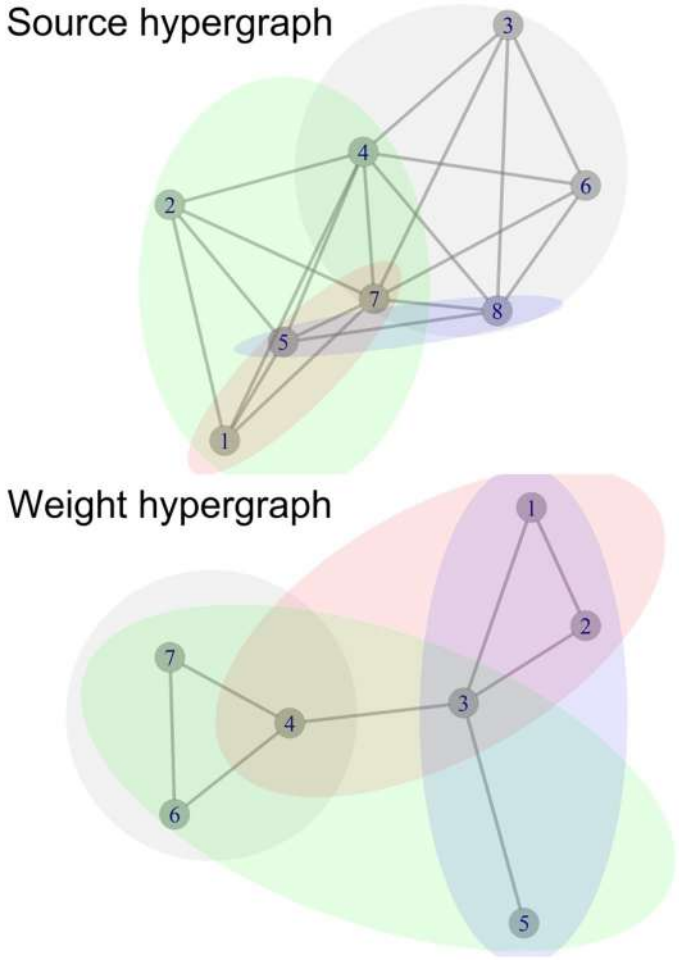


Figure 1. Independent component analysis matrix sparseness structure interpretation as a hypergraph. The source matrix sparseness (top) represents K molecular components (hyperedges) associated with different gene (node) combinations that may represent different molecular mechanisms or their alterations. The weight matrix sparseness (bottom) defines different profiles defined by individual combinations of these molecular components to characterize patient disease heterogeneity.

Although this approach is suitable to give an insight into mechanisms common to all individuals, it is known in oncology that the tumor may result from different molecular mechanisms (or different alterations/state of the same mechanism) across different patients that complicate the development of drugs applicable to broad cancer populations. A natural way to consider this inter-individual heterogeneity with LVM is to impose a second sparseness structure to the weight matrix, associating the individuals to the components. Again, the corresponding sparseness structure may be considered as a hypergraph (Figure 1), and in which this time, each hyperedge represents the subset of individuals presenting the corresponding molecular mechanisms (see Figure 1).

As the number of molecular mechanisms is unknown (but likely large), and the number of their possible alterations may grow rapidly with the number of individuals, one would prefer not to fix the dimension of the model but to infer the number of components present in the studied population. These 2 modeling perspectives may be considered simultaneously using the beta-Bernoulli process (BBP)[9] as a prior on the hypergraph matrix of the weight matrix.[6,10] The rationale behind this approach is to consider a prior on the infinite-dimensional model space assigning only a finite number of 1 in the hypergraph matrix almost surely (therefore a finite number of sparse components) in a finite sample.[11] In other words, this approach considers that there is an infinite number of molecular alterations, but only a subset is present in our finite sample. This approach has the appealing property to allow the model's complexity to grow with the number of observations under the regularization of the BBP hyperparameters.[10-12]

While previously mentioned works focused on sparse coding of the weight matrix or of the source matrix, we propose to impose sparseness on the mixture weight matrix and on the source matrix. To our knowledge, this is the first study imposing this double sparseness in an infinite-dimensional model and proposing an optimization procedure to improve the reconstruction of the underlying latent structures. However, the interpretation of the resulting components remains complex and hazardous because they may represent different molecular mechanisms/pathways or their different alterations among the patients. Instead of precisely

characterizing these molecular mechanisms, we propose to identify "alterations" from a baseline molecular profile. Our choice was motivated by the assumption that the differential disease progression or drug resistance results in a mixture of multiple molecular alterations, which could be different between the patients. We imposed this baseline constraint by enforcing the first component of the weight matrix to be a vector of ones, that is, all individuals are associated with this component, representing the molecular background of the population (that we named the baseline molecular profile). We also consider a noise component (such as the noisy ICA[13]) to capture the specific individual background (shared with no other individuals) and measurement error.

In this work, we first assessed the ability of our isgICA to reconstruct the weight and the source sparseness structures through a simulation study. We compared our approach for the identification of the number of components and for the reconstruction of the matrices to state-of-the-art algorithms. Finally, we applied our method to model the gene expression heterogeneity in a large gene expression dataset of tumors from breast cancer patients included in clinical trials of anthracycline-based chemotherapy and illustrate the relevance of this algorithm to blindly identify relevant known breast cancer gene expression signatures.

Methods

Infinite sparse graphical independent component analysis (isgICA)

Let N the number of individuals, P the number of genes and K be the number of latent components. The noisy independent component analysis aims to decompose a data matrix $\mathbf{X} \in \mathbb{R}^{P\times N}$ into the product of 2 matrices plus residual noise as follows:

$$\mathbf{X} = \mathbf{\Phi W} + \mathbf{E}, \tag{1}$$

where $\boldsymbol{W} \in \mathbb{R}^{K\times N}$ denotes the weight matrix, $\mathbf{\Phi} \in \mathbb{R}^{P\times K}$ denotes the unobserved source matrix, and $\boldsymbol{E} \in \mathbb{R}^{P\times N}$ denotes the additive Gaussian noise matrix. As sparseness is a form of non-Gaussianity, it imposes a shrinkage prior on $\mathbf{\Phi}$ (see the complete model formulation paragraph for prior details), favoring the component independence and interpretability of the components. An alternative approach to introduce sparseness may be to consider a binary matrix $\mathbf{\Theta}$ representing the sparseness structure of $\mathbf{\Phi}$, that is, $\mathbf{X} = (\mathbf{\Theta} \circ \mathbf{\Phi})\mathbf{W} + \mathbf{E}$, with $\circ$ representing the elementwise product. This equation corresponds to the graphical sparse factor model formulation of Yoshida and West,[2] which has inspired the name of our approach.

We used this approach to impose sparseness on $\mathbf{W}$, allowing allocation of a subset of the K components to each individual. Considering the sparse binary matrix $\mathbf{Z} \in \mathbb{1}^{K\times N}$, the model is:

$$\mathbf{X} = \Phi\left(\mathbf{W} \circ \mathbf{Z}\right) + \mathbf{E}. \tag{2}$$

As the number of unobserved molecular mechanisms is unknown but likely high, we consider a nonparametric ICA with an infinite number of components (ie, K=∞). The beta-Bernoulli process (BBP) is a suitable nonparametric prior for binary matrix ($\mathbf{Z}$) with an infinite number of rows (or columns), providing a finite number of non-zero rows almost surely in the case of a finite sample. The nonparametric nature of this approach allows the model's complexity to grow with the data (ie, K increases with N), which is an appealing property to infer the number of molecular components present in the studied population, which should increase with the number of individuals.

Baseline profile and isgICA model

We define a baseline profile as a non-sparse latent component, that is, a component associated with all individuals; that is, $\mathbf{Z}$ is defined by as $\mathbf{Z} = \left[\boldsymbol{Z}_0, \boldsymbol{Z}^*\right]$ where $\mathbf{Z}_0 = [1,\ldots,1]_N$, and $\mathbf{Z}^*$ is drawn from the beta-Bernoulli process. We called the sparse binary components of $\mathbf{Z}^*$ the baseline profile alterations. The dimension of the corresponding baseline profile isgICA matrices are $\mathbf{Z} \in \mathbb{1}^{K\times(N+1)}$, $\mathbf{W} \in \mathbb{R}^{(K+1)\times N}$, and $\mathbf{\Phi} \in \mathbb{R}^{P\times(K+1)}$.

Complete model formulation

We considered conjugate priors for the elements of the model matrices, which allow for posterior analytical calculation and straightforward inference. The graphical representation of the proposed model is illustrated in Figure 2.

The complete model is expressed as:

$$\begin{aligned}
\boldsymbol{X} &\sim \mathcal{N}\left(\mathbf{\Phi}\left(\mathbf{W} \circ \mathbf{Z}\right), \mathrm{diag}\left(\tau_{E_1}^{-1}, \ldots, \tau_{E_P}^{-1}\right)\right) \\
\tau_{E_j} &\sim \mathrm{Gamma}\left(g, h\right) \\
\mathbf{\Phi}_{j,k} &\sim \mathcal{N}\left(0, \tau_{\Phi_{j,k}}^{-1}\right) \\
\tau_{\Phi_{j,k}} &\sim \mathrm{Gamma}\left(c_{j,k}, d_{j,k}\right) \\
W_{k,i} &\sim \mathcal{N}\left(0, \tau_W^{-1}\right) \\
\tau_W &\sim \mathrm{Gamma}\left(e, f\right) \\
Z^*_{k^*,\boldsymbol{i}} &\sim \mathrm{Bernoulli}\left(\pi_{k^*}\right) \\
\boldsymbol{Z}_0 &= [1,\ldots,1]_N \\
\pi_{k^*} &\sim \mathrm{Beta}\left(\frac{\alpha}{\mathrm{K}}, \frac{\beta\left(\mathrm{K}-1\right)}{\mathrm{K}}\right)
\end{aligned} \tag{3}$$

where $\mathrm{i} = 1,\ldots,\mathrm{N}, \mathrm{j} = 1,\ldots,\mathrm{P}, \mathrm{k} = 0,\ldots,\mathrm{K},$ and $\mathrm{k}^* = 1,\ldots,\mathrm{K}.$ These priors apply regularization on the elements of $\mathbf{W}$ and $\mathbf{\Phi}$ using an automatic relevance determination (ARD) prior. Because strongly regularized elements are closed, but not equal to zero, we use the term pseudo-sparseness to distinguish this structure to the stricter sparseness imposed to $\mathbf{W}$ by $\boldsymbol{Z}$. Considering $c = d = 1$, the combination defines a super-Gaussian prior over the $\mathbf{\Phi}$ elements, favoring source sparseness, and thus, independence. Considering gamma distribution for the priors over τ_W, it corresponds to the Bayesian ridge prior for all the elements of the weight matrix $\mathbf{W}$.

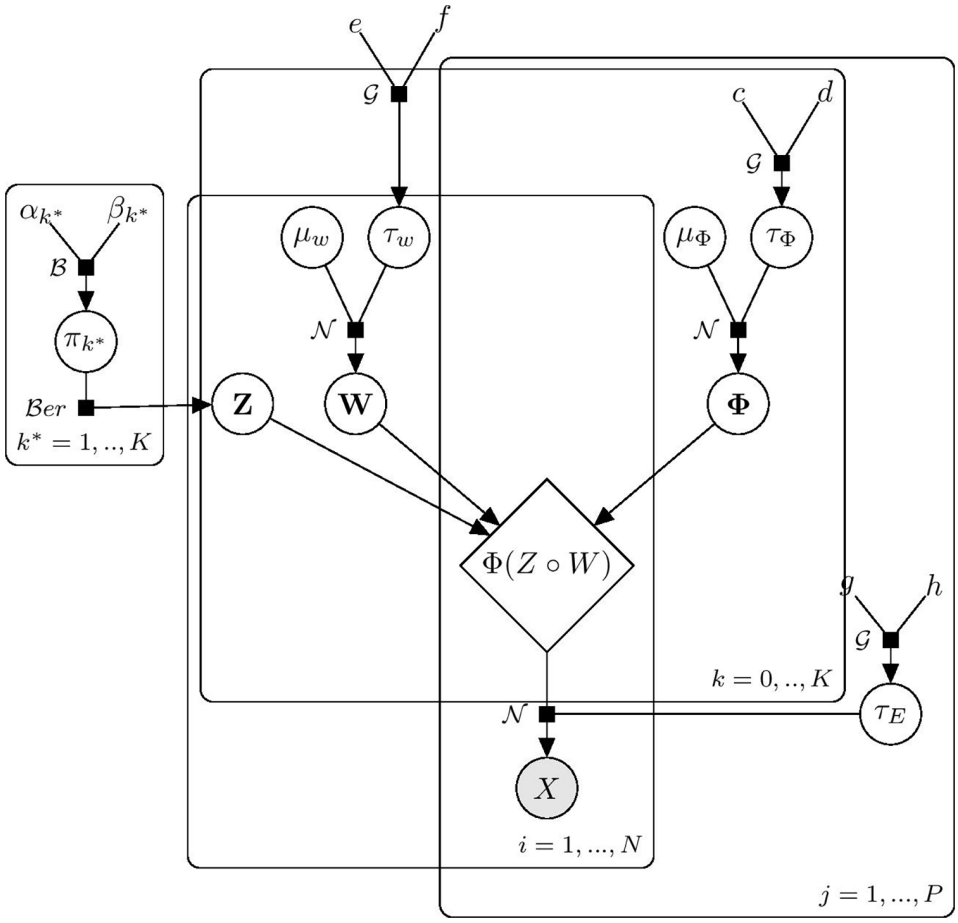


Figure 2. Graphical model representation of the infinite sparse independent component analysis. Observed variables are denoted with shaded nodes, while unobserved variables are shown as white nodes.
Abbreviations: B, Beta distribution; Ber, Bernoulli distribution; G, Gamma distribution; N, Normal distribution.

Consistent estimation of the noise precisions (τ_E), essential to determine the optimal number of latent components, is particularly challenging in a high dimensional setting and is a current hot research topic for matrix factorization noisy models (probabilistic PCA, FA, noisy ICA).[14] Our empirical results from a simulation study confirmed this theoretical statement, highlighting that the model overfits the data, decomposing a part of the noise variance as additional irrelevant components (results not shown). To alleviate this issue, we standardized the input genes and fixed the noise precisions to $\tau_E = 1$, that is, to the variance of the centered and scaled gene expressions. This constraint indirectly regularizes the weight and source parameters, allowing identification of only strong signals, but may lead to false negatives in the hypergraph matrix identification (see results section).

Parameter inference and hyperparameter tuning

As the posterior computation using MCMC is notoriously slow when the number of parameters is high, we used variational Bayesian (VB) inference under the mean-field assumption to approximate the true posterior distribution.[15,16] We derived the variational evidence lower bound of the likelihood (ELBO) and the variational parameter update equations in Appendix A1.

For a computational purpose, we used the truncated beta process for the inference, with a maximum number of components noted K_{max}.[17] For all simulations and data analysis, we considered the prior hyperparameter values: $K_{max} = 100$, $\alpha = 1$, $c = d = 1$, $e = f = 10^{-6}$.

Due to the lack of a simple analytical form of the conjugacy between the prior of the beta distribution and the beta distribution for its moments, we tuned the β hyperparameter of the BBP using Bayesian optimization using the R package ParBayesianOptimization[18] based on 6 initialization evaluations and 24 epochs (total of 30 evaluations). Considering $\alpha = 1$, we reparametrized β as $\mu = \frac{1}{1+\beta}$ according to Ferrari and Cribari-Neto,[19] which have support in the interval [0; 1] ($\mu = 0$ corresponding to $\beta = +\infty$ and $\mu = 1$ to $\beta = 0$), to avoid restraining the support of $\beta \in [0; +\infty]$ with an a priori maximum value for the range of BO evaluation points.

Standard whitening ICA

We evaluated the ability of our method to reconstruct the matrix sparseness structures from simulated data sets in comparison with state-of-the-art algorithms. Due to computational issues for our larger scenario (time and/or memory), we used the standard

Table 1. Simulation scenarios. N, P, and σ_E^2 are the number of individuals, the number of genes and the noise variance respectively.

	N = 100	N = 500
K = 10 simulated components		
P = 500	$\sigma_E^2 = \{0.5, 1, 1.5, 2\}$	$\sigma_E^2 = \{0.5, 1, 1.5, 2\}$
P = 1000	$\sigma_E^2 = \{0.5, 1, 1.5, 2\}$	$\sigma_E^2 = \{0.5, 1, 1.5, 2\}$
P = 5000	$\sigma_E^2 = \{0.5, 1, 1.5, 2\}$	$\sigma_E^2 = \{0.5, 1, 1.5, 2\}$
K = 30 simulated components		
P = 500	$\sigma_E^2 = \{0.5, 1, 1.5, 2\}$	$\sigma_E^2 = \{0.5, 1, 1.5, 2\}$
P = 1000	$\sigma_E^2 = \{0.5, 1, 1.5, 2\}$	$\sigma_E^2 = \{0.5, 1, 1.5, 2\}$
P = 5000	$\sigma_E^2 = \{0.5, 1, 1.5, 2\}$	$\sigma_E^2 = \{0.5, 1, 1.5, 2\}$

whitening ICA model implemented in the *ica*[20] and the *fastICA*[21] R packages, and pre-selected the components with eigenvalue higher than one. The eigenvalue criteria specifies that only components with eigenvalues larger than 1 should be preserved since each component should explain the variance of a single variable.

Results

The simulations, the parameter optimization, and the data visualization were performed using R software (version 3.6.0). The R codes and data are available on https://github.com/Oncostat/isgICA .

Synthetic data

Scenarios. We simulated synthetic datasets from equation (3), according to different scenarios (see Table 1) with $N = \{100, 500\}$ individuals, $P = \{500, 1000, 5000\}$ genes, and $K = \{10, 30\}$ components. The structure of Z was randomly generated to contain approximately 35% of ones. We considered 4 noise parameters ($\sigma_E^2 = \{0.5, 1, 1.5, 2\}$) to assess the impact of the signal-to-noise ratio.

For all scenarios, the elements of the source matrix $\mathbf{\Phi}$ were generated for each component from Gaussian distributions with variance equal to one, and different means (from −3 to 3) to evaluate if the model may identify components with specific patterns, such as mainly positive or negative values, or both (for means closed to 0). Random blocks were generated to assign sparseness structure to this matrix presented by the Figure 5). The elements of the weight matrix W were drawn from a standard Gaussian distribution (mean = 0, variance = 1). We simulated 10 data sets for each scenario.

Performance criteria

As the standard ICA, our model is identifiable up to scaling, sign reversion, and column permutation.[22] To evaluate the model reconstruction, we aligned the estimated components to the simulated ones using as a distance the mean of the absolute Pearson correlation coefficients of each pair of the column of the simulated source matrix $\mathbf{\Phi}$ and non-zero estimated columns of $\widehat{\mathbf{\Phi}}$. For the component i of the simulated $\mathbf{\Phi}$ and the component j of the estimated $\widehat{\mathbf{\Phi}}$, this distance is estimated by $\left| \frac{cov\left(\Phi_{.,i}, \widehat{\Phi}_{.,j}\right)}{\sigma_{\Phi_{.,i}}, \sigma_{\widehat{\Phi}_{.,j}}} \right|$, which is invariant to the scaling and sign reversion. We used the Hungarian algorithm for an efficient re-ordering from this distance, using the HungarianSolver function of the RcppHungarian[23] R package.

The mean of the absolute Pearson correlation coefficients between the simulated $\mathbf{\Phi}$ and $\mathbf{Z°W}$ and the column ordered $\widehat{\mathbf{\Phi}}$ and $\widehat{\mathbf{Z°W}}$ was presented to assess the reconstruction of the source and weight matrix respectively.

According to the component ordering defined previously, the reconstruction of the sparseness structure of the weight matrix (except the baseline profile, ie, $\mathbf{Z^*}$) was assessed with the accuracy criterion, defined by: $\frac{True\ ones + True\ zeros}{N \times K} \in [0, 1]$.

Latent structure reconstruction

We first evaluated the ability of the algorithms to identify the number of components (Figure 3 and Table 2). The isgICA recovered the exact number of latent components in the majority of the simulations, but it underestimated the number in the lower dimension scenarios (P = 500), especially when the number of observations was low with respect to the number of component (N = 100, K = 30). This behavior was slightly more apparent when the noise variance was increased. The number of components selected with the classical whitening ICA using the eigenvalue method increased quickly with the dimension (P/N ratio) to the maximal

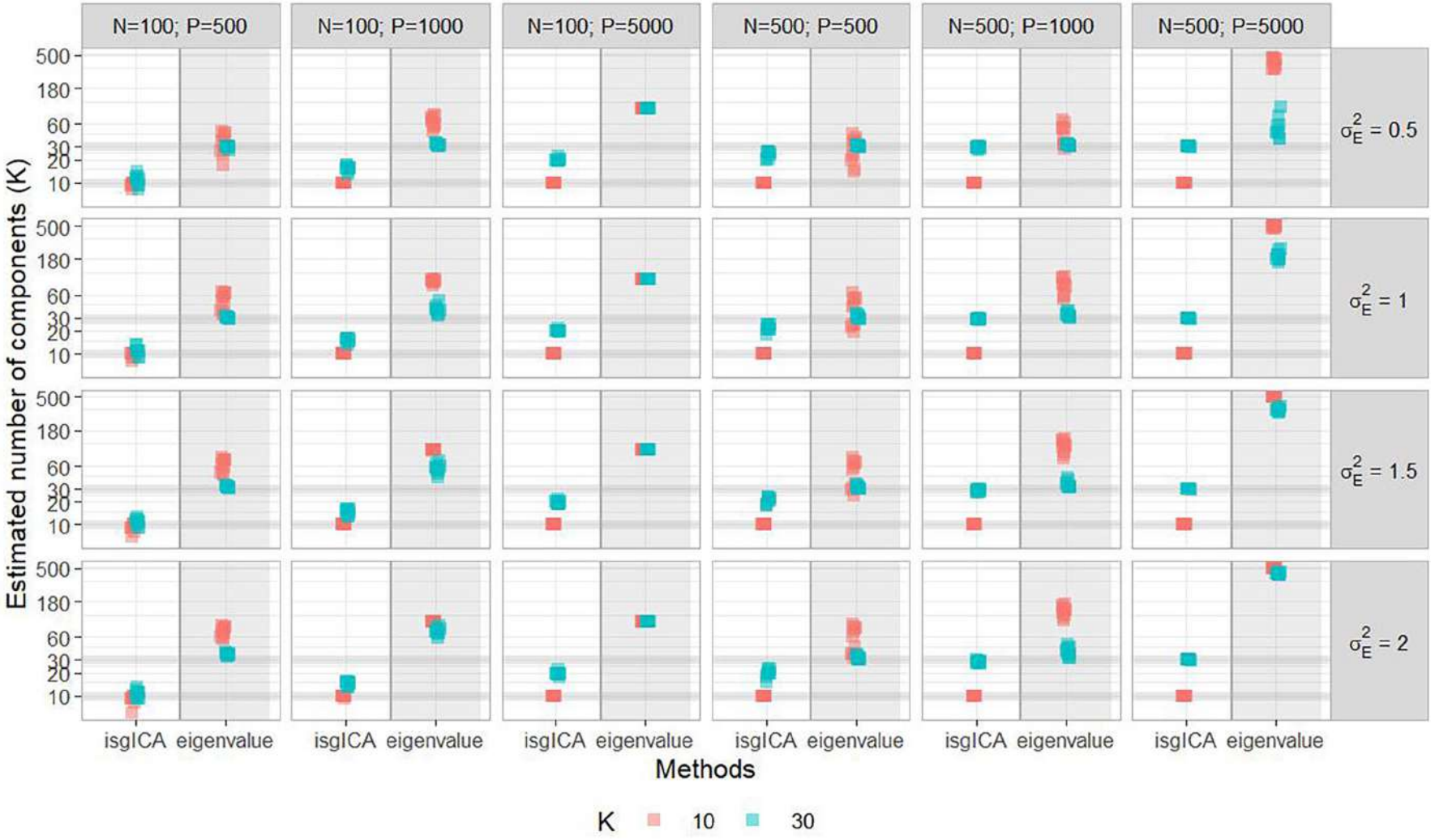


Figure 3. Reconstruction of the number of latent components between isgICA and the standard ICA using the eigenvalue criteria with 10 (red) or 30 (blue) simulated components (K). Each row corresponds to the scenarios with different noise variances (σ_E^2) and each column corresponds to different dimensions (N individuals and P genes), for 10 simulations in each scenario.

Table 2. Number of identified components (median [2.5%-97.5%] percentiles) using the eigenvalue method for the standard ICA and by the isgICA. N, P, and σ_E^2 are the number of individuals, the number of genes, and the noise variance respectively.

MODEL	σ_E^2	N=100			N=500		
		P=500	P=1000	P=5000	P=500	P=1000	P=5000
Eigen-value method		10 simulated components					
	0.5	35 [19, 48]	66 [50, 79]	99 [99, 99]	28 [14, 43]	48 [29, 68]	412 [320, 457]
	1.0	54 [33, 66]	92 [82, 99]	99 [99, 99]	42 [20, 62]	82 [54, 106]	496 [470, 499]
	1.5	66 [46, 77]	99 [96, 99]	99 [99, 99]	56 [26, 76]	114 [79, 140]	499 [499, 499]
	2.0	74 [56, 84]	99 [99, 99]	99 [99, 99]	67 [33, 88]	139 [102, 167]	499 [499, 499]
		30 simulated components					
	0.5	30 [27, 30]	31 [30, 34]	99 [99, 99]	30 [29, 32]	32 [30, 33]	48 [38, 95]
	1.0	31 [28, 32]	38 [32, 49]	99 [99, 99]	30 [29, 34]	32 [30, 38]	196 [167, 243]
	1.5	32 [30, 34]	56 [43, 70]	99 [99, 99]	31 [30, 35]	33 [31, 42]	338 [313, 371]
	2.0	36 [33, 39]	72 [60, 85]	99 [99, 99]	32 [30, 35]	38 [32, 48]	424 [404, 447]
isgICA method		10 simulated components					
	0.5	10 [8, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]
	1.0	10 [8, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]
	1.5	10 [7, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]
	2.0	9 [6, 10]	10 [9, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]	10 [10, 10]
		30 simulated components					
	0.5	11 [8, 14]	16 [13, 17]	20 [19, 22]	25 [20, 26]	29 [27, 30]	30 [29, 30]
	1.0	11 [9, 13]	16 [13, 16]	20 [19, 22]	22 [19, 24]	28 [27, 30]	30 [29, 30]
	1.5	11 [9, 13]	15 [13, 16]	20 [19, 22]	22 [18, 24]	28 [27, 30]	30 [29, 30]
	2.0	11 [9, 13]	15 [13, 16]	20 [18, 22]	20 [16, 23]	28 [27, 30]	30 [29, 30]

number of components allowed by this approach (minimum between N-1 and P-1). It also slightly increased when the noise variance was raised in our algorithm.

Due to this over-decomposition using the eigenvalue method in the majority of the scenarios, the calculation of the reconstruction criteria for the *ica* and *fastICA* was not possible. In this case, we selected the 10 estimated components (or 30, according to the scenario) for which the source components were the most correlated to the simulated ones. We also performed an oracle sensitivity analysis, fixing a priori the number of components to the simulated ones for these 2 approaches, in order to have a comparison of their reconstruction to the blind reconstruction of the isgICA (see results in Appendix A2, Figures A1 and A2)

The isgICA outperformed the *ica* for the reconstruction of the source matrix (**Φ**) when the number of components was estimated using the eigenvalue method (Figure A1). This difference was less clear for the *fastICA* (correlation to simulated source matrix equal to .818 [.191, .934] (median [2.5%-97.5%] percentiles) for isgICA, .767 [.513,.982] for *fastICA*, and .189 [.0.28,.834] for *ica*), especially due to the lower performance of the isgICA in the scenarios with the lowest dimensions (N = 100/P = 500, and N = 500/P = 500). The performance of all methods decreased with the increase of the noise variance, but the isgICA was less impacted for high dimensional scenarios. The oracle *fastICA* and *ica* models presented in the majority of the scenarios a performance similar to the isgICA, excepting in the scenarios where the isgICA underperformed, that is, in the case of low dimension.

For all the scenarios, the isgICA outperformed the other methods for the reconstruction of the weight sparse matrices **Z°W** (mean absolute Pearson correlations equal to .968 [.225, .999] (median [2.5%-97.5%] percentiles) for isgICA, .453 [.194, .972] for the *fastICA*, .515 [.199, .983] for the *ica*), except for the scenarios with N = 100/500, P = 500 and K = 30. This result was explained by the over-decomposition using the eigenvalue method. Considering the oracle *fastICA* and *ica* models, the reconstruction performances were similar for all methods.

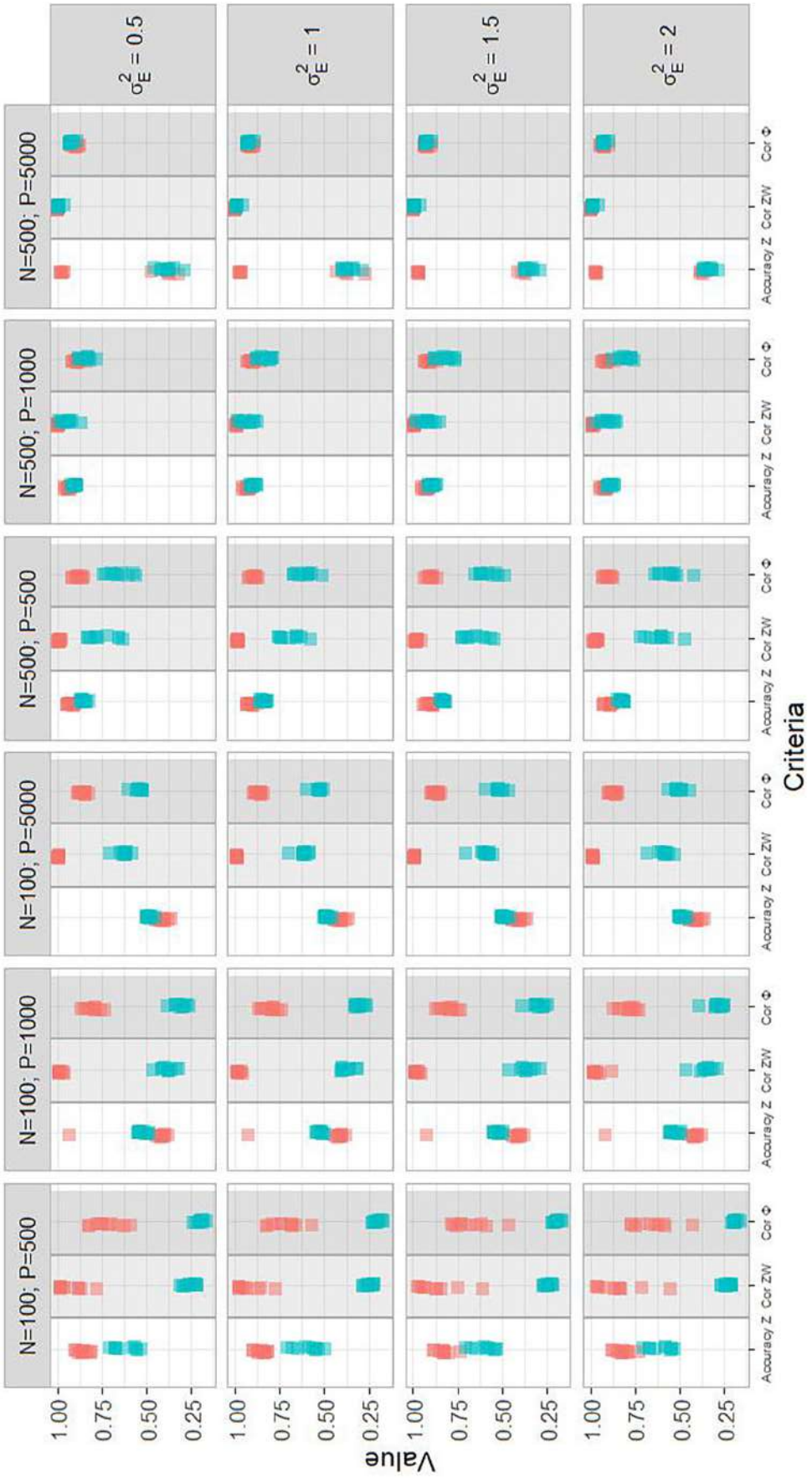


Figure 4. Reconstruction of the weight sparseness structure (**Z**), of the weight sparse matrix (**Z**°**W**) and sources matrix (**Φ**) according to different noise variances in rows with 10 (red) and 30 (blue) simulated components (K) and different dimensions (N individuals and P genes) in columns: the accuracy of the reconstruction of **Z** (first criteria), the mean absolute correlation of **Z**°**W** (second criteria) and the mean absolute correlation of **Φ** (third criteria), for 10 simulations in each scenario.

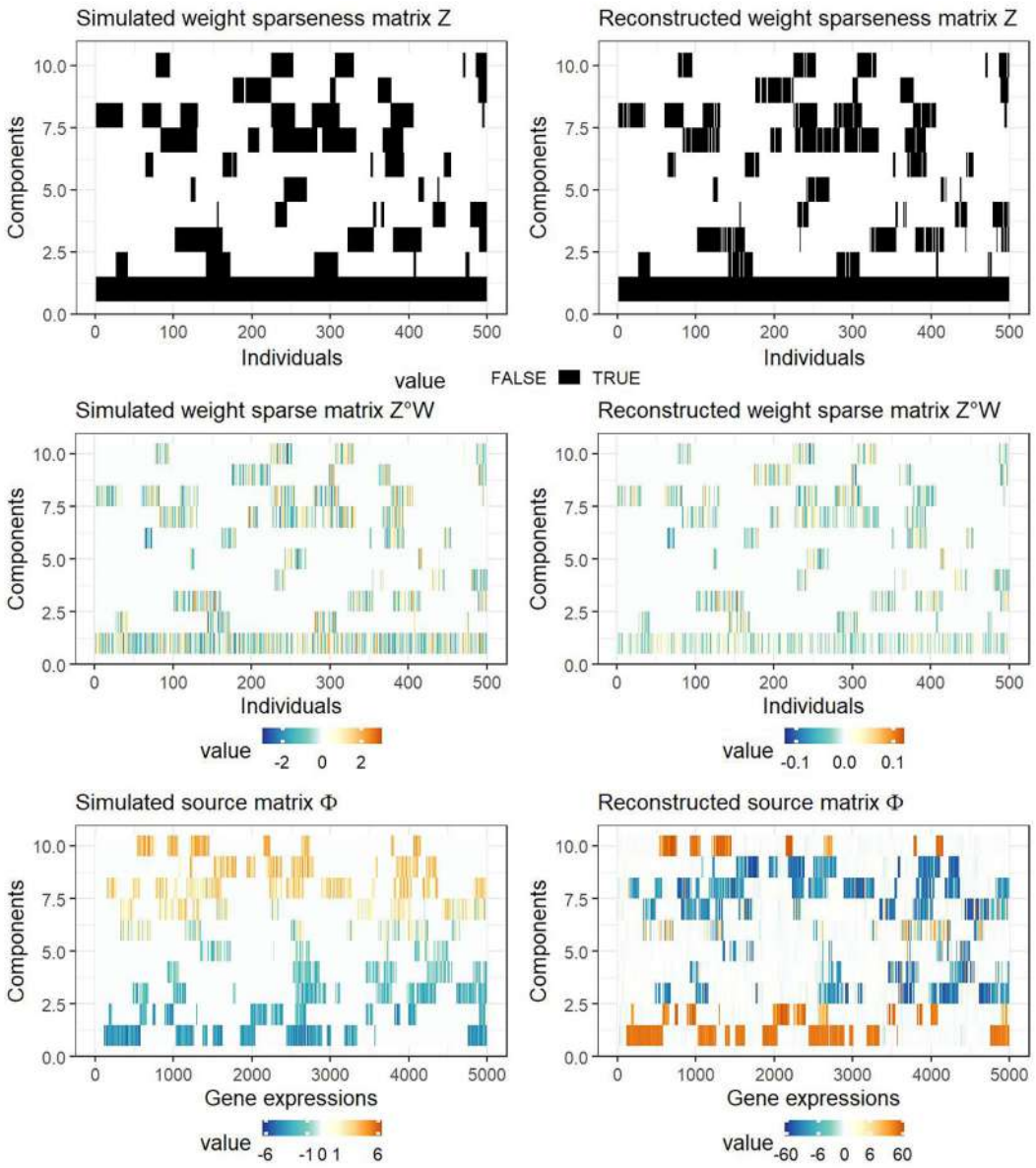


Figure 5. Visual representation of the reconstruction of the weight sparseness structure (Z, top row, accuracy of .982) sparse weight structure (**W**∘**Z**, middle row, correlation of .999) and sources(**Φ**, bottom row, correlation of .873) from a diagonal structure and with N = 500, P = 5000, K = 10 and $\sigma_E^2 = 0.5$.

The isgICA results are summarized in the Figure 4. The ability of the isgICA to reconstruct the sparseness structure of the weight matrix (**Z**) was accurate in the majority of the scenarios (accuracy >.8), but decreased when the number of parameters increased (ie, increasing P or K) to reach the simulated 35% of ones in the non-zeros components, corresponding to the accuracy of all-ones Z matrices. However, the good reconstruction of the weight sparse matrix **(Z°W)** in the high dimension scenarios indicates that this decrease of the accuracy for the strict sparseness is counterbalanced by the pseudo-sparseness induced by the ARD prior on **W**.

To illustrate the reconstruction ability of the isgICA, the Figure 5 shows the reconstruction of the sparse weight matrix (**Z°W**) and the source matrix (**Φ**) for N = 500, P = 5000, K = 10 and $\sigma_E^2 = 0.5$. In this example, our approach was able to identify the correct number of non-zero components and reconstruct the accuracy of the weight sparseness matrix of .982, the sparse weight matrix with a mean absolute correlation of .999 and the source matrix with a mean absolute correlation of .873. As illustrated in the second row of the Figure 5, the ICA-based algorithm suffers from the standard ICA identifiability issues for the source matrix: the sign of the elements of some components may be reversed regarding the simulated one, and they may present higher values (scaling and sign identifiability issues). However, the ranking of the simulated and estimated values was highly correlated, allowing interpretation of the higher values of the source matrix as the most contributing genes to the components.

It can be noted that the *ica* method did not converge for 2.9% (14/480) simulations of the high dimension scenario (N = 500, P = 5000, K = 10) due to the large number of components identified using the eigenvalue method. Therefore, the result of this method is slightly overoptimistic and should be interpreted accordingly (including for the context of a real application).

The gain of precision for the (blind) matrix reconstruction and convergence rate for the isgICA relatively to the other methods came at the cost of a larger computational time relatively to the eigenvalue method (few minutes to several hours), that increases quickly with the dimension (Appendix A2, Figure A3).

Application: Early breast cancer data

We applied our method to publicly available gene expression data obtained from tumor biopsies in 614 breast cancer patients that were included in clinical trials of anthracycline-based chemotherapy,[24,25] available in the *biospear* R package.[26] The expression data of 22 277 probes (Affymetrix array) was preprocessed via frozen robust multiarray[27] and cross-platform normalization.[28] Probes were filtered if the interquartile range $\leqslant 1$. The remaining 1689 probes were standardized and then filtered with the package *jetset*[29] to retain a single probe by gene, resulting in a final dataset including the expression of 1063 genes. As in Belhechmi et al,[30] we mapped to probes to 3 molecular signatures with a prognostic effect in early breast cancer (Immune System, Proliferation, and Stroma invasion[31]) and one without (SRC activation signature[31]), all the other probes were categorized as "Others".

Figure 6 presents the hypergraph matrix of the individual heterogeneity. The model identified 22 non-zero components (including the baseline profile).

To investigate the molecular relevance of these results, we overcame the problems of sign and scale reversion by ranking the absolute source element values amongst each component to identify the most contributive genes to each identified molecular alterations. Figure 7 shows the distribution of the absolute values of the source matrix elements of each component according to the different breast cancer signatures. The proliferation-based signature, immune-system signature, and stroma-related signature seemed to be related to the components 2/5, 4/7/8/15, and 3, respectively. The SRC, which was picked as a "negative control" signature in Belhechmi et al,[30] was not straightforward to map to a particular molecular component.

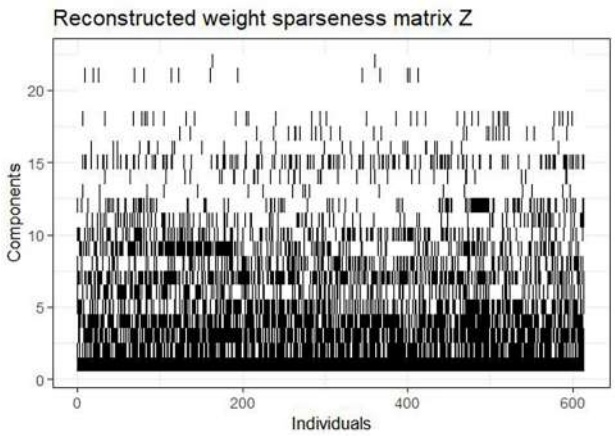


Figure 6. Hypergraph matrix of the individual heterogeneity (weight sparseness structure) extracted from the breast cancer dataset using the infinite sparse graphical independent component analysis with baseline profile. The model identified 22 components (including the baseline profile).

Discussion

In this paper, we proposed a novel approach to characterize inter-patient heterogeneous molecular mechanisms. To our knowledge, this is the first approach that assumes that the molecular profile of each patient is a mixture of different molecular components, which can be shared with the other patients. We modeled these components as alterations from a baseline molecular component shared by all individuals, representing the mechanisms common to all patients, while the noise captures the individual molecular background. Assuming that each molecular component represents alterations of a molecular pathway or a group of related pathways, this approach may help us to understand molecular mechanisms and identify potential targets for drug development.

We illustrated the concept using a gene expression dataset of breast cancer tumor samples from patients included in clinical trials of anthracycline-based chemotherapy. The correspondence of the identified molecular profiles with known molecular pathways that play a prognostic role in early breast cancer (proliferation, immune system, and stroma pathway) suggests that our approach may help characterize the molecular context of particular subpopulations.

In the simulation study, our method was capable of blindly identifying the true number of components and their (sparseness) structures, up to scaling and sign reversion, which are well-known identifiability issues in standard ICA. To alleviate these issues, we proposed to use the absolute values of the source elements to identify the most contributive genes to each component for molecular interpretation. Comparing to 2 other popular ICA algorithms (fastICA and ica), our model better reconstructed blindly the number of components and the weight and source matrices, and had a similar performance when we a priori fixed the true number of components in these 2 algorithms.

Our algorithm was able to provide a better reconstruction performance of the weight sparse matrix **(Z°W)** than these 2 algorithms, even when the number of components was fixed a priori to the true number in the simulations. Our algorithm was also less sensitive to the increase of the dimension and the random noise variance. However, its lower performance in the lowest dimension scenarios suggests that the regularization may be too strong. Due to the well-known underestimation of the noise variance in high dimensional sparse models,[32] we fixed it to 1 (ie, the variance of the classically standardized X if all component elements are equal to 0) to avoid the over-decomposition of the variance that results in an excess of components (not shown). This choice has an impact on the estimates because the noise variance is theoretically lower than 1 if the variance of X is explained by some components. That induces a too low signal-to-noise ratio resulting in posterior means of the component elements close to zero[32] (ie, over-regularization), and therefore in an underestimation of the number of components. By contrast, the increase of the dimension decreases the influence of the noise variance, resulting in increasing of the posterior mean of the parameters that escape to the "spike" regularization pattern **(Z)** in the higher dimension scenarios (as

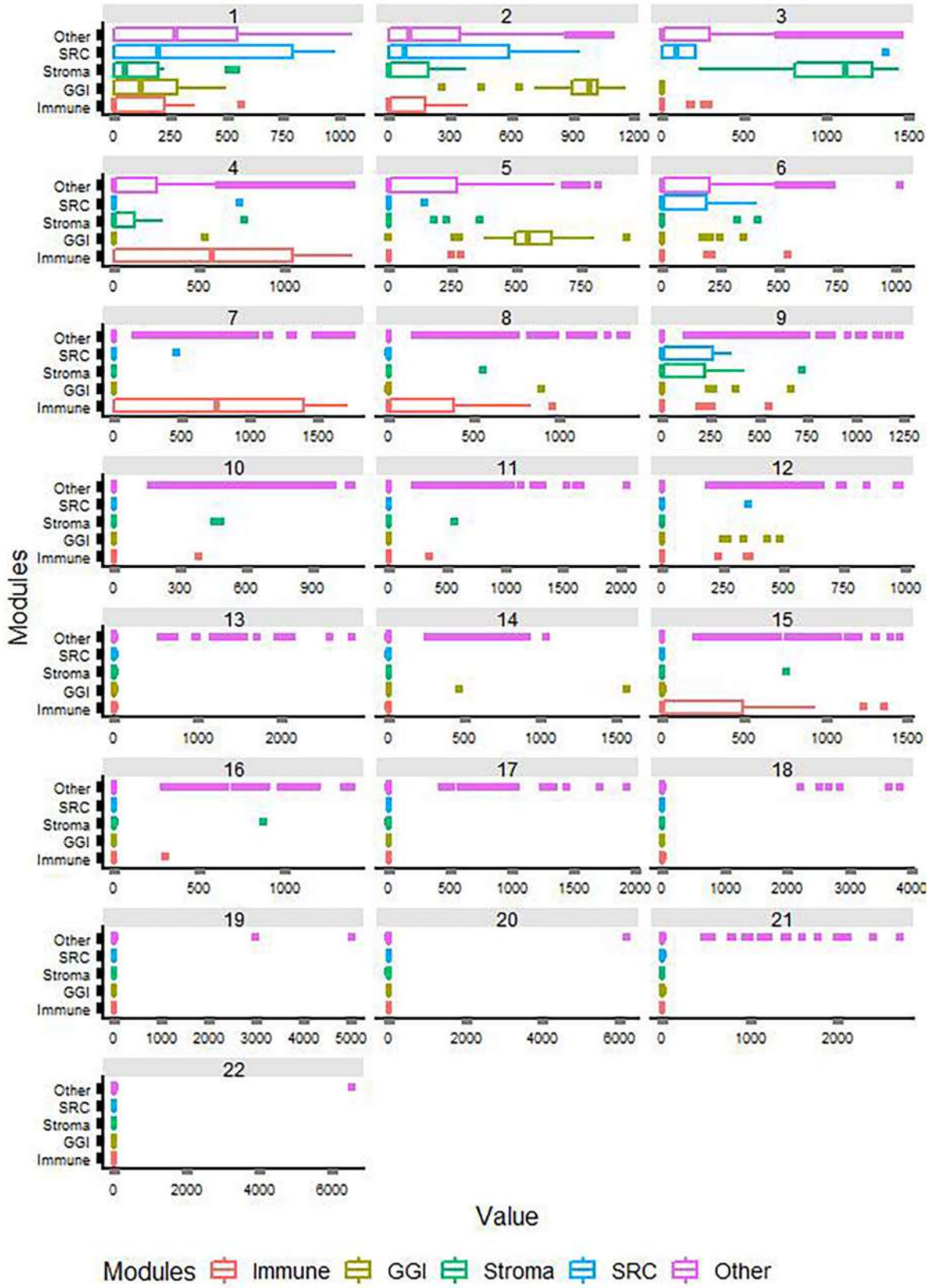


Figure 7. Distribution of the absolute value of the source matrix elements associated with the probes of genes associated with 3 molecular signatures with a prognostic effect in early breast cancer (GGI, Immune2, and Stroma2) and one without (SRC activation signature). All the other genes were categorized as "Others". A proliferation-related signature, immune-related signature, and stroma-related signature seemed to be related to components 2/5, 4/7/8/15, and 3, respectively. The SRC was not related to one specific component.

reflected by the all-ones **Z** matrix posterior). However, the isgICA presents good performances in these settings as these parameters were regularized by the "slab" pattern (ARD prior), resulting in the observed pseudo-sparseness. In further research, the use of independent priors could be explored to alleviate the underestimation of the noise variance. But this approach based on non-conjugate priors requires complex algorithms for inference which could increase the computational time for a high dimensional data set.

As illustrated by our benchmark, the performance of our algorithm came at the cost of an important computational time that could be a practical limitation. A first next step will be to re-implement the current algorithm with GPU computation to scale-up to large datasets.

We showed that this approach is able to identify blindly components deviating from a baseline profile. Future research will focus on improvements for their identifiability and interpretability, including the integration of additional external information. We expect that some mixture weight components can reconstruct observed individual variables not considered by the model. It could be possible to extend this model, fixing the elements of some weight components to the values of observed individual variables, which may be relevant to explain the gene expressions (eg, gender, age, tumor stage). Beyond the adjustment for known characteristics, this extension could be used to perform differential analysis adjusted for unobserved individual characteristics. Moreover, while the bulk sequencing data results of a mixture of several elements (not only tumor cells, but also healthy tissue[33] or tumor microenvironment cells[34]), other sources of data such as reference molecular profiles could be used to improve the identifiability and interpretability of the components. Another interesting way to integrate external information is to consider the patient characteristics as explanatory variable of the sparseness structure of the source matrix to model different states of the graph, as suggested by Wang et al.[35]

The joint modeling of our isgICA and a clinical outcome, such as patient survival, could be of particular interest for precision medicine, favoring the identification of independent molecular profiles more specific to the patient prognostic. This extension will support the estimation of component/treatment interaction in the survival model to highlight pathways related to treatment response for the precision medicine context.

Finally, as proposed in the wide literature of omics data deconvolution methods, our approach may also be extended to other non-Gaussian omics data, such as count (raw RNAseq, proteomics) or binary (mutations) data using different link functions.

Conclusion

We developed an isgICA model with a baseline profile to characterize blindly the individual heterogeneity from this baseline profile in a high-dimensional setting. This approach illustrates a novel concept for the identification of composite molecular profiles which could be key to understanding the different mechanisms of disease and identify potential targets to develop new treatments.

Author Contributions

SLR developed the model, and takes responsibility for the R programs and the accuracy of the data analysis. SLR drafted the manuscript; DD and SM contributed to a critical revision of the manuscript for important intellectual content, supervised the study equally, and gave final approval. All of the authors read and approved the final manuscript.

Data Availability

The breast cancer gene expression dataset is publicly available in the biospear R package. All code and associated data for the infinite sparse graphical independent component analysis with baseline profile is available on https://github.com/Oncostat/isgICA.

ORCID iD

Sarah-Laure Rincourt https://orcid.org/0000-0003-0367-3254

REFERENCES

1. Cunningham JP, Ghahramani Z. Linear dimensionality reduction: survey, insights, and generalizations. *J Mach Learn Res*. 2015;16:2859-2900.
2. Yoshida R, West M. Bayesian learning in sparse graphical factor models via variational mean-field annealing. *J Mach Learn Res*. 2010;11:1771-1798.
3. Feng S, Heath E, Jefferson B, et al. Hypergraph models of biological networks to identify genes critical to pathogenic viral response. *BMC Bioinformatics*. 2021;22:287.
4. West M, Nevins JR, Marks JR, Spang R, Zuzan H. Bayesian factor regression models in the "large p, small n" paradigm. In: Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith AFM, West M. eds. *Bayesian Statistics*. Vol. 7. Oxford University Press; 2003:723-732.
5. Knowles D, Ghahramani Z. Nonparametric Bayesian sparse factor models with application to gene expression modeling. *Ann Appl Stat*. 2011;5:1534-1552.
6. Griffiths TL, Ghahramani Z. Infinite latent feature models and the Indian buffet process. In: *Proceedings of the 18th International Conference on Neural Information Processing Systems. NIPS-05*. MIT Press; 2005:475-482.
7. Kabán A. On Bayesian classification with Laplace priors. *Pattern Recognit Lett*. 2007;28:1271-1282.
8. Carvalho CM, Polson NG, Scott JG. Handling sparsity via the horseshoe. *J Mach Learn Res*. 2009;5:73-80.
9. Hjort NL. Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann Stat*. 1990;18:1259-1294.
10. Paisley J, Carin L. Nonparametric factor analysis with beta process priors. In: *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, Montreal, Quebec, Canada, Association for Computing Machinery; 2009; 777–784. doi:10.1145/1553374.1553474
11. Hjort NL, Holmes C, Müller P, Walker SG, eds. *Bayesian Nonparametrics*. Cambridge University Press; 2010.
12. Gershman SJ, Blei DM. A tutorial on Bayesian nonparametric models. *J Math Psychol*. 2012;56:1-12.
13. Hyvarinen A. Gaussian moments for noisy independent component analysis. *IEEE Signal Process Lett*. 1999;6:145-147.
14. Bouveyron C, Latouche P, Mattei P. Exact dimensionality selection for Bayesian PCA. *Scand J Stat*. 2020;47:196-211.
15. Beal MJ. *Variational Algorithms for Approximate Bayesian Inference*. 2003. https://cse.buffalo.edu/faculty/mbeal/papers/beal03.pdf
16. Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: a review for statisticians. *J Am Stat Assoc*. 2017;112:859-877.
17. Paisley J, Jordan MI. A constructive definition of the beta process. Published online April 3, 2016. Accessed June 25, 2021. http://arxiv.org/abs/1604.00685
18. Samuel W. ParBayesianOptimization: Parallel Bayesian Optimization of hyperparameters. Published online 2020. https://cran.r-project.org/web/packages/ParBayesianOptimization/index.html
19. Ferrari S, Cribari-Neto F. Beta regression for modelling rates and proportions. *J Appl Stat*. 2004;31:799-815.
20. Helwig NE. ica: Independent Component Analysis. Published online 2018. https://CRAN.R-project.org/package=ica
21. Marchini JL, Heaton C, Ripley BD. fastICA: FastICA algorithms to perform ICA and projection pursuit. Published online 2019. https://CRAN.R-project.org/package=fastICA
22. Sokol A, H. Maathuis M, Falkeborg B. Quantifying identifiability in independent component analysis. *Electron J Stat*. 2014;8:1438-1459.

23. Kuhn HW. The Hungarian method for the assignment problem. *Nav Res Logist Q*. 1955;2:83-97.
24. Desmedt C, Di Leo A, de Azambuja E, et al. Multifactorial approach to predicting resistance to anthracyclines. *J Clin Oncol*. 2011;29:1578-1586.
25. Hatzis C, Pusztai L, Valero V, et al. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA*. 2011;305:1873-1881.
26. Ternès N, Rotolo F, Michiels S. biospear: an R package for biomarker selection in penalized Cox regression. *Bioinformatics*. 2018;34:112-113.
27. McCall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis (fRMA). *Biostatistics*. 2010;11:242-253.
28. Shabalin AA, Tjelmeland H, Fan C, Perou CM, Nobel AB. Merging two gene-expression studies via cross-platform normalization. *Bioinformatics*. 2008; 24:1154-1160.
29. Li Q, Birkbak NJ, Gyorffy B, Szallasi Z, Eklund AC. Jetset: selecting the optimal microarray probe set to represent a gene. *BMC Bioinformatics*. 2011;12:474.
30. Belhechmi S, Bin R, Rotolo F, Michiels S. Accounting for grouped predictor variables or pathways in high-dimensional penalized Cox regression models. *BMC Bioinformatics*. 2020;21:277.
31. Ignatiadis M, Singhal SK, Desmedt C, et al. Gene modules and response to neoadjuvant chemotherapy in breast cancer subtypes: a pooled analysis. *J Clin Oncol*. 2012;30:1996-2004.
32. Moran GE, Ročková V, George EI. Variance prior forms for high-dimensional Bayesian variable selection. Published online January 9, 2018. Accessed April 29, 2022. http://arxiv.org/abs/1801.03019
33. Petralia F, Wang L, Peng J, Yan A, Zhu J, Wang P. A new method for constructing tumor specific gene co-expression networks based on samples with tumor purity heterogeneity. *Bioinformatics*. 2018;34:i528-i536.
34. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling tumor infiltrating immune cells with CIBERSORT. *Methods Mol Biol*. 2018;1711:243-259.
35. Wang Z, Baladandayuthapani V, Kaseb AO, et al. Bayesian edge regression in undirected graphical models to characterize interpatient heterogeneity in cancer. *J Am Stat Assoc*. 2022;0:1-14.
36. Chen B, Chen M, Paisley J, et al. Bayesian inference of the number of factors in gene-expression analysis: application to human virus challenge studies. *BMC Bioinformatics*. 2010;11:552.

Appendix

A1. Variational equations

Inspired by Chen et al,[36] we used coordinate ascent algorithm to minimize the evidence lower bound with a mean field approximation. The update equations of the variational parameters are described below, where $i = 1,\ldots,N, j = 1,\ldots,P, k = 0,\ldots,K_{max}$ and $k^* = 1,\ldots,K_{max}$.

Variational update of $\langle z^*_{k^*,i}\rangle$:

$$\begin{aligned} q\left(z^*_{k^*,i}|-\right) &= Bernoulli\left(z^*_{k^*,i};\pi_{k^*}\right) \\ &= \frac{q\left(z_{k^*,i}=1|-\right)}{q\left(z_{k^*,i}=1|-\right)+q\left(z_{k^*,i}=0|-\right)} \end{aligned} \tag{A1-1}$$

with the symbol $\langle\bullet\rangle$ defining the expectation of the argument, with

$$\begin{aligned} q(z^*_{k^*,i}=1|\mathbf{X},\mathbf{Z}_{-\mathbf{k}^*,\mathbf{i}},\mathbf{W},\mathbf{\Phi},\tau_E) &\propto \exp\geq\Big(\left\langle \ln(\pi_{k^*})\right\rangle \\ &-\frac{1}{2}\Big(\left\langle\Phi_{k^*}\right\rangle^T \operatorname{diag}(\langle\tau_E\rangle)\left\langle\Phi_{k^*}\right\rangle\left\langle w_{k^*,i}\right\rangle^2 \\ &-2\left\langle\Phi_{k^*}\right\rangle^T \operatorname{diag}(\langle\tau_E\rangle)X_i^{-k^*}\left\langle w_{k^*,i}\right\rangle\Big)\Big) \\ with\ X^{-k^*}_{j,i} &= x_{j,i}-\sum_{l=0,l\neq k^*}^{K}\left\langle\Phi_{j,l}\right\rangle\left\langle z_{l,i}\right\rangle\left\langle w_{l,i}\right\rangle \end{aligned} \tag{A1-2}$$

and

$$\begin{aligned} q(z_{k^*,i}=0 \quad |\mathbf{X},\mathbf{Z}_{-\mathbf{k}^*,\mathbf{i}},\mathbf{W},\mathbf{\Phi},\tau_E) \\ \propto \exp\Big(\left\langle ln\left(1-\pi_{k^*}\right)\right\rangle\Big); \end{aligned} \tag{A1-3}$$

Variational update of $\langle\pi_{k^*}\rangle$:

$$\begin{aligned} q\left(\pi_{k^*}|-\right) &= Beta\left(\pi_{k^*,i};\alpha'_{k^*},\beta'_{k^*}\right) \\ \left\langle\alpha'_{k^*}\right\rangle &= \sum_{i=1}^{N}\left\langle z_{k^*,i}\right\rangle+\frac{\alpha}{K_{max}} \\ \left\langle\beta'_{k^*}\right\rangle &= N+\frac{\beta\left(K_{max}-1\right)}{K_{max}}-\sum_{i=1}^{N}\left\langle z_{k^*,i}\right\rangle; \end{aligned} \tag{A1-4}$$

Variational update of $\left\langle\Phi_{j,k}\right\rangle$:

$$\begin{aligned} q\left(\Phi_{j,k}|-\right) &= \mathcal{N}\left(\Phi_{j,k};\mu_{\Phi_{j,k}},\tau^{-1}_{\Phi_{j,k}}\right) \\ \left\langle\tau_{\Phi_{j,k}}\right\rangle &= \sum_{i=1}^{N}\left\langle\tau_{E_j}\right\rangle\left\langle w_{k,i}\right\rangle^2\left\langle z_{k,i}\right\rangle^2+\left\langle\tau_{\Phi_{j,k}}\right\rangle \\ \left\langle\mu_{\Phi_{j,k}}\right\rangle &= \left\langle\tau_{\Phi_{j,k}}\right\rangle^{-1}\left(\sum_{i=1}^{N}\left\langle\tau_{E_j}\right\rangle\left\langle w_{k,i}\right\rangle\left\langle z_{k,i}\right\rangle X^{-k}_{j,i}\right); \end{aligned} \tag{A1-5}$$

Variational update of $\langle W_i\rangle$:

$$\begin{aligned} q\left(W_i|-\right) &= \mathcal{N}\left(W_i;\mu_{W_i},\tau^{-1}_{W_i}\right) \\ \left\langle\tau_{W_i}\right\rangle &= \left(\left\langle\mathbf{\Phi}^T\right\rangle\circ\tilde{\boldsymbol{Z}}_{\boldsymbol{i}}\right)\operatorname{diag}\left(\left\langle\tau_E\right\rangle\right) \\ &\quad \left(\left\langle\mathbf{\Phi}\right\rangle\circ\tilde{\mathbf{Z}}_{\mathbf{i}}^{\mathbf{T}}\right)+\left\langle\tau_W\right\rangle I_K \\ \left\langle\mu_{W_i}\right\rangle &= \left\langle\tau^{-1}_{W_i}\right\rangle\left(\left\langle\mathbf{\Phi}\right\rangle\circ\tilde{\mathbf{Z}}_{\mathbf{i}}\right)\operatorname{diag}\left(\left\langle\tau_E\right\rangle\right)x_i \end{aligned} \tag{A1-6}$$

with $\tilde{Z}_i := \left[\langle z_i\rangle,\ldots,\langle z_i\rangle\right]$, $\langle z_i\rangle$ vector repeated K_{max} times;

Variational update of $\left\langle\tau_{\Phi_{j,k}}\right\rangle$:

$$\begin{aligned} q\left(\tau_{\Phi_{j,k}}|-\right) &= \text{Gamma}\left(\tau_{\Phi_{j,k}};c_{j,k},d_{j,k}\right) \\ \left\langle c_{j,k}\right\rangle &= c_0+\frac{1}{2} \\ \left\langle d_{j,k}\right\rangle &= d_0+\frac{1}{2}\left\langle\Phi_{\text{j,k}}\right\rangle^2; \end{aligned} \tag{A1-7}$$

Variational update of $\left\langle\tau_W\right\rangle$:

$$\begin{aligned} p\left(\tau_W|-\right) &= \text{Gamma}\left(\tau_W;e,f\right) \\ \left\langle e\right\rangle &= e_0+\frac{NK_{max}}{2} \\ \left\langle f\right\rangle &= f_0+\frac{1}{2}\sum_{i=1}^{N}\left\langle w_i^T\right\rangle\left\langle w_i\right\rangle. \end{aligned} \tag{A1-8}$$

A2. Comparison between the isgICA and the whitening ICA models

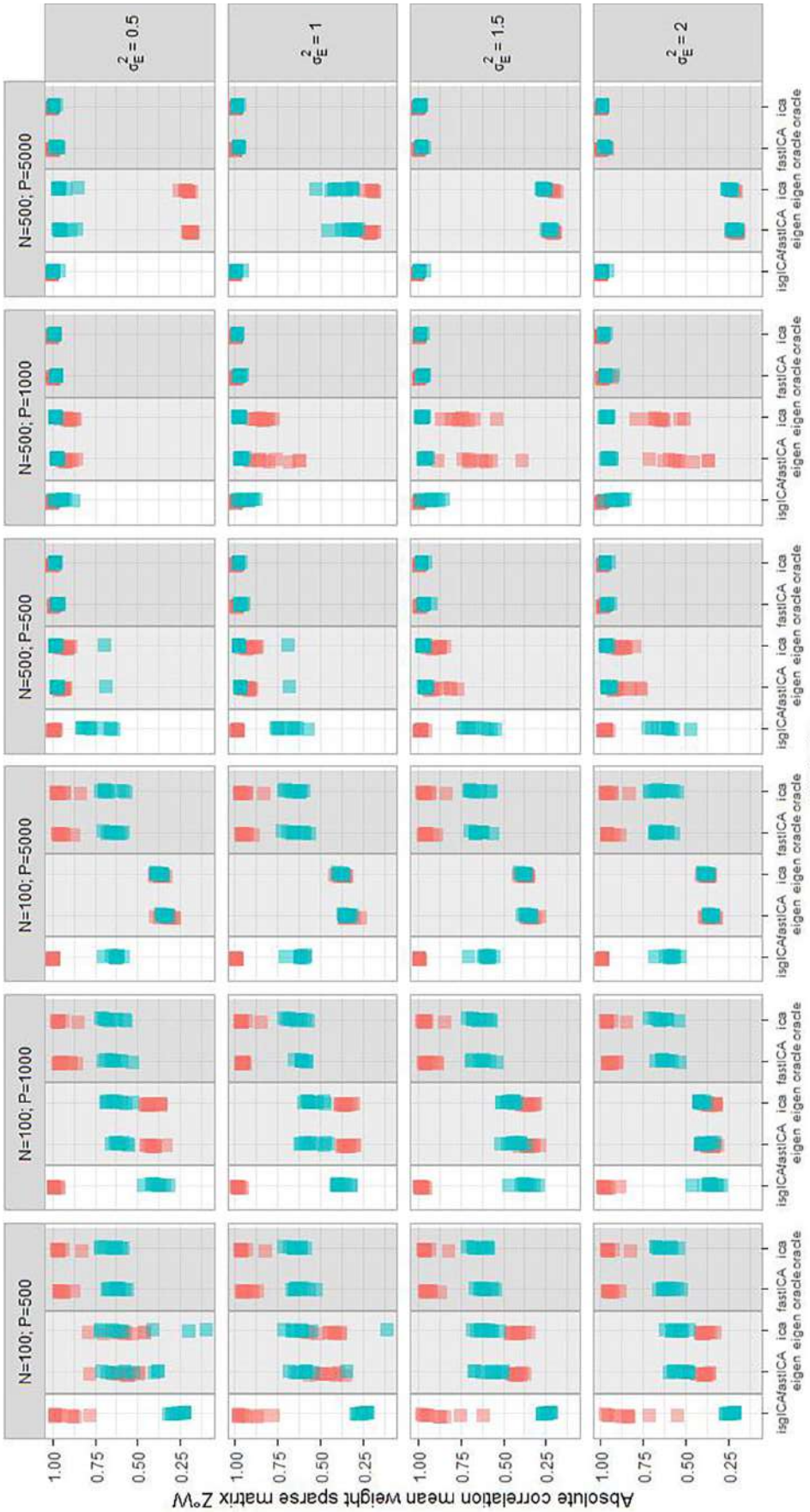


Figure A1. Performance of the reconstruction of the weight sparse matrix **Z°W** of the isgICA, *ica*, and *fastICA*, for the different scenarios with 10 (red) or 30 (blue) simulated components (K), for different noise variances (σ_E^2) in rows, and different dimensions (N individuals and P genes) in columns, for 10 simulations in each scenario.

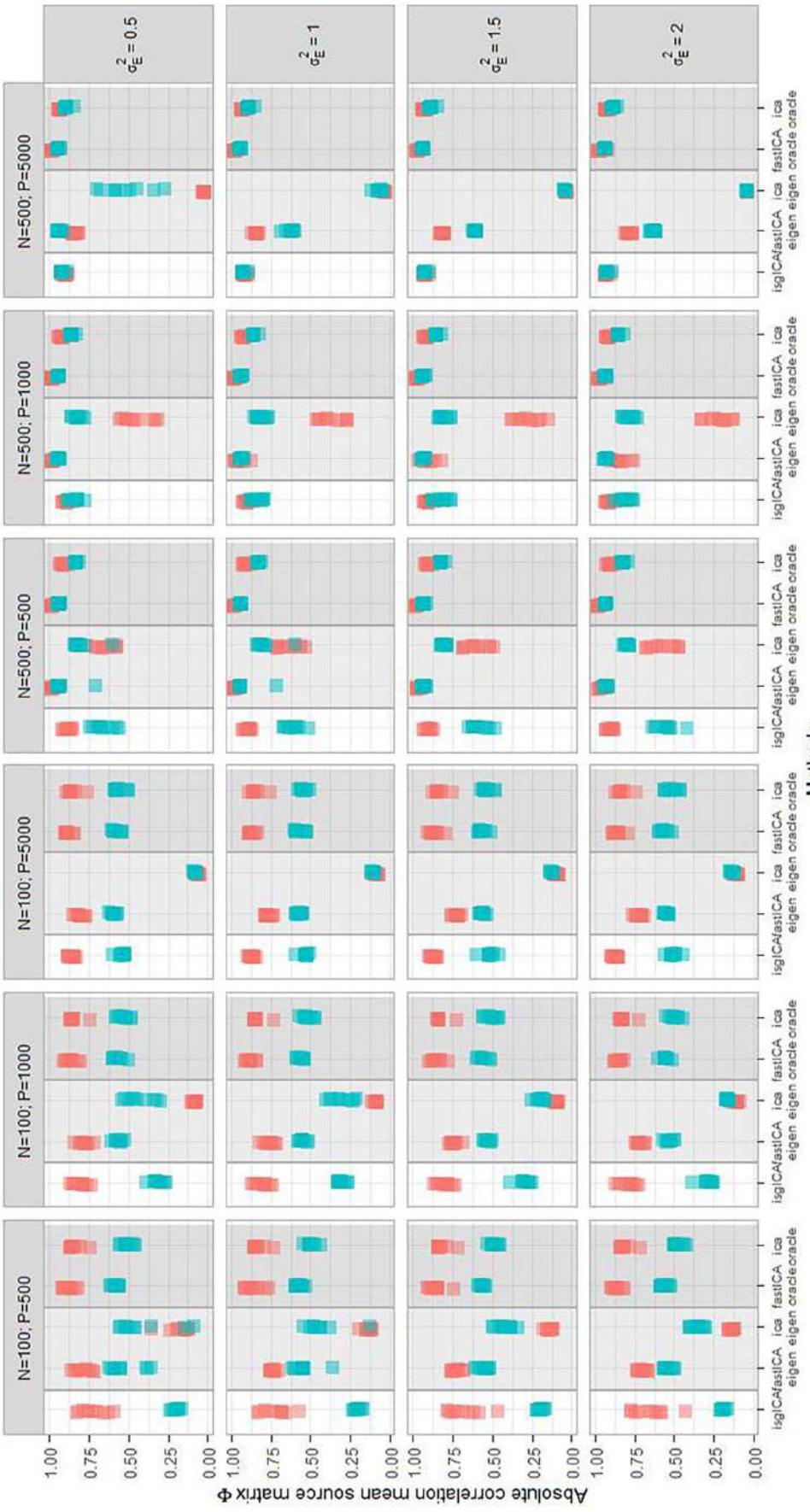


Figure A2. Performance of he reconstruction of the pseudo-sparseness of the source matrix **Φ** of the isgICA, *ica*, and *fastICA*, for the different scenarios with 10 (red) or 30 (blue) simulated components (*K*), for different noise variances (σ_E^2) in rows, and different dimensions (N individuals and P genes) in columns, for 10 simulations in each scenario.

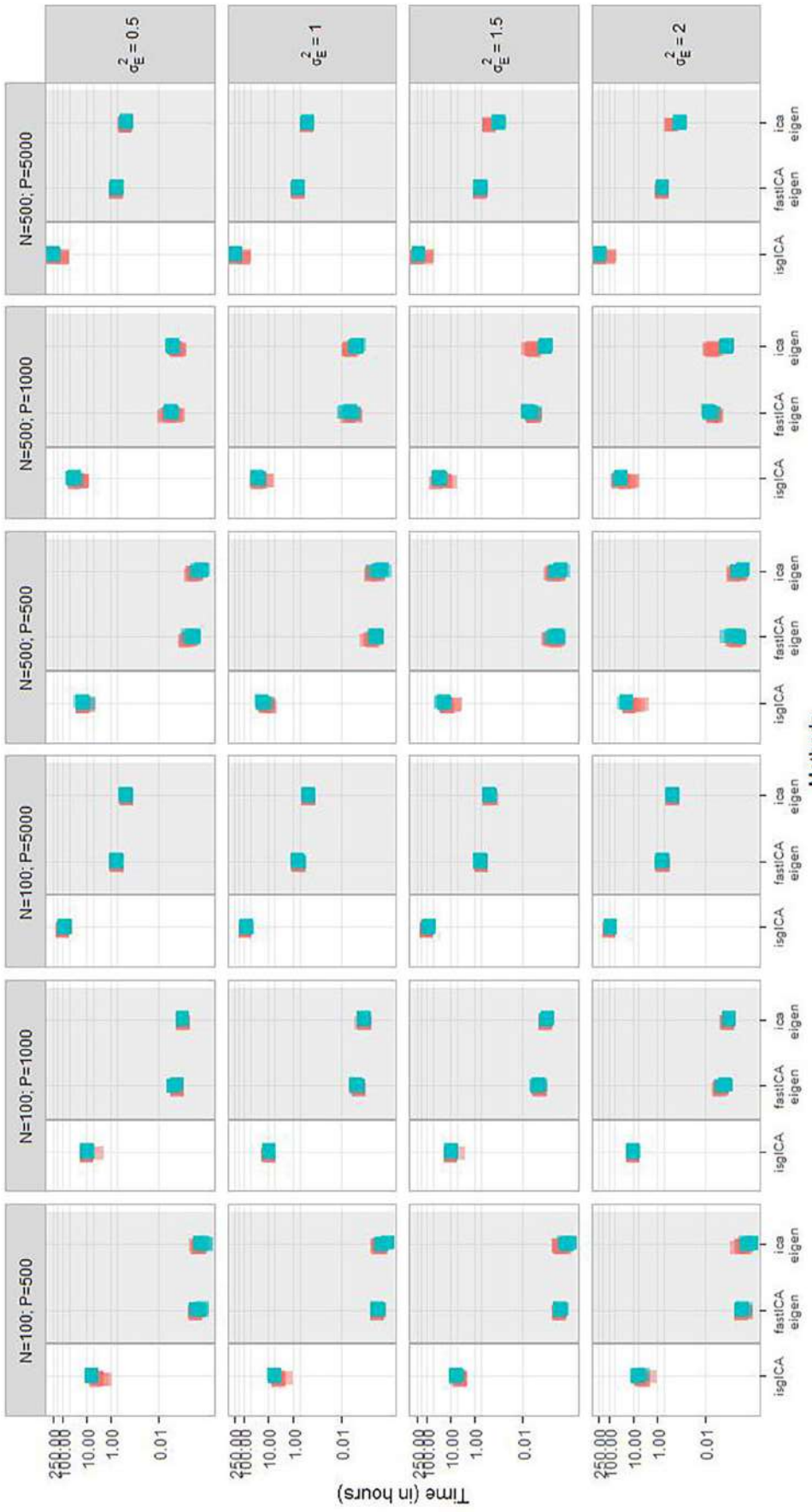


Figure A3. Computational times of the *ica*, *fastICA*, and isglCA for the different scenarios with 10 (red) or 30 (blue) simulated components (K), different noise variances (σ_E^2) in rows, and different dimensions (N individuals and P genes) in columns, for 10 simulations in each scenario.

Annexe D

Article 2

Journal of Bioinformatics and Computational Biology

A nonparametric Bayesian joint model for latent individual molecular profiles and survival in oncology

Sarah-Laure Rincourt *

Oncostat U1018, Inserm, University Paris-Saclay, labelled Ligue Contre le Cancer, Villejuif, France.

Stefan Michiels

Oncostat U1018, Inserm, University Paris-Saclay, labelled Ligue Contre le Cancer, Villejuif, France.

Office of Biostatistics and Epidemiology, Gustave Roussy, University Paris-Saclay, Villejuif, France.

stefan.michiels@gustaveroussy.fr

Damien Drubay

Oncostat U1018, Inserm, University Paris-Saclay, labelled Ligue Contre le Cancer, Villejuif, France.

Office of Biostatistics and Epidemiology, Gustave Roussy, University Paris-Saclay, Villejuif, France.

Correspondence: damien.drubay@gustaveroussy.fr

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

The development of prognostic molecular signatures considering the inter-patient heterogeneity is a key challenge for the precision medicine. We propose a joint model of this heterogeneity and the patient survival, assuming that tumor expression results from a mixture of a subset of independent signatures. We deconvolute the omics data using a non-parametric independent component analysis with a double sparseness structure for the source and the weight matrices, corresponding to the gene-component and individual-component associations, respectively. In a simulation study, our approach identified the correct number of components and reconstructed with high accuracy the weight (>0.85) and the source (>0.75) matrices sparseness. The selection rate of components with high-to-moderate prognostic impacts was close to 95%, while the weak impacts were selected with a frequency close to the observed false positive rate (<25%). When applied to the expression of 1063 genes from 614 breast cancer patients, our model identified 15 components, including six associated to the patient survival, and related to three known prognostic pathways in early breast cancer (i.e., immune system, proliferation, and stromal invasion). The proposed algorithm provides a new insight into the individual molecular heterogeneity that is associated with patient prognosis to better understand the complex

*

tumor mechanisms.

Keywords: Independent component analysis; survival; prognosis.

1. Introduction

Identifying the molecular biomarkers that are associated with a patient's prognosis is essential in stratified medicine. Even though several individual biomarkers (e.g., mutations, copy number alterations or gene expression values) with strong effects on the survival of cancer patients have been identified in the past, the synergy of the weaker statistical signals of the different biomarkers involved in intricate molecular pathways is more complex to identify.

Non-observed structures in omics data are usually identified using latent variable models [35,11](LVM) for dimension reduction, including principal component analysis (PCA), independent component analysis (ICA), or factor analysis (FA). Similar to the graphical factor model [39], the interpretability of the components can be enforced by introducing sparseness to limit the contribution of the biomarkers to a subset of components. This allows the identified components to be interpreted as pathways or molecular signatures [19,24,37]. The most popular methods (e.g., the PCA, standard and sparse partial least square (sPLS), or PLS differential analysis (sPLS-DA)) are implemented in the R package [32], which is dedicated to the analysis of (multi)omics data.

In oncology, different tumors of the same organ can result from various molecular mechanisms in different patients, which complicates the development of stratified or targeted treatment. Paisley and Carin [24,10] proposed a sparse Bayesian nonparametric graphical factor model that imposes sparseness on the weight matrix instead of on the source matrix, and therefore allows us to consider that each source is allocated to only a subset of individuals and may represent an inter-individual heterogeneity in the involved latent process in the observed omics data structure. Inspired by these works, we propose a new LVM with double sparseness [31], in which we first consider a sparse component matrix to identify potential molecular signatures. This sparseness is induced by a non-Gaussian prior (i.e. the standard Bayesian ICA), favoring the independence of the components and limiting the rotation identifiability issues. Second, we considered a sparse weight matrix, assuming that these biological mechanisms are different molecular alterations that are not shared by all individuals. Inspired by Chen et al. [10], we used the beta-Bernoulli process [16] to simultaneously infer the number of components and the sparseness from an infinite component model (considering that there is an infinity of components but only a subset is present in the observed individuals). We introduced a component that is shared by all individuals, which represents the population molecular background. This allows us to interpret the sparse components as molecular alterations that are shared by several individuals. We then added a noise term to capture the individual molecular specificities.

This algorithm, and the ones previously mentioned, are dedicated to unsuper-

vised tasks or classification and are not intended to identify prognostic molecular signatures. A large number of works have used a 2-step procedure [28,38,2,14,27], as follows: i) fit an unsupervised LVM on the omics data (e.g., PCA, ICA, etc.); and ii) use the individual weights as explanatory variables in a survival regression. Given the independent inference processes of these two steps, the survival information is not considered in the LVM parameter estimations, which are thus not optimized for the prognostic task.

Joint modeling approaches have been proposed to simultaneously integrate the omics data and the survival data for the inference of the latent variables, such as the PLS-survival model [3,5]. Several versions of the factor analysis-survival joint model are also available in the Bayesian framework [22,9,26,23]. In [6], the authors integrate the survival response as a covariable (without taking into account the censored data).

To capture the individual patient's heterogeneity, the current paper proposes a joint model of survival analysis and the proposed infinite sparse graphical ICA (isgICA), which is referred to SisgICA. Assuming that the proposed molecular components may impact on the patient's survival, we considered a model that shares the binary latent space of the beta-Bernoulli process between the isgICA and a Weibull accelerated failure time (AFT) model (i.e., the sparseness structure of the weight matrix is used as a set of explanatory variables of the survival submodel). This model simultaneously uses the information of both sources (i.e., survival and gene expression) to define sparse structures that would allow us to deconvolute the gene expression and explain the survival. The performance of the SisgICA model is evaluated on synthetic data sets and a gene expression data set of breast cancer patients who were included in a clinical trial of anthracycline-based chemotherapy. This paper is organized as follows. Section 2 presents the submodels of the joint model. Section 3 presents the model's performance in the synthetic and real data sets. We then discuss our results and conclude in Sections 4 and 5, respectively. The appendix provides an update of the model.

2. Materials and methods

2.1. *Latent sparse component submodel: infinite sparse graphical independent component analysis (isgICA)*

We developed an infinite independent component analysis with a double sparseness constraint, as follows: a sparse source matrix that defines the components corresponding to molecular signatures and considering a sparse weight matrix that defines the allocation of the patients to the different components.

Let N denote the number of individuals, P denote the number of genes, and K denote the number of latent components (sources). The isgICA aims to decompose a data matrix $\boldsymbol{X} \in \mathbb{R}^{P \times N}$ into the product of three matrices plus residual noise, as follows:

$$\boldsymbol{X} = \boldsymbol{\Phi}(\boldsymbol{W} \circ \boldsymbol{Z}) + \boldsymbol{E}, \tag{1}$$

where $\boldsymbol{W} \in \mathbb{R}^{K \times N}$ denotes the weight matrix, $\boldsymbol{\Phi} \in \mathbb{R}^{P \times K}$ denotes the unobserved source matrix, the sparse binary matrix is $\boldsymbol{Z} \in \mathbb{1}^{K \times N}$, $\boldsymbol{E} \in \mathbb{R}^{P \times N}$ denotes an additive Gaussian noise matrix, and $\circ$ represents the element-wise product. The sparseness of the source matrix is favored when using the automatic relevance determination (ARD) prior (i.e., a super-Gaussian prior over the Φ elements), which increases the component independence for interpretability. Similarly, we use the ARD prior for the weight matrix W, where we consider a common prior for all of the elements of W, corresponding to an L2 regularization (i.e., Bayesian ridge).

Given that the number of unobserved molecular mechanisms is unknown but is likely to be high, we consider a nonparametric ICA with an infinite number of components (i.e., K = ∞). We used the beta-Bernoulli process (BBP), which is a nonparametric prior, for the binary matrix (Z), which enforces the sparseness of a weight matrix with an infinite number of components. The posterior almost surely provides a finite number of non-zero components in the case of a finite number of observations. The nonparametric nature of this approach allows the model's complexity to grow with the data (i.e., K increases with N). This process is appealing because it infers the number of molecular components that are present in the study population. Because other components are present in the population, those that are not observed in our sample could be observed if we expand our sample. We imposed a non-sparse latent component (i.e., a component associated with all individuals) to represent a *baseline profile*. This allows us to interpret the other components as molecular alterations from the ground-truth expression that is common to all individuals. In other words, Z is defined as $Z = [Z_0, Z^*]$, where $Z_0 = [1, \ldots, 1]_N$, and Z^* is drawn from the beta-Bernoulli process. We call the sparse binary components of Z^* the baseline profile alterations.

The consistent estimation of the noise precision (τ_E), which is essential to limit the model overfit in the infinite dimensional settings, is particularly challenging and is currently a popular research topic for noisy matrix factorization models (e.g., probabilistic PCA, FA, noisy ICA) [8]. To alleviate this issue, we standardized the input variables and fixed the noise precision to $\tau_E = 1$ (i.e., to the variance of the centered and scaled gene expressions). This constraint indirectly regularizes the weight and source parameters, which allows us to identify only strong signals but may lead to false negatives in the Z matrix identification.

2.2. *Parametric survival submodel: Weibull AFT regression*

We assumed that the components are a denoised presence of a molecular alteration k and that its allocation to a patient i, represented by $Z_{k,i}$, could be associated with survival. Therefore, we assumed that the gene expressions X and the patient survival times T are independent, conditionally to Z. We focused on the Weibull accelerated failure time (AFT) model, which corresponds to a *log-linear model* of

the logarithm of the survival time Y of the form [21]:

$$\begin{aligned} &T \sim \text{Weibull}(Z^T\beta, \tfrac{1}{\sigma_Y}), \\ &\log T = Y = Z^T\beta + \sigma_Y\epsilon \sim \text{Gumbel}(Z^T\beta, \sigma_Y), \\ &\epsilon = \tfrac{log(T) - Z^T\beta}{\sigma_Y} \sim \text{Gumbel}(\log(Z^T\beta), \tfrac{1}{\sigma_Y}), \end{aligned} \tag{2}$$

where $\beta \in \mathbb{R}^{K\times 1}$ are the regression parameters, σ_Y is the scale parameter, and $\epsilon \in \mathbb{R}^{N\times 1}$ are the error terms.

Censoring is the main characteristic of the survival data, and the true survival time T is usually not observed for all patients. Instead, we observe the time until the last follow-up information of the patient, T^*, and the patient status δ at this time ($\delta_i = 0$ if censored, $\delta_i = 1$ if the event of interest is observed for the i-th patient). The likelihood of the AFT model is:

$$\begin{aligned} L(\beta, \sigma_Y; Y) &= \textstyle\prod_{i=1}^{N} f_i(y_i)^{\delta_i} S_i(y_i)^{1-\delta_i} \\ &= \textstyle\prod_{i=1}^{N} \left[\frac{1}{\sigma_Y}\exp(\epsilon)\right]^{\delta_i} \exp(-\exp(\epsilon)). \end{aligned} \tag{3}$$

Because the allocation to the baseline profile (i.e. logarithm of the baseline hazard) is equal to one for all observations, it is confounded with the model intercept and the alterations from the baseline molecular profile correspond to explanatory covariables of the model. We considered a multivariate normal prior for the regression parameters, and a Wishart prior for the precision matrix. We define the complete joint SisgICA model and its priors in the appendix A.1.

2.3. *Parameter inference and hyperparameter tuning*

Given that the posterior computation using MCMC is notoriously slow when the number of parameters is high, we used variational Bayesian inference (VI) under the mean field assumption to approximate the true posterior distribution [7]. We derived the variational evidence lower bound of the joint likelihood (ELBO) and the variational parameter update equations in appendix A.2. Because there is no closed form to approximate the survival submodel parameter posteriors, we used black box variational inference [29,18] with the tensorflow software [1].

For computational purposes, we used the truncated beta process for the inference, with a maximum number of components noted K_{max} [25]. For all of the simulations and data analysis, we considered the isgICA prior hyperparameter values as $K_{max} = 100, a = 1, c = d = 1, e = f = 10^{-6}$. We fixed the Wishart prior hyperparameter values for the survival parameters as $\kappa_0 = 1; \nu_0 = K; V_0 = 0.1 \times \mathbb{I}_K; S_0 = 0.1 \times \mathbb{I}_K$.

Due to the lack of a simple analytical form of the conjugacy between the prior of the beta distribution and the beta distribution for its moments, we tuned the b hyperparameter of the BBP using Bayesian optimization (BO) and implemented in the R package ParBayesianOptimization [33], based on six initialization evaluations and nine epochs (total of 15 evaluations). Considering $a = 1$, we reparametrized b as $\mu = 1/(1 + b)$ according to Ferrari and Cribari-Neto [13], which has support in

the interval [0; 1] ($\mu = 0$ corresponding to $b = +\infty$ and $\mu = 1$ to $b = 0$), to avoid restraining the support of $b \in [0; +\infty]$ with an a priori maximum value for the range of BO evaluation points.

3. Results

The simulations, the parameter optimization, and the data visualization were performed using R software (version 3.6.0). The R codes and data are available at `https://github.com/Oncostat/SisgICA`.

3.1. *Synthetic data*

3.1.1. *Simulation set up and evaluation criteria*

We simulated synthetic data sets from Eq. (A.1), with N=600 subjects and P=1200 variables, to mimic our breast cancer application (which is described below). We considered four scenarios using two censoring rates (30% and 70%) for 10 and 30 components. We simulated 100 data sets for each scenario.

For all scenarios, the elements of the source matrix Φ were generated for each component from Gaussian distributions with a variance that is equal to one, and different means (from -3 to 3) to evaluate if the model may identify components with different intensities. Random blocks were generated to assign a sparsity structure to this matrix, which can be easily visualized (see Fig. 6). The elements of the weight matrix W were drawn from a standard Gaussian distribution (mean = 0, variance = 1). The sparse binary structure Z was randomly generated to contain approximately 35% of ones. We considered the noise parameter to be equal to $\sigma_E^2 = 1$.

Our model is identifiable up to scaling, sign reversion, and column permutation [34], as the standard ICA. We used the mean of the absolute Pearson correlation coefficients of each pair of simulated and non-zero estimated components as a distance to order the columns of the estimated source matrix $\widehat{\Phi}$ and the rows of the estimated sparsity structure of the weight matrix $\widehat{Z}$. For each component i of the simulated Φ and component j of the estimated $\widehat{\Phi}$, this distance is estimated by $\left|\frac{\text{COV}(\Phi_{\cdot,i},\widehat{\Phi_{\cdot,j}})}{\sigma_{\Phi_{\cdot,i}}\sigma_{\widehat{\Phi_{\cdot,j}}}}\right|$, which is invariant to the scaling and sign reversion. To quantify the reconstruction of the source matrix, the mean of the absolute Pearson correlation coefficients $\overline{|\text{Cor}_\Phi|}$ between the simulated Φ and the column ordered $\widehat{\Phi}$ was provided. According to the component ordering that was defined previously, the reconstruction of the sparseness structure of the weight matrix (except the baseline profile; i.e., Z^*) was assessed with the accuracy criterion, which is defined by: $\frac{\text{True ones} + \text{True zeros}}{N \times K} \in [0, 1]$.

Using Z to join the isgICA and the AFT model, we can interpret the parameters of the regression while avoiding the sign and scale issues. We only consider six non-null parameters in the vector b, which is associated to the latent component in the Weibull submodel in Eq. (2), with arbitrary values set to $[-1, -0.5, -0.1, 0.1, 0.5, 1]$.

This allows us to consider signatures with weak (0.1) to strong (1) prognostic impact. The shape of the real and censored Weibull distributions were fixed to 2 and the censored scale was adapted to ensure the censoring rate.

Three criteria are proposed to evaluate the reconstruction of the parameters, as follows: the selection, the coverage and the bias. The selection of parameters was defined by the 95% credibility interval, not including 0. This criterion is important when the number of variables is large and the number of false positives is considered to have a higher priority over the false negatives. The 95% coverage was defined as the proportion of times, when the simulation is repeated, that the theoretical value of the parameter is within the 95% credibility interval. The third criterion was the bias, which is defined by the difference between the estimate and the theoretical parameter value.

3.1.2. *Latent structure reconstruction and survival parameter selection*

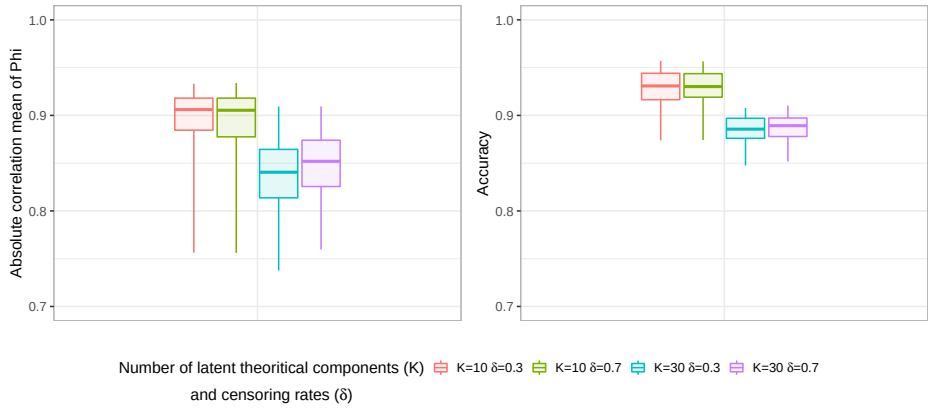


Fig. 1. Performance criteria according to different censoring rates and number of components of the isgICA submodel of the joint SisgICA model. The boxplots are defined by the median, the quartiles and the percentile interval [2.5% - 97.5%]: the reconstruction of the accuracy (left-hand column) and of the absolute correlation mean of Φ (right-hand column).

The robustness of the algorithm to reconstruct the matrices of interest is visualized in Fig. 1, and is summarized in Table 1. In all of the simulated data sets that simulated 10 components, the algorithm recovered the exact number of latent components and provided a good reconstruction of the Z matrix (accuracy $>$ 0.930 [0.873, 0.957](median [2.5 - 97.5%] percentiles)) and Φ matrix (mean absolute correlation $>$ 0.906 [0.755, 0.934]), whatever the censoring rate. As for the simulation for 30 latent components, the algorithm provided a good reconstruction of the number of latent components 29 [27-30] and of the reconstruction of Z (0.887 [0.849, 0.909]) and Φ (0.849 [0.743, 0.910]) matrices, regardless of the censoring rate.

K_{th}	censoring rate$_{th}$	K	Accuracy	$\overline{\lvert \mathrm{Cor}_{\Phi} \rvert}$
10	0.3	10 [10, 10]	0.931 [0.874, 0.957]	0.906 [0.756, 0.933]
10	0.7	10 [10, 10]	0.930 [0.874, 0.957]	0.905 [0.756, 0.934]
10	all	10 [10, 10]	0.930 [0.873, 0.957]	0.906 [0.755, 0.934]
30	0.3	29 [26.475, 30]	0.886 [0.848, 0.908]	0.841 [0.738, 0.909]
30	0.7	29 [27.475, 30]	0.889 [0.852, 0.910]	0.852 [0.760, 0.909]
30	all	29 [26.975, 30]	0.887 [0.849, 0.909]	0.849 [0.743, 0.910]

Table 1. Performance criteria according to different censoring rates and number of components of the isgICA submodel of the joint SisgICA model with the median percentile interval [2.5% - 97.5%]: the reconstruction of the number of components (third column), the accuracy (fourth column) and of the absolute correlation mean of Φ (fifth column).

We evaluated the probability of the selection (Fig. 2, first row), the coverage (Fig. 2, second row), and the bias of the survival parameters (Fig. 2, third row). The non-null parameters were selected with a probability close to one for the stronger prognostic effects ($|\beta| = 1$), and decreased with the intensity of the effect. The probability of selection of the smaller prognostic effects ($|\beta| = 0.1$) was close to those of the true null parameters ($\approx 25\%$). This low selection rate for the null parameters can be explained by a low bias and a high coverage probability. The results in Fig. 7 and Fig. 8 show the regularization of the survival prior that with the increased component number or censoring rate, the non-null parameters were underestimated with bias increasing as a function of the magnitude of the prognostic effect, which came with a decreasing coverage and selection.

3.2. *Application: early breast cancer data*

We applied our method to publicly available gene expression data that were obtained from the tumor biopsies in 614 breast cancer patients that were included in clinical trials of anthracycline-based chemotherapy [12,15], which are available in the biospear R package [36]. The standardized 1,689 probes were filtered with the jetset package [20] to retain a single probe by gene, which resulted in a final data set including the expression of 1,063 genes. As in Belhechmi et al. [4], we mapped to probes of three molecular signatures with a prognostic effect in early breast cancer (i.e., Immune System, Proliferation and Stroma invasion [17]) and one without (i.e., SRC activation signature [17]). All of the other probes were categorized as "Other."

Fig. 3 presents the binary sparse matrix of the individual heterogeneity and the sparseness of the source matrix. The model identified 15 non-zeros components (including the baseline profile) with six non-null parameters, which are summarized in Table 2.

To investigate potential biological relevance of the molecular components, Fig. 4 shows the distribution of the absolute values (to alleviate the sign identifiability

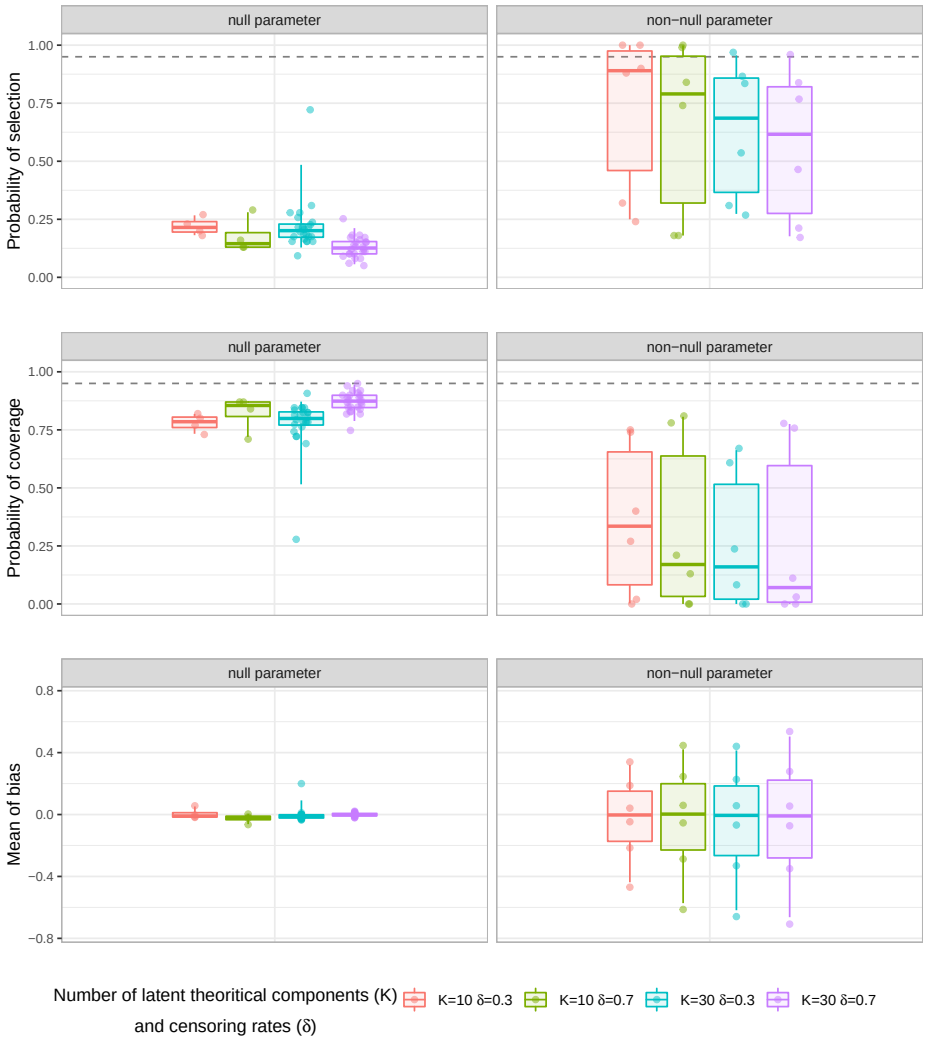


Fig. 2. Performance criteria according to different censoring rates and number of components of the Weibull-AFT submodel of the joint SisgICA model. The boxplots are defined by the median, the quartiles and the percentile interval [2.5% - 97.5%]: the selection (top row), the coverage (middle row) and the bias (bottom row).

issue) of the source matrix elements of each component according to the different breast cancer signatures selected by the survival submodel; for the complete figure, see Fig. 9. The proliferation-based signature, immune-system signature, and stroma-related signature seemed to be strongly related to the alteration components 1, 3, and 2, respectively. The SRC, which was picked as a "negative control" signature in Belhechmi et al [4], was not straightforward to map to a particular molecular

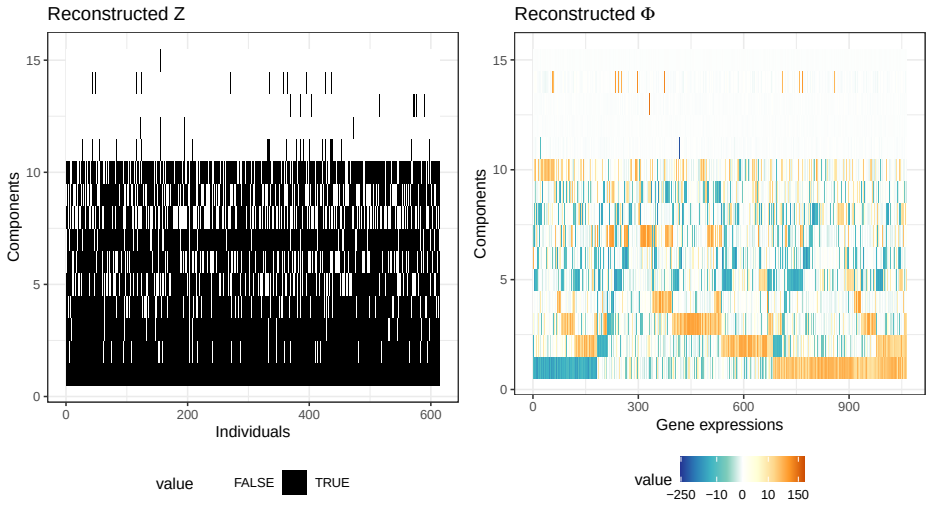


Fig. 3. Binary sparse matrix of the individual heterogeneity (weight sparsity structure, left-hand column) and source matrix components (right-hand column) extracted from the breast cancer data set using the joint SisgICA model. The model identified 15 components (including the baseline profile) with six non-null parameters that were significantly associated with the survival submodel (the baseline, and the alteration components 1, 2, 3, 4 and 6).

Components	Parameter Estimate	Hazard ratio
Baseline profile	0.605 [0.428, 0.782]	
Alteration component 1	0.372 [0.187, 0.557]	1.451 [1.206, 1.745]
Alteration component 2	0.412 [0.231, 0.593]	1.509 [1.259, 1.809]
Alteration component 3	0.327 [0.143, 0.510]	1.386 [1.154, 1.665]
Alteration component 4	0.355 [0.166, 0.544]	1.426 [1.181, 1.723]
Alteration component 6	0.256 [0.067, 0.445]	1.292 [1.069, 1.560]

Table 2. Parameter estimates of the survival submodel of the joint SisgICA model: mean credible interval [2.5% - 97.5%]

component.

4. Discussion

In this paper, we proposed a joint SisgICA model of patient survival and tumor gene expression to identify a subset of molecular signatures that may be associated with a patient's prognosis. We assumed that each tumor is a mixture of a specific subset of a limited number of molecular signatures, which allows us to model the individual molecular alterations of a baseline population profile. To our knowledge, this is the first approach that addresses this question and it could provide a new

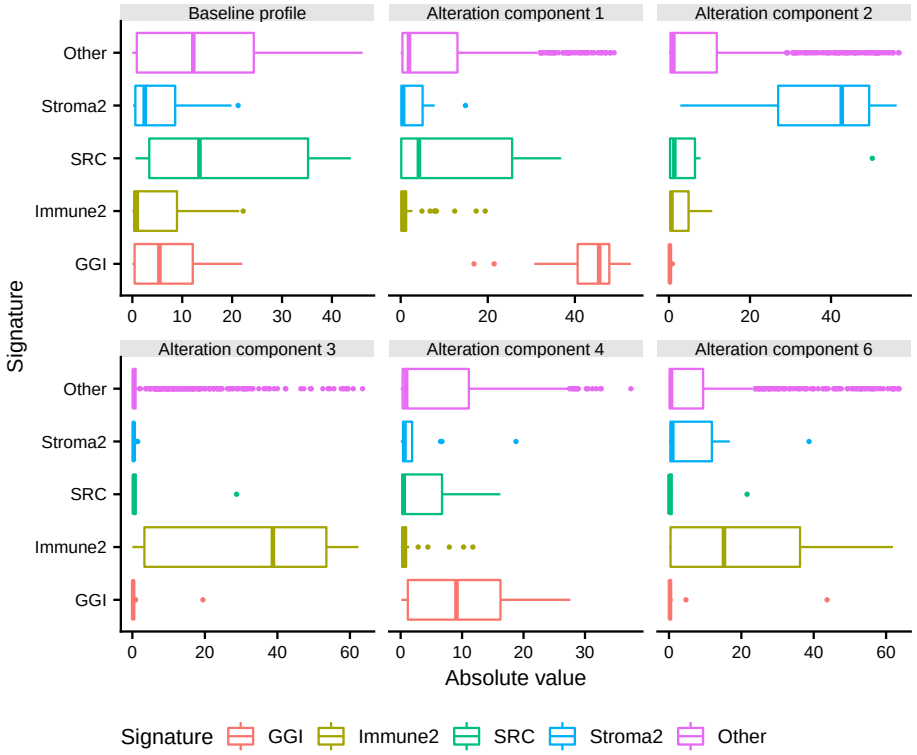


Fig. 4. Distribution of the absolute value of the 6 of 15 source matrix component elements that are significantly associated to the survival data, and associated with the probes of genes belonging to three molecular signatures with a prognostic effect in early breast cancer (i.e., GGI, Immune2, and Stroma2) and one without (i.e., SRC activation signature). All the other genes were categorized as "Other." A proliferation-related signature, immune-related signature, and stroma-related signature seemed to be related to alteration components 1, 3, and 2, respectively. The SRC was not related to one specific component.

vision to understand inter-patient heterogeneity.

Based on a nonparametric prior, our algorithm is able to identify the number of components that are present in the observed sample from an hypothetical infinite number in our simulations. Considering a double sparsity constraint for a noisy ICA, we were able to accurately re-identify the individual heterogeneity structure (accuracy $> 85\%$) in the synthetic data sets. While the baseline profile seemed to capture the molecular background that is present in the population, the noise term captures the individual specificity and the identified sparse components capture shared molecular signatures.

Due to the well-known identifiability issues of the ICA (i.e., scaling and sign reversion), the source matrix is not directly interpretable. Therefore, we used the

absolute values of the source elements to interpret the components. Identifying the most contributing genes to each component allows us to describe any potential molecular mechanisms that may be related to each of these components. Using rank based statistics (to ignore the scaling issue) on the absolute value of the source matrix elements (to alleviate the sign reversion issue), we showed that our algorithm was able to reconstruct the gene contribution to the components (correlation of the estimated components to the simulated ones $> 75\%$). Moreover, the regularization applied to this matrix using a shrinkage prior allowed us to recover a pseudo-sparseness structure that is close to the simulated one, which helps to improve the interpretation of the components.

Using the structure of the allocation of the components to the individuals (Z) as a set of explanatory variables of the survival sub-model, we showed that our algorithm was able to identify prognostic components with high to moderate effects. Due to the regularization of the prior that we considered to limit the false positive rate to an reasonable value ($< 25\%$), the components with small prognostic impacts were not selected. The high censoring rate slightly increased this regularization, which increased the true positive rate and the false positive rate, but it does not significantly impact the reconstruction of the matrices of the isgICA submodel. Increasing the number of components performed a similar role on the regularization.

We illustrated this approach using gene expression data that are measured in tumors from breast cancer patients treated with anthracycline-based chemotherapy who were included in clinical trials. This method was able to identify particular components with a high prognostic impact that are associated to known molecular pathways with strong prognostic value in early breast cancer (i.e., proliferation, immune system and stromal pathway [17]). The non-selected components were not related to the existing molecular prognostic pathways, which supports our assumption that the algorithm can recognize cancer mechanisms.

The next major perspective of this work will be to include the standard clinicopathological factors in the model (e.g., age, cancer stage, etc.) and the treatments to identify molecular signatures associated with differential survival benefit after particular treatments. One option would be to use causal inference methods to provide a more interpretable model and limit identifiability issues. Finally, as proposed in the wide literature of deconvolution methods for omics data, our approach may also be extended to other non-Gaussian omics data, such as count (raw RNAseq, proteomics) or binary (mutations) data using different link functions, and to multi-omics, performing a multiblock analysis that shares the same latent process or a hierarchical latent structure [30].

5. Conclusion

In this paper, we proposed a joint SisgICA model that is based on gene expression and patient survival to identify the components that are related to different molecular mechanisms that are involved in the cancer mechanisms and the patient

prognostic. Considering that each tumor does not result from only one molecular signature but results from a patient-specific mixture of several molecular signatures, our approach may contribute to the understanding of molecular mechanisms.

Appendix A.

A.1. *Joint model: SisgICA priors*

We suppose the independence of X and Y conditional on Z. Using Z to joint the isgICA and the survival model, the baseline profile is confounded with the intercept of the survival submodel and the alterations for the baseline profile corresponded to explanatory covariables. The complete model is defined by Eq. (A.1) and Fig. 5 with the distribution abbreviations: $\mathcal{B}$, Beta distribution; $\mathcal{B}er$, Bernoulli distribution; $\mathcal{G}$, Gamma distribution; $\mathcal{N}$, Normal distribution; $\mathcal{MN}$, Multinormal distribution; $\mathcal{W}e$, Weibull distribution; $\mathcal{W}is$ Wishart distribution:

$$
\begin{array}{l}
X = \Phi(W \circ Z) + E \\
X_i \sim \mathcal{N}(\Phi(W \circ Z), \mathrm{diag}(\tau_{E_1}^{-1}, ..., \tau_{E_p}^{-1})) \\
\Phi_{j,k} \sim \mathcal{N}(0, \tau_{\phi_{j,k}}^{-1}) \\
W_{k,i} \sim \mathcal{N}(0, \tau_W^{-1} I_K) \\
\tau_{\phi_{j,k}} \sim \mathcal{G}(c_{j,k}, d_{j,k}) \\
\tau_W \sim \mathcal{G}(e, f) \\
\\
\log T = Y = Z^T\beta + \sigma_Y \epsilon \\
Y \sim \mathrm{Gumbel}(Z^T\beta, \sigma_Y) \\
\beta \sim \mathcal{MN}(\mu_\beta, (\kappa_0 \Lambda_\beta)^{-1}) \\
\Lambda_\beta \sim \mathcal{W}is(\nu_{\Lambda_\beta}, V_{\Lambda_\beta}) \\
\sigma_Y \sim \mathcal{LN}(m_{\sigma_Y}, s^2_{\sigma_Y}) \\
\\
Z_{k^*,i} \sim \mathcal{B}e(\pi_k) \\
\pi_{k^*} \sim \mathcal{B}(\frac{a}{K}, \frac{b(K-1)}{K}) \\
Z_{0,} = [1, .., 1]_N
\end{array}
\tag{A.1}
$$

where i = 1, ..., N; j = 1, ..., P; k = 0, ..., K; and k*= 1,...,K.

A.2. *Variational SisgICA equations*

Inspired by [10], we used a coordinate ascent algorithm to minimise the evidence lower bound with a mean field approximation. The updated equations of the variational parameters are described below, where i=1,. . . ,N,j=1,. . . ,P,k=0,. . . ,K_{max} and k*=1,. . . ,K_{max}.

We define the lower bound for the evidence based on the independence of X and

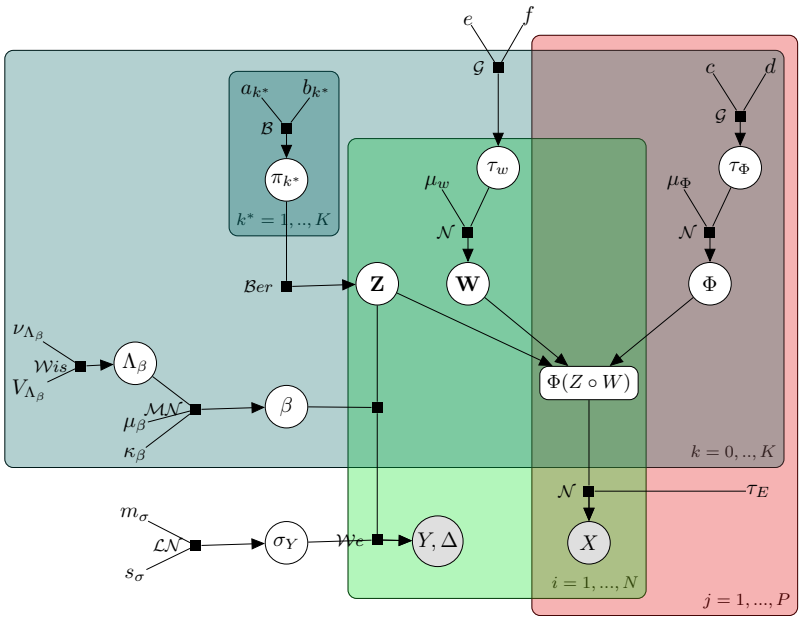


Fig. 5. **Graphical representation of the joint SisgICA model.** The observed variables are denoted with shaded nodes, while unobserved variables are shown as white nodes. Abbreviation: $\mathcal{B}$, Beta distribution; $\mathcal{B}er$, Bernoulli distribution; $\mathcal{G}$, Gamma distribution; $\mathcal{N}$, Normal distribution; $\mathcal{MN}$, Multinormal distribution; $\mathcal{W}e$, Weibull distribution; and $\mathcal{W}is$, Wishart distribution.

Y conditional on Z :

$$\begin{aligned} ELBO(q) &= \mathbb{E}_q[\log(p(X,Y|\theta)] - KL(q(\theta)|p(\theta)) \\ &= \underbrace{\mathbb{E}_q[\log(p(X|\theta_X)] + \mathbb{E}_q[\log(p(Y|\theta_Y)]}_{\text{X independent at Y conditional on Z}} - KL(q(\theta_X)|p(\theta_X)) - KL(q(\theta_Y)|p(\theta_Y)) \\ &= \underbrace{\mathbb{E}_q[\log(p(X|\theta_X)] - KL(q(\theta_X)|p(\theta_X))}_{\text{isgICA submodel}} + \underbrace{\mathbb{E}_q[\log(p(Y|\theta_Y)] - KL(q(\theta_Y)|p(\theta_Y))}_{\text{Survival submodel}} \end{aligned} \tag{A.2}$$

We define the joint distribution as the factorized variational distribution (mean-field assumption) :

$$p(Z, X, \Delta, Y) = p(Z|\pi)p(X|Z)p(Y|\Delta, Z) \tag{A.3}$$

Variational update of $<z^*_{k^*,i}>$:

$$\begin{aligned} q(z^*_{k^*,i}|-) &= Bernoulli(z^*_{k^*,i}; \pi_{k^*}) \\ &= \frac{q(z_{k^*,i}=1|-)}{q(z_{k^*,i}=1|-)+q(z_{k^*,i}=0|-)} \end{aligned} \tag{A.4}$$

with the symbol $< \bullet >$ define the expectation of the argument;

with

$$\begin{aligned} q(z^*_{k^*,i} = 1|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z_{-k^*,i}},\boldsymbol{W},\boldsymbol{\Phi},\tau_E) \\ \propto \exp(<\ln(\pi_{k^*})> \quad -\tfrac{1}{2}(<\Phi_{k^*}>^T diag(\tau_E)<\Phi_{k^*}><w_{k^*,i}>^2 \\ -2<\Phi_{k^*}>^T diag(\tau_E)\mathbf{X}_i^{-k^*}<w_{k^*,i}>)) \\ +\delta_i(<\epsilon_i> - \log<\sigma_Y>) - exp(<\epsilon_i>)) \\ \text{with } \mathbf{X}_{j,i}^{-k^*} = x_{j,i} - \sum_{l=0,l\neq k^*}^{K} <\Phi_{j,l}><z_{l,i}><w_{l,i}> \end{aligned} \tag{A.5}$$

and

$$\begin{aligned} q(z_{k^*,i} = 0|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z_{-k^*,i}},\boldsymbol{W},\boldsymbol{\Phi},\tau_E) \\ \propto \exp(<\ln(1-\pi_{k^*})>) \end{aligned} \tag{A.6}$$

Variational update of $<\pi_{k^*}>$:

$$\begin{aligned} q(\pi_{k^*}|-) = Beta(\pi_{k^*,i}; a'_{k^*}, b'_{k^*}) \\ <a'_{k^*}> = \sum_{i=1}^{N} <z_{k^*,i}> + \tfrac{a}{K_{max}} \\ <b'_{k^*}> = N + \tfrac{b(K_{max}-1)}{K_{max}} - \sum_{i=1}^{N} <z_{k^*,i}>; \end{aligned} \tag{A.7}$$

Variational update of $<\Phi_{j,k}>$:

$$\begin{aligned} q(\Phi_{j,k}|-) = \mathcal{N}(\Phi_{j,k}; \mu_{\Phi_{j,k}}, \tau^{-1}_{\Phi_{j,k}}) \\ <\tau_{\Phi_{j,k}}> = \sum_{i=1}^{N} \tau_{E_j} <w_{k,i}>^2 <z_{k,i}>^2 + <\tau_{\Phi_{j,k}}> \\ <\mu_{\Phi_{j,k}}> = <\tau_{\Phi_{j,k}}>^{-1} \Big(\sum_{i=1}^{N} \tau_{E_j} <w_{k,i}><z_{k,i}> \mathbf{X}_{j,i}^{-k} \Big); \end{aligned} \tag{A.8}$$

Variational update of $<W_i>$:

$$\begin{aligned} q(W_i|-) = \mathcal{N}(W_i; \mu_{W_i}, \tau^{-1}_{W_i}) \\ <\tau_{W_i}> = (\tau_E)(<\Phi> \circ \widetilde{Z}_i^T) + <\tau_W> I_K \\ <\mu_{W_i}> = <\tau^{-1}_{W_i}>(<\Phi> \circ \widetilde{Z}_i) diag(\tau_E) x_i \\ \text{with } \widetilde{Z}_i := [<z_i>, ..., <z_i>] \ , \ <z_i> \text{vector repeated K times}; \end{aligned} \tag{A.9}$$

Variational update of $<\tau_{\Phi_{j,k}}>$:

$$\begin{aligned} q(\tau_{\Phi_{j,k}}|-) = Gamma(\tau_{\Phi_{j,k}}; c_{j,k}, d_{j,k}) \\ <c_{j,k}> = c_0 + \tfrac{1}{2} \\ <d_{j,k}> = d_0 + \tfrac{1}{2} <\Phi_{j,k}>^2; \end{aligned} \tag{A.10}$$

Variational update of $<\tau_W>$:

$$\begin{aligned} p(\tau_W|-) = Gamma(\tau_W; e, f) \\ <e> = e_0 + \tfrac{NK_{max}}{2} \\ <f> = f_0 + \tfrac{1}{2} \sum_{i=1}^{N} <w_i^T><w_i>. \end{aligned} \tag{A.11}$$

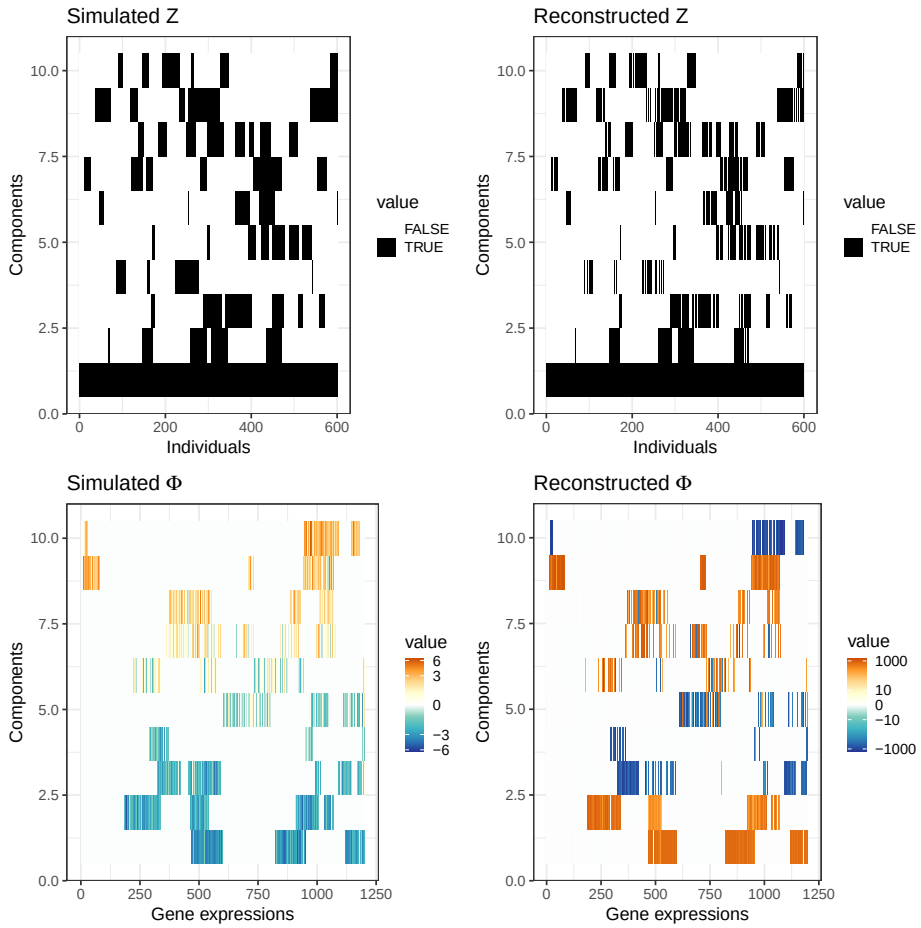


Fig. 6. Visual representation of the reconstruction of the weight sparsity structure (Z, top row) and sources (Φ, bottom row) in the scenario $\delta = 0.7$ and $K_{\text{th}} = 10$. These reconstructions correspond to an accuracy of 0.962 for the reconstruction of Z, and a correlation of 0.900 for the reconstruction of Φ.

A.3. *Visual representation of the reconstructed Z and Φ in the scenario $\delta = 0.7$ and $K_{th} = 10$*

A.4. *Relation between the bias and the coverage / selection in the synthetic data*

A.5. *Complete distribution of the absolute value of the source matrix elements in the early breast cancer data*

References

1. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia

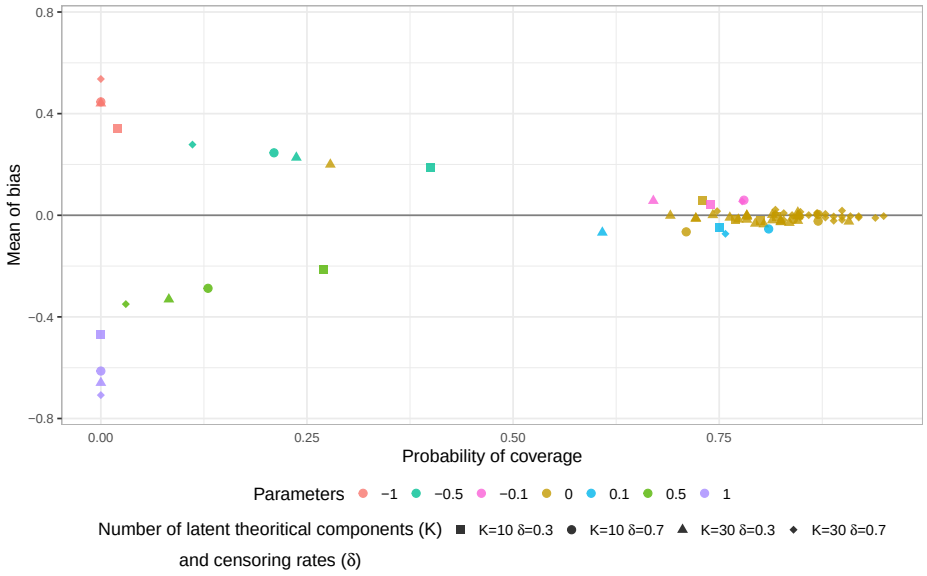


Fig. 7. Visual representation of the relation of the bias and the coverage according to the scenarios with censoring rates $\{0.3, 0.7\}$ and $K_{\text{th}} = \{10, 30\}$.

Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mané D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X, *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015, software available from tensorflow.org.

2. Baek S, Ho YY, Ma Y, Using sufficient direction factor model to analyze latent activities associated with breast cancer survival, *Biometrics* **76**(4):1340–1350, 2020. doi:10.1111/biom.13208, URL `https://doi.org/10.1111/biom.13208`.
3. Bastien P, Bertrand F, Meyer N, Maumy-Bertrand M, Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data, *Bioinformatics* **31**(3):397–404, 2014. doi:10.1093/bioinformatics/btu660.
4. Belhechmi S, Bin RD, Rotolo F, Michiels S, Accounting for grouped predictor variables or pathways in high-dimensional penalized Cox regression models, *BMC Bioinformatics 2020 21:1* **21**(1):1–20, 2020. doi:10.1186/S12859-020-03618-Y, URL `https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03618-y`.
5. Bertrand F, Bastien P, Meyer N, Maumy-Bertrand M, plsrcox, cox-models in a high dimensional setting in r, *Proceedings of User2014!*, p. 177, 2014.
6. Bhattacharya A, Dunson DB, Sparse bayesian infinite factor models, *Biometrika* , 2011. doi:10.1093/biomet/asr013.
7. Blei DM, Kucukelbir A, McAuliffe JD, Variational inference: a review for statisticians, *Journal of the American Statistical Association* **112**(518):859–877, 2017. doi:10.1080/01621459.2017.1285773.
8. Bouveyron C, Latouche P, Mattei PA, Exact dimensionality selection for Bayesian PCA, *Scandinavian Journal of Statistics* **47**(1):196–211, 2020. doi:10.1111/sjos.12424.
9. Cai J, Liang C, Bayesian analysis of semi-parametric cox models with latent variables,

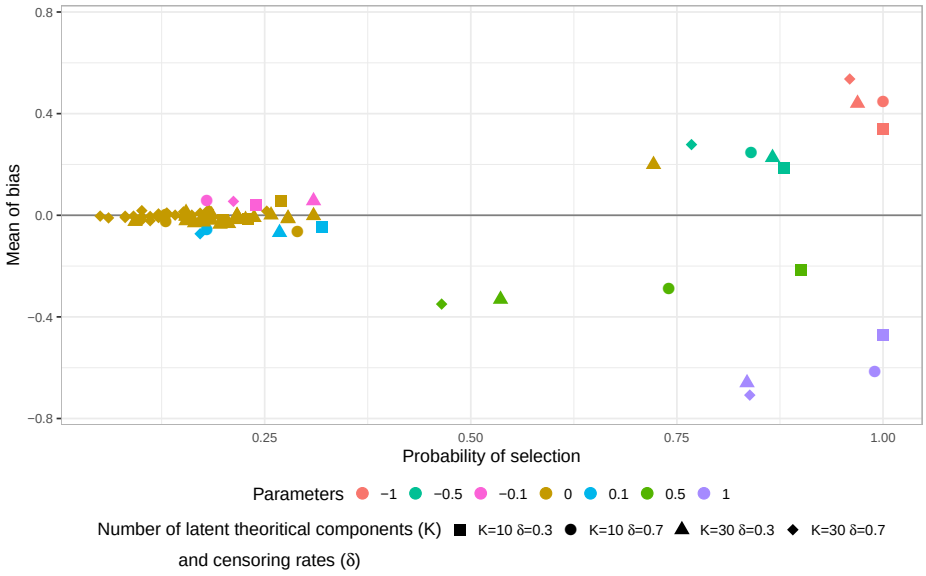


Fig. 8. Visual representation of the relation of the bias and the selection according to the scenarios with censoring rates $\{0.3, 0.7\}$ and $K_{\text{th}} = \{10, 30\}$.

Stat Methods Med Res **28**(7):2150–2164, 2019.

10. Chen B, Chen M, Paisley J, Zaas A, Woods C, Ginsburg GS, Hero A, Lucas J, Dunson D, Carin L, Bayesian inference of the number of factors in gene-expression analysis: application to human virus challenge studies, *BMC Bioinformatics* **11**(1):552, 2010. doi:10.1186/1471-2105-11-552.
11. Cunningham JP, Ghahramani Z, Linear dimensionality reduction: Survey, insights, and generalizations, *Journal of Machine Learning Research* **16**(89):2859–2900, 2015, URL `http://jmlr.org/papers/v16/cunningham15a.html`.
12. Desmedt C, Di Leo A, de Azambuja E, Larsimont D, Haibe-Kains B, Selleslags J, Delaloge S, Duhem C, Kains JP, Carly B, Maerevoet M, Vindevoghel A, Rouas G, Lallemand F, Durbecq V, Cardoso F, Salgado R, Rovere R, Bontempi G, Michiels S, Buyse M, Nogaret JM, Qi Y, Symmans F, Pusztai L, D'Hondt V, Piccart-Gebhart M, Sotiriou C, Multifactorial approach to predicting resistance to anthracyclines, *Journal of Clinical Oncology* **29**(12):1578–1586, 2011. doi:10.1200/JCO.2010.31.2231, URL `http://www.ncbi.nlm.nih.gov/pubmed/21422418http://ascopubs.org/doi/10.1200/JCO.2010.31.2231`.
13. Ferrari S, Cribari-Neto F, Beta regression for modelling rates and proportions, *Journal of Applied Statistics* **31**(7):799–815, 2004. doi:10.1080/0266476042000214501, URL `https://doi.org/10.1080/0266476042000214501`.
14. Gal J, Bailleux C, Chardin D, Pourcher T, Gilhodes J, Jing L, Guigonis JM, Ferrero JM, Milano G, Mograbi B, Brest P, Chateau Y, Humbert O, Chamorey E, Comparison of unsupervised machine-learning methods to identify metabolomic signatures in patients with localized breast cancer, *Computational and Structural Biotechnology Journal* **18**:1509–1524, 2020. doi:10.1016/j.csbj.2020.05.021, URL `https://doi.org/10.1016/j.csbj.2020.05.021`.

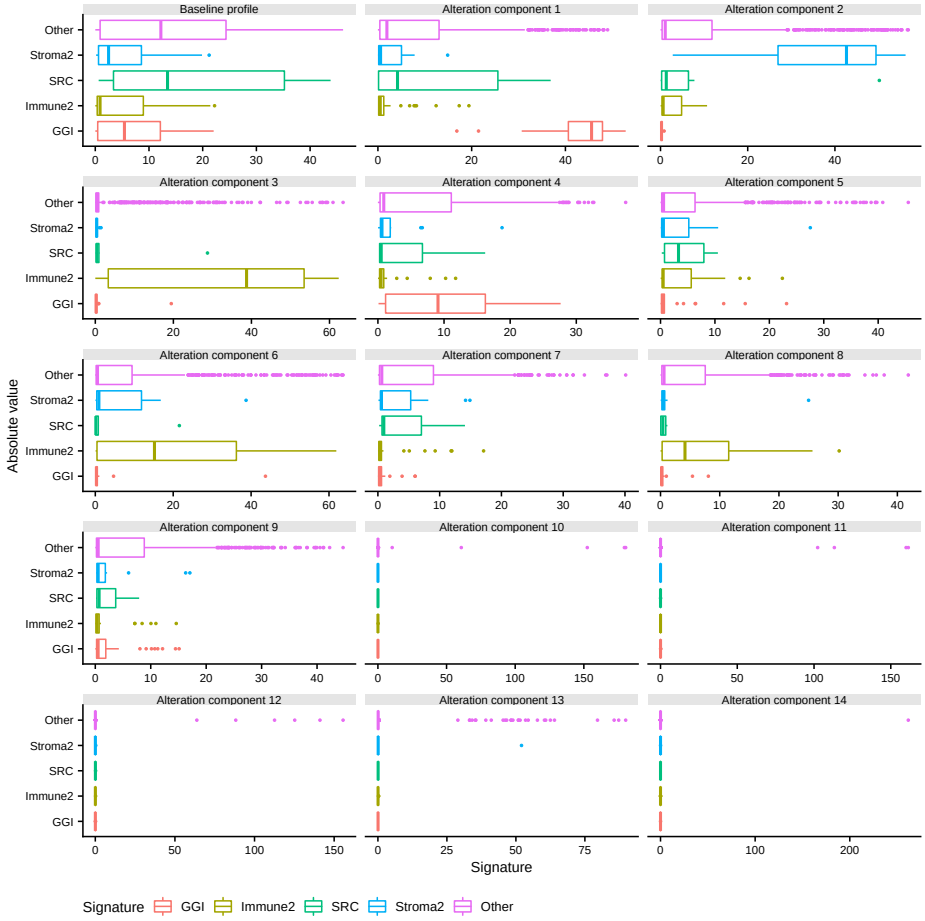


Fig. 9. Distribution of the absolute value of the 15 source matrix component elements that are associated with the probes of genes belonging to three molecular signatures with a known prognostic effect in early breast cancer (GGI, Immune2, and Stroma2) and one without (SRC activation signature). All of the other genes were categorized as "Other."

15. Hatzis C, Pusztai L, Valero V, Booser DJ, Esserman L, Lluch A, Vidaurre T, Holmes F, Souchon E, Wang H, Martin M, Cotrina J, Gomez H, Hubbard R, Chacón JI, Ferrer-Lozano J, Dyer R, Buxton M, Gong Y, Wu Y, Ibrahim N, Andreopoulou E, Ueno NT, Hunt K, Yang W, Nazario A, DeMichele A, O'Shaughnessy J, Hortobagyi GN, Symmans WF, A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer, *JAMA* **305**(18):1873, 2011. doi:10.1001/jama.2011.593, URL `http://www.ncbi.nlm.nih.gov/pubmed/21558518http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5638042http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2011.593`.
16. Hjort NL, Nonparametric Bayes estimators based on beta processes in models for

life history data, *The Annals of Statistics* **18**(3):1259–1294, 1990. doi:10.1214/aos/1176347749, URL http://projecteuclid.org/euclid.aos/1176347749.

17. Ignatiadis M, Singhal SK, Desmedt C, Haibe-Kains B, Criscitiello C, Andre F, Loi S, Piccart M, Michiels S, Sotiriou C, Gene modules and response to neoadjuvant chemotherapy in breast cancer subtypes: a pooled analysis, *Journal of Clinical Oncology* **30**(16):1996–2004, 2012. doi:10.1200/JCO.2011.39.5624, URL http://ascopubs.org/doi/10.1200/JCO.2011.39.5624.
18. Knoblauch J, Jewson J, Damoulas T, *Generalized variational inference: three arguments for deriving new posteriors*, 2019.
19. Knowles DA, Ghahramani Z, Infinite sparse factor analysis and infinite independent components analysis., *ICA 2007*, Lecture Notes in Computer Science, Vol. 4666, Springer, pp. 381–388, 2007. ISBN 978-3-540-74493-1, URL http://dblp.uni-trier.de/db/conf/ica/ica2007.html#KnowlesG07.
20. Li Q, Birkbak NJ, Gyorffy B, Szallasi Z, Eklund AC, Jetset: selecting the optimal microarray probe set to represent a gene, *BMC Bioinformatics* **12**(1):474, 2011. doi:10.1186/1471-2105-12-474, URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-474.
21. Liu E, Lim K, Using the weibull accelerated failure time regression model to predict time to health events, *bioRxiv* , 2018. doi:10.1101/362186.
22. Liu M, Sun J, Herazo-Maya JD, Kaminski N, Zhao H, Joint models for time-to-event data and longitudinal biomarkers of high dimension, *Statistics in Biosciences* **11**(3):614–629, 2019. doi:10.1007/s12561-019-09256-0, URL https://doi.org/10.1007/s12561-019-09256-0.
23. McCurdy SR, Molinaro A, Pachter L, *A latent variable model for survival time prediction with censoring and diverse covariates*, 2017.
24. Paisley J, Carin L, Nonparametric factor analysis with beta process priors, *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, 2009. ISBN 9781605585161. doi:10.1145/1553374.1553474.
25. Paisley J, Jordan MI, *A constructive definition of the beta process*, 2016.
26. Pan D, Kang K, Wang C, Song X, Bayesian proportional hazards model with latent variables, *Statistical Methods in Medical Research* **28**(4):986–1002, 2019. doi:10.1177/0962280217740608, pMID: 29226787.
27. Park M, Kim D, Moon K, Park T, Integrative analysis of multi-omics data based on blockwise sparse principal components, *International Journal of Molecular Sciences* **21**(21):8202, 2020. doi:10.3390/ijms21218202, URL https://doi.org/10.3390/ijms21218202.
28. Peterson DR, Constructing multivariate prognostic gene signatures with censored survival data, in *Methods in Molecular Biology*, , Springer New York, pp. 85–101, 2013. doi:10.1007/978-1-60327-337-4_6, URL https://doi.org/10.1007/978-1-60327-337-4_6.
29. Ranganath R, Gerrish S, Blei DM, *Black Box Variational Inference*, 2013.
30. Ray P, Zheng L, Lucas J, Carin L, Bayesian joint analysis of heterogeneous genomics data, *Bioinformatics* **30**(10):1370–1376, 2014. doi:10.1093/bioinformatics/btu064, URL https://doi.org/10.1093/bioinformatics/btu064.
31. Rincourt SL, Michiels S, Drubay D, Complex disease individual molecular characterization using infinite sparse graphical independent component analysis, *Cancer informatics* , in-revision 2022.
32. Rohart F, Gautier B, Singh A, Lê Cao KA, mixomics: An r package for 'omics feature selection and multiple data integration, *PLOS Computational Biology* **13**(11):1–19, 2017. doi:10.1371/journal.pcbi.1005752.

33. Samuel Wilson, ParBayesianOptimization: parallel Bayesian optimization of hyperparameters, *https://cranr-projectorg/package=parbayesianoptimization* , 2020.
34. Sokol A, Maathuis MH, Falkeborg B, Quantifying identifiability in independent component analysis, *Electronic Journal of Statistics* **8**:1438–1459, 2014. doi:10.1214/14-EJS932.
35. Székely GJ, Rizzo ML, Bakirov NK, Measuring and testing dependence by correlation of distances, *Annals of Statistics* **35**(6):2769–2794, 2007. doi:10.1214/009053607000000505, URL `https://projecteuclid.org/euclid.aos/1201012979`.
36. Ternès N, Rotolo F, Michiels S, biospear: an R package for biomarker selection in penalized Cox regression, *Bioinformatics (Oxford, England)* **34**(1):112–113, 2018. doi:10.1093/bioinformatics/btx560, URL `https://doi.org/10.1093/bioinformatics/btx560`.
37. Vidaurre D, van Gerven MAJ, Bielza C, Larrañaga P, Heskes T, Bayesian sparse partial least squares, *Neural Computation* **25**(12):3318–3339, 2013. doi:10.1162/neco_a_00524, URL `https://doi.org/10.1162/neco_a_00524`.
38. Witten DM, Tibshirani R, Survival analysis with high-dimensional covariates, *Statistical Methods in Medical Research* **19**(1):29–51, 2009. doi:10.1177/0962280209105024, URL `https://doi.org/10.1177/0962280209105024`.
39. Yoshida R, West M, Bayesian learning in sparse graphical factor models via variational mean-field annealing, *Journal of Machine Learning Research* **11**:1771–1798, 2010.

université PARIS-SACLAY | ÉCOLE DOCTORALE Santé Publique (EDSP)

Titre : Intégration de données omiques pour la modélisation de l'impact de l'hétérogénéité inter-tumorale dans la survie de patients atteints de cancer

Mots clés : Hétérogénéité tumorale ; Analyse en composantes indépendantes ; survie ; pronostic ; variational Bayes ; données de grande dimension

Résumé : Le développement de signatures moléculaires pronostiques basées sur des biomarqueurs omiques tenant compte de l'hétérogénéité inter-patients est un défi majeur pour la médecine de précision. L'objectif de cette thèse était de proposer de nouvelles approches pour l'identification de ces mécanismes tumoraux.

Dans un premier travail, nous avons déconvolué les données omiques à l'aide d'une analyse Bayésienne en composantes indépendantes Bayésienne non-paramétrique avec une double structure de parcimonie pour les matrices des sources et des poids, correspondant respectivement aux associations gène-composante et individu-composante.

Nous avons ensuite proposé ensuite un modèle conjoint de cette hétérogénéité et de la survie de patients, en supposant que l'expression tumorale résulte d'un mélange d'un sous-ensemble de signatures indépendantes.

Les algorithmes proposés fournissent de nouvelles perspectives sur l'hétérogénéité moléculaire individuelle et sur l'association avec le pronostic du patient afin de mieux comprendre les mécanismes tumoraux complexes.

Title : Omics integration to model the impact of inter-tumoral heterogeneity on the survival of patients with cancer

Keywords : Tumor heterogeneity ; Independent component analysis ; survival ; prognosis ; variational Bayes ; High dimensional data

Abstract : The development of prognostic molecular signatures based on omic biomarkers considering the inter-patient heterogeneity is a key challenge in precision medicine. The objective of this thesis was to propose new approaches for the identification of these tumor mechanisms.

In the first axis, we deconvoluted the omics data using a Bayesian non-parametric independent component analysis with a double sparseness structure for the source and the weight matrices, corresponding to the gene-component and individual-component associations, respectively.

In the second axis, we proposed a joint model of this heterogeneity and the patient survival, assuming that tumor expression results from a mixture of a subset of independent signatures.

The proposed algorithms provide new insights into the individual molecular heterogeneity, and with regards to patient prognosis to better understand the complex tumor mechanisms.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France