N^o D'ORDRE : 8889

UNIVERSITÉ DE PARIS SUD
U.F.R. SCIENTIFIQUE D'ORSAY

THÈSE

présentée pour obtenir le titre de

DOCTEUR EN SCIENCES DE L'UNIVERSITÉ PARIS XI

Spécialité : Mathématiques

par

Fanny VILLERS

Sujet de la thèse :

Tests et sélection de modèles pour l'analyse de données protéomiques et transcriptomiques.

Rapporteurs :	M.	Christophe	AMBROISE
	Mme	Béatrice	LAURENT

Soutenue le 12 décembre 2007 devant le jury composé de :

M.	Christophe	AMBROISE	Rapporteur
M.	Jean-François	CHICH	Examinateur
Mme	Sylvie	HUET	Co-directrice de thèse
M.	Pascal	MASSART	Directeur de thèse
Mme	Sophie	SCHBATH	Présidente du jury

Remerciements

Tout d'abord un grand merci à Sylvie pour m'avoir encadrée pendant toute la durée de cette thèse. Sylvie, merci pour ton aide, ta disponibilité et ta bonne humeur.

Je tiens également à remercier Pascal qui est à l'origine de cette rencontre. Pascal, merci de m'avoir fait confiance et encouragée à entreprendre cette thèse.

Je remercie Christophe et Béatrice qui m'ont fait l'honneur et le plaisir de rapporter mon travail. Et merci à Sophie et Jean-François pour avoir participé à mon jury.

Un grand merci à Nicolas sans qui cette thèse ne serait pas ce qu'elle est.

Un grand merci également à Brigitte pour son aide si précieuse et à Christophe G. pour avoir relu si attentivement mon travail.

Je voudrais également remercier toute l'unité MIA pour m'avoir accueillie chaleureusement tout au long de ces années. En particulier merci à Eric pour son aide informatique et pour avoir été mon chauffeur pendant ces trois ans. Merci à Suzanne pour son aide en LaTeX et à Elisa pour ses relectures. Merci également à Liliana et Olivier pour leurs réponses et à Patricia, Nicole et Rosa pour leur efficacité. Et à tous les membres de l'unité que je ne cite pas, merci pour votre gentillesse et tout simplement pour avoir été là.

Je n'oublie évidemment pas Anne, Valérie, Etienne, Fadia, Aude et Najat. Merci pour votre amitié et votre soutien.

Pour finir, merci à Sam pour s'être si bien occupé de moi lorsque je n'avais pas le temps de m'occuper de lui. Et merci à mon frère et mes parents pour leur confiance inconditionnelle et leur fierté.

Table des matières

Introduction

L'étude de l'expression des gènes fait appel à deux approches : d'une part l'analyse du transcriptome constitué par l'ensemble des ARN_m présents dans une cellule dans une situation donnée, et d'autre part l'analyse du protéome représenté par les protéines qui sont codées par ces ARN_m. Les données protéomiques et transcriptomiques sont caractérisées par un grand nombre de variables et en général peu d'observations. Cette thèse a pour but de proposer des méthodes statistiques adaptées pour traiter ces données.

La première partie concerne l'analyse de données protéomiques. Parmi les différentes étapes de l'analyse de données protéomiques, nous nous sommes focalisés sur l'analyse différentielle du protéome. Il s'agit de comparer la quantité de protéine exprimée, ou abondance, selon différentes conditions expérimentales. Ces conditions expérimentales peuvent représenter différents milieux de cultures, ou des patients sains et malades, etc. L'analyse différentielle repose le plus souvent sur l'analyse de données issues d'images de gels d'électrophorèse 2D.

La modélisation statistique des observations issues des gels d'E2D est complexe et nécessite la mise en oeuvre de méthodes statistiques non triviales pour de non statisticiens : planification d'expériences, validation de modèle, procédures de tests multiples, etc. Le Chapitre 1 présente ces méthodes statistiques de façon à les rendre accessibles aux collègues biologistes.

L'analyse de ces données soulève également des problèmes de statistique ouverts ou non résolus de façon complétement satisfaisante. Pour le statisticien, l'analyse différentielle revient à détecter les composantes non nulles de l'espérance d'un vecteur gaussien de grande dimension. Dans le cas où on désire comparer simultanément un grand nombre de conditions expérimentales (au moins 3), les composantes de ce vecteur gaussien ne sont plus indépendantes. Nous proposons d'estimer le nombre de composantes non nulles d'un vecteur gaussien dont la structure de covariance est connue (à une constante près) à l'aide d'une méthode de "sélection de modèles" reposant sur un critère des moindres carrés pénalisé. Ce travail est décrit dans le Chapitre 2.

Les deuxième et troisième parties concernent la recherche de liens fonctionnels entre entités biologiques (gènes, protéines, petites molécules, etc.). Récemment a émergé l'utilisation des modèles graphiques gaussiens pour décrire les réseaux génétiques. Un graphe est constitué d'un ensemble de sommets et d'un ensemble d'arêtes ou liaisons entre les sommets. Les modèles graphiques gaussiens forment un outil permettant d'analyser et de visualiser des dépendances conditionnelles entre variables aléatoires. Ils sont fréquemment utilisés pour des études financières ou sociologiques.

Dans l'étude des données génomiques, la particularité est que l'on dispose de peu d'expériences par rapport au nombre de gènes. Récemment plusieurs méthodes ont été proposées dans ce cadre de données de grandes dimensions, dans le but d'inférer des réseaux géniques à partir de l'observation de l'expression de gènes. Nous disposons de peu de résultats théoriques pour ces méthodes, ou alors de résultats asymptotiques et donc peu informatifs sur leur comportement en situation réelle. Dans le domaine des données post-génomiques il est par ailleurs difficile de valider une méthode statistique par les connaissances biologiques. Nous avons donc entrepris un travail de simulation afin d'évaluer le domaine de validité de plusieurs méthodes d'estimation de graphe. Ce travail est décrit dans le Chapitre 3.

Dans certains cas les biologistes ont une bonne connaissance des relations directes entre les gènes et/ou les protéines. Sur la base d'observation de l'expression de ces entités, il peut s'avérer intéressant de tester si le graphe qui s'en déduit est correct. Nous proposons en collaboration avec Nicolas Verzelen[1], une méthode permettant de construire un test de validation de graphe. Cela revient à tester une hypothèse linéaire dans un modèle de régression multivariée, dont les variables explicatives sont aléatoires. Ce travail est décrit dans les Chapitres 4 et 5.

Sélection de modèles pour l'analyse différentielle d'images d'électrophorèse

L'électrophorèse bi-dimensionnelle

L'analyse différentielle du protéome a pour but de comparer les expressions des protéines correspondant à deux ou plusieurs conditions expérimentales différentes. Les données sur lesquelles elle porte sont souvent issues de l'électrophorèse 2D. L'E2D est une méthode permettant de séparer les protéines selon leur charge électrique et leur masse moléculaire. Les données issues de l'E2D se présentent sous forme d'images. Un exemple d'image de gel d'électrophorèse 2D est donné dans la figure 1.

Les logiciels d'analyse d'images utilisés par les biologistes permettent de détecter les spots sur les gels d'E2D, de les apparier et d'en évaluer le volume. Le volume du spot caractérise l'abondance de la protéine. Chaque spot est identifié par un numéro et caractérisé par les observations du volume sur chaque gel. L'analyse différentielle consiste à détecter les spots dont le volume diffère selon les conditions expérimentales. En terme statistique, il s'agit de tester simultanément un grand nombre d'hypothèses : pour chaque spot, on teste l'hypothèse d'absence de différence d'expression. Ce problème a déjà été bien étudié pour l'analyse différentielle du transcriptome, et les différentes étapes de l'analyse statistique de ces données sont maintenant bien maitrisées. Mais leur application aux données provenant des gels d'E2D n'est pas triviale. En effet, les données étant issues d'une analyse d'image complexe présentent une grande variabilité ; le nombre de données manquantes est grand ; le nombre de répétitions est faible ; certaines observations ne sont pas pertinentes pour des raisons techniques. Face à ces difficultés, les logiciels commerciaux utilisés par

[1] Université Paris Sud

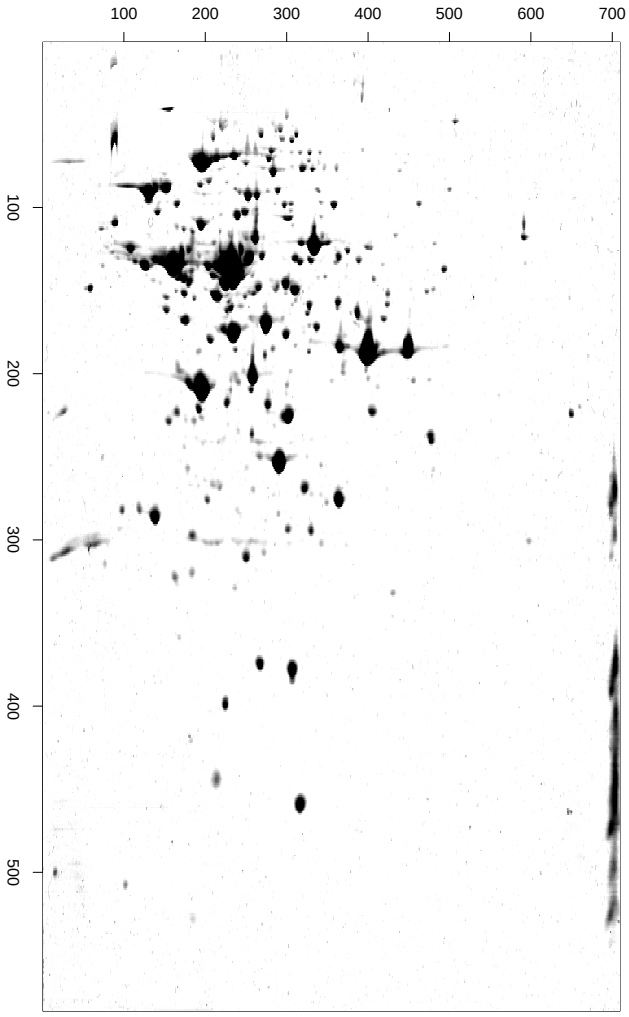

FIG. 1 – Séparation et quantification des protéines

les biologistes pour l'analyse différentielle du protéome ne sont en général pas complétement satisfaisants. En collaboration avec l'unité MIA-Jouy, je me suis impliquée dans un travail de mise à disposition de méthodes permettant d'analyser au mieux ces données. Le Chapitre 1 correspond à un article publié dans Journal of Chromatography qui présente les méthodes statistiques pertinentes pour l'analyse de ces données. La première étape concerne la planification d'expérience afin d'éviter les biais d'expérimentation. La deuxième étape concerne l'analyse différentielle proprement dite. Nous considérons deux approches possibles : une approche spot par spot dans laquelle on compare pour chaque spot les différences d'expression en ne tenant compte que des informations pour ce spot ; une approche globale qui tient compte des observations pour tous les spots, dans laquelle on étudie les différences d'interaction des spots entre les conditions deux à deux dans un modèle d'analyse de variance. Nous soulignons l'importance d'analyses préliminaires pour supprimer les spots non pertinents, valider le modèle statistique et mettons en évidence l'importance de la modélisation de la variance et du traitement des données manquantes.

Analyse différentielle basée sur l'étude des différences d'interaction

Pour détecter les spots dont l'expression diffère selon la condition, nous considérons le modèle d'analyse de variance proposé dans le chapitre 1. Le modèle considéré, pour l'observation Y_{jcg} (une transformation convenable du volume) du spot j sur le gel g sous la condition c, est le suivant :

$$Y_{jcg} = \mu + \alpha_j + \beta_c + \gamma_{jc} + \delta_{cg} + \sigma\varepsilon_{jcg}$$

où μ un effet moyen, α_j l'effet du spot j, β_c l'effet de la condition c, δ_{cg} l'effet du gel g de la condition c et γ_{jc} le terme d'interaction entre le spot j et la condition c. Les variables aléatoires ε_{jcg} sont supposées indépendantes, de loi normale centrée et de variance 1. Le nombre de spots j varie de 1 à J, le nombre de conditions c de 1 à C, et le nombre d'observations g du spot j sous la condition c varie de 1 à G (nous nous sommes placés dans le cas d'un plan équilibré, c'est-à-dire que chaque spot est observé sur les G gels de chaque traitement).

Si ce modèle est validé (abscence d'interactions spot-gel, variance homogène) l'analyse différentielle repose sur l'étude des différences d'interaction :

$$F = (F_{jcc'}) = (\gamma_{jc} - \gamma_{jc'}) \qquad j \in \{1, \ldots, J\},\ \ c < c' \in \{1, \ldots, C\}^2$$

Notons X l'estimateur du vecteur F donné par l'analyse de variance.

Dans le cas où l'on veut comparer plus de deux traitements simultanément les composantes du vecteur X ne sont pas indépendantes. Nous considérons donc le modèle suivant :

$$X = F + \sigma P_n \varepsilon$$

où $n = JC(C-1)/2$ est la dimension des vecteurs F et X, ε est de loi normale centrée de matrice de covariance la matrice identité, P_n est connue et σ inconnue.

L'analyse différentielle consiste alors en la détection de composantes non nulles dans l'espérance d'un vecteur gaussien dont les composantes ne sont pas indépendantes.

Des méthodes basées sur des procédures de tests multiples et ne nécessitant pas d'hypothèse particulière sur la structure de dépendance des statistiques de test ont été proposées comme par exemple les procédures de Bonferroni, Benjamini et Yekuteli [3], et Benjamini et Liu [2]. Des études de simulations montrent que ces méthodes sont conservatives, c'est à dire que le nombre de composantes non nulles non détectées par la méthode de test est grand.

Nous proposons dans le Chapitre 2 une approche basée sur les méthodes d'estimation par sélection de modèles. Nous considérons une collection $\mathcal{M}_n$ de sous ensembles d'indices m de $\{1,\ldots,n\}$ de cardinal D_m. Pour chaque sous ensemble $m=(i_i\ldots,i_{D_m})$, nous considérons $\mathcal{F}_m$ l'espace engendré par $(e_{i_1},\ldots,e_{i_{D_m}})$ où les vecteurs $(e_1)_{i=1,\ldots,n}$ sont les vecteurs de la base canonique de $\mathbb{R}^n$. Nous obtenons donc une collection de sous espaces vectoriels $\{\mathcal{F}_m, m\in\mathcal{M}_n\}$. Nous estimons F sur $\mathcal{F}_m$ en minimisant sur $\mathcal{F}_m$ le critère empirique des moindres carrés noté γ_n et défini pour tout $T\in\mathbb{R}^n$ par $\gamma_n(T)=||X-T||_n^2$. Nous obtenons ainsi une collection d'estimateurs $\{\hat{F}_m, m\in\mathcal{M}_n\}$. Finalement le meilleur estimateur est $\tilde{F}=\hat{F}_{\hat{m}}$ où $\hat{m}$ est le modèle obtenu par minimisation d'un critère des moindres carrés pénalisé.

Notons Π_m la projection orthogonale sur l'espace $\mathcal{F}_m$ et M le maximum des valeurs propres de P_nP_n'. A l'aide d'un résultat établi par Y. Baraud[1], nous obtenons que pour un modèle m de $\mathcal{M}_n$, une pénalité de la forme :

$$pen(m)\geq\frac{\sigma^2}{n}\left\{c_1 Trace(\Pi_m P_n P_n')+c_2 M Log(\frac{n}{D_m})D_m\right\}$$

permet de garder sous contrôle le risque quadratique de l'estimateur $\tilde{F}$.

Comme P_nP_n' est connue, car issue de l'analyse de variance, nous pouvons simplifier la pénalité qui devient :

$$pen(m)=\frac{\sigma^2}{n}\left\{c_1\frac{1}{G}(2-\frac{2}{J})+c_2\frac{C}{G}Log(\frac{n}{D_m})\right\}D_m$$

où J,C,G sont respectivement les nombres de spots, de conditions et de gels et où $n=JC(C-1)/2$

L'estimateur pénalisé dépend du choix des constantes c_1 et c_2 et de la variance σ^2. Nous procédons en deux étapes pour obtenir les valeurs de ces constantes dans la fonction de pénalité.

Nous supposons tout d'abord que la variance du bruit est connue et nous déterminons les valeurs optimales des constantes c_1 et c_2. Nous choisissons les valeurs c_1 et c_2 qui minimisent, uniformément en J, C et F, le rapport entre le risque de l'estimateur $\tilde{F}(c_1,c_2)$ et le plus petit des risques des estimateurs de la collection $\{\hat{F}_m, m\in\mathcal{M}_n\}$. Nous estimons ce rapport de risques par la méthode de Monte -Carlo pour une collection de vecteurs F et différentes valeurs de J et de C.

[1] communication personnelle

Nous supposons ensuite que la variance est inconnue. Plutôt que d'utiliser un estimateur de la variance, nous considérons la variance comme une constante, et nous cherchons à estimer la fonction de pénalité à partir des données. Nous utilisons une méthode heuristique proposée par Birgé et Massart [4] et qui a été mise en oeuvre par Lebarbier [6] pour la détection de ruptures dans un modèle de régression. Nous écrivons la pénalité sous cette forme :

$$pen_\alpha(m) = \alpha f_n(D_m)$$

où

$$\alpha = \sigma^2$$

et f_n est une fonction convenablement choisie. Nous mettons alors en oeuvre sur des simulations la méthode heuristique pour estimer α et évaluons la performance de l'estimateur obtenu.

Estimation de graphe

Cette partie, presentée dans le Chapitre 3, concerne les modèles graphiques gaussiens, fréquemment utilisés pour inférer des réseaux géniques à partir de l'observation de l'expression de gènes.

Un graphe est constitué d'un ensemble de sommets qui correspondent aux gènes ou protéines, et d'un ensemble d'arêtes ou liaisons entre les sommets. Soit $X = (X_1, \ldots, X_p)$ une collection de variables aléatoires indexées par les sommets du graphe. Un modèle graphique gaussien suppose que X suit une loi normale multivariée. On définit le graphe de concentration comme un graphe dans lequel il existe une arête entre les sommets a et b si X_a et X_b sont des variables aléatoires non indépendantes conditionnellement à toutes les autres variables $X_{-\{a,b\}} = \{X_c, c \in \{1, \ldots, p\} \setminus \{a, b\}\}$, c'est à dire dans le cas gaussien, si le coefficient de corrélation partielle $\mathrm{Corr}(X_a, X_b | X_{-\{a,b\}})$ est non nul. Lorsque la matrice de covariance de X est inversible, une arête entre les sommets a et b correspond à une composante (a, b) non nulle dans la matrice de précision définie comme l'inverse de la matrice de covariance.

Dans le cas où le nombre d'observations est grand devant le nombre de variables, de nombreux travaux traitant le problème de l'estimation de la structure de dépendance des variables ont été publiés. Les méthodes proposées reposent sur l'estimation de la matrice de précision. Ces méthodes ne peuvent s'adapter au cas où le nombre d'observations est inférieur au nombre de variables, ce qui est en général le cas dans l'étude des données génomiques où l'on dispose de peu d'expériences par rapport au nombre de gènes.

Récemment plusieurs méthodes ont été proposées dans ce cadre. Schäfer et Strimmer [9] proposent d'estimer la matrice des corrélations partielles en utilisant un pseudo inverse et du bagging pour réduire la variance de l'estimateur. Wille et Bühlmann [10] proposent d'approcher le graphe de concentration par le graphe de covariance 0-1 construit à partir des covariances marginales et des covariances conditionnelles d'ordre 1. Meinshausen et

Bühlmann [7] proposent d'estimer les voisins de tous les sommets du graphe avec l'estimateur lasso, les voisins d'un sommet étant les sommets qui lui sont connectés, puis d'en déduire une estimation du graphe.

En collaboration avec l'équipe de MIA-Jouy, j'ai mis en oeuvre ces méthodes et les ai comparées par simulation. On dispose en effet de peu de résultats théoriques, ou alors de résultats qui sont asymptotiques et donc peu informatifs sur les méthodes en situation réelle. Or ce qui intéresse les biologistes est de connaitre le comportement de ces méthodes pour un nombre p de gènes, et n observations. On évalue la capacité de ces méthodes à estimer le "vrai" graphe en calculant par simulation la puissance et le taux d'arêtes détectées à tort.

Nous avons comparé ces méthodes en simulant différentes structures de graphe : graphe dans lequel les arêtes sont réparties uniformément et graphe avec des groupes de sommets très connectés et d'autres peu connectés. En effet dans les réseaux observés dans la réalité, les sommets sont fréquemment structurés en groupe ayant des propriétés de connectivité différentes. Nous nous sommes donc appuyés sur le modèle de mélange pour les graphes aléatoires proposé par J.J. Daudin, F. Picard, et S.Robin [5].

Nous avons ensuite appliqué ces différentes méthodes sur un jeu de données réelles. Pour pouvoir juger de la qualité d'inférence de ces méthodes, nous avons utilisé des données pour lesquelles le vrai graphe est connu. Ces données produites par Sachs et al. [8] sont des données d'expression de protéines qui sont impliquées dans un réseau bien étudié dans la littérature. Les relations entre ces protéines sont donc bien connues, ce qui nous permet de juger et de comparer les résultats obtenus par les méthodes d'estimation. L'autre intérêt de ce jeu de données est que le nombre d'observations n pour les protéines est très grand, ce qui est rarement le cas dans l'étude de données postgénomiques. Cela nous a permis d'étudier l'influence du nombre d'observations sur les méthodes d'estimation. Le graphe impliquant les protéines étant connu, nous avons évalué et comparé la capacité des méthodes à estimer ce graphe en calculant par simulation la puissance des méthodes en fonction du nombre d'observations.

La comparaison objective de plusieurs méthodes d'estimation pour différentes structures de graphe, et l'évaluation de leurs performances dans des situations pouvant s'approcher de la réalité est un travail novateur que nous soumettrons prochainement pour publication.

Test de validation de graphe

Cette partie présente un travail réalisé en collaboration avec Nicolas Verzelen[1].

Dans la partie précédente, le but était d'inférer des réseaux géniques à partir de l'observation de l'expression de gènes. Dans cette partie l'objectif est différent. Nous proposons un

[1]Université Paris Sud

test de validation de graphe. Il s'agit à partir d'un jeu de données et d'un graphe proposé par le biologiste de tester que l'on n'a pas oublié d'arêtes entre des sommets.

Plus précisément, nous nous donnons un n- échantillon d'un vecteur aléatoire gaussien $X = (X_1, \ldots, X_p)$, et un graphe $\mathcal{G} = (\Gamma, E)$, où $\Gamma = \{1, \ldots, p\}$ est l'ensemble des sommets du graphe $\mathcal{G}$ et E l'ensemble des arêtes. Lorsque la structure d'indépendance conditionnelle de X est représentée par le graphe $\mathcal{G}$, c'est à dire qu'une arête entre deux sommets a et b dans le graphe $\mathcal{G}$ correspond à un coefficient de corrélation partielle $\mathrm{Corr}(X_a, X_b | X_{-\{a,b\}})$ non nul entre les variables X_a et X_b, nous disons que X est un modèle graphique gaussien en accord avec le graphe $\mathcal{G}$. Dans cette partie, nous construisons à partir d'un n- échantillon du vecteur X, un test de l'hypothèse "X est un modèle graphique gaussien en accord avec le graphe $\mathcal{G}$ ". Nous appellons un tel test *test de graphe*.

Pour cela nous considérons chaque sommet du graphe $\mathcal{G}$ et construisons pour chacun de ces sommets ce que nous appelons un *test de voisinage*. Soit $a \in \Gamma$ un sommet du graphe $\mathcal{G}$. Nous définissons son voisinage noté $ne(a)$ comme étant l'ensemble des sommets $b \in \Gamma \setminus \{a\}$ tel qu'il y ait une arête entre a et b, et notons $\overline{ne}(a) := ne(a) \cup \{a\}$. Pour chaque sommet a, nous construisons un test de l'hypothèse : " X_a est indépendant de $\{X_i, i \in \Gamma \setminus \overline{ne}(a)\}$ conditionnellement à $\{X_i, i \in ne(a)\}$".

Comme X suit une loi normale, X_a peut se décomposer de la manière suivante :

$$X_a = \sum_{b \in \Gamma \setminus \{a\}} \theta_b^a X_b + \epsilon_a$$

où θ^a est le vecteur de $\mathbb{R}^{p-1}$ des coefficients de la régression et où ϵ_a est une variable aléatoire gaussienne centrée et indépendante de $X_{-a} = \{X_b, b \in \Gamma \setminus \{a\}\}$. Le vecteur θ^a est déterminé par l'inverse de la matrice de variance-covariance de X et le voisinage $ne(a)$ du sommet a correspond alors à l'ensemble des sommets $b \in \Gamma \setminus \{a\}$ tels que le coefficient θ_b^a soit non nul. Tester l'hypothèse nulle "X_a est indépendant de $X_{\Gamma \setminus \overline{ne}(a)}$ conditionnellement à $X_{ne(a)}$" est alors équivalent à tester l'hypothèse nulle

$$H_{0,a} : "\theta^a_{\Gamma \setminus \overline{ne}(a)} = 0" \quad \text{contre l'alternative} \quad H_{1,a} : "\theta^a_{\Gamma \setminus \overline{ne}(a)} \neq 0"$$

Le test de voisinage revient donc à tester une hypothèse linéaire dans un modèle de régression multivariée, dont les variables explicatives sont aléatoires.

La procédure pour le test de voisinage est décrite dans le chapitre 4. La procédure est similaire à celle proposée par Baraud et al [1] dans le cadre d'une régression en design fixe. Soit un sommet $a \in \Gamma$. Nous considérons une collection finie $\mathcal{M}_a$ de sous ensembles m de $\Gamma \setminus \overline{ne}(a)$ et une collection $\{\alpha_m, m \in \mathcal{M}_a\}$ de nombres dans $]0, 1[$. Pour chacun de ces sous ensembles m, nous considérons le test de Fisher de niveau α_m de l'hypothèse nulle :

$$H_{0,a} : "\theta_{\Gamma \setminus \overline{ne}(a)} = 0" \quad \text{contre l'alternative} \quad H_{1,a,m} : "\theta_{\Gamma \setminus \overline{ne}(a)} \neq 0 \text{ and } \theta_{\Gamma \setminus (\overline{ne}(a) \cup m)} = 0"$$

et nous rejetons l'hypothèse nulle $H_{0,a}$ si nous la rejetons pour l'un de ces tests. Dans le Chapitre 4, nous décrivons cette procédure de tests multiples et discutons du choix de la

collection $\{\alpha_m, m \in \mathcal{M}_a\}$. Nous proposons de choisir ces poids α_m selon deux procédures différentes. La première correspond à une procédure de Bonferroni et est donc conservative. Nous proposons alors une seconde procédure qui permet d' obtenir un test de niveau exact. Dans ce Chapitre 4 sont également présentés des résultats théoriques sur la puissance du test et sur des vitesses minimax qui ont été établis par Nicolas Verzelen. Dans la dernière section de ce chapitre, nous présentons une étude de simulations menée dans le but d'illustrer les résultats théoriques établis dans les sections précédentes. En particulier nous montrons que le test de voisinage est facile à implémenter et comparons par simulations les deux procédures proposées pour le choix de la collection $\{\alpha_m, m \in \mathcal{M}_a\}$.

Dans le Chapitre 5, nous déduisons de ce test de voisinage le test de graphe. Dans ce chapitre nous illustrons et discutons de l'application de nos procédures sur des données simulées et des données réelles.

Nous mettons en oeuvre les tests de voisinage et de validation de graphe en simulant des graphes pertinents pour les réseaux géniques et nous discutons du choix de la collection de modèles $\mathcal{M}_a$.

Nous appliquons ensuite notre test de validation sur des données réelles. Nous reprenons les données produites par Sachs et al. [8]. Dans le Chapitre 3 nous avons mis en oeuvre des méthodes d'estimation sur ces données. Nous utilisons alors le test de graphe comme un outil pour valider les graphes estimés. Puis nous profitons du grand nombre d'observations de ce jeu de données pour étudier l'influence du nombre d'observations sur la puissance du test.

Ces deux chapitres 4 et 5 seront soumis prochainement à publication.

References

[1] Baraud, Y., Huet, S., and Laurent, B. Adaptative tests of linear hypotheses by model selection. *Annals of Statistics 31*, 1 (2003), 225–251.

[2] Benjamini, Y., and Liu, W. A distribution-free multiple-test procedure that controls the false discovery rate. *Research Paper* (1999).

[3] Benjamini, Y., and Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist. 29(4)* (2001), 1165–1188.

[4] Birgé, L., and Massart, P. Minimal penalties for gaussian model selection. *Probab. Theory Related Fields 134* (2006).

[5] Daudin, J. J., Picard, F., and Robin, S. A mixture model for random graphs. Tech. Rep. RR-5840, INRIA, Rapport de Recherche, 2006.

[6] Lebarbier, E. Detecting multiple change-points in the mean of a gaussian process by model selection. *Signal Processing 85* (2005), 717–736.

[7] Meinshausen, N., and Bühlmann, P. High dimensional graphs and variable selection with the Lasso. *Annals of Statistics 34*, 3 (2006), 1436–1462.

[8] Sachs, K., Perez, O., D.Pe'er, Lauffenburger, D. A., and Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science 308* (2005), 523–529.

[9] Schäfer, J., and Strimmer, K. An empirical bayes approach to inferring large-scale gene association nerworks. *Bioinformatics 21* (2005), 754–764.

[10] Wille, A., and Bühlmann, P. Low-order conditional independence graphs for inferring genetic networks. *Stat. Appl. Genet. Mol. Biol. 5* (2006).

Première partie

Sélection de modèles pour l'analyse différentielle d'images d'électrophorèse

Chapter 1

Statistics for proteomics: experimental design and 2-DE differential analysis

Jean-François Chich[1], Olivier David[2], Fanny Villers[2], Brigitte Schaeffer[2], Didier Lutomski[3] and Sylvie Huet[2]

Cet article est paru dans *Journal of Chromatography B* en 2007.

Abstract

Proteomics relies on the separation of complex protein mixtures using bidimensional electrophoresis. This approach is largely used to detect the expression variations of proteins prepared from two or more samples. Recently, attention was drawn on the reliability of the results published in literature. Among the critical points identified were experimental design, differential analysis and the problem of missing data, all problems where statistics can be of help. Using examples and terms understandable by biologists, we describe how a collaboration between biologists and statisticians can improve reliability of results and confidence in conclusions.

1.1 Introduction

The term "proteomics" appeared in 1995 [54, 56]. The primary goal of this new discipline was to study the protein complement of a genome but it rapidly appeared that this task was far from reach even if some ambitious and international initiatives, like HUPO (Human Proteome Organisation) founded in 2001 [27], were undertaken.

[1]INRA, Jouy-en-Josas, Biologie Physico-Chimique des Prions, VIM

[2]INRA, Jouy-en-Josas, Mathématiques et Informatique Appliquées MIA

[3]Laboratoire de Biochimie des Protéines et Protéomique, UMR CNRS BioMoCeTi

Proteomics relies mainly on the separation of a complex mixture of proteins by bidimensional electrophoresis (2-DE), mass measurement of peptides generated after spot proteolysis by mass spectrometry and search in databases.

Constant technical improvements were performed over the years, in particular accuracy and easiness of use of mass spectrometers and database enrichment. However, numerous publications are now considered of questionable quality. To improve overall quality and results reliability, four weak points to consider were identified recently [55]. These points concern experimental design, analysis of protein abundance data, confidence in protein identification by mass spectrometry and analytical incompleteness. The third point has been already discussed [46, 29] and we will turn toward the three other points.

The role and importance of experimental design were described for transcriptomics but less frequently for proteomics. While proteomics cannot usually handle as much data as transcriptomics, the importance of experimental design should be emphasized. We show here how statistics can help to define suitable experimental designs, using classical knowledge on this subject [15, 18] and knowledge issued from transcriptomics [14, 34, 58]. We show that establishing an experimental design in a dynamic collaboration between biologists and statisticians is useful to forecast sampling or experimental biases. In particular, experimental design allows to limit systematic errors, to improve precision of subsequent statistical tests and contributes thus to reduce the number of false positives.

Differential analysis of spot volume is generally handled by using commercial software packages (see article by R. Joubert-Caron in the same issue) that propose statistical tools to help to conclude on the significance of variation. Probably most biologists consider that these packages are sufficient for their purpose. We show here that statisticians propose powerful tools that can be used for improving data analysis. These tools can help to rationalise the decisions on significance and can draw attention on the imperfections of gels, spot mismatches and other artifacts of 2-DE gels.

Moreover, analytical incompleteness is encountered in proteomics, especially in spots missing (missing data) on one or more gels coming from the same series. The particular and difficult problem of these missing data on 2-DE gels is generally due to experimental problems and must be taken into account in the statistical analysis. However, the questions raised by these problem are not trivial and are discussed.

Lastly, we emphasize the interest of collaborations between biologists and statisticians at different levels of proteomics experiments, in order to draw the most robust conclusions from experimental results.

1.2 Experimental designs

1.2.1 A practical example : cell line proteomics

In this part, an example corresponding to a question of proteomics applied to cell biology is described. The situation presented here can be easily applied to other situations. Cell lines are largely used as biological tools to study effects of an infection, the effect(s) of a drug and so on. These immortalized cells are easy to grow and are a useful source of large amounts of proteins needed for proteomic studies.

However, proteomics on these cells is subject to variabilities that must be taken into account when possible. The first variability is clonal drift due to the number of passages

of cells (a technique for diluting cells that enables them to be kept alive and growing under cultured conditions for extended periods of time), acquired chronic contaminations or metabolic modifications due to culture media. Unfortunately, this variability cannot be easily considered since cell lines are heterogeneous (see for example, in the prion field [6, 37]) from passage to passage. Thus, cloning artifacts can induce false conclusions.

Another source of variability can be due to small biological variations (cell growth variability etc). If a researcher wants to study the effect of a drug on a cell line, he adds the drug in a serie of flasks and a placebo in another. However, differences observed between flasks for the same treatment account for a variability that can be taken into account by experimental design and statistical method(s).

The second variability is a technical one. It involves the preparation of cells and the protein solubilisation prior to the separation by 2-DE: for example, if cells are washed using a cold ice buffer, it is likely that they will express chaperones that might interfere with the studied phenomenon. A variability can be due also to the apparatus used for cell culture (variable heat or humidity of an incubator) or used for protein separation. One aim of statistical methods presented in this article is to take account of these sources of variation.

A study using cell lines, was undertook by two of us [12] and it is used as a basis to present ideas underlying experimental design. Two cell lines were used to study prion infection. The first one is GT1-7, a subclone of GT1, a highly differentiated hypothalamic cell line displaying a number of neuron functions [42]. The second line results from infection of GT1-7 clone with the Chandler strain of prion (ScGT1-7: Sc for scrapie, the prion disease of sheep) [51]. Though GT1 and ScGT1 were described to be in a steady state after respectively 12 passages [42] and 55 passages post-infection [51], clonal drift could occur in cell populations grown in so different laboratories for years and generate variability. In order to study how prion infection affects cellular metabolism, a proteomic approach was made on both cell lines.

1.2.2 Two-phase experiment

Our *two-phase experiment* was performed as schematised in Figure 1.1A. The first phase consists of the cell cultures. GT1-7 and ScGT1-7 cells were grown separately for several passages in order to obtain an amount of proteins compatible with a proteomic analysis. The second phase consists of protein extraction from these cells and to separate them by 2-DE.

The objective of this two-phase experiment was to compare protein abundance according to two conditions. These conditions are defined as healthy (H), for GT1-7, and infected (Sc) for ScGT1-7. In the first phase, cell cultures, sample preparation and/or pooling represent the *biological phase*. Separation of proteins by 2-DE, organization of gel runs and staining represent the *technical phase*. Technical variability due to the electrophoresis apparatus was considered to be non significant because running conditions (current, buffer and temperature...) were controlled; thus six 2-DE (IEF then SDS-PAGE) were run for H (represented by “batch 1 ”), followed by six 2-DE for Sc (“batch 2 ”).

After silver staining, gels were scanned and classically analysed using the software ImageMaster 4.01 (GE-Healthcare Bioscience). After spot volume measurements and matching, differential analysis (Section 3) was performed and spots showing significant abun-

dancy variations (`http://genome.jouy.inra.fr/gt1`) between H and Sc were identified by mass spectrometry. The expression of the proteins corresponding to these spots can be considered to be specifically affected by prion chronic infection. They can be potential markers of prion disease or targets for drug therapy.

Thus, variability arises generally from both phases, calling for rational implementation of work plans [9, 10]. It should be emphasized that establishing an experimental design allows bias reduction and increased confidence in experimental results.

1.2.3 Constructing experimental blocks or *blocking*

If the researcher suspects variability during gel runs for example, he can account for this variability by constructing *blocks*. A block is a set of experimental materials considered as consistent. The objective of blocking is to make the comparison between observed conditions, with little as possible dependent on artefacts or heterogeneities (differences between gels, etc...) and as much as possible dependent on the differences the researcher needs to characterize (effect of infection in the example). Thus, blocking takes into account heterogeneities known or suspected from the beginning of the experiment and improves the precision of the following statistical analysis. When blocking is used, both the blocks and the allocation of conditions to blocks have to be chosen (see [15, 18] for more information on blocking).

An example of blocking is shown in Figure 1.1B. Cultures are performed as described in the previous schema. Variabilities being suspected between "batch 1 "and "batch 2 ", the researcher constructs one block that will run in "batch 1 "and a second block that will run in "batch 2 ". Each block is composed of 3 gels (isoelectric focusing followed by sodium dodecyl sulfate polyacrylamide gel electrophoresis) with proteins from H (sample 1) and 3 gels with proteins from Sc (sample 2). If there is a heterogeneity due to the electrophoresis apparatus, the researcher will be able to identify it by comparing gels loaded with sample 1 from both batches and gels loaded with sample 2 also from both batches. Moreover, the researcher will be able both to quantify this effect and to improve precision in differential analysis. The remaining variabilities are due to differences between H and Sc that the researcher wants to characterize.

The technique of blocking for the technical phase was described for transcriptomics, to take into account the heterogeneities linked to arrays and dyes in microarrays experiments and complicated designs were proposed [14, 34, 58, 35]. Pairing is a particular blocking with blocks of size 2 that was described for transcriptomics but also for 2D difference gel electrophoresis (2D-DIGE) experiments [33]. Blocking can also be used in the biological phase [3].

1.2.4 Randomization

Because it is difficult for a researcher to identify all sources of variation using his judgement or experience, the use of randomization is a common practice. Randomization was recommended for microarray experiments [14, 34]. To our knowledge, randomization for proteomics was described only for experimental design in mass spectrometry [29].

We will consider again the previously described experiment but with unknown heterogeneities both in biological and in technical phases. This experiment thus calls for a randomization for both phases [10]. The schema of the experimental design is shown on

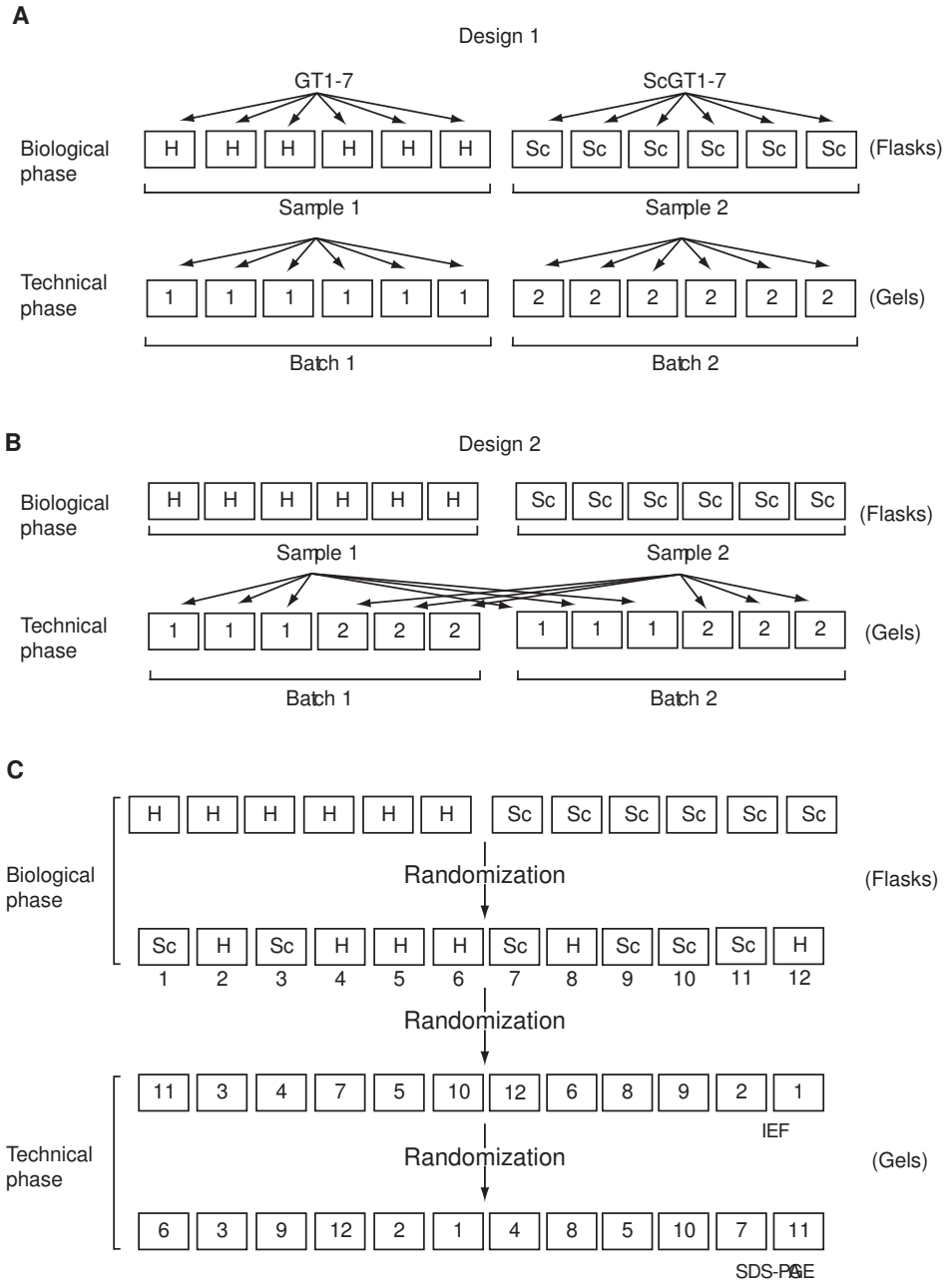

Figure 1.1: *Examples of experimental designs. Squares correspond to flasks in the biological phase and to gels in the technical phase. H denotes healthy GT1-7 cells and Sc denotes ScGT1-7 (chronically infected GT1-7 cells).* ***A****: unrandomized design for experiment on cell lines. Healthy cell (H) and infected cell (Sc) were cultured in several flasks. Samples were pooled and 2-DE were performed.* ***B****: unrandomized design with "blocks ", the heterogeneities suspected in the technical phase can be taken into account in differential analysis.* ***C****: example of a randomized design without blocking. Numbers denote sample landmarks. GT1-7 cells (H or Sc) were cultured in different flasks. Heterogeneities are suspected in biological and technical phases and a randomization is performed for both phases. The effect of suspected potential biases is eliminated.*

Figure 1.1C. The incubator is supposed to be heterogeneous (heat, CO_2 distribution etc.) and a randomization is performed before placing the culture flasks with landmarks (six "H "and six "Sc ") inside. Cells are cultured and proteins are extracted individually from each flask. To avoid biases due to for example, preferential current run, differences in strip or gel batches, buffer composition, randomization can be performed at the different levels of the technical phase : a) allocation of proteic samples to strips, b) strip placement in IEF apparatus, c) strip deposit at the top of second dimension gels, after IEF, d) gel placement in migration cuve. Subsequent gel staining, image analysis and statistical treatment are performed as usual. The example presented here is a simple randomization. However, designs with blocks can also be randomized [15, 18].

In brief, randomization reduces systematic errors when comparing the conditions and estimating the precision of the results.

1.2.5 Replication

Although randomization and/or blocking allow control of extraneous variables, the result of a single 2-DE experiment is not satisfactory due to the intrinsic variability of the method. In proteomics, variability can be found in biological phase as well as in technical phase. Variability was estimated to be high in 2-DE [13] and it can lie in the number of spots detected, on the variance of the spot volume measured (discussed in Section 3) by software analysis. The authors showed also that a fully manual analysis is more reliable than a fully automated one. However, replication in both phases can be problematic due to the low amount of initial sample or due to low protein content in a sample (biopsy for example). Some tools are available to estimate the experiment precision *a priori* on the basis of the researcher objectives (`http://www.emphron.com/`).

In order to illustrate the concept of replication, we present 3 examples, shown in Figure 1.2. The first design (Figure 1.2A) has one biological replication (1 flask for H, 1 flask for Sc) and six technical replications (six 2-DE per flask). Its drawback is that biological variance cannot be estimated. The differential analysis will be based on the technical variance only and the precision of analysis will be over-estimated. This situation can increase the number of false positives. Similarly, if protein extracts from several flasks are pooled into a single sample for each condition (as in Figure 1.1A), the differential analysis will also be based on technical variance only and the number of false positives may increase. The use of several pools per condition avoids this problem (see [14, 34] for a discussion on pooling). The second design (Figure 1.2B) shows three biological replications (3 flasks for H and 3 for Sc) and two technical replications (two 2-DE per flask). The third design has six biological replications and only one technical replication. For both designs, six 2-DE were made for each H and Sc. Both designs have thus the same ability to account for the technical variance. However, the third design (Figure 1.2C) has more biological replications and the subsequent analysis will be more accurate.

Cell lines do not give the opportunity to perform *true* biological independent replications because all cells derive from a unique cell. However, several factors (clonal drift, culture media modifications, *de novo* latent infections etc.) can induce, at least theoretically, variations that can be assessed by replication of flask cultures without pooling. True biological replications can be envisaged using primary culture cells [16] because these cultures can be designed to be as independent as possible from each other.

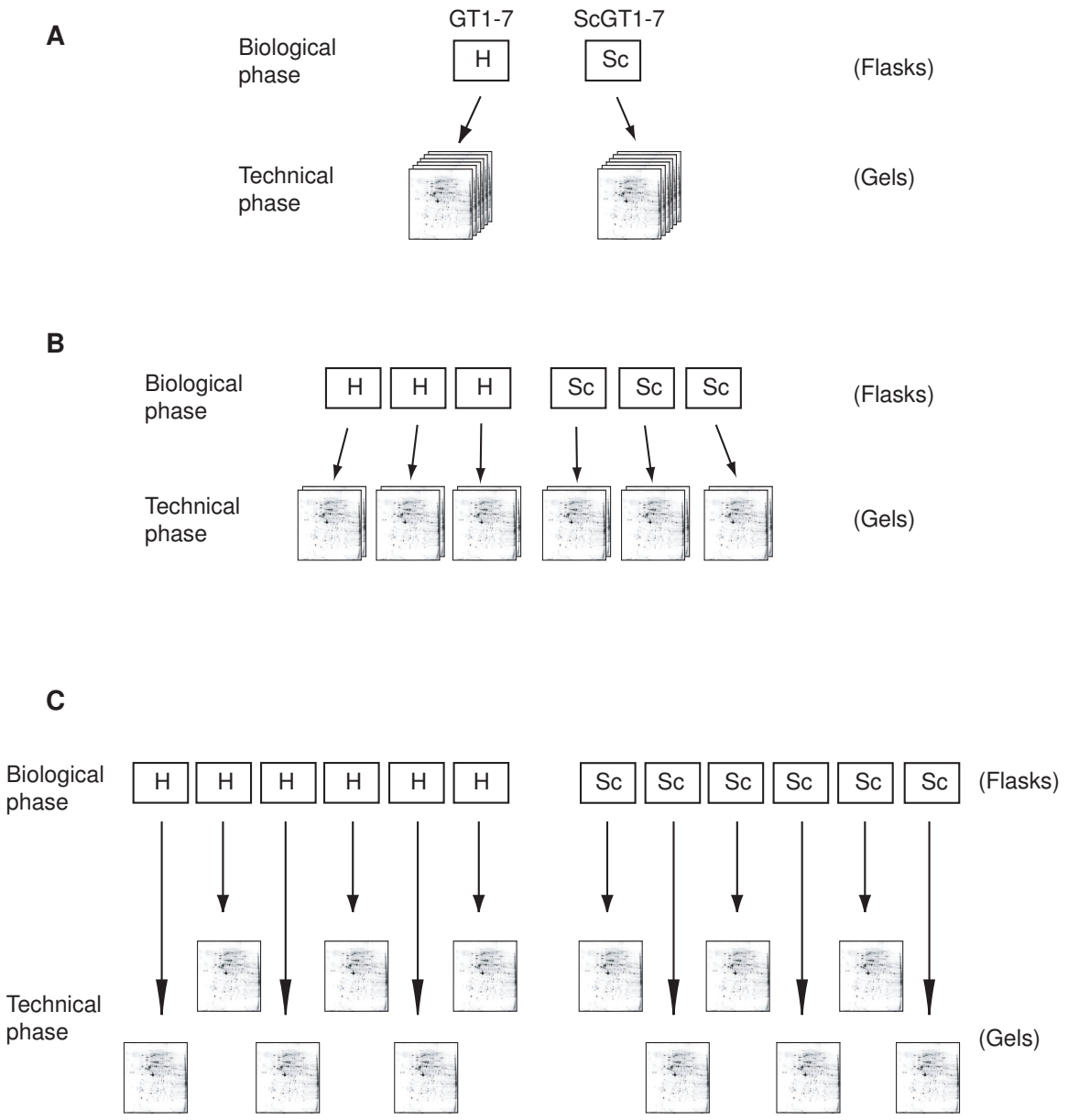

Figure 1.2: *Example of experimental designs using replication. Replication should take into account technical limits and the fiability of expected results.* ***A****: Biological phase without replication and technical phase with six replications for each sample.* ***B****: Biological phase with three replications and technical phase with 2 replications for each sample.* ***C****: Biological phase with six replications and technical phase without replication.*

Replications are necessary to assess and increase the precision of the subsequent analysis results. From a statistical point of view, biological replications are more efficient than technical replications. In practice, biologists use technical replications due to the well known variability of the 2-DE technique. Literature studying the experimental design of microarray experiments emphasizes using replications to assess and control experimental variability [14, 34, 58]. However, these recommandations are discussed since a system of unreplicated experiments was described [58]. For proteomics, discussions on the number of observations are available in [33, 44].

1.2.6 Discussion

In summary, experimental design is an important part of biological experiments, especially in proteomics where technical methods are long, difficult with intrinsic technical variability. The situation of two-phase experiment is often encountered in proteomics [3] and transcriptomics [34] and experimental design offers several tools that can be useful to minimise the effect of variabilities on the results. The choice of a suitable experimental design can improve the reliability of true positives detection and reduces the number of false positives in the subsequent differential analysis.

While general guidelines can be drawn from the examples shown here, it should be kept in mind that establishing an experimental design is a compromise between the availability of biological material, the technical difficulties of the approach and the reliability of the expected results.

1.3 Differential analysis

The aim of the differential analysis is to detect the proteins whose abundance differs according to the condition. In statistical terms, this comes to test simultaneously a large number of hypotheses: for each spot j, we have to test the hypothesis H_j that the spot volume does not differ according to the condition, or in other words that the corresponding protein is not variant in abundance. This problem has been extensively studied for the determination of the differentially expressed genes in microarray experiments [4, 23, 21, 45, 53, 2]. Nevertheless, the adaptation of these works to data coming from 2-DE is not direct for the following reasons:

- the data present a great variability due to the complexity of the image analysis [25, 48, 47, 20, 39].
- the number of missing data is large, up to 50-60% [49]
- generally the replication number is small, between 3 and 6.
- some observations are irrelevant. This is the case when the image analysis process detects spots in trails or overlapping spots.

The statistical analysis provides a list of variant spots, using a procedure that is based on a statistical model and on the data. The parameters controlling the procedure, for example the probability of deciding wrongly that a spot is variant, are estimated. In reality,

because some observations are irrelevant, this list is only a list of potentially interesting spots and must be carefully examined before validation.

Detection of pertinent spots is thus an iterative procedure between the researcher and the results of the statistical analysis. Nevertheless, providing methodological tools for detecting irrelevant observations and variant spots, may help to best analyze the data.

The testing procedure consists of choosing a *test statistic* and deciding if the hypothesis H_j is rejected or not. These problems are treated in section 1.3.1. The next question to consider is *what does the procedure control ?* This is the object of section 1.3.2.

In practice, preliminary analysis is necessary in order to verify that the statistical model is consistent with the data. Another important question is *which strategy for missing data ?* All this will be discussed in section 1.3.3.

1.3.1 Statistical models and testing methods

Statistical problems involved in gel analysis have been discussed [49, 20, 11, 26, 43]. In this paper we consider two approaches, the spot by spot approach (or univariate analysis) where the test statistic for testing the hypothesis H_j is based on the data observed for the spot j only, and the global approach (or multivariate analysis) where the test of H_j is based on the results of an analysis of variance considering all the observations together. The methods taking into account the experimental design, for example blocking, are mentioned in Section 3.1.3.

Let us denote by Y_{jcg} the response for spot j, under condition c, on gel g. The response is the percentage of volume on gel g or a suitable transformation of spot volume, that is a transformation for which the statistical assumptions needed for applying the methods described below will be reasonably satisfied. The problem of choosing a transformation is discussed in section 3.3.3.

1.3.1.1 Spot by spot analysis

In the spot by spot analysis, we assume that the Y_{jcg}'s are distributed as Gaussian independant variables with mean m_{jc} and variance σ_j^2. Testing H_j comes to test that the m_{jc}'s are all equal using the classical Fisher or Student statistics test, see Table 1.1. Several variants of Student statistics have been proposed [22]. Non parametric tests can also be applied such as the Mann-Whitney (or Kolmogorov) test. If we denote by F_c the distribution function of the observations of spot j under condition c, the Mann-Whitney test consists in testing that the distributions F_c are identical. It does not need to assume Gaussian distribution.

1.3.1.2 Global analysis

In the global analysis, we start with an ANOVA (analysis of variance) model, where the response Y_{jcg} is modeled as follows:

$$Y_{jcg} = (\mathrm{G})_g + (\mathrm{SpC})_{jc} + \mathrm{E}_{jcg}$$

where $(\mathrm{G})_g$ is the effect of gel g (the part of variability due to gel g in the response Y), and $(\mathrm{SpC})_{jc}$ is the spot×condition effect defined as the effect of spot j under condition c on the response Y. The random errors E_{jcg} are distributed as centered Gaussian independant

variables with the same variance σ^2. Testing H_j comes to test that the differences between the spot×condition effects $(\mathrm{SpC})_{jc}$ are zero using the Student or Fisher statistic, see Table 1.2.

1.3.1.3 Analyses with block effects

The block effects in the experimental design have to be taken into account in the modeling [35, 34, 9, 10, 3]. Let us consider the example given by Design 2 of Figure 1.1B. Let Y_{jcag} be the response of spot j under condition c measured on gel g in experimental apparatus a. In the spot by spot approach, for each spot j, we consider the following ANOVA model,

$$Y_{jcag} = (\mathrm{A})_a + (\mathrm{C})_c + (\mathrm{AC})_{ac} + \mathrm{E}_{jcag}$$

where $(\mathrm{A})_a$ is the mean effect of apparatus a, $(\mathrm{C})_c$ the mean effect of condition c, and $(\mathrm{AC})_{ac}$ is an apparatus×condition effect. The last effect is called an interaction effect meaning that the condition effect may differ according to the apparatus. For each spot j the model parameters are estimated, and testing H_j comes to test that the $(\mathrm{C})_c$'s are all equal. In the global approach model, the response Y_{jcag} is modeled as follows:

$$Y_{jcag} = (\mathrm{G})_g + (\mathrm{SpA})_{ja} + (\mathrm{SpC})_{jc} + (\mathrm{SpAC})_{jac} + \mathrm{E}_{jcag}$$

where $(\mathrm{G})_g$ is the gel effect, $(\mathrm{SpA})_{ja}$ is the effect of spot j observed in apparatus a, $(\mathrm{SpAC})_{jac}$ is an apparatus×condition×spot effect, and $(\mathrm{SpC})_{jc}$ is the the effect of spot j under condition c. As before, testing H_j comes to test that the differences between the spot×condition effects are zero.

1.3.1.4 Decision rules

The decision rule for rejecting H_j is the following: for each spot j, we calculate the pvalue p_j, defined as the probability for rejecting H_j when H_j is true. The hypothesis H_j is rejected when p_j is small. Therefore the set of variant spots corresponds to the smallest pvalues. For example, we can choose to reject H_j when p_j is smaller than 5%.

1.3.2 Controlling the testing procedure

The question that arises is *how many errors are we doing when testing J hypotheses ?* We commit an error in two situations:

1. when we decide that a spot is a variant when it is not. Such an error leads to a false positive. The number of such errors is denoted FPos.

2. when we decide that a spot is not a variant when it is. Such an error leads to a false negative.

The control and elimination of false positives is important in order to avoid drawing false conclusions, particularly when the conclusions are the starting point of a new costly experiment.

If we run the testing procedure with $\alpha = 0.05$, then we expect up to 5% of the total number of spots to be variants by chance alone. In other words, if the differential analysis is performed with $J = 1000$, then we expect up to 50 spots to be wrongly detected as variants. Such a control is not acceptable. Two methods described below overcomes this problem.

The family-wise error rate or FWER The FWER is defined as the probability of having at least one false positive. It can be shown that if each hypothesis H_j is rejected when $p_j \leq \alpha$, then FWER $\leq \alpha J$. Choosing $\alpha = 0.05/J$ leads to FWER ≤ 0.05. This choice of α is known as the Bonferroni correction. This procedure allows very few occurrences of false positives, but makes the decision rule that a spot is differentially expressed very strict.

The false discovery rate or FDR If R denotes the number of rejected hypotheses, the FDR is defined as the expected value of the ratio FPos$/R$ when R is positive. Controlling the FDR at level 0.05 means that up to 5% of spots among the spots detected as variants, are identified by chance. This procedure proposed by Benjamini and Hochberg [5] is detailed in Figure 1.3. Several variants and improvements of this procedure have been proposed [22, 2, 52].

Which one to choose ? The choice between FDR or FWER procedure should be made on the basis of the aim of the research. If the differential analysis is a work whose objective is to list potential proteins involved in a physiological process, FDR method provides a reliable tool. If the objective is to determine if a protein is a potential biomarker, according to [55], false positives must be totally eliminated and FWER method should be preferred. However, in this case, further investigation is needed after this step to validate the biomarker.

1.3.3 Preliminary analyses

1.3.3.1 Removing irrelevant data

Testing simultaneously a large number of hypotheses has a cost: the larger the J, the more the procedure is strict. Therefore retaining in the differential analysis spots for which the observations are not relevant may compromise the differential analysis for the other spots. The amount of protein quantified in each spot can be computed when the spots are correctly detected by the image analysis software, but 2-DE images present smears and trails corresponding to migration artifacts. Those spots are frequently located near the left or right side of the image, corresponding to zones of accumulation of protein not within the pH gradient used, and around over abundant proteins, such as actin or tubulin for example. To improve the analysis, those spots are deleted, see Figure 1.4.

1.3.3.2 Checking gel replications within conditions

The experiment can be used for differential analysis if within each condition, the gels can be viewed as replications of the same observation. However, the classification of gels into conditions may be uncertain because of biological or technical variability in the experiment.

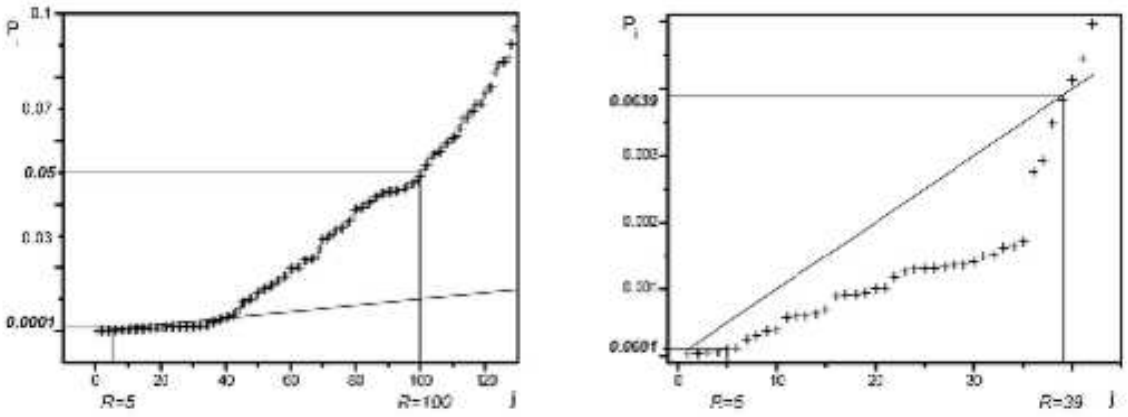

Figure 1.3: *Decision rules. Once the pvalues p_j are calculated, it remains to define a threshold, such that the hypothesis H_j is rejected as soon as p_j is smaller than the threshold. Let us formulate the problem in another way by considering the set of ordered pvalues into ascending order, $p_{(1)} < p_{(2)} < \ldots < p_{(J)}$, and a set of thresholds that may depend on j denoted τ_j. The number of rejected hypotheses H_j, R, is defined as the largest j such that $p_{(j)} \leq \tau_j$. Finally, we reject H_j if $p_j \leq \tau_R$. If all the $p_{(j)}$'s satisfy $p_{(j)} > \tau_j$, then $R = 0$: none of the hypotheses H_j is rejected.* ♦ *If τ_j is constant and equal to α, then R is simply the number of pvalues that are smaller than α. We can take $\alpha = 0.05$, or applied the Bonferroni correction with $\alpha = 0.05/J$.* ♦ *The method proposed by Benjamini and Hochberg takes $\tau_j = 0.05j/J$. Then H_j is rejected if $p_j \leq 0.05R/J$. They have shown that the FDR is controlled as follows:* $\mathrm{FDR} \leq 0.05T/J$ *where T is the number of spots that are not differentially expressed. These methods are illustrated by the graphics of pvalues $p_{(j)}$ in ascending order as function of j, for $j = 1, \ldots, 125$ on the left and $j = 1, \ldots, 42$ on the right. These data are coming from a simulated example with $J = 500$. The number of rejected H_j equals 100 if $\tau_j = 0.05$, 5 if $\tau_j = 0.05/500$ and 39 if the Benjamini and Hochberg's method is used with $\tau_j = 0.05j/J$*

Data-mining methods are suitable for checking this assumption, considering the gels as the experimental units (the cases) and the spots as the variables. Unsupervised methods such as Principal Component Analysis (PCA) or hierarchical clustering, ignores the condition under which the gels were observed. Their aim is to discover structures from the evidence of the data matrix alone. If the structure proposed by the analysis consists of splitting the gels into conditions, then we are allowed to use the data set for differential analysis (see paper by R. Joubert-Caron in the same issue). If not, such an analysis may give information on what is going wrong in the data set.

Some of these methods such as PCA, cannot be used when many spots are missing, particularly in the context of 2-DE gels, because they are not missing at random. It is then possible to run the method only on spots observed on all the gels. Another possibility is to attribute values to missing data. This point will be discussed in Section 1.3.4.

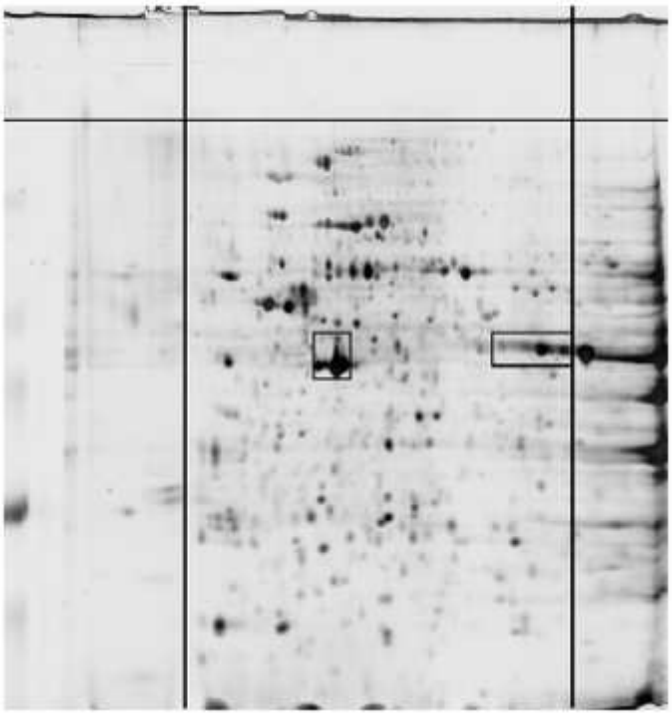

Figure 1.4: *Removing irrelevant spots. Spots near the left and right sides of the gel and near the top are deleted as well as spots located near actin and tubulin that are overabundant.*

1.3.3.3 Choice of a suitable transformation of the observed volumes

The testing procedures presented in Section 1.3.1 rely on statistical assumptions that should be checked.

The global approach assumes that there exists a suitable transformation of the observed volumes such that an ANOVA model is appropriate for modeling the data. The spot by spot approach assumes that there exists a transformation such that the gels within

a condition are replications of the same observation, and the variance of the resulting response does not depend on the condition c. If these assumptions are not satisfied, then the testing procedures are false, that is, the calculation of the pvalues is no longer valid. Usually the gel effect is eliminated by calculating for each spot the percentage of volume on the gel. Then the Student test or the Mann-Whitney test is used for testing H_j for each spot j. However, it has been observed that the larger the spot, the larger the variance [40, 26]. It is therefore worthwhile to look for a transformation in order to stabilize the variance. This heterogeneity in the data variability is mainly due to a scale phenomenon, well-known when the observation (the spot volume) is a count (number of pixel $\times$ intensity)

In practice the problem is to find a transformation T of the volumes V_{jcg} or the percentage of volumes on each gel $\%V_{jcg}$, such that the transformed data $Y_{jcg} = T(V_{jcg})$ or $T(\%V_{jcg})$ satisfy the assumptions of Section 1.3.1. In some cases the logarithmic transformation is applied with success. In other cases, other transformations are more appropriate. The Box-Cox method allows to estimate an optimal transformation from the data [26, 24]. Other normalization methods based on the data have been proposed [43, 30]. In any case, graphics and statistical analyses are useful for detecting the presence of structures in the variance of the data [26, 1]. Precisely let us denote by R_{jcg} the residuals defined as $R_{jcg} = Y_{jcg} - \widehat{Y}_{jcg}$, where $\widehat{Y}_{jcg}$ is the predicted value for spot j on gel g under condition c: in the spot by spot approach, $\widehat{Y}_{jcg}$ is simply equal to $Y_{jc\cdot}$; in the global approach, $\widehat{Y}_{jcg} = \widehat{(\mathrm{G})}_g + \widehat{(\mathrm{SpC})}_{jc}$, where $\widehat{(\mathrm{G})}_g$ and $\widehat{(\mathrm{SpC})}_{jc}$ are respectively the estimated gel and spot$\times$condition effects. The residuals are estimating the random errors E_{jcg}. If the chosen model is correct, then their distribution is nearly the same than the errors distribution. Therefore, structures in the variance of the observations may be detected for example by examining graphics of residuals versus the predicted values, or the position on the gel. If such structures exist, they can be taken into account in the global ANOVA model. Moreover, looking carefully at spots j whose absolute residuals $|R_{jcg}|$ or empirical variances are very high, may reveal problems during the image analysis process, as mismatching for example. It gives the opportunity to correct the data if necessary.

A residual analysis for studying the variability of data coming from example of Section 1.2 is shown at Figure 1.5. The graphic of residuals versus the predicted values shows that the residuals are increasing with the spot volume. The optimal transformation for stabilizing the variance is estimated by the Box-Cox method: we found $T(\%V) = (\%V)^{1/3}$. For that example, we did not found that the data variability was depending on the spots position on the gel.

Other sources of heterogeneity may exist in the data, and the distribution of residuals may be much more spread out than the gaussian distribution though no particular structure was detected in the variance of the observations. Some authors [36, 21] proposed in the context of differential analysis of gene expression, to use bootstrap methods to address the problem of non Gaussian distribution of the test statistic. Nevertheless it should be noted that bootstrap method is not well adapted to the spot by spot approach because of the small number of replications. Moreover applying bootstrap methods for the differential analysis of 2-DE in a global ANOVA model, leads to heavy computation. Indeed, because of missing data, the algorithm for estimating the parameters is time consuming.

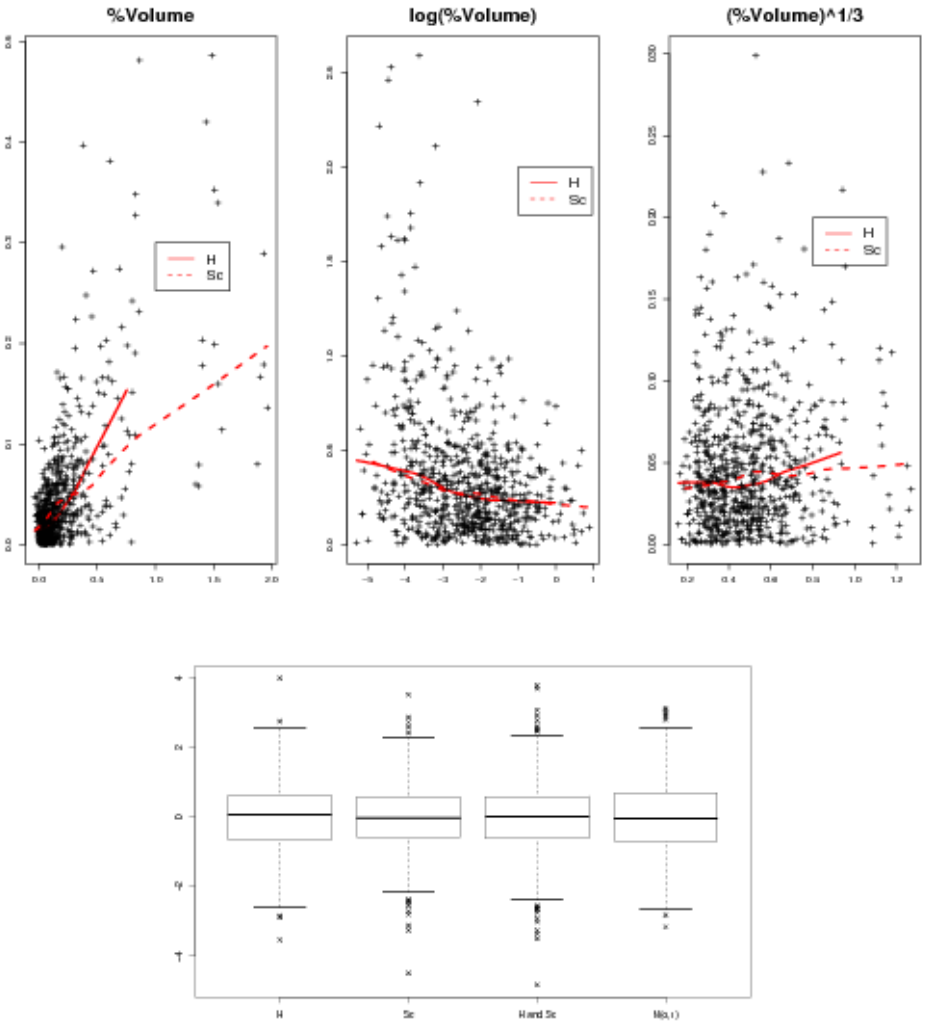

Figure 1.5: *Graphics for studying the data variability. The two first graphics represent the absolute values of the residuals* $R_{jcg} = Y_{jcg} - \widehat{Y}_{jcg}$ *versus the predicted values* $\widehat{Y}_{jcg}$ *for the data coming from Example of Section 1.2. Only spots whose volume ratio was greater than 2 were considered in the differential analysis. On the left, the* Y_{jcg} *are the percentage of volumes on each gel:* $Y_{jcg} = \%V_{jcg}$*. The lines are smoothed fits of the data, one for each condition. They clearly show that the residuals are increasing with the mean. The logarithmic transformation* $Y_{jcg} = \log(\%V_{jcg})$*, see the graphic on the middle, inverts the tendancy: the smoothed fits of the residuals are decreasing functions of the predicted values. On the right,* $Y_{jcg} = (\%V_{jcg})^{1/3}$*. This power transformation allows to stabilize the variance of the observations, the smoothed fits being nearly horizontal. The last graphic represents the distribution of the standardized residuals after the power transformation. The standardized residuals should be distributed as independant Gaussian variables with mean 0 and variance 1. The two first boxplots consider the residuals under each condition. They do not show any particular difference between the conditions. The third boxplot represents the distribution of all the residuals, and the last one the distribution of* n *simulated Gaussian (0,1) variates, where* n *is the total number of residuals. Looking at these graphics, there is no reason to suspect that the responses* $Y_{jcg} = (\%V_{jcg})^{1/3}$ *are not Gaussian distributed with the same variance.*

1.3.4 Strategy for missing data

Missing data cannot be ignored in differential analysis of 2-DE, because they affect a large number of spots, and because the lack of observation may be due to proteins variant in abundance.

Several reasons lead to missing observations, for example the actual absence of a given protein, or a mismatching. In some cases, it is possible to guess the reason. For example, when the spot is not observed on any gels within a condition, the protein may be absent. But generally it is hazardous to interpret missing data without a tedious inspection of the data, spot by spot.

The usual testing procedures used for the differential analysis does not need a complete data set, but they need a minimum number of observations for each spot, at least one observation for each spot under each condition. If we use the spot by spot approach, at least three observations for each spot for the comparison of two conditions (see Table 1.1A) are needed. But, as it is shown in Tables 1.3 and 1.4, one should prefer to have at least 5 or 6 observations for each spot.

Some authors proposed to set the missing data to the value 0, or to the lowest observed value in the data set [1]. Such a procedure assumes that all missing data are due to lack of protein. Others proposed to replace the missing data for one spot on one gel by the mean of the observations for this spot. More sophisticated methods have been proposed as the k-nearest neighbour method [32, 31]. Nevertheless they are not adapted to the case where values are missing on all the gels corresponding to one condition.

Another solution is to replace missing data by some simulated values, for example by drawing Gaussian variables with mean m and variance s^2. The values of m and s^2 may be chosen with the help of the data. For example, m is the smallest or one of the smallest observed values as the 0.025 quantile of the data, and s^2 is the median of the empirical variances calculated for each spot. The question is now *how many missing data must be replaced by simulation ?* One possibility is to simulate missing data in order to get the minimum number of observations required for the statistical analysis. At the opposite end we could simulate data wherever they are missing. The risk is then to bias the differential analysis by introducing additional information possibly erroneous.

What is a good strategy for missing data in 2-DE analysis is an open question that needs further work.

1.3.5 Discussion

Whatever the approach chosen, spot by spot or global approach, it is always advantageous to carry out preliminary analyses as described in Section 1.3.3. The differential analysis of 2-DE gels is an iterative process. The statistical analysis will provide a list of differentially expressed spots in terms of protein abundance, based on a decision rule strongly dependant on the data variability and on the number of replications, see Table 1.4. This list will be confirmed or rejected by the researcher. If several spots are rejected, it may be worthwhile to suppress these spots from the data and go back to the beginning of the analysis.

One customary practice is to retain among spots detected as variants those that are biologically significant, see [43]. For example, the 2-fold change rule is applied: it consists of keeping spots whose volume ratio is greater than 2. Another practice is to do the differential analysis only with spots whose volume ratio is greater than 2. Let us clarify

that from a statistical point of view, those rules have no meaning. In practice, spots whose volume ratio is smaller than 2 may be observed with great precision and with a large number of replications, and thus may be detected as variant. Suppression of those spots before the differential analysis may lead to eliminate variant spots. Statistics cannot decide what is biologically pertinent or not, but can propose objective methods based on the data, to suggest both what could be interesting, and what should be moved aside or corrected.

Let us now discuss the choice between the spot by spot and the global approaches.

- The spot by spot analysis does not need sophisticated software, and is proposed by the software packages used for image analysis. It is thus very attractive. Nevertheless, as each test uses information coming from only one spot, a large number of replications are necessary, see Table 1.4. In section 1.3.1 we assumed that the variance of the observations for one spot was identical under both conditions. This assumption could be relaxed and the test statistic adapted to the case where the variance is dependant on the condition. Therefore, the number of replications by condition should be large enough to estimate properly the variance of the test statistic, see Table 1.3.

  The Mann-Whitney test is attractive because it does not assume Gaussian distribution but it is based on the ranks of the observations rather than on the observations. It lacks power when the number of observations is small [53]: a minimum of 7 replications by condition is needed according to [43].

  Whatever the test statistic, it is assumed that for each condition, and spot, the observations are replications. Therefore, the data normalization, as suppressing the gel effect, and more generally the block effects, must be done before the testing procedure. However, it should be noticed that including additional effects reduces degrees of freedom in the Student statistic.

- The global analysis uses information from all the data for testing each hypothesis H_j. The gel effects on the mean response, denoted $(\mathrm{G})_g$ in section 1.3.1, are estimated together with the spot×condition effects, denoted $(\mathrm{SpC})_{jc}$. The variance has been assumed the same for all spots, but this assumption may be weakened by taking into account information on the variance structure. For example, the variance may depend on the condition, or on the spot localization on the image, or on the spot. Because of missing data, a statistical software, such as R `cran.r-project.org` or SAS `www.sas.com` is needed.

Let us finally underline that detecting *significant* differences in protein abundance relies on a statistical procedure that compares the differences of observed spot volumes to their variability. Therefore, the experimental design must guarantee the possibility to estimate properly this variability. Variability in the data may come from the biological and technical phases. Replications in the biological phase may be difficult to obtain in some situations, as for example when sample are taken on people or animals. In the technical phase, 3 or 4 replications in most proteomics studies should be possible. The statistician has to take into account these situations, to propose suitable statistical methods, as for example

methods based on global ANOVA models, and to precise the limits in which the results can be handled.

1.4 Conclusion

Accurate differential analysis of proteomic data outcomes of rigorously designed experiments and produces reliable results. This dynamic interaction requires a close interdisciplinary collaboration at every step of the project and is beneficial for both biologists and statisticians. Further investigations using the results issued from such a collaboration can be considered with increased confidence. Statistical tools such as discriminate analysis, regression methods or supervised classification [8, 19, 28, 59, 7, 17] can be further applied to accurately discriminate the status of unknown samples, normal or pathologic for instance. The interaction schema between statisticians and biologists is particularly important for the detection of differentially expressed proteins involved in pathologies since it can lead to the discovery of biomarker candidates. Another field of collaboration between both disciplines is the search for functional molecular (proteins only or proteins and mRNAs etc.) networks. The aim of this approach is to establish the relationships existing between the different cellular actors in order to (re)-construct a causality network. Statistical methods in this field are under development and numerous fundamental mathematical researches are actively in progress [50, 57, 38, 41]. It should be emphasized that the interactions between mathematicians, statisticians and biologists are not limited for providing increased confidence in biological results; they allow the delineation of new areas where collaborative research is needed.

Tables

Table 1.1 : Spot by spot analysis: test of H_j

A Comparison of two conditions

Let us denote $Y_{jc\cdot}$ the empirical mean of the Y_{jcg}, n_{jc} the number of observations of spot j under condition c, and SCR_{jc} the residual sum of squares defined as follows:

$$\mathrm{SCR}_{jc} = \sum_{g=1}^{n_{jc}} (Y_{jcg} - Y_{jc\cdot})^2.$$

If $n_{j1} + n_{j2} \geq 3$, the test statistic for testing H_j: "$m_{j1} = m_{j2}$" against "$m_{j1} \neq m_{j2}$" is defined as follows :

$$S_j = \frac{|Y_{j1\cdot} - Y_{j2\cdot}|}{\sqrt{\frac{\mathrm{SCR}_{j1}+\mathrm{SCR}_{j2}}{n_{j1}+n_{j2}-2}\left(\frac{1}{n_{j1}} + \frac{1}{n_{j2}}\right)}}.$$

Let us denote by Z_d a Student variable with d degrees of freedom. If H_j is true, then S_j is distributed as $|Z_{n_{j1}+n_{j2}-2}|$ and the pvalue p_j is defined as

$$p_j = \mathrm{pr}\left(\left|Z_{n_{j1}+n_{j2}-2}\right| > S_j\right)$$

B Comparison of C conditions, $C \geq 3$. If n_j, the total number of observations for spot j, is greater than $C + 1$, the test statistic for testing H_j: "$m_{j1} = \ldots = m_{jC}$" against the alternative that there exists two conditions c, c' such that "$m_{jc} \neq m_{jc'}$" is defined as follows :

$$S_j = \frac{n_j - C}{C - 1} \frac{\sum_{c=1}^{C} n_{jc}(Y_{jc\cdot} - Y_{j\cdot\cdot})^2}{\sum_{c=1}^{C} \mathrm{SCR}_{jc}}$$

where $Y_{jc\cdot}$, n_{jc} and SCR_{jc} are defined as above, and $Y_{j\cdot\cdot}$ is the mean of the responses for spot j.

Let us denote by F_{d_1,d_2} a Fisher variable with d_1 and d_2 degrees of freedom. If H_j is true, then S_j is distributed as F_{C-1,n_j-C} and the pvalue p_j is defined as

$$p_j = \mathrm{pr}\left(F_{C-1,n_j-C} > S_j\right)$$

Table 1.2 : Global ANOVA approach: test of H_j

Assume that we compare two conditions. The test statistic S_j, is based on the least squares estimators of the differences $(\text{SpC})_{j1} - (\text{SpC})_{j2}$'s divided by their estimated standard errors. S_j is distributed as $|Z_{n-GC-(J-1)(C-1)}|$, where $n = \sum_{j=1}^{J} n_j$ is the total number of observations, J the number of spots, G the number of gels for each condition.

The pvalues p_j are defined as in Table 1.1.

Table 1.3 : Effect of the sample size on the estimated variance variability

Let $X_1, \ldots, X_n$ be n independant Gaussian observations with mean m and variance σ^2. The empirical variance s^2 defined as follows

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - X_\cdot)^2 \,, \text{ where } X_\cdot = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

is an unbiased estimator of the variance σ^2. Its coefficient of variation is equal to

$$\text{CV}(s^2) = 100 \frac{\text{standard-error}(s^2)}{\text{mean}(s^2)} = 100 \sqrt{\frac{2}{n-1}}.$$

It follows immediately that if $n = 3$, $\text{CV}(s^2) = 100\%$, if $n = 9$, $\text{CV}(s^2) = 50\%$

Let us now apply these results to the Student test used for testing H_j in the spot by spot approach. For the sake of simplicity, assume that $n_{j1} = n_{j2} = n_j/2$. The denominator of S_j is the square-root of the estimated variance of the difference $Y_{j1} - Y_{j2}$. More precisely, the variance of $Y_{j1} - Y_{j2}$ is estimated by $S_j^2 = 4s_j^2/n_j$ where s_j^2 is the empirical variance. Using the results given above, we get that the coefficient of variation of S_j^2 equals $100\sqrt{2/(n_j - 2)}$. For example, $n_j = 6$ leads to $\text{CV}(4s_j^2/n_j) = 71\%$, $n_j = 12$ leads to 45%. These simple calculations show the importance of the number of replications in the differential analysis.

Table 1.4 : Spot by spot approach and Student statistic: variations of the pvalues as function of the estimated standard-error and the number of replications

Let us consider a differential analysis of 2-DE comparing two conditions, based on a spot by spot approach, where the number of spots J is equal to 500. Suppose that

- after the logarithmic transformation of the volume percentages, the responses are Gaussian distributed with the same variance,
- using a Bonferroni procedure that controls the FWER at 5%, 5 spots were detected as variant. Precisely, the hypothesis H_j was rejected if the pvalue was lower than 0.0001.
- using the procedure controlling the FDR at 5%, we found 39 variant spots. In that case, the hypothesis H_j was rejected if the pvalue was lower than 0.0039.

Let us consider spots for which the means difference $\delta_{12} = |Y_{j1\cdot} - Y_{j2\cdot}|$ equals $\log(2)$. According to our experience when analyzing 2-DE data, the estimated standard-error of these δ_{12} may vary between 0.05 and 1. The table below gives the pvalues corresponding to $\delta_{12} = \log(2)$ for several values of their estimated standard-errors, denoted s.e.(δ_{12}) and several values of the number of replications. It is assumed that the number of replications is the same under each conditions: $n_{j1} = n_{j2} = n_j/2$.

	$n_j = 4$	$n_j = 6$	$n_j = 8$	$n_j = 10$	$n_j = 12$
s.e.$(\delta_{12}) = 0.05$	0.0052	< 0.0001	< 0.0001	< 0.0001	< 0.0001
s.e.$(\delta_{12}) = 0.1$	0.020	0.00011	< 0.0001	< 0.0001	< 0.0001
s.e.$(\delta_{12}) = 0.2$	0.074	0.013	0.0027	0.00059	0.00013
s.e.$(\delta_{12}) = 0.5$	0.30	0.16	0.097	0.059	0.037
s.e.$(\delta_{12}) = 1$	0.55	0.44	0.36	0.30	0.26

This table highlights that the decision rule is strongly dependant on the variability of the data and the number of replications.

References

[1] Almeida, J., Stanislaus, R., Krug, E., and Arthur, J. Normalization and analysis of residual variation in two-dimensional gel electrophoresis for quantitative differential proteomics. *Proteomics 5* (2005), 1242–1249.

[2] Aubert, J., Bar-Hen, A., Daudin, J.-J., and Robin, S. Determination of the differentially expressed genes in microarray experiments using local fdr. *BMC Bioinformatics 5* (2004), article 125.

[3] Bahrman, N., Gouis, J. L., Negroni, L., Amilhat, L., Leroy, P., Lainé, A.-L., and Jaminon, O. Differential protein expression assessed by two-dimensional gel electrophoresis for two wheat varieties grown at four nitrogen levels. *Proteomics 4*, 3 (2004), 709–719.

[4] Baldi, P., and Long, A. A bayesian framework for the analysis of microarray expression data; regularized t-test and statistical inferences of gene changes. *Bioinformatics 17* (2001), 509–519.

[5] Benjamini, Y., and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B 57* (1995), 289–300.

[6] Bosque, P., and Prusiner, S. Cultured cell sublines highly susceptible to prion infection. *Journal of Virology 74* (2000), 4377–86.

[7] Boulesteix, A., and Tutz, G. Identification of interaction patterns and classification with applications to microarray data. *Computational Statistics and Data Analysis 50*, 3 (2006), 783–802.

[8] Boulesteix, A. L., Tutz, G., and Strimmer, K. A cart-based approach to discover emerging patterns in microarray data. *Bioinformatics 19*, 18 (2003), 2465–2472.

[9] Brien, C. Analysis of variance tables based on experimental structure. *Biometrics 39*, 1 (1983), 53–59.

[10] Brien, C., and Bailey, R. Multiple randomizations. *Journal of the Royal Statistical Society: Series B 68*, 4 (2006), 571–609.

[11] Chang, J., Remmen, H. V., Ward, W., Regnier, F., Richardson, A., and Cornell, J. Processing of data generated by 2-dimensional gel electrophoresis for statistical analysis: missing data, normalization, and statistics. *Journal of Proteome Research 3* (2004), 1210–1218.

[12] Chich, J.-F., Schaeffer, B., Bouin, A.-P., Mouthon, F., Labas, V., Larramendy, C., Deslys, J.-P., and Grosclaude, J. Functional blocks identified by proteomics are similarly impaired by an anti-prion drug and by prion infection in a neuronal cell model. in preparation.

[13] Choe, L., and Lee, K. Quantitative and qualitative measure of intralaboratory two-dimensional protein gel reproducibility and the effects of sample preparation, sample load, and image analysis. *Electrophoresis 24* (2003), 3500–7.

[14] Churchill, G. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics Supplement 32* (2002), 490–495.

[15] Cox, D. *Planning of experiments*. Wiley series in Probability and Mathematical Statistics. John Wiley, 1958.

[16] Cronier, S., Laude, H., and Peyrin, J. Prions can infect primary cultured neurons and astrocytes and promote neuronal cell death. *Proceedings of the National Academy of Sciences of the United States of America 101* (2004), 12271–6.

[17] Dai, J., Lieu, L., and Rocke, D. Dimension reduction for classification with gene expression microarray data. *Statistical Applications in Genetics and Molecular Biology 5*, 1 (2006), article 6.

[18] Dean, A., and Voss, D. *Design and analysis of experiments*. Springer, New York, 1999.

[19] Dettling, M., and Bühlmann, P. Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis 90*, 1 (2004), 106–131.

[20] Dowsey, A., Dunn, M., and Yang, G.-Z. The role of bioinformatics in two-dimensional gel electrophoresis. *Proteomics 3* (2003), 1567–1596.

[21] Dudoit, S., Shaffer, J. P., and Boldrick, J. C. Multiple hypothesis testing in microarray experiments. *Statistical Science 18* (2003), 71–103.

[22] Dudoit, S., van der Laan, M., and Pollard, K. Multiple testing. part i. single-step procedures for control of general type i error rates. *Statistical Applications in Genetics and Molecular Biology 3*, 1 (2004), article 13.

[23] Dudoit, S., Yang, Y., Callow, M., and Speed, T. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Journal of the American Statistical Association 74* (2002), 829–836.

[24] Durbin, B., Hardin, J., Hawkins, D., and Rocke, D. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics 18*, Suppl 1 (2002), S105–S110.

[25] Gustafsson, J., Blomberg, A., and Rudemo, M. Warping two-dimensional electrophoresis gel images to correct for geometric distortions of the spot pattern. *Electrophoresis 23* (2002), 1731–1744.

[26] Gustafsson, J., Ceasar, R., Glasbey, C., Blomberg, A., and Rudemo, M. Statistical exploration of variation in quantitative two-dimensional gel electrophoresis data. *Proteomics 4* (2004), 3791–3799.

[27] Hanash, S. Samir hanash discusses how hupo aims to globalize proteomics research. *Drug Discovery Today 7*, 15 (2002), 797–801.

[28] Hastie, T., and Tibshirani, R. Efficient quadratic regularization for expression arrays. *Biostatistics 5*, 3 (2004), 329–340.

[29] Hu, J., Coombes, K., Morris, J., and Baggerly, K. The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. *Briefings in Functional Genomics and Proteomics 3* (2005), 322–31.

[30] Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., and M.Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics 18*, Suppl 1 (2002), S96–S104.

[31] Jung, K., Gannoun, A., Sitek, B., Apostolov, O., Schramm, A., Meyer, H., Stuhler, K., and Urfer, W. Statistical evaluation of methods for the analysis of dynamic protein expression data from a tumor study. *REVSTAT Statistical Journal 4* (2006), 67–80.

[32] Jung, K., Gannoun, A., Sitek, B., Meyer, H., Stuhler, K., and Urfer, W. Analysis of dynamic protein expression data. *REVSTAT Statistical Journal 3* (2005), 99–111.

[33] Karp, N., and Lilley, K. Maximising sensitivity for detecting changes in protein expression: Experimental design using minimal cydyes. *Proteomics 5*, 12 (2005), 3105 – 3115.

[34] Kerr, M. Design considerations for efficient and effective microarray studies. *Biometrics 59*, 4 (2003), 822–828.

[35] Kerr, M., and Churchill, G. Experimental design for gene expression microarrays. *Biostatistics 2* (2001), 183–201.

[36] Kerr, M., Martin, M., and Churchill, G. Analysis of variance for gene expression microarray data. *Journal of Computational Biology 7* (2000), 819–837.

[37] Klöhn, P.-C., Stoltze, L., Flechsig, E., Enari, M., and Weissmann, C. A quantitative, highly sensitive cell-based infectivity assay for mouse scrapie prions. *Proceedings of the National Academy of Sciences of the United States of America 100* (2003), 11666–71.

[38] Li, H., and Gui, J. Gradient directed regularization for sparse gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics 7*, 2 (2006), 302–317.

[39] Liu, B.-F., Sera, Y., Matsubara, N., Otsuka, K., and Terabe, S. Signal denoising and baseline correction by discrete wavalet transform for microchip capillary electrophoresis. *Electrophoresis 24* (2003), 3260–3265.

[40] Malone, J., McGarry, K., and Bowerman, C. Performing trend analysis on spatio-temporal proteomics data using differential ratio data mining. In *6th EPSRC PREP Conference* (2004).

[41] Meinshausen, N., and Bühlmann, P. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34*, 3 (2006).

[42] Mellon, P., Windle, J., Goldsmith, P., Padula, C., Roberts, J., and Weiner, R. Immortalization of hypothalamic gnrh neurons by genetically targeted tumorigenesis. *Neuron 5* (1990), 1–10.

[43] Meunier, B., Bouley, J., Piec, I., Bernard, C., Picard, B., and Hocquette, J. Data analysis methods for detection of differential protein expression in two-dimensional gel electrophoresis. *Analytical Biochemistry 340* (2005), 226–230.

[44] Molloy, M., Brzezinski, E., Hang, J., McDowell, M., and VanBogelen, R. Overcoming technical variation and biological variation in quantitative proteomics. *Proteomics 3*, 10 (2003), 1912–1919.

[45] Reiner, A., Yekutieli, D., and Benjamini, Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics 19* (2003), 368–375.

[46] Riter, L., Vitek, O., Gooding, K., Hodge, B., and Julian, R. Statistical design of experiments as a tool in mass spectrometry. *Journal of Mass Spectrometry 40* (2005), 565–79.

[47] Rogers, M., Graham, J., and Tonge, R. Statistical models of shape for the analysis of protein spots in two-dimensional electrophoresis gel images. *Proteomics 3* (2003), 887–896.

[48] Rogers, M., Graham, J., and Tonge, R. Using statistical image models for objective evaluation of spot detection in two-dimensional gels. *Proteomics 3* (2003), 879–886.

[49] Roy, A., Seillier-Moiseiwitsch, F., Lee, K., Hang, Y., Marten, M., and Raman, B. Analyzing two-dimensional gel images. *Chance 16* (2003), 13–18.

[50] Schäfer, J., and Strimmer, K. An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics 21*, 6 (2005), 754–764.

[51] Schatzl, H., Laszlo, L., Holtzman, D., Tatzelt, J., DeArmond, S., Weiner, R., Mobley, W., and Prusiner, S. A hypothalamic neuronal cell line persistently infected with scrapie prions exhibits apoptosis. *Journal of Virology 71* (1997), 8821–31.

[52] Storey, J., and Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America 100* (2003), 9440–9445.

[53] Wang, S., and Ethier, S. A generalized likelihood ratio test to identify differentially expressed genes from microarray data. *Bioinformatics 20* (2004), 100–104.

[54] Wasinger, V., Cordwell, S., Cerpa-Poljak, A., Yan, J., Gooley, A., Wilkins, M., Duncan, M., Harris, R., Williams, K., and Humphery-Smith, I. Progress with gene-product mapping of the mollicutes: Mycoplasma genitalium. *Electrophoresis 16* (1995), 1090–4.

[55] Wilkins, M., Appel, R., Eyk, J. V., Chung, M., Gorg, A., Hecker, M., Huber, L., Langen, H., Link, A., Paik, Y., Patterson, S., Pennington, S., Rabilloud, T., Simpson, R., Weiss, W., and Dunn, M. Guidelines for the next 10 years of proteomics. *Proteomics 6*, 1 (2006), 4–8.

[56] Wilkins, M., Sanchez, J., Gooley, A., Appel, R., Humphery-Smith, I., Hochstrasser, D., and Williams, K. Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnology and Genetic Engineering Reviews 13* (1996), 19–50.

[57] Wille, A., and Bühlmann, P. Low-order conditional independence graphs for inferring genetic networks. *Statistical Applications in Genetics and Molecular Biology 5*, 1 (2006), 1–32.

[58] Yang, Y., and Speed, T. Design issues for cDNA microarray experiments. *Nature Reviews Genetics 3* (2002), 579–588.

[59] Zhu, J., and Hastie, T. Classification of gene microarrays by penalized logistic regression. *Biostatistics 5*, 3 (2004), 427–443.

Chapitre 2

Sélection de modèles pour l'analyse différentielle d'images d'électrophorèse

L'analyse différentielle du protéome a pour but de comparer les expressions des protéines correspondant à deux ou plusieurs conditions expérimentales différentes. Pour détecter les spots dont le volume diffère selon les conditions, nous étudions les différences d'interaction des spots entre les conditions deux à deux dans un modèle d'analyse de variance. L'analyse différentielle revient alors à détecter les composantes non nulles de l'espérance d'un vecteur gaussien de grande dimension. Dans le cas où on désire comparer simultanément plus de 3 conditions expérimentales, les composantes de ce vecteur gaussien ne sont plus indépendantes. Nous proposons d'estimer le nombre de composantes non nulles d'un vecteur gaussien dont la structure de covariance est connue (à une constante près) à l'aide d'une méthode de sélection de modèles reposant sur un critère des moindres carrés pénalisé.

2.1 Introduction

Reprenant l'approche globale proposée au Chapitre 1, nous modélisons les observations à l'aide d'un modèle d'analyse de la variance de la façon suivante :

$$Y_{jcg} = \mu + \alpha_j + \beta_c + \gamma_{jc} + \delta_{cg} + \sigma\varepsilon_{jcg} \tag{2.1}$$

où Y_{jcg} est l'observation du spot j, sous la condition c et sur le gel g, μ un effet moyen, α_j l'effet du spot j, β_c l'effet de la condition c, δ_{cg} l'effet du gel g de la condition c et γ_{jc} le terme d'interaction entre le spot j et la condition c. Les variables aléatoires ε_{jcg} sont supposés indépendantes, de loi normale centrée et de variance 1. Le nombre de spots j varie de 1 à J, le nombre de conditions c de 1 à C, et le nombre d'observations g du spot j sous la condition c varie de 1 à G. Nous nous plaçons en effet dans le cadre d'un plan équilibré, sans données manquantes, c'est-à-dire que l'on observe les J spots sur les G gels de chaque condition.

Si le modèle donné par l'équation (2.1) est validé, la différence d'expression du spot j entre les conditions c et c' est mesurée par la différence $\gamma_{jc} - \gamma_{jc'}$. Nous cherchons à détecter

les spots j pour lesquels les différences d'interaction $\gamma_{jc} - \gamma_{jc'}$ sont fortement positives ou négatives.

2.2 Présentation du modèle

Soit $\gamma = (\gamma_{jc})$ avec $j \in \{1, \ldots, J\}$ et $c \in \{1, \ldots, C\}$ le vecteur des interactions et $\hat{\gamma}$ son estimateur donné par l'analyse de variance. Nous notons F le vecteur des différences d'interactions :

$$F = (F_{jcc'}) = (\gamma_{jc} - \gamma_{jc'}) \qquad j \in \{1, \ldots, J\},\ \ c < c' \in \{1, \ldots, C\}^2$$

et $X = (\hat{\gamma}_{jc} - \hat{\gamma}_{jc'})_{jcc'}$ son estimateur. Les vecteurs F et X sont de dimension $n = JC(C-1)/2$ et nous notons F_i et X_i, pour $i \in \{1, \ldots, n\}$ leur n composantes.

La structure de covariance de X est connue car issue de l'analyse de variance. Cette structure dépend du nombre de conditions que l'on compare.

2.2.1 Comparaison de $C = 2$ conditions

Dans le cas où l'on compare $C = 2$ conditions, n est égal à J et pour tout $i \in \{1, \ldots, n\}$:

$$\mathrm{Var}(X_i) = \mathrm{Var}(\hat{\gamma}_{j1} - \hat{\gamma}_{j2}) = \frac{\sigma^2}{G}(2 - \frac{2}{J})$$

$$\mathrm{Cov}(X_i, X_{i'}) = -2\frac{\sigma^2}{JG}$$

Lorsque le nombre de spots J est grand, on peut donc négliger les corrélations entre les X_i, $i \in \{1, \ldots, n\}$. Le modèle de régression est donc :

$$X = F + \tau\varepsilon$$

avec ε de loi normale centrée et de matrice de covariance la matrice identité I_n, et $\tau = \frac{\sigma^2}{G}(2 - \frac{2}{J})$.

Dans ce cadre d'un modèle gaussien dont les erreurs sont indépendantes, plusieurs méthodes ont déjà été proposées pour détecter les composantes non nulles de F.

Si on considère une approche test, détecter les composantes non nulles de F revient à tester simultanément un grand nombre d'hypothèses : pour chaque $i \in \{1, \ldots, n\}$, nous testons l'hypothèse H_i :" $F_i = 0$". Benjamini et Hochberg [1] ont proposé une procédure qui permet de contrôler le taux de faux positifs.

Si on considère une approche estimation, il s'agit d'estimer le nombre de composantes non nulles de F. Ce problème a déjà fait l'objet de plusieurs travaux. L'approche proposée par Birgé et Massart [4] basée sur la vraisemblance pénalisée peut être appliquée. Les travaux de Huet [6], puis Baraud et al. [9] considèrent le cas où la variance est inconnue.

2.2.2 Comparaison de C conditions avec $C > 2$

Dans le cas où l'on compare plus de 2 conditions, les corrélations entre les X_i ne peuvent plus être négligées. En effet,

$$\mathrm{Var}(\hat{\gamma}_{jc} - \hat{\gamma}_{jc'}) = \frac{\sigma^2}{G}(2 - \frac{2}{J})$$

$$\mathrm{Cov}(\hat{\gamma}_{jc} - \hat{\gamma}_{jc'}, \hat{\gamma}_{jc'} - \hat{\gamma}_{jc''}) = -\frac{\sigma^2}{G} + \frac{\sigma^2}{JG}$$

$$\mathrm{Cov}(\hat{\gamma}_{jc} - \hat{\gamma}_{jc'}, \hat{\gamma}_{jc} - \hat{\gamma}_{jc''}) = \frac{\sigma^2}{G} - \frac{\sigma^2}{JG}$$

$$\mathrm{Cov}(\hat{\gamma}_{jc} - \hat{\gamma}_{jc'}, \hat{\gamma}_{j'c} - \hat{\gamma}_{j'c'}) = -2\frac{\sigma^2}{JG}$$

Les termes en $1/G$, où G est le nombre de gels ne peuvent être négligés car en général G est petit, et les composantes du vecteur gaussien X ne sont donc plus indépendantes. Nous considérons alors le modèle suivant :

$$X = F + \sigma P_n \varepsilon \tag{2.2}$$

où ε est de loi normale centrée et de matrice de covariance la matrice identité I_n, et $\sigma^2 P_n P_n'$ est la matrice de variance-covariance de X. La matrice $P_n P_n'$ est connue. L'analyse de variance fournit un estimateur de σ^2. Cependant, le modèle d'analyse de variance donné par l'équation (2.1) suppose que les observations sont de même variance et que les gels sont des répétitions. Pour ne pas accorder trop de confiance à ce modèle, nous supposons dans la suite que σ^2 est inconnue.

Si le nombre J de spots est grand, les termes $\mathrm{Cov}(\hat{\gamma}_{jc} - \hat{\gamma}_{jc'}, \hat{\gamma}_{j'c} - \hat{\gamma}_{j'c'})$ peuvent être négligés. La matrice $P_n P_n'$ est alors diagonale par blocs. Nous avons montré que chaque bloc de $P_n P_n'$ a $C - 1$ valeurs propres non nulles, égales à C/G.

Dans la suite nous nous plaçons dans le modèle défini par l'équation (2.2). L'analyse différentielle consiste alors en la détection de composantes non nulles dans l'espérance d'un vecteur gaussien dont les composantes ne sont pas indépendantes.

Des méthodes basées sur des procédures de tests multiples et ne nécessitant pas d'hypothèse particulière sur la structure de dépendance des statistiques de test ont été proposées par Benjamini et Yekuteli [3] et Benjamini et Liu [2]. Ces méthodes présentées dans la Section 2.3, sont en fait conservatives.

Dans le cadre de méthode d'estimation, nous utilisons un résultat théorique établi par Yannick Baraud et proposons dans la Section 2.4 un critère de choix de modèle basé sur la vraisemblance pénalisée, tenant compte de la dépendance du modèle.

2.3 Approche tests d'hypothèses multiples

La procédure proposée par Benjamini et Yekuteli [3], et celle proposée par Benjamini et Liu [2], sont des extensions de la procédure de Benjamini et Hochberg [1], qui permettent de s'affranchir de l'hypothèse d'indépendance des statistiques de tests.

Ce sont deux procédures qui majorent le FDR, et qui sont valables quelle que soit la loi jointe des statistiques de tests. Le FDR est l'espérance du taux de faux positifs définis de la façon suivante :

$$FDR = E(\frac{\text{nombre de } H_i \text{ rejetées à tort}}{\text{nombre de } H_i \text{ rejetées}} 1_{\text{ nb de } H_i \text{ rejetées } \geq 1})$$

Procédure de Benjamini et Yekuteli :

– Soit

$$u_i = \frac{i}{n(1 + \frac{1}{2} + \ldots + \frac{1}{n})}\alpha$$

pour $i \in \{1, \ldots, n\}$

La procédure consiste à :

– Classer les p-values associées aux tests d'hypothèse H_i pour $i \in \{1, \ldots, n\}$.

$$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$$

– Rejeter les hypothèses $H_{(1)}, \ldots, H_{(k)}$, où :

$$k = \max(i \in \{1, \ldots, n\}, p_{(i)} \leq u_i).$$

Cette procédure permet de contrôler le FDR : $FDR \leq \alpha$

Procédure de Benjamini et Liu

– Soit

$$d_i = \min(1, \frac{n}{(n - i + 1)^2}\alpha)$$

pour $i \in \{1, \ldots, n\}$

La procédure consiste à :

– Classer les p-values associées aux tests d'hypothèse H_i pour $i \in \{1, \ldots, n\}$.

$$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$$

– Rejeter les hypothèses $H_{(1)}, \ldots, H_{(k-1)}$, où :

$$k = \min(i \in \{1, \ldots, n\}, p_{(i)} > d_i).$$

Cette procédure permet aussi de contrôler le FDR : $FDR \leq \alpha$

Ces procédures peuvent être utilisées pour l'analyse différentielle avec un nombre de traitements supérieur à deux, puisqu'elles ne nécessitent pas d'hypothèses particulières sur la structure de dépendance des statistiques de tests. Cependant ces procédures sont très conservatives. En fait le FDR peut être très inférieur à α ce qui a pour conséquence de diminuer la puissance du test. La correction de Bonferroni permet de contrôler le FWER, c'est-à-dire la probabilité de rejeter une hypothèse à tort. Elle est également valable quelle que soit la structure de dépendance entre les statistiques de tests. En pratique, elle peut s'avérer moins conservatives que les deux méthodes précédemment citées.

Dans la suite nous avons voulu exploiter le fait que la matrice $P_n P_n$' est connue, et nous avons considéré une approche basée sur des méthodes d'estimation par sélection de modèles.

2.4 Approche sélection de modèles

Le modèle considéré est de la forme :

$$X = F + \sigma P_n \epsilon,$$

où ε est de loi normale centrée et de matrice de covariance la matrice identité I_n. Nous notons $\mathcal{F}$ l'ensemble des vecteurs de $\mathbb{R}^n$ auquel appartient le vecteur inconnu F. Notre but est d'estimer le nombre de composantes non nulles de F. Dans cette section, nous décrivons les étapes de la sélection de modèle et nous donnons la "bonne" forme de la pénalité.

2.4.1 Principe

2.4.1.1 Collection de modèles

Soit m un sous ensemble de $\{1, \ldots, n\}$ de cardinal D_m. Nous considérons la collection $\mathcal{M}_n$ de tous les sous ensembles m de $\{1, \ldots, n\}$ de cardinal plus petit que D_{max} pour un certain D_{max}.

$$\mathcal{M}_n = \{m \subset \{1, \ldots, n\}, D_m \leq D_{max}\}$$

Pour chaque sous ensemble $m = (i_i \ldots, i_{D_m})$ de $\{1, \ldots, n\}$ de cardinal D_m, nous considérons $\mathcal{F}_m$ l'espace engendré par $(e_{i_1}, \ldots, e_{i_{D_m}})$ où les vecteurs $(e_1)_{i=1,\ldots,n}$ sont les vecteurs de la base canonique de $\mathbb{R}^n$. Nous obtenons donc une collection de sous espaces vectoriels de $\mathcal{F}$:

$$\{\mathcal{F}_m, m \in \mathcal{M}_n\}.$$

2.4.1.2 Définition de l'estimateur

Nous estimons F sur le sous espace $\mathcal{F}_m$ par minimisation d'une fonction de contraste. Nous considérons le contraste empirique qui est le critère des moindres carrées défini pour tout $T \in \mathcal{F}$ par :

$$\gamma_n(T) = ||X - T||_n^2$$

où on note $||.||_n^2$ la norme euclidienne de $\mathbb{R}^n$ renormalisée par n et définie pour tout $T \in \mathcal{F}$ par :

$$||T||_n^2 = \frac{\sum_{i=1}^n T_i^2}{n}$$

L'estimateur de F sur le sous espace linéaire $\mathcal{F}_m$, noté $\hat{F}_m$ est celui qui minimise le contraste empirique γ_n sur $\mathcal{F}_m$. Cet estimateur est défini par

$$\hat{F}_m = \text{argmin}_{\ T\in\mathcal{F}_m} ||X - T||_n^2$$

Soit Π_m la projection orthogonale sur l'espace $\mathcal{F}_m$. Alors l'estimateur $\hat{F}_m$ est défini par :

$$\hat{F}_m = \Pi_m X.$$

Nous obtenons donc une collection d'estimateurs de F :

$$\{\hat{F}_m, m \in \mathcal{M}_n\}.$$

2.4.1.3 Risque de l'estimateur

Pour mesurer la qualité de l'estimateur, nous cherchons à voir s'il est proche du vecteur F. Nous associons pour cela à chaque estimateur un risque. Le risque de l'estimateur $\hat{F}_m$ est défini par :

$$\mathbb{E}_F[||F - \hat{F}_m||_n^2].$$

Proposition 1. *L'estimateur $\hat{F}_m$ de F vérifie*

$$\mathbb{E}_F[||F - \hat{F}_m||_n^2] = ||F - \Pi_m F||_n^2 + \frac{Tr(\Pi_m P_n P_n')}{n}\sigma^2 \tag{2.3}$$

où Tr est l'opérateur Trace.

Preuve. Par définition de $\hat{F}_m = \Pi_m X$ et le théorème de Pythagore,

$$||F - \hat{F}_m||_n^2 = ||F - \Pi_m F||_n^2 + \sigma^2 ||\Pi_m P_n \epsilon||_n^2.$$

Comme $n||\Pi_m P_n \epsilon||_n^2 = \epsilon'(\Pi_m P_n)'(\Pi_m P_n)\epsilon$, alors $n\mathbb{E}_F[||\Pi_m P_n \epsilon||_n^2] = \mathbb{E}_F[\epsilon' P_n' \Pi_m P_n \epsilon] = Tr(P_n' \Pi_m P_n)$ car ε est de loi normale centrée et de matrice de covariance la matrice identité I_n.

Dans le cadre de l'analyse différentielle d'images d'électrophorèses, la matrice $P_n P_n'$ est connue. En effet, la structure de covariance de X est connue à σ^2 près car issue de l'analyse de variance et nous avons vu dans la Section 2.2.2, que les termes diagonaux de la matrice $P_n P_n'$ sont tous égaux à $\frac{1}{G}(2 - \frac{2}{J})$, noté dans la suite w.

Le risque de l'estimateur $\hat{F}_m$ se décompose donc en la somme de deux termes :

$$\mathbb{E}_F[||F - \hat{F}_m||_n^2] = ||F - \Pi_m F||_n^2 + \frac{D_m}{n} w\sigma^2 \tag{2.4}$$

Le premier terme est un terme de biais : il représente l'erreur d'approximation de F sur le sous-espace linéaire $\mathcal{F}_m$. Le second terme est un terme de variance : il représente l'erreur d'estimation dans $\mathcal{F}_m$. Ce terme est proportionnel à la dimension du modèle m.

L'objectif est de sélectionner le "meilleur" estimateur parmi la collection $(\hat{F}_m)_{m\in\mathcal{M}_n}$. L'estimateur idéal est celui qui réalise le plus petit des risques de la collection d'estimateurs. D'après la forme du risque défini en (2.4), plus la dimension D_m du modèle m est grande, plus le terme de biais diminue mais plus le terme de variance augmente. Le modèle idéal est donc celui qui réalise le meilleur compromis entre le biais et la variance.

Comme le risque dépend du vecteur inconnu F, l'estimateur idéal dépend lui aussi de F et ne peut donc pas être utilisé comme estimateur. L'objectif de la sélection de modèle est donc de construire un critère qui sélectionne à partir des données, un modèle $\hat{m}$ tel que l'estimateur $\hat{F}_{\hat{m}}$ ait un risque aussi proche que possible de l'infimum des risques.

2.4.1.4 Sélection de modèles

Soit $pen : \mathcal{M}_n \to \mathbb{R}^+$ une fonction de pénalité. L'estimateur du minimum de contraste pénalisé est défini par :

$$\tilde{F} = \hat{F}_{\hat{m}}$$

avec $\hat{m}$ qui minimise sur la collection de modèles $\mathcal{M}_n$ le critère pénalisé suivant :

$$crit(m, pen) = \gamma_n(\hat{F}_m) + pen(m) = ||X - \hat{F}_m||_n^2 + pen(m)$$

Le choix d"un "bon" estimateur de F est alors équivalent au choix d'une "bonne" fonction de pénalité *pen*.

2.4.2 Forme de la pénalité

Dans cette partie, nous utilisons un résultat établi par Y. Baraud[1] pour obtenir la "bonne" forme de la pénalité.

Donnons tout d'abord plusieurs notations utilisées dans la suite du chapitre. Les valeurs propres de la matrice $\Pi_m P_n P_n' \Pi_m$ sont notées $(\lambda_1(m), \ldots, \lambda_n(m))$. Nous définissons alors :

$$S^2(m) = \sum_{i=1}^{n} \lambda_i^2(m)$$

$$\text{et } M(m) = \max_i \{\lambda_i(m)\}.$$

Les valeurs propres de la matrice $P_n P_n'$ sont notées $(\lambda_1, \ldots, \lambda_n)$ et nous définissons :

$$M = \max_i \{\lambda_i\}.$$

[1] communication personnelle

2.4.2.1 Résultat établi par Y.Baraud

Y. Baraud a établi un résultat dans le cadre de l'estimation de l'espérance d'un vecteur gaussien lorsque les erreurs sont dépendantes. Ce résultat donne une minoration de la pénalité de façon à contrôler le risque de l'estimateur pénalisé. De ce résultat nous déduisons le théorème suivant :

Theorème 1. *Soit* $\beta \in]0,1[$. *Soient* $(L_m)_{m\in\mathcal{M}_n}$ *des poids strictement positifs tels que* $\Sigma_n := \sum exp(-L_m D_m) < \infty$. *Si* $\forall m \in \mathcal{M}_n$,

$$pen(m) \geq \frac{2}{1+\beta}\left\{Tr(\Pi_m P_n P_n') + \beta\frac{S^2(m)}{M} + 2S(m)\sqrt{L_m Dm} + 2M(m)L_m D_m\right\}\frac{\sigma^2}{n}$$

$$+\frac{1+\beta}{\beta}ML_m D_m\frac{\sigma^2}{n} \tag{2.5}$$

Alors

$$\mathbb{E}_F[||F-\tilde{F}||_n^2] \leq C_1(\beta)\inf_m\left\{||F-\Pi_m F||_n^2 + pen(m)\right\} + C_2(\beta)M\Sigma_n\frac{\sigma^2}{n}$$

avec $C_1(\beta) = (1+\beta)/(1-\beta)$ *et* $C_2(\beta) = (3+6\beta+\beta^2)/(\beta(1-\beta))$

La preuve de ce théorème est donnée dans la section 2.7. Dans la suite de cette section, nous simplifions la forme de la pénalité donnée par (2.5).

Corollaire 2. *Soit* $\beta \in (0,1)$ *et* $\eta \in (0,M)$. *Si* $\forall m \in \mathcal{M}_n$,

$$pen(m) \geq \frac{\sigma^2}{n}\left\{(2+\frac{2\eta}{1+\beta})Tr(\Pi_m P_n P_n') + (\frac{2}{1+\beta}(2+\frac{1}{\eta}) + \frac{1+\beta}{\beta})ML_m D_m\right\} \tag{2.6}$$

Alors

$$\mathbb{E}_F(||F-\tilde{F}||_n^2) \leq C_1(\beta)\inf_m\left\{||F-\Pi_m(F)||_n^2 + pen(m)\right\} + C_2(\beta)M\Sigma_n\frac{\sigma^2}{n} \tag{2.7}$$

Preuve. Ce corollaire est déduit du Théorème 1 par les deux inégalités suivantes :

$$S^2(m) \leq M(m)Tr(\Pi_m P_n P_n') \leq MTr(\Pi_m P_n P_n'), \tag{2.8}$$

et

$$\forall\eta \in (0,M), 2S(m)\sqrt{L_m D_m} \leq \frac{\eta}{M}S^2(m) + \frac{M}{\eta}L_m D_m \tag{2.9}$$

cette dernière inégalité provenant de :

$$\forall\alpha, y, z > 0, \quad 2yz \leq \alpha y^2 + \frac{1}{\alpha}z^2 \tag{2.10}$$

Les poids L_m sont liés à la complexité des modèles que l'on considère. Ils sont pris égaux à $\log(n/D_m)$ comme proposé par Birgé et Massart. A la différence du théorème établi par Birgé et Massart dans la cadre d'un modèle gaussien dans lequel les erreurs sont indépendantes, la fonction de pénalité dépend du modèle m mais pas uniquement via sa dimension D_m. En effet apparait dans la pénalité le terme $Tr(\Pi_m P_n P_n')$. La pénalité pour le modèle m dépend donc aussi de la composition de la matrice $P_n P_n'$ avec la projection sur l'espace $\mathcal{F}_m$.

2.4.2.2 Forme de la pénalité

Dans cette section, nous appliquons le Corollaire 2 dans le but d'estimer le nombre de composantes non nulles de F. La matrice $P_n P_n'$ étant connue, la valeur du maximum M des valeurs propres de la matrice $P_n P_n'$, et la trace de la matrice $\Pi_m P_n P_n'$ sont alors connues.

Le maximum des valeurs propres de $P_n P_n'$ vaut C/G. Les termes diagonaux de la matrice $P_n P_n'$ sont tous égaux à $w = \frac{1}{G}(2 - \frac{2}{J})$, et la trace de $\Pi_m P_n P_n'$ vaut donc wD_m.

La forme de la fonction de pénalité devient donc plus simplement :

$$pen(m) = \frac{\sigma^2}{n}\left\{c_1 w + c_2 \frac{C}{G} L_m\right\} D_m \tag{2.11}$$

où c_1 et c_2 sont deux constantes.

Cette fonction de pénalité dépend donc de deux constantes c_1 et c_2 dont les valeurs optimales sont inconnues et de la variance du bruit σ^2 qui est aussi inconnue. Dans la Section 2.5, nous considérons tout d'abord le cas où la variance du bruit est connue et nous calibrons de façon optimale les constantes c_1 et c_2. Puis dans la Section 2.6, nous abordons le problème de l'estimation de la variance.

2.5 Calibration des constantes de la pénalité

La pénalité est de la forme

$$pen(m) = \frac{\sigma^2}{n}\left\{c_1 w + c_2 \frac{C}{G} L_m\right\} D_m$$

Dans ce chapitre, nous supposons la variance du bruit σ^2 connue. La pénalité, et donc l'estimateur pénalisé $\tilde{F} = \hat{F}_{\hat{m}}$, dépendent du choix des constantes c_1 et c_2. Notre objectif est d'obtenir des valeurs optimales pour les constantes c_1 et c_2, c'est-à-dire des valeurs qui mènent à de "bons" estimateurs.

L'estimateur idéal est celui qui réalise le plus petit des risques de la collection d'estimateurs $\{\hat{F}_m, m \in \mathcal{M}_n\}$, appelé oracle et noté $\mathcal{O}_{(J,C)}(F, \sigma^2)$:

$$\mathcal{O}_{(J,C)}(F, \sigma^2) = \inf_{m \in \mathcal{M}_n} \mathbb{E}_F[\|F - \widehat{F}_m\|_n^2]$$

Cet estimateur idéal est inaccessible car dépend du vecteur inconnu F.

L'estimateur pénalisé $\tilde{F}$ est "bon" si son risque est aussi proche que possible de l'oracle. Notons $\mathcal{R}_{(J,C)}(\tilde{F}, \sigma^2, \text{pen})$ le risque associé à l'estimateur pénalisé $\tilde{F}$. Nous évaluons la performance de l'estimateur $\tilde{F}$par le rapport suivant :

$$\frac{\mathcal{R}_{(J,C)}(\tilde{F}, \sigma^2, \text{pen})}{\mathcal{O}_{(J,C)}(F, \sigma^2)} \tag{2.12}$$

Nous cherchons alors des constantes c_1 et c_2 optimales pour tout vecteur F et tout couple (J, C). Ces constantes sont choisies pour minimiser le rapport de risques défini en (2.12), uniformément en F, J et C. Elles sont obtenues par une étude de simulations.

2.5.1 Procédure de simulation

La procédure de simulations consiste à calculer pour un vecteur F et le couple (J, C) donnés, le rapport de risques défini par :

$$\frac{\mathcal{R}_{(J,C)}(\tilde{F}, \sigma^2, c_1, c_2)}{\mathcal{O}_{(J,C)}(F, \sigma^2)} \tag{2.13}$$

Nous commencons par décrire le calcul de ce rapport de risque pour un vecteur F et un couple (J, C) donné. Puis nous décrivons la procédure pour une collection de vecteurs F et une collection de couples (J, C). Dans la Section 2.5.2 nous donnons les paramètres de notre étude de simulations, en particulier la collection de vecteurs F et de couples (J, C) considérés et les différentes valeurs de c_1 et c_2 utilisées.

Calcul du rapport de risque pour un vecteur F et un couple (J, C) donnés

Nous calculons tout d'abord l'oracle :

$$\begin{aligned}
\mathcal{O}_{(J,C)}(F, \sigma^2) &= \inf_{m \in \mathcal{M}_n} \mathbb{E}_F[\|F - \widehat{F}_m\|_n^2] \\
&= \inf_{m \in \mathcal{M}_n} \left\{ ||F - \Pi_m F||_n^2 + \frac{D_m}{n} w\sigma^2 \right\} \\
&= \inf_{m \in \mathcal{M}_n} \frac{1}{n} \left\{ \sum_{i \notin m} F_i^2 + D_m w\sigma^2 \right\} \\
&= \inf_{D=0,\ldots,D_{max}} \frac{1}{n} \left\{ \sum_{i=D+1}^{n} F_{(i)}^2 + Dw\sigma^2 \right\}
\end{aligned}$$

lorsque les F^2 sont rangés par ordre décroissant : $F_{(1)}^2 \geq F_{(2)}^2 \geq \ldots \geq F_{(n)}^2$ et où nous rappelons que D_{max} est la dimension maximale de nos modèles.

Nous calculons ensuite le risque de l'estimateur $\tilde{F}$. Celui ci noté $\mathcal{R}_{(J,C)}(\tilde{F}, \sigma^2, c_1, c_2)$ est défini par :

$$\mathcal{R}_{(J,C)}(\tilde{F}, \sigma^2, c_1, c_2) = \mathbb{E}_F[\|F - \hat{F}_{\hat{m}}\|_n^2]$$

où $\hat{m}$ minimise le critère pénalisé :

$$crit(m, pen) = \gamma_n(\hat{F}_m) + pen(m) = ||X - \hat{F}_m||_n^2 + pen(m) \tag{2.14}$$

Ce risque est estimé par simulations par la méthode de Monte-Carlo.

- Nous simulons N_{sim} échantillons $X^s, s \in \{1, \ldots, N_{sim}\}$ de loi normale d'espérance F et de variance $\sigma^2 P_n P_n'$.
- Pour chaque échantillon nous calculons $\hat{m}^s(c_1, c_2)$ qui minimise le critère de vraisemblance pénalisée (2.14).
- Nous estimons alors $\mathcal{R}_{(J,C)}(\tilde{F}, \sigma^2, c_1, c_2)$ par :

$$\frac{1}{N_{sim}} \sum_{s=1}^{N_{sim}} ||F - \hat{F}_{\hat{m}^s(c_1,c_2)}||_n^2$$

Pour chaque vecteur F, chaque couple (J, C) et chaque constante c_1 et c_2, nous calculons alors le rapport de risque défini par l'Equation (2.13).

Procédure pour une collection de vecteurs F et de couples (J, C)

Nous cherchons ensuite les constantes c_1 et c_2 optimales pour tout F et tout couple (J, C). La procédure se décompose en deux étapes.

- Nous fixons tout d'abord la valeur du couple (J, C) et cherchons les valeurs optimales de c_1 et c_2 uniformément en F. Pour cela, nous considérons un ensemble fini de vecteurs F, nous appliquons la procédure décrite précedemment pour chacun des vecteurs et nous considérons le rapport suivant :

$$r_{(J,C)}(c_1, c_2) = \sup_F \frac{\mathcal{R}_{(J,C)}(\tilde{F}, \sigma^2, c_1, c_2)}{\mathcal{O}_{(J,C)}(F, \sigma^2)}$$

- Nous effectuons la même procédure pour plusieurs valeurs du couple (J, C). Nous notons $\mathcal{J}$ l'ensemble des valeurs choisies pour J et $\mathcal{C}$ l'ensemble des valeurs choisies pour C.

Nous disposons donc d'une famille de valeurs

$$\{r_{(J,C)}(c_1, c_2), c_1, c_2 > 0, J \in \mathcal{J}, C \in \mathcal{C}\}$$

Pour en extraire les constantes optimales c_1 et c_2, nous effectuons une étude graphique.

Nous cherchons à savoir si ces constantes optimales sont indépendantes de C et sinon de quelle manière elles en dépendent. Pour cela nous fixons une valeur de C. Nous fixons ensuite la constance c_1 et traçons les graphiques des fonctions :

$$c_2 \to r_{(J,C)}(c_1, c_2)$$

pour les différentes valeurs de J. Nous procèdons à cette étude pour différentes valeurs de c_1.

A c_1 et J fixés, la constante optimale de c_2 est donnée par :

$$c_2^*(J, C, c_1) = \arg\min_{c_2} r_{(J,C)}(c_1, c_2)$$

Pour chaque valeur de C, nous prenons alors comme valeur optimale pour c_1, la valeur qui stabilise $c_2^*(J, C, c_1)$ en J. Soit $c_1^*(C)$ cette valeur. Nous prenons alors comme valeur optimale de c_2, la valeur $c_2^*(C) = c_2^*(J, C, c_1^*(C))$ qui est stable pour toutes les valeurs de J.

Nous avons donc pour chaque valeur C les constantes optimales $c_1^*(C)$ et $c_2^*(C)$. Nous regardons alors si ces constantes dépendent de C.

2.5.2 Plan de simulation

Nous avons considéré les ensembles de valeurs suivantes :

- différentes valeurs de J : $\mathcal{J} = \{50, 100, 500, 1000\}$
- différentes valeurs de C : $\mathcal{C} = \{2, 3, 4, 5\}$
- une variance de bruit $\sigma^2 = 1$
- 6 valeurs de c_1 : $c_1 \in \{0, 2, 3, 4, 5, 8\}$, et nous faisons varier c_2 à partir de 0 par pas de 0.1.
- Nsim= 100
- Collection de vecteurs F :
  Dans le cadre de l'analyse différentielle, le vecteur F correspond au vecteur des différences d'interactions :

$$F = (F_{jcc'}) = (\gamma_{jc} - \gamma_{jc'}) \qquad j \in \{1, \ldots, J\},\ c < c' \in \{1, \ldots, C\}^2 \qquad (2.15)$$

  Nous allons choisir le vecteur γ et nous en déduirons F. Choisir γ revient à choisir le nombre de composants non nulles de γ, noté k_0, leurs valeurs et leurs localisations (c'est-à-dire les valeurs de j et c correspondant aux composantes non nulles). Nous considérons pour chaque couple (J, C) les valeurs suivantes pour k_0 :

$$k_0 = 0, 1, [\frac{JC}{100}], [\frac{JC}{40}]$$

  où [.] correspond à la partie entière. Ces valeurs ont été choisies de manière à avoir au maximum 8 à 10% de valeurs non nulles pour F.
  Pour chaque valeur k_0 non nulle, nous simulons un vecteur γ avec k_0 composantes non nulles choisies aléatoirement, et prises constantes égales à γ_0. Nous obtenons un vecteur F avec k_{0_F} composantes non nulles par l'équation (2.15). Les valeurs non nulles de F dépendent du nombre k_0 de composantes non nulles pour γ et de leurs

emplacements dans le vecteur γ. Nous avons choisi γ_0 égal à $5/\sqrt{2}$ de manière à ce que le rapport signal/bruit de la plus petite valeur non nulle de F (valeur non nulle de F divisée par l'ecart-type de X) soit égal à 2.5.

- Nous choisissons comme dimension maximale D_{max} pour les modèles m de la collection $\mathcal{M}_n$ la valeur :

$$[\frac{JC}{4}] + [\frac{JC}{8}]$$

En effet avec les valeurs de k_0 choisies, nous avons vérifié que le nombre de composantes non nulle dans le vecteur F est plus petit que $[JC/4]$.

2.5.3 Résultats

Dans cette section, nous analysons les résultats des simulations obtenus pour calculer les constantes optimales c_1^* et c_2^*.

Les graphes des fonctions

$$c_2 \to r_{(J,C)}(c_1, c_2)$$

sont représentés respectivement dans les figures (2.1),(2.2), (2.3), (2.4), (2.5) et (2.6) pour $C = 2, 3, 4, 5, 6, 7$ avec les différentes valeurs de J et de c_1 définis dans la Section 2.5.2.

Pour chaque valeur de J et de C, nous évaluons le minimum $c_2^*(J, C, c_1)$ des fonctions $r_{(J,C)}(c_1, c_2)$ à c_1 fixée. Pour chaque valeur de C, nous prenons pour c_1^* la valeur qui stabilise $c_2^*(J, C, c_1)$ en J, et obtenons ainsi une constante optimale $c_1^*(C)$ et la valeur optimale $c_2^*(C) = c_2^*(J, C, c_1^*(C))$ associée.

Nous obtenons : $c_1^*(C) = 4$ pour $C = 2$, et $c_1^*(C) = 3$ pour les autres valeurs de C.

Cependant, dans le cas $C = 2$, pour $c_1 = 3$ le minimum $c_2^*(J, 2, 3)$ semble également stable quand J est grand, et sa valeur vaut $c_2^*(J, 2, 3) = 2.1$

Nous choisissons alors une constante optimale c_1^* indépendante de C et égale à 3.
Voici les valeurs optimales $c_2^*(C)$ associées :

- $C = 2$ $\quad c_2^*(2) = 2.1$
- $C = 3$ $\quad c_2^*(3) = 1.4$
- $C = 4$ $\quad c_2^*(4) = 1.1$
- $C = 5$ $\quad c_2^*(5) = 0.8$
- $C = 6$ $\quad c_2^*(6) = 0.7$
- $C = 7$ $\quad c_2^*(7) = 0.6$

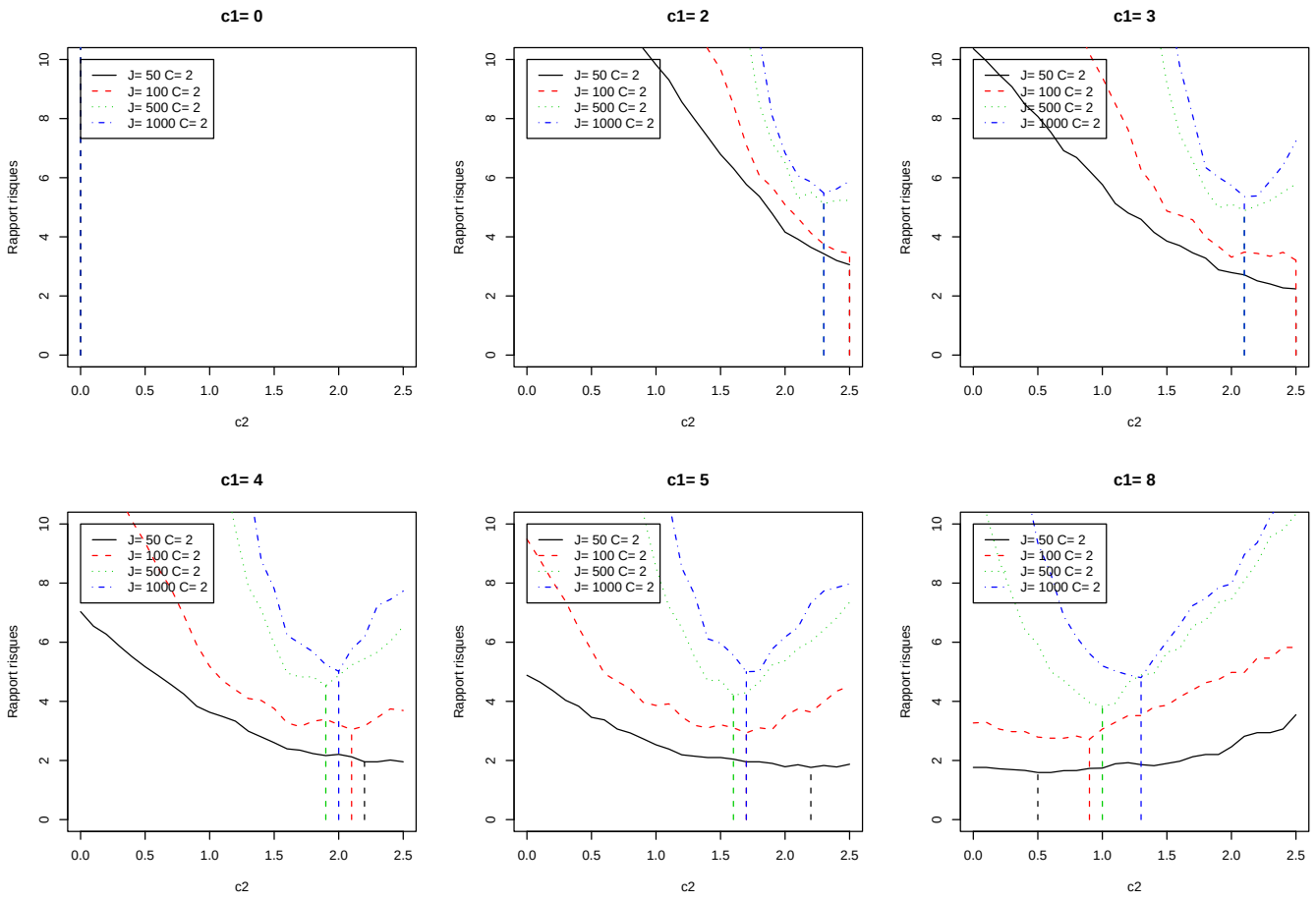

FIG. 2.1 – graphe de la fonction $c_2 \to r_{(J,C)}(c_1, c_2)$ pour C=2. Le choix de $c_1 = 4$ semble stabiliser au mieux le minimum $c_2^*(J, 2, 4)$ pour toutes les valeurs de J, et la valeur $c_2^*(J, 2, 4)$ est proche de 2 (comprise entre 1.9 et 2.2)

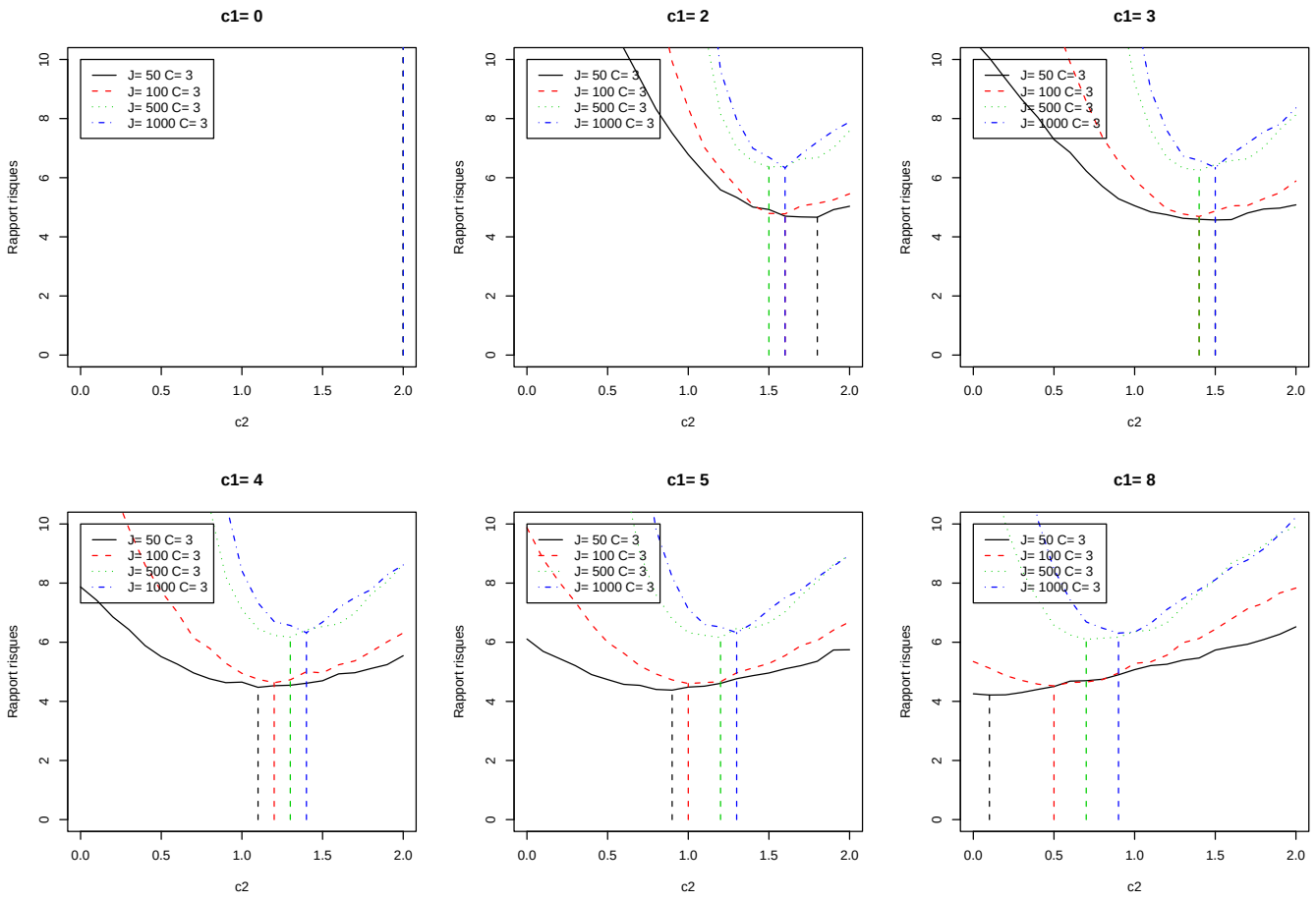

FIG. 2.2 – graphe de la fonction $c_2 \to r_{(J,C)}(c_1, c_2)$ pour C=3. Le choix de $c_1 = 3$ semble stabiliser au mieux le minimum $c_2^*(J, 3, 3)$ pour toutes les valeurs de J, et la valeur $c_2^*(J, 3, 3)$ est proche de 1.4 (égale à 1.4 ou 1.5)

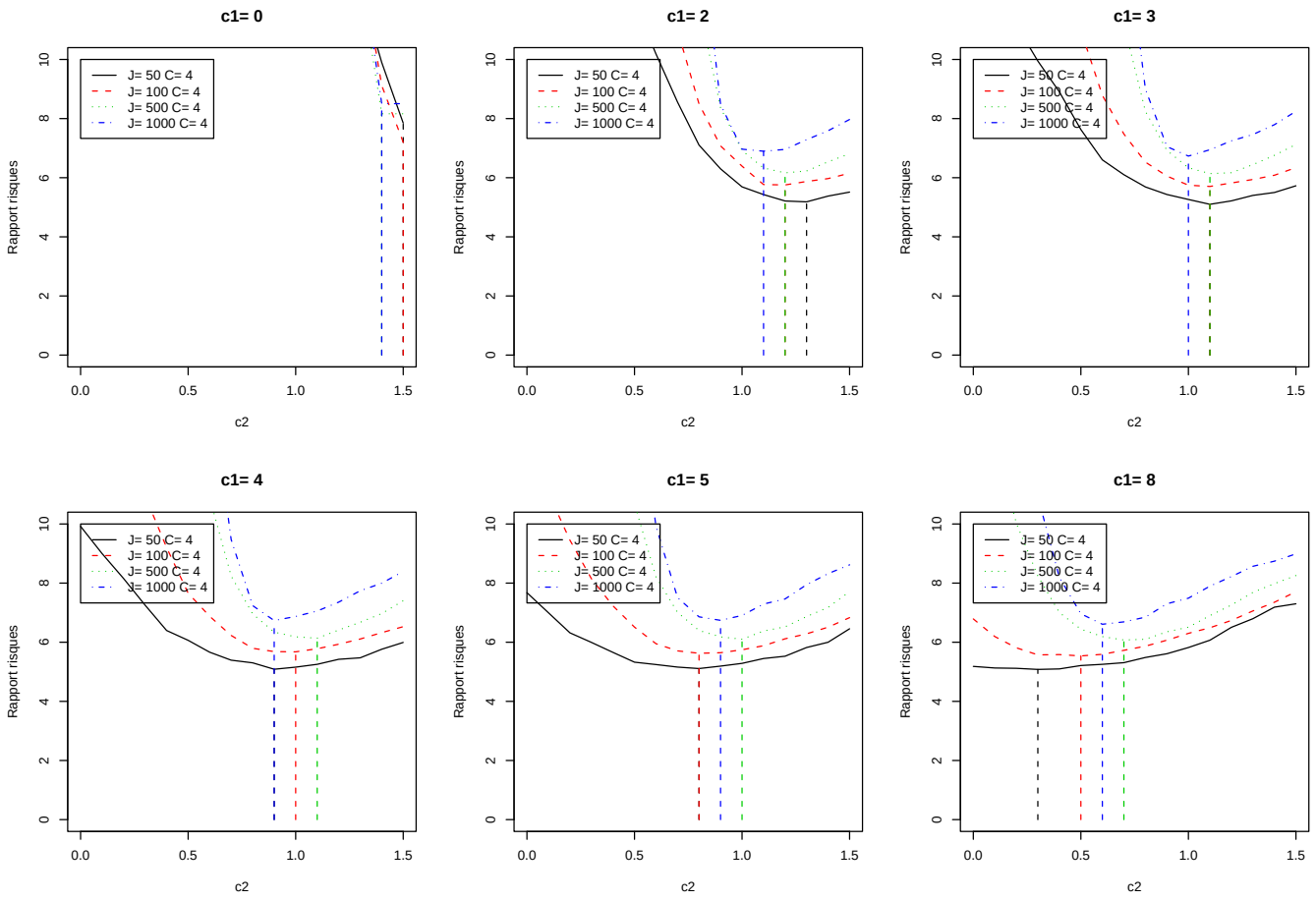

FIG. 2.3 – graphe de la fonction $c_2 \to r_{(J,C)}(c_1, c_2)$ pour C=4. Le choix de $c_1 = 3$ semble stabiliser au mieux le minimum $c_2^*(J, 4, 3)$ pour toutes les valeurs de J, et la valeur $c_2^*(J, 4, 3)$ est proche de 1.1 (égale à 1 ou 1.1)

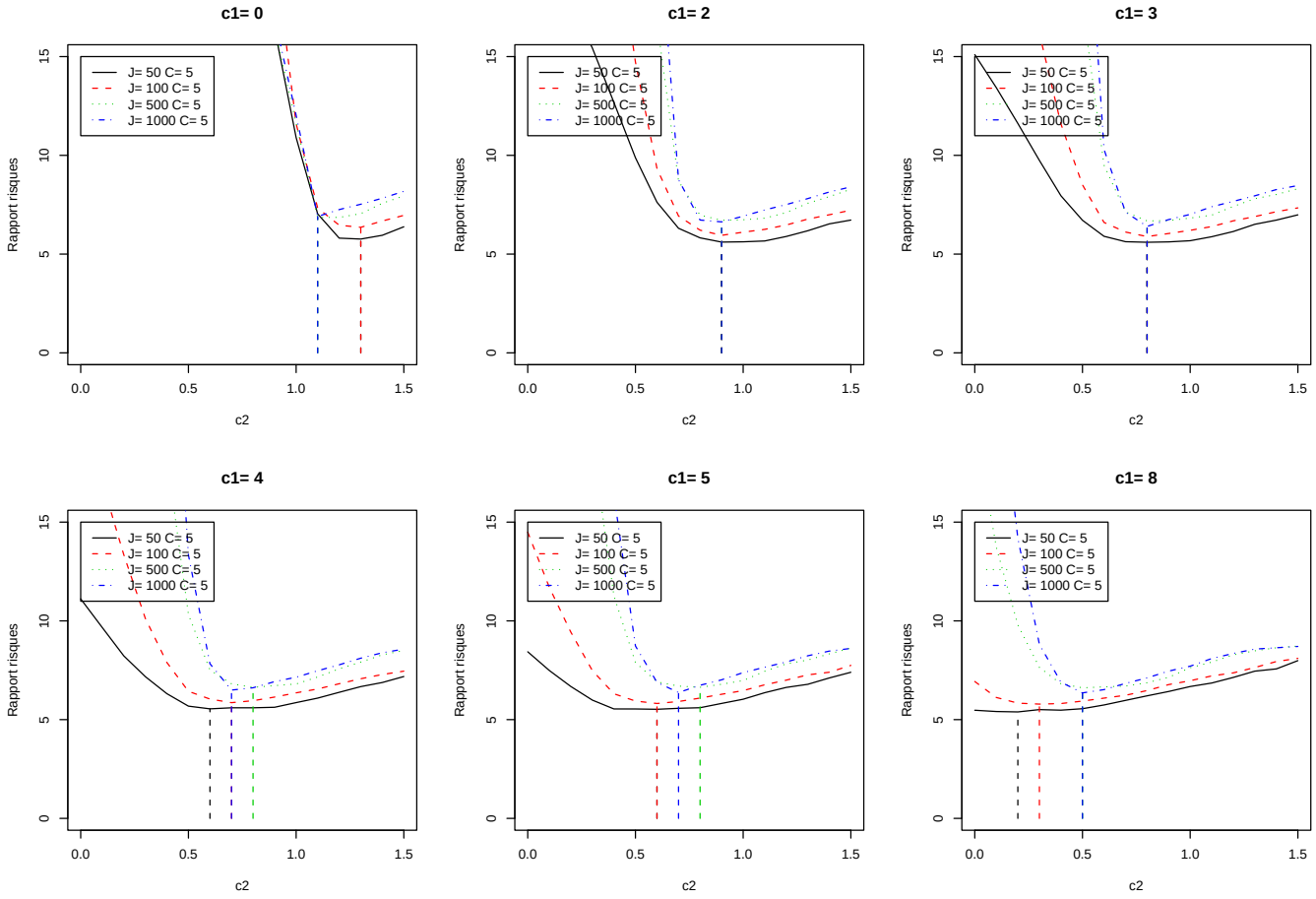

FIG. 2.4 – graphe de la fonction $c_2 \to r_{(J,C)}(c_1, c_2)$ pour C=5. Le choix de $c_1 = 3$ semble stabiliser au mieux le minimum $c_2^*(J, 5, 3)$ pour toutes les valeurs de J, et la valeur $c_2^*(J, 5, 3)$ est proche de 0.8

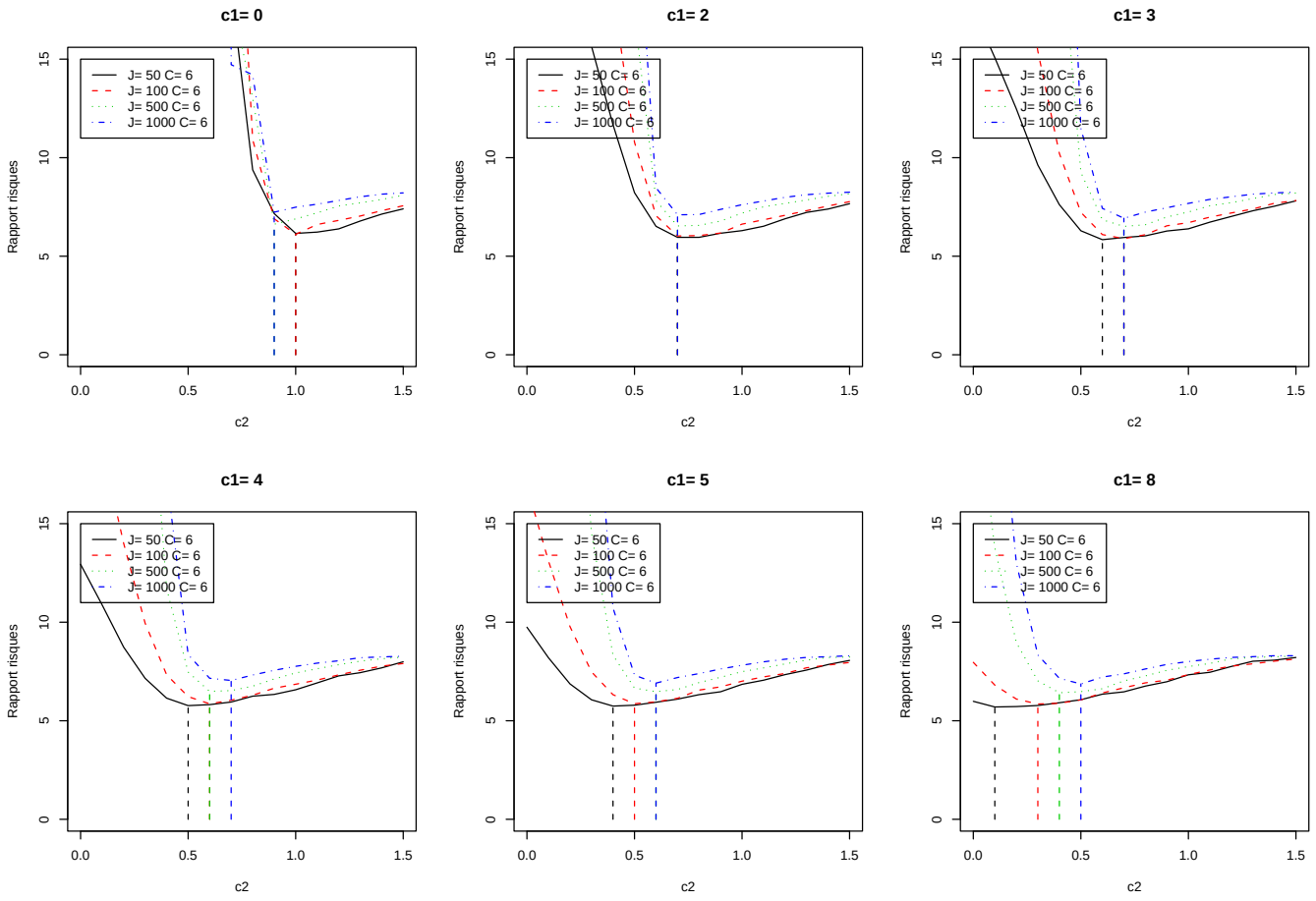

FIG. 2.5 – graphe de la fonction $c_2 \to r_{(J,C)}(c_1, c_2)$ pour C=6. Le choix de $c_1 = 3$ ou 2 semble stabiliser au mieux le minimum $c_2^*(J, 6, 3)$ pour toutes les valeurs de J, et la valeur $c_2^*(J, 6, 3)$ est proche de 0.7

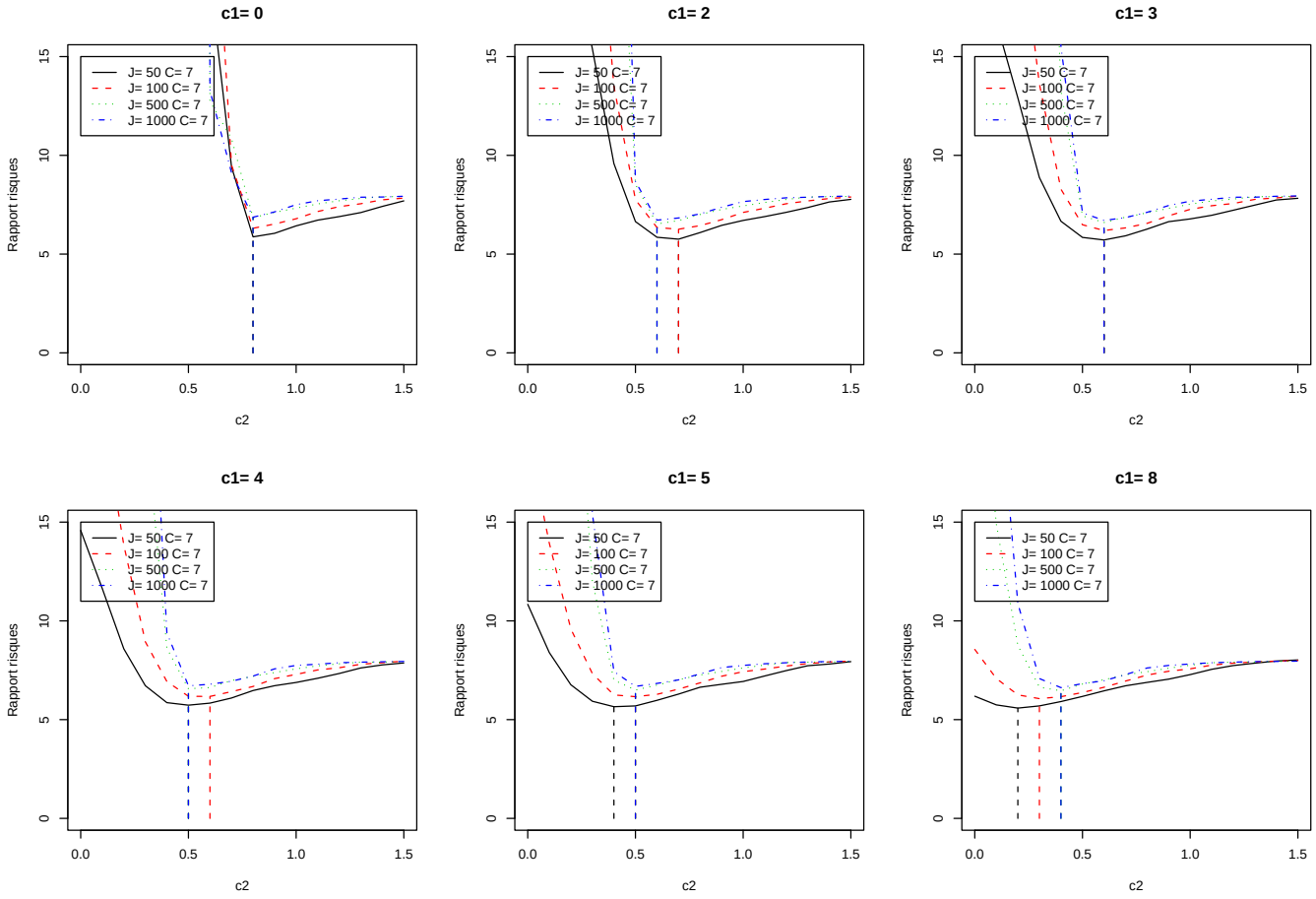

FIG. 2.6 – graphe de la fonction $c_2 \to r_{(J,C)}(c_1, c_2)$ pour C=7. Le choix de $c_1 = 3$ semble stabiliser au mieux le minimum $c_2^*(J, 7, 3)$ pour toutes les valeurs de J, et la valeur $c_2^*(J, 7, 3)$ est proche de 0.6

La valeur $c_2^*(C)$ n'est donc pas indépendante de la valeur de C. Par contre nous constatons que le produit de C et de $c_2^*(C)$ reste à peu près constant et égal à 4.2 pour toutes les valeurs de C

Nous choisissons une fonction de pénalité de la forme :

$$pen(m) = \frac{\sigma^2}{Gn}\left\{c_1^*(2-\frac{2}{J}) + c_2^*(C)CL_m\right\}D_m$$

où $n = JC(C-1)/2$, les poids $L_m = \log(n/D_m)$ et comme constantes optimales $c_1^* = 3$ et $c_2^*(C)$ telle que pour tout C, le produit $c_2^*(C)C$ reste constant égal à 4.2. La constante optimale $c_2^*(C)$ est donc choisie égale à $4.2/C$ pour chaque valeur de C.

2.6 Utilisation d'une méthode heuristique

Dans la Section 2.5, nous avons supposé que la variance du bruit était connue afin de calibrer les constantes de la pénalité de façon optimale. Nous supposons maintenant que cette variance est inconnue. Plutôt que d'utiliser un estimateur de la variance, nous considérons la variance comme une constante, et nous cherchons à estimer la fonction de pénalité à partir des données. Nous utilisons une méthode heuristique proposée par Birgé et Massart [5] et qui a été mise en oeuvre par Lebarbier [8] pour la détection de ruptures dans un modèle de régression.

2.6.1 L'heuristique de pente

Nous écrivons la pénalité sous la forme générale suivante : pour $m \in \mathcal{M}_n$,

$$pen_\alpha(m) = \alpha f_n(D_m)$$

où

$$f_n(D) = \frac{D}{n}\left\{\frac{c_1^*}{c_2^*(C)C}(2-\frac{2}{J}) + Log(\frac{n}{D})\right\}$$

avec comme constantes optimales, $c_1^* = 3$ et $c_2^*(C)$ telle que pour tout C, le produit $c_2^*(C)C$ reste constant égal à 4.2.

Le critère pénalisé est alors défini par :

$$crit_\alpha(m) = \gamma_n(\hat{F}_m) + \alpha f_n(D_m)$$

L'idée de l'heuristique de pente est que si l'on ne pénalise pas assez on choisit systématiquement des modèles de grandes dimensions. La pénalité minimale correspond à la plus petite pénalité avec laquelle le modèle sélectionné est de dimension raisonnable. Pour estimer la pénalité minimale à partir des données, l'idée est de faire varier la constante de pénalité α par petit pas en partant de 0, et de sélectionner pour chaque valeur de α le modèle $\hat{m}(\alpha)$ qui minimise le critère pénalisé. On repère ensuite $\hat{\alpha}$ tel que la dimension

$D_{\hat{m}(\alpha)}$ est grande si $\alpha < \hat{\alpha}$ et raisonnable si $\alpha \geq \hat{\alpha}$. En pratique, les dimensions des modèles $D_{\hat{m}(\alpha)}$ restent assez élevées, puis chutent brusquement vers une petite valeur quand α atteint un certain seuil $\hat{\alpha}$. La pénalité minimale est alors définie par :

$$\hat{\alpha} f_n(D_m).$$

La pénalité optimale est alors choisie, comme conseillé par Birgé et Massart, comme valant deux fois la pénalité minimale. Le critère pénalisé est donc défini par :

$$crit(m) = \gamma_n(\hat{F}_m) + 2\hat{\alpha} f_n(D_m)$$

et l'estimateur final est $\tilde{F} = \hat{F}_{\hat{m}}$ où $\hat{m}$ minimise le critère pénalisée parmi tous les modèles m de $\mathcal{M}_n$.

En pratique, nous faisons varier α par petit pas en partant de 0 et traçons les fonctions : $\alpha \to D_{\hat{m}(\alpha)}$. La valeur $\hat{\alpha}$ est associée au plus grand saut de dimensions.

Si le saut maximal de dimension est atteint pour plusieurs valeurs α, nous prenons pour $\hat{\alpha}$ la plus petite valeur de α, comme l'a proposé Lebarbier dans [8]. Nous nous intéressons en effet au premier moment où il y a un changement brusque de dimensions.

2.6.2 Application

Dans cette section nous mettons en oeuvre l'heuristique de pente sur des simulations.

2.6.2.1 Paramètres de simulation

Dans l'étude de simulations, nous choisissons les paramètres suivants :

$$J = 100; C = 3; k_0 = \{0, 1, [\frac{JC}{100}], [\frac{JC}{40}]\}; \gamma_0 = \{5/\sqrt{2}, 5\sqrt{2}\}$$

Pour chaque valeur k_0 non nulle, nous simulons un vecteur γ avec k_0 composantes non nulles, choisies aléatoirement, égales à γ_0. Nous avons considéré deux valeurs pour γ_0 de manière à ce que le rapport signal/bruit de la plus petite valeur non nulle de F soit égal à 2.5 quand $\gamma_0 = 5/\sqrt{2}$ et à 5 quand $\gamma_0 = 5\sqrt{2}$. Pour chaque valeur de k_0 et de γ_0, les vecteurs F correspondants sont obtenus par l'équation (2.15). Nous obtenons ainsi une collection de vecteurs F. Pour chaque vecteur F, nous simulons 100 vecteurs gaussiens $X^s, s \in \{1, \ldots, 100\}$ d'espérance F et de variance $P_n P_n'$.

Nous faisons varier α à partir de 0 par pas de 0.1. Pour chaque échantillon X^s, nous calculons $\hat{m}^s(\alpha)$ qui minimise parmi tous les modèles m de $\mathcal{M} = \{m \in \{1, \ldots, n\}, D_m \leq D_{max}\}$ le critère de vraisemblance pénalisée :

$$crit_\alpha(m) = \gamma_n(\hat{F}_m) + \alpha f_n(D_m) = ||X - \hat{F}_m||_n^2 + \alpha f_n(D_m) = ||X - \Pi_m X||_n^2 + \alpha f_n(D_m)$$

Pour chaque simulation, nous repérons la valeur $\hat{\alpha}^s$ associé au plus grand saut de dimension.

Dans la Section 2.6.2.2 nous illustrons ce saut de dimension pour une simulation, et montrons l'influence du choix de D_{max} sur l'estimation de α et donc de l'estimateur final. Pour cela nous présentons les résultats pour une grande valeur de $D_{max} = [JC/5]$. Avec les valeurs de k_0 choisies et pour $J = 100$ et $C = 3$, le nombre de composantes non nulles dans le vecteur F est plus petit que $[JC/10]$. Ainsi il suffit de considérer des modèles de dimension au plus $[JC/10]$. En choisissant $D_{max} = [JC/5]$, notre objectif est de montrer les conséquences du choix d'une trop grande valeur pour D_{max}.

L'estimateur final est $\tilde{F}^s = \hat{F}_{\hat{m}^s(2\hat{\alpha}^s)}$

Dans la Section 2.6.2.3, nous évaluons les performances de l'estimateur $\tilde{F}$, en présentant, pour chaque valeur de k_0 et de γ_0 :

1. le rapport entre le risque associé à l'estimateur $\tilde{F}$ et l'oracle, le risque associé à l'estimateur $\tilde{F}$ étant estimé par simulations par la méthode de Monte-Carlo. Nous renvoyons pour plus de détails à la Section 2.5.1.
2. la moyenne sur les 100 simulations du taux de faux positifs défini par :

$$\frac{\sum_{i\in\hat{m}(2\hat{\alpha})} 1_{F_i=0}}{D_{\hat{m}(2\hat{\alpha})}} 1_{D_{\hat{m}(2\hat{\alpha})}>0}$$

3. la distribution des dimensions $D_{\hat{m}(2\hat{\alpha})}$ pour les 100 simulations.

2.6.2.2 Choix de la dimension maximale

Dans la figure 2.7 nous présentons pour une simulation, le graphe de $\alpha \to D_{\hat{m}(\alpha)}$ pour $k_0 = 1$, $\gamma_0 = 5\sqrt{2}$ et $D_{max} = [JC/5]$. La valeur $k_0 = 1$ correspond à $k_{0_F} = 6$ composantes non nulles dans le vecteur F.

Sur le graphe de la figure 2.7, nous observons plusieurs sauts de grande dimension. Le plus grand saut de dimensions va être différent selon la valeur de D_{max} que nous choisissons. Si nous prenons $D_{max} = [JC/5] = 60$ alors $\hat{\alpha} = 1.6$ et le modèle sélectionné est de dimension 9, tandis que si $D_{max} = [JC/10] = 30$ alors $\hat{\alpha} = 2$ et le modèle sélectionné est de dimension 6. En fait, lorsque $D_{max} = [JC/5]$, trois composantes sont détectées à tort comme non nulles.

Nous retrouvons donc ce qu'avait mis en évidence Lebarbier dans le cadre de détection de rupture dans un modèle de régression : le choix de D_{max} peut jouer un rôle dans l'estimation de α et donc dans la sélection de l'estimateur pénalisé. Si l'on choisit D_{max} trop grand, des sauts maximaux sont observés pour des valeurs de α petites. La pénalité sera alors trop petite pour pénaliser correctement le critère et un modèle de trop grande dimension sera sélectionné.

Remarque : Même avec une grande pénalité, nous ne sélectionnons pas un modèle de dimension nulle. Cela vient du fait que la valeur de γ_0 est élevée et que deux des composantes de F ont un rapport signal/bruit égal à 10.

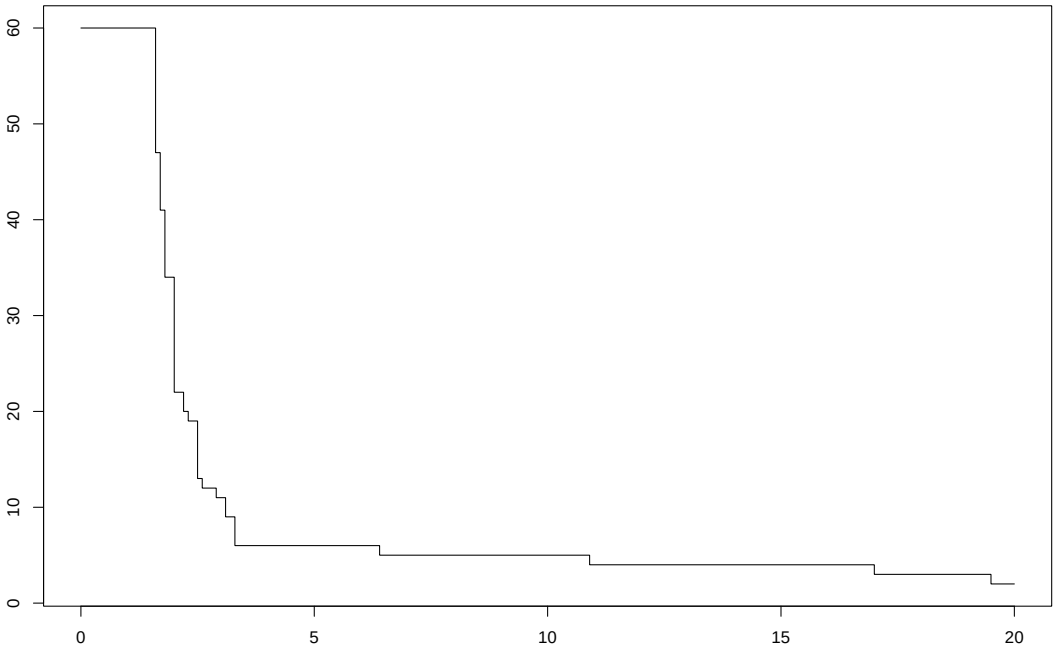

FIG. 2.7 – graphe de $\alpha \to D_{\hat{m}(\alpha)}$ pour $k_0 = 1$, $\gamma_0 = 5\sqrt{2}$ et $D_{max} = [JC/5]$

2.6.2.3 Performance de l'estimateur

Dans les Tables 2.1, 2.2 et 2.3 nous présentons pour chaque valeur de k_0, le nombre correspondant k_{0_F} de composantes non nulles pour le vecteur F, les rapports entre le risque de l'estimateur sélectionné et l'oracle, la moyenne du taux de faux positifs et la distribution de $D_{\hat{m}(2\hat{\alpha})}$.

Dans la Table 2.1 sont présentés les résutats pour $\gamma_0 = 5/\sqrt{2}$ et $D_{max} = JC/5$.

Lorsque $k_{0_F} = 0$, la méthode détecte très peu de faux positifs.

Lorsque $k_{0_F} > 0$, la méthode a tendance à sous estimer le nombre de composantes non nulles de F. En fait, lorsque $\gamma_0 = 5/\sqrt{2}$, le rapport signal/bruit d'un grand nombre de composantes de F est faible, de l'ordre de 2.5. Ces valeurs sont du même ordre de grandeur que le bruit et la méthode a des difficultés à les détecter. Par contre les valeurs qui ont un rapport signal/bruit plus grand sont bien détectées.

De plus le taux de faux positifs est élevé. Sur plusieurs simulations où un modèle de trop grande dimension a été sélectionnée, la méthode détecte des composantes non nulles à tort. Nous avons vu dans la section précédente que si l'on choisit D_{max} trop grand, des sauts maximaux de dimension peuvent être observés pour des valeurs de α trop petites et entrainer la sélection d'un modèle de dimension trop grande. C'est ce qui se passe sur ces

simulations, comme nous le confirment les résultats de la table (2.1), obtenus en diminuant la valeur de D_{max}.

Dans la Table 2.2 nous présentons les résutats pour $\gamma_0 = 5/\sqrt{2}$ et D_{max} égal à $JC/10$.

Nous observons à nouveau que la méthode sous-estime le nombre de composantes non nulles k_{0_F}. Les taux de faux positifs sont faibles dans la Table 2.2, ce qui nous confirme que les taux de faux positifs élevés dans la Table 2.1, sont dus au choix d'une valeur D_{max} trop grande.

Pour vérifier que la méthode sous-estime k_{0_F} car ne détecte pas les valeurs ayant un faible rapport signal buit, nous présentons dans la Table 2.3 les résutats pour $\gamma_0 = 5\sqrt{2}$. Le rapport signal/bruit de la plus petite valeur non nulle de F est alors égale à 5. La valeur D_{max} est choisi égale à $JC/10$ d'après les résultats précédents.

Commentons tout d'abord les résultats pour k_{0_F} égal à $0, 6$ et 12. Les rapports des risques entre l'estimateur sélectionné et l'oracle sont plus petits dans la Table 2.3 que dans la Table 2.2. De plus les taux de faux positifs sont faibles et les distributions de $D_{\hat{m}(2\hat{\alpha})}$ sont très proches du nombre k_{0_F} de composantes non nulles de F. La méthode est donc satisfaisante.

Commentons maintenant les résultats pour $k_{0_F} = 24$. Le rapport des risques est beaucoup plus grand que pour $k_{0_F} = 0, 6$ ou 12, et la méthode sous-estime le nombre de composantes non nulles de F. Si nous choisissons D_{max} plus grand, égal à $[JC/5]$, alors le rapport des risques diminue, vaut 1.67 et le taux de faux positifs vaut 6%, ce qui est plus satisfaisant.

La méthode est donc satisfaisante lorsque le pourcentage de valeurs non nulle est faible. La difficulté en pratique reste de choisir la dimension maximale D_{max}.

2.6.2.4 Comparaison par simulations avec la méthode de sélection de modèle à variance connue

Dans certaines applications, la variance de l'erreur peut être estimée avec précision et donc supposée connue. Il est donc intéressant d'évaluer la performance de l'estimateur proposé dans le cas où la variance est connue.

Nous comparons par simulations les performances de la méthode de sélection de modèles avec la pénalité calibrée dans la Section 2.5, lorsque α est connue et lorsque α est estimée avec l'heuristique de pente. La procédure de simulations est la même que celle présentée en Section 2.6.2.1. Nous évaluons pour chacune des deux méthodes :

1. le rapport entre le risque associé à l'estimateur pénalisé $\tilde{F} = \hat{F}_{\hat{m}}$ et l'oracle.
2. la moyenne sur les 100 simulations du taux de faux positifs
3. la distribution des dimensions $D_{\hat{m}^s}$ pour les 100 simulations.

Dans la Table 2.4, nous présentons les résultats de la méthode à variance connue pour $\gamma_0 = 5\sqrt{2}$. Les résultats pour $\gamma_0 = 5\sqrt{2}$ lorsque α est estimée à partir des données sont présentés dans la Table 2.3.

La méthode de sélection de modèles à variance connue a tendance à surestimer le nombre de composantes non nulles de F, et le taux de faux positifs est élevé. Ainsi, dans le

k_0	0	1	3	7
k_{0_F}	0	6	12	24
$\mathcal{R}_{r/o}$	0.002*	5.06	4.05	3.68
t_f	0.020	0.101	0.096	0.048

Distribution de $D_{\hat{m}(2\hat{\alpha})}$ pour $k_{0_F} = 0$

$D_{\hat{m}(2\hat{\alpha})}$	$= 0$	> 0
	98	2

Distribution de $D_{\hat{m}(2\hat{\alpha})}$ pour $k_{0_F} = 6$

$D_{\hat{m}(2\hat{\alpha})}$	< 2	$\in [2, 4]$	$\in [5, 7]$	> 7
	6	78	12	4

Distribution de $D_{\hat{m}(2\hat{\alpha})}$ pour $k_{0_F} = 12$

$D_{\hat{m}(2\hat{\alpha})}$	< 5	$\in [5, 10]$	$\in [11, 13]$	> 13
	3	76	16	5

Distribution de $D_{\hat{m}(2\hat{\alpha})}$ pour $k_{0_F} = 24$

$D_{\hat{m}(2\hat{\alpha})}$	< 10	$\in [10, 15]$	$\in [16, 20]$	> 20
	2	56	40	2

TAB. 2.1 – Cas $\gamma_0 = 5/\sqrt{2}$ et $D_{max} = JC/5$. Dans le premier tableau, les deux premières lignes donnent le nombre de composantes non nulles respectivement dans les vecteurs γ et F ; la troisième ligne correspond au risque de l'estimateur sélectionné lorsque $k_0 = 0$ (*) et au rapport entre le risque de l'estimateur sélectionné et l'oracle lorsque $k_0 > 0$; la dernière ligne correspond à la moyenne du taux de faux positifs. Les 4 tableaux suivants donnent la distribution de $D_{\hat{m}(2\hat{\alpha})}$ pour les différentes valeurs de k_{0_F}.

k_0	0	1	3	7
k_{0_F}	0	6	12	24
$\mathcal{R}_{r/o}$	0.00*	4.50	4.01	4.91
t_f	0.000	0.011	0.012	0.003

Distribution de $D_{\hat{m}(2\hat{\alpha})}$ pour $k_{0_F} = 0$

$D_{\hat{m}(2\hat{\alpha})}$	$= 0$	> 0
	100	0

Distribution de $D_{\hat{m}(2\hat{\alpha})}$ pour $k_{0_F} = 6$

$D_{\hat{m}(2\hat{\alpha})}$	< 2	$\in [2, 4]$	$\in [5, 7]$	> 7
	13	85	2	0

Distribution de $D_{\hat{m}(2\hat{\alpha})}$ pour $k_{0_F} = 12$

$D_{\hat{m}(2\hat{\alpha})}$	< 5	$\in [5, 10]$	$\in [11, 13]$	> 13
	5	95	0	0

Distribution de $D_{\hat{m}(2\hat{\alpha})}$ pour $k_{0_F} = 24$

$D_{\hat{m}(2\hat{\alpha})}$	< 10	$\in [10, 15]$	$\in [16, 20]$	> 20
	23	76	1	0

TAB. 2.2 – Cas $\gamma_0 = 5/\sqrt{2}$ et $D_{max} = JC/10$. Dans le premier tableau, les deux premières lignes donnent le nombre de composantes non nulles respectivement dans les vecteurs γ et F ; la troisième ligne correspond au risque de l'estimateur sélectionné lorsque $k_0 = 0$ (*) et au rapport entre le risque de l'estimateur sélectionné et l'oracle lorsque $k_0 > 0$; la dernière ligne correspond à la moyenne du taux de faux positifs. Les 4 tableaux suivants donnent la distribution de $D_{\hat{m}(2\hat{\alpha})}$ pour les différentes valeurs de k_{0_F}.

k_0	0	1	3	7
k_{0_F}	0	6	12	24
$\mathcal{R}_{r/o}$	0.00*	1.98	1.66	7.74
t_f	0.000	0.027	0.022	0.002

Distribution de $D_{\hat{m}(2\hat{\alpha})}$ pour $k_{0_F} = 0$

$D_{\hat{m}(2\hat{\alpha})}$	$= 0$	> 0
	100	0

Distribution de $D_{\hat{m}(2\hat{\alpha})}$ pour $k_{0_F} = 6$

$D_{\hat{m}(2\hat{\alpha})}$	< 5	$= 5$	$= 6$	$= 7$	> 7
	1	16	68	12	3

Distribution de $D_{\hat{m}(2\hat{\alpha})}$ pour $k_{0_F} = 12$

$D_{\hat{m}(2\hat{\alpha})}$	< 11	$= 11$	$= 12$	$= 13$	> 13
	2	18	62	13	5

Distribution de $D_{\hat{m}(2\hat{\alpha})}$ pour $k_{0_F} = 24$

$D_{\hat{m}(2\hat{\alpha})}$	< 23	$= 23$	$= 24$	$= 25$	> 25
	59	19	19	3	0

TAB. 2.3 – Cas $\gamma_0 = 5\sqrt{2}$ et $D_{max} = JC/10$. Dans le premier tableau, les deux premières lignes donnent le nombre de composantes non nulles respectivement dans les vecteurs γ et F ; la troisième ligne correspond au risque de l'estimateur sélectionné lorsque $k_0 = 0$ (*) et au rapport entre le risque de l'estimateur sélectionné et l'oracle lorsque $k_0 > 0$; la dernière ligne correspond à la moyenne du taux de faux positifs. Les 4 tableaux suivants donnent la distribution de $D_{\hat{m}(2\hat{\alpha})}$ pour les différentes valeurs de k_{0_F}.

k_0	0	1	3	7
k_{0_F}	0	6	12	24
$\mathcal{R}_{r/o}$	0.003*	2.53	2.27	2.12
t_f	0.010	0.088	0.116	0.123

Distribution de $D_{\hat{m}(2\hat{\alpha})}$ pour $k_{0_F} = 0$

$D_{\hat{m}(2\hat{\alpha})}$	$= 0$	> 0
	99	1

Distribution de $D_{\hat{m}(2\hat{\alpha})}$ pour $k_{0_F} = 6$

$D_{\hat{m}(2\hat{\alpha})}$	< 5	$= 5$	$= 6$	$= 7$	> 7
	1	7	54	18	20

Distribution de $D_{\hat{m}(2\hat{\alpha})}$ pour $k_{0_F} = 12$

$D_{\hat{m}(2\hat{\alpha})}$	< 11	$= 11$	$= 12$	$= 13$	> 13
	0	1	27	28	44

Distribution de $D_{\hat{m}(2\hat{\alpha})}$ pour $k_{0_F} = 24$

$D_{\hat{m}(2\hat{\alpha})}$	< 23	$= 23$	$= 24$	$= 25$	> 25
	0	0	8	12	80

TAB. 2.4 – Méthode de sélection avec $\sigma^2 = 1$ connue. Cas $\gamma_0 = 5\sqrt{2}$ et $D_{max} = JC/10$. Dans le premier tableau, les deux premières lignes donnent le nombre de composantes non nulles respectivement dans les vecteurs γ et F ; la troisième ligne correspond au risque de l'estimateur sélectionné lorsque $k_0 = 0$ (*) et au rapport entre le risque de l'estimateur sélectionné et l'oracle lorsque $k_0 > 0$; la dernière ligne correspond à la moyenne du taux de faux positifs. Les 4 tableaux suivants donnent la distribution de $D_{\hat{m}(2\hat{\alpha})}$ pour les différentes valeurs de k_{0_F}.

cas que nous avons considéré où le rapport signal/bruit est élevé, la pénalité s'avère trop faible pour pénaliser correctement le critère et nous sélectionnons des modèles de trop grande dimension.

Pour $k_0 = 0,6$ et 12, les rapports de risque sont plus petits avec la méthode où la variance est estimée à partir des données. Par contre pour $k_0 = 24$, la méthode à variance connue a un plus petit risque. Lorsque le rapport signal/bruit est grand, il est donc plus pénalisant pour le risque d'oublier des vrais positifs que de détecter des faux positifs.

A partir de cette comparaison, nous pouvons faire trois remarques. Tout d'abord, la pénalité proposée dans la Section 2.5 permet de contrôler un risque. Elle n'a pas pour but de contrôler le taux de faux positifs, ce qui explique qu'il peut éventuellement être grand. De plus, en prenant deux fois la pénalité minimale, on choisit en général une pénalité plus grande que lorsque la variance est connue. Finalement, l'heuristique de pente permet d'estimer σ^2, mais elle semble aussi permettre si besoin d'ajuster les constantes c_1 et c_2. Les constantes c_1^* et c_2^* ont en effet été choisies de façon optimales pour différentes valeurs de J, C et F, mais il se peut que pour un couple (J, C) et un vecteur F particulier, ces constantes ne soient pas les plus adéquates même si elles sont proches des optimales.

2.7 Preuve du théorème 1

Pour alléger les calculs, on écrit $X = F + P_n\epsilon$ avec $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.
On notera aussi $< . >_n$ le produit scalaire associé à $||.||_n$. Dans cette preuve, nous utilisons plusieurs fois l'inégalité suivante :

$$\forall \alpha, y, z > 0, \quad 2yz \leq \alpha y^2 + \frac{1}{\alpha} z^2. \tag{2.16}$$

Par définition de γ_n, pour tout $T \in \mathbb{R}_n$,

$$||F - T||_n^2 = \gamma_n(T) + 2 < T - X, P_n\epsilon >_n + ||P_n\epsilon||_n^2,$$

d'où pour tout $m \in \mathcal{M}_n$,

$$||F - \tilde{F}||_n^2 - ||F - \Pi_m F||_n^2 = \gamma_n(\tilde{F}) - \gamma_n(\Pi_m F) + 2 < \tilde{F} - \Pi_m F, P_n\epsilon >_n . \tag{2.17}$$

En remarquant que $\gamma_n(\Pi_m F) = ||X - \Pi_m F||_n^2 = \gamma_n(\hat{F}_m) + ||\Pi_m P_n\epsilon||_n^2$, et en l'introduisant dans l'équation(2.17) , nons obtenons :

$$||F - \tilde{F}||_n^2 \leq ||F - \Pi_m F||_n^2 + 2 < \tilde{F} - \Pi_m F, P_n\epsilon >_n + pen(m) - pen(\hat{m}) - ||\Pi_m P_n\epsilon||_n^2$$

puisque par définition de $\tilde{F}$, $\gamma_n(\tilde{F}) + pen(\hat{m}) \leq \gamma_n(\hat{F}_m) + pen(m)$.

Comme $\tilde{F} = \Pi_{\hat{m}} X = \Pi_{\hat{m}} F + \Pi_{\hat{m}} P_n\epsilon$, cette équation devient :

$$||F - \tilde{F}||_n^2 \leq ||F - \Pi_m F||_n^2 + 2 < \Pi_{\hat{m}} F - F, P_n\epsilon >_n + 2 < F - \Pi_m F, P_n\epsilon >_n$$

$$+2||\Pi_{\hat{m}} P_n\epsilon||_n^2 + pen(m) - pen(\hat{m}) - ||\Pi_m P_n\epsilon||_n^2. \tag{2.18}$$

Posons :

$$u_m = \begin{cases} \frac{\Pi_m F - F}{||\Pi_m F - F||_n} & \text{si } \Pi_m F \neq F, \\ 0 & \text{sinon.} \end{cases}$$

Soit $\alpha \in]0,1[$ un nombre qui sera précisé plus tard.
D'après l'inégalité (2.16), nous avons :

$$\begin{aligned} 2| < \Pi_{\hat{m}} F - F, P_n\epsilon >_n | &= 2||\Pi_{\hat{m}} F - F||_n < u_{\hat{m}}, P_n\epsilon >_n \\ &\leq \alpha||\Pi_{\hat{m}} F - F||_n^2 + \frac{1}{\alpha} < u_{\hat{m}}, P_n\epsilon >_n^2 \\ &\leq \alpha||F - \tilde{F}||_n^2 - \alpha||\Pi_{\hat{m}} P_n\epsilon||_n^2 + \frac{1}{\alpha} < u_{\hat{m}}, P_n\epsilon >_n^2, \end{aligned}$$

en utilisant à nouveau que $\tilde{F} = \Pi_{\hat{m}} X = \Pi_{\hat{m}} F + \Pi_{\hat{m}} P_n\epsilon$.

L'inégalité (2.18) devient alors :

$$(1-\alpha)||F - \tilde{F}||_n^2 \leq ||F - \Pi_m F||_n^2 - \alpha||\Pi_{\hat{m}} P_n\epsilon||_n^2 + \frac{1}{\alpha} < u_{\hat{m}}, P_n\epsilon >_n^2$$

$$+2 < F - \Pi_m F, P_n\epsilon >_n +2||\Pi_{\hat{m}} P_n\epsilon||_n^2 + pen(m) - pen(\hat{m}) - ||\Pi_m P_n\epsilon||_n^2.$$

D'où en majorant grossièrement $-||\Pi_m P_n\epsilon||_n^2$ par 0, nous obtenons :

$$(1-\alpha)||F - \tilde{F}||_n^2 \leq ||F - \Pi_m F||_n^2 + pen(m) + 2 < F - \Pi_m F, P_n\epsilon >_n$$

$$+(2-\alpha)||\Pi_{\hat{m}} P_n\epsilon||_n^2 + \frac{1}{\alpha} < u_{\hat{m}}, P_n\epsilon >_n^2 - pen(\hat{m}). \tag{2.19}$$

Maintenant posons pour tout $m \in \mathcal{M}_n$:

$$np_1(m) = \left\{ Tr(\Pi_m P_n P_n') + \beta\frac{S^2(m)}{M} + 2S(m)\sqrt{L_m D m} + 2M(m)L_m D_m \right\} \sigma^2 \tag{2.20}$$

$$np_2(m) = 2\sigma^2 M L_m D_m. \tag{2.21}$$

Alors l'hypothèse sur la pénalité devient :

$$pen(m) \geq (2-\alpha)p_1(m) + \frac{1}{\alpha}p_2(m) \quad \forall m \in \mathcal{M}_n$$

avec α choisi tel que $(2-\alpha)(1+\beta) = 2$.

L'inéquation (2.19) devient alors :

$$(1-\alpha)||F - \tilde{F}||_n^2 \leq ||F - \Pi_m F||_n^2 + pen(m) + 2 < F - \Pi_m F, P_n\epsilon >_n$$

$$+(2-\alpha)\{||\Pi_{\hat{m}} P_n\epsilon||_n^2 - p_1(\hat{m})\}_+$$

$$+\frac{1}{\alpha}\{< u_{\hat{m}}, P_n\epsilon >_n^2 - p_2(\hat{m})\}_+$$

avec $x_+ = \max(x, 0)$.

D'où, comme $E(< F - \Pi_m F, P_n\epsilon >_n) = 0$, nous en déduisons :

$$E(||F - \tilde{F}||_n^2) \leq \frac{1}{1-\alpha}[||F - \Pi_m F||_n^2 + pen(m)$$

$$+(2-\alpha)E\left(\{||\Pi_{\hat{m}} P_n\epsilon||_n^2 - p_1(\hat{m})\}_+\right)$$

$$+\frac{1}{\alpha}E\left(\{< u_{\hat{m}}, P_n\epsilon >_n^2 - p_2(\hat{m})\}_+\right)]. \tag{2.22}$$

Nous avons alors deux termes que l'on souhaite contrôler : $E\left(\{||\Pi_{\hat{m}} P_n\epsilon||_n^2 - p_1(\hat{m})\}_+\right)$ et $E\left(\{||\Pi_{\hat{m}} P_n\epsilon||_n^2 - p_1(\hat{m})\}_+\right)$.

1. Contrôle du premier terme

Lemme (obtenu par Laurent et Massart [7]) :
Soit $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ un vecteur aléatoire de $\mathbb{R}^n$, où les ϵ_i sont des variables gaussiennes i.i.d de variance σ^2, et soit A une matrice de $\mathbb{M}_n(\mathbb{R})$. Soit $\lambda_1, \ldots, \lambda_n$ l'ensemble des valeurs propres de la matrice AA'. On note $S^2 = \sum_{i=1}^n \lambda_i^2$ et $M = \max_i(\lambda_i)$. Soit $\chi^2(A) = ||A\epsilon||_n^2$. On a alors pour tout $x > 0$

$$\mathbb{P}(\chi^2(A) \geq Tr(AA')\sigma^2 + 2\sigma^2\sqrt{S^2 x} + 2\sigma^2 M x) \leq e^{-x}.$$

Par définition de $p_1(m)$, donnée par (2.20),

$$\mathbb{P}(||\Pi_{\hat{m}} P_n\epsilon||_n^2 - p_1(\hat{m}) \geq (2 + \frac{1}{\beta})M\frac{\sigma^2}{n}\xi)$$

$$= \mathbb{P}(n||\Pi_{\hat{m}} P_n\epsilon||_n^2 \geq \sigma^2 Tr(\Pi_m P_n P_n') + \sigma^2\beta\frac{S^2(m)}{M} + 2\sigma^2 S(m)\sqrt{L_m D m}$$

$$+2\sigma^2 M(m) L_m D_m + 2M\sigma^2\xi + \frac{1}{\beta}M\sigma^2\xi)$$

$$\leq \mathbb{P}(n||\Pi_{\hat{m}} P_n\epsilon||_n^2 \geq \sigma^2 Tr(\Pi_m P_n P_n') + \sigma^2\beta\frac{S^2(m)}{M} + 2\sigma^2 S(m)\sqrt{L_m D m}$$

$$+2\sigma^2 M(m) L_m D_m + 2M(m)\sigma^2\xi + \frac{1}{\beta}M\sigma^2\xi) \qquad \text{car } M(m) \leq M$$

$$= \mathbb{P}(n||\Pi_{\hat{m}} P_n \epsilon||_n^2 \geq \sigma^2 Tr(\Pi_m P_n P_n') + \sigma^2 \beta \frac{S^2(m)}{M} + 2\sigma^2 S(m)\sqrt{L_m D m}$$

$$+2\sigma^2 M(m)(L_m D_m + \xi) + \frac{1}{\beta} M\sigma^2 \xi). \tag{2.23}$$

Nous obtenons d'après l'inégalité (2.16) :

$$2S(m)\sqrt{L_m D_m + \xi} \leq 2S(m)\sqrt{L_m D_m} + 2S(m)\sqrt{\xi} \leq 2S(m)\sqrt{L_m D_m} + \frac{\beta}{M} S^2(m) + \frac{M}{\beta}\xi. \tag{2.24}$$

D'où en notant $x_m(\xi) = L_m D_m + \xi$ et d'après (2.23) et (2.24) :

$$\mathbb{P}(||\Pi_m P_n \epsilon||_n^2 - p_1(m) \geq (2 + \frac{1}{\beta}) M \frac{\sigma^2}{n} \xi)$$

$$\leq \mathbb{P}(n||\Pi_m P_n \epsilon||_n^2 \geq \sigma^2 Tr(\Pi_m P_n P_n') + 2\sigma^2 S(m)\sqrt{x_m(\xi)} + 2\sigma^2 M(m) x_m(\xi))$$

$$\leq e^{-L_m D_m} e^{-\xi} \qquad \text{d'après le lemme.}$$

En intégrant en ξ nous obtenons :

$$E\left(\{||\Pi_m P_n \epsilon||_n^2 - p_1(m)\}_+\right) \leq (2 + \frac{1}{\beta}) M \frac{\sigma^2}{n} e^{-L_m D_m}. \tag{2.25}$$

2. Contrôle du deuxième terme

Pour tout $m \in \mathcal{M}_n, \sqrt{n} < u_m, P_n \epsilon >_n = \sqrt{n} < P_n' u_m, \epsilon >_n$ suit une loi normale centrée de variance $||P_n' u_m||_n^2$. Pour chaque $t > 0$ la déviation standard d'une variable gaussienne permet d'écrire que

$$\mathbb{P}(\sqrt{n}| < P_n' u_m, \epsilon >_n | \geq t\sigma ||P_n' u_m||_n) \leq e^{-t^2/2}.$$

Donc pour $t^2 = 2x_m(\xi) = 2(L_m D_m + \xi)$

$$\mathbb{P}(n| < P_n' u_m, \epsilon >_n |^2 \geq 2\sigma^2 x_m(\xi) ||P_n' u_m||_n^2) \leq e^{-L_m D_m} e^{-\xi}. \tag{2.26}$$

D'où d'après la définition (2.21) de $p_2(m)$ et comme $||P_n' u_m||_n^2 \leq M$,

$$\mathbb{P}(| < u_m, P_n \epsilon >_n |^2 - p_2(m) \geq 2M \frac{\sigma^2}{n} \xi)$$

$$= \mathbb{P}(n| < P_n' u_m, \epsilon >_n |^2 \geq 2\sigma^2 x_m(\xi) M)$$

$$\leq \mathbb{P}(n| < P_n' u_m, \epsilon >_n |^2 \geq 2\sigma^2 x_m(\xi)||P_n' u_m||_n^2) \qquad \text{car } ||P_n' u_m||_n^2 \leq M$$

$$\leq e^{-L_m D_m} e^{-\xi} \qquad \text{d'après (2.26),}$$

ce qui donne en intégrant en ξ,

$$E\left(\{< u_m, P_n\epsilon >_n^2 - p_2(m)\}_+\right) \leq 2M\frac{\sigma^2}{n}e^{-L_m D_m}. \tag{2.27}$$

3. Conclusion

Comme

$$E\left(\{||\Pi_{\hat{m}} P_n\epsilon||_n^2 - p_1(\hat{m})\}_+\right) \leq \sum_{m\in\mathcal{M}_n} E\left(\{||\Pi_m P_n\epsilon||_n^2 - p_1(m)\}_+\right)$$

et

$$E\left(\{< u_{\hat{m}}, P_n\epsilon >_n^2 - p_2(\hat{m})\}_+\right) \leq \sum_{m\in\mathcal{M}_n} E\left(\{< u_m, P_n\epsilon >_n^2 - p_2(m)\}_+\right),$$

nous obtenons alors d'après (2.22), (2.25) et (2.27) :

$$E(||F - \tilde{F}||_n^2) \leq \frac{1}{1-\alpha}\left(||F - \Pi_m F||_n^2 + pen(m) + (2-\alpha)(2+\frac{1}{\beta})M\frac{\sigma^2}{n}\Sigma_n + \frac{1}{\alpha}2M\frac{\sigma^2}{n}\Sigma_n\right)$$

Comme α est choisi tel que $(2-\alpha)(1+\beta) = 2$, nous obtenons :

$$E(||F - \tilde{F}||_n^2) \leq \frac{1+\beta}{1-\beta}[||F - \Pi_m F||_n^2 + pen(m)] + \frac{3+6\beta+\beta^2}{\beta(1+\beta)}M\frac{\sigma^2}{n}\Sigma_n$$

et le résultat en prenant l'inf sur les sous ensembles $m \in \mathcal{M}_n$.

Références

[1] Benjamini, Y., and Hochberg, Y. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B 57* (1995), 289–300.

[2] Benjamini, Y., and Liu, W. A distribution-free multiple-test procedure that controls the false discovery rate. *Research Paper* (1999).

[3] Benjamini, Y., and Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist. 29(4)* (2001), 1165–1188.

[4] Birgé, L., and Massart, P. Gaussian model selection. *J. Eur. Math. Soc. (JEMS) 3* (2001), 203–268.

[5] Birgé, L., and Massart, P. Minimal penalties for gaussian model selection. *Probab. Theory Related Fields 134* (2006).

[6] Huet, S. Model selection for estimating the non zero components of a Gaussian vector. *ESAIM Probab. Stat. 10* (2006), 164–183.

[7] Laurent, B., and Massart, P. Adaptive estimation of a quatratic functional by model selection. *Annals of Statistics 28* (2000), 1302–1338.

[8] Lebarbier, E. Detecting multiple change-points in the mean of a gaussian process by model selection. *Signal Processing 85* (2005), 717–736.

[9] Y. Baraud, C. Giraud, S. H. Gaussian model selection with unknown variance. *Annals of Statistics* (2007).

Deuxième partie

Estimation de graphes

Chapter 3

Assessing the validity domains of graphical gaussian models in order to infer relationships among components of complex biological systems

Abstract

Gene or protein expression data may be used for inferring relations in gene or protein association networks. Different statistical approaches suitable for the case of a large number of genes, have been recently proposed. We compare their performances for detecting the true relations between the cellular components, on the basis of a wide-ranging simulation study. We then apply these new statistical approaches to a biological data set.

3.1 Introduction

Biological systems involve complex cellular processes built up from physical and functional interactions between molecular entities (genes, proteins, small molecules,...). Thus, to understand how these processes are regulated, it is necessary to study the behavior of the molecular machinery. Recently, biotechnological developments were focused on the characterization and the quantification of cellular system components leading to produce a huge amount of various data. So, one of major challenges is nowadays to understand from these data, how molecular entities interact i.e. what the functional links are, in the context of a whole system. To this end, several mathematical and computational approaches are developing. Most of them, such as kernel-based methods (Okamoto et al. [7], Yellaboina et al. [13]) imply a learning phase and so need *a priori* knowledge. Other methods, based on correlations or clustering, can reveal proximities between variables but do not bring to light the direct or functional links. Lately, Werhli and Husmeier [11] proposed to use Bayesian networks combining expression data with multiple sources of prior knowledge for

reconstructing gene regulatory networks.

A valuable complement to all of these methods is graphical Gaussian modeling (GGM) that can infer direct relations between variables from a set of repeating observations of these variables without any *a priori* knowledge. Graphical modeling is the use of a graph to represent a model. A graph is a set of nodes and edges which can be represented as a graphic for a visual study or as a matrix for computer processing. Graphical modeling is based on the conditional independence concept. In other words, a direct relation between two variables exists if those two variables are conditionally dependent given all remaining variables. In the Gaussian setting, a direct relation between two variables corresponds to a non-zero entry in the partial correlation matrix. As the estimate of the partial correlation matrix is related to the inverse of the empirical covariance matrix, a direct relation between two variables also corresponds to a non-zero entry in the inverse of the covariance matrix.

Graphical models are classically used when the number of observations, denoted n, is larger than the number of variables, denoted p. This is generally the case in financial or sociological studies where surveys concern few variables and a lot of observations. But it is not the case in the post-genomic context where each experiment is costly in time and money. So the number of repetitions is limited; moreover, each experiment generates numerous data. Then the data set structure, $p \gg n$, does not match with the assumptions of the graphical modeling classical approach and the empirical covariance matrix cannot be inverted. Over the last years, some mathematical and computational researches were developed for surrounding that dimensionality problem. Thus, Schäfer and Strimmer [10, 9], Wille and Bühlmann [12], Li and Gui [4], Meinshausen and Bühlmann [6] have proposed various methods, all based on the fact that the number of direct relations between two variables is very small in regards to the number of possible relations, $p(p-1)/2$.

The purpose of our study is to determine the validity domain of some methods recently proposed. The core of this chapter is divided in three parts. The first one describes the statistical methodology involved in Schäfer and Strimmer [9], Wille and Bühlmann [12] and Meinshausen and Bühlmann [6] approaches. The second part presents simulations carried out with each of these three methods, under different conditions of dataset structure. The third part illustrates the interest of the graphical Gaussian modeling with an application to flow cytometry data produced by Sachs et al [8]. In the conclusions we discuss the performance of each method and we bring some recommendations according to their validity domain.

3.2 Statistical methods

Let $\Gamma = \{1, \ldots, p\}$ be the set of nodes of the graph. The p nodes of the graph are identified with p Gaussian random variables. Let us denote by $\mathbf{X} = (X_1, \ldots, X_p)^T$, a p random vector distributed as a multivariate Gaussian $\mathcal{N}(0, \Sigma)$. For m a subset of $\{1, \ldots, p\}$ with cardinality $|m| \leq p-2$, we denote by Γ^{-m} the set of nodes that are not in m, $\Gamma^{-m} = \Gamma \setminus m$, and by $\mathbf{X}^{-m}$ the $p - |m|$ random vector whose components are the variables X_c, where $c \in \Gamma^{-m}$. There exists an edge between nodes a and b if and only if, the random variates X_a and X_b are not independent conditionnally to $\mathbf{X}^{-\{a,b\}}$. In other words, assuming that the matrix Σ is nonsingular, there exists an edge between nodes a and b if and only if the component (a, b) of the concentration matrix $K = \Sigma^{-1}$ is non zero. These graphs are called concentration graphs or full conditionnal independence graphs. The statistical challenge is

to detect the edges in the graph on the basis of a n-sample from a multivariate distribution $\mathcal{N}(0, \Sigma)$. For each $i = 1, \ldots, n$ we denote by $\mathcal{X}_i = (X_{i1}, \ldots, X_{ip})$ the observations obtained at experiment i. When the number of observation n is large enough, at least $n \geq p+1$, in order to guarantee that the sample covariance matrix S is nonsingular, several methods have been proposed. A detailed review can be found in a recent paper by Drton and Perlman [3]. However, when the interest lies on genomic networks, we are generally dealing with data where the number of variables p is large and the number of experiments n is small. Several methods have been proposed recently in that context. Schäfer and Strimmer [10, 9] proposed bagging methods in order to stabilize the estimator either of P, the correlation matrix associated to Σ, or Π the partial correlation matrix. Then they estimate the probability of an edge between two nodes (a, b) by estimating the density of the estimated partial correlation coefficient. Wille and Bühlmann [12] proposed to estimate a lower-order conditionnal independence graph instead of the concentration graph and to use a multiple testing procedure for detecting edges. Another different approach is proposed by Meinshausen and Bühlmann [6]. The neighbors of each node are estimated using a model selection procedure based on the LASSO method. The probability of falsy joining distinct connectivity components in the graph is controlled.

In the next section we briefly describe these methods, specifying their theoretical properties if any.

3.2.1 Estimating the concentration matrix

Schäfer and Strimmer consider the problem of obtaining accurate and reliable estimates of the covariance matrix Σ or its inverse K. They propose several strategies and we retain the bagging approach for our simulation study as it was recommended by the authors in [9]. Precisely, they proposed to estimate the partial correlation coefficients using the Moore-Penrose pseudo-inverse adding bootstrap aggregation (bagging) in order to reduce the variance of the estimator. For each bootstrap sample $\mathcal{X}^*$, the empirical correlation matrix $\widehat{P}^*$ is calculated. The bagged estimator is the empirical mean of the $\widehat{P}^*$'s from the bootstrap samples. The matrix Π is estimated by $\widehat{\Pi}^{\text{bagged}}$, the pseudo inverse of the bagged correlation matrix estimator. It remains to define a decision rule for detecting the significant components of Π. Schäfer and Strimmer assume that the distribution of the $\widehat{\Pi}^{\text{bagged}}_{a,b}$'s is known up to some parameters that are estimated. They deduce from this estimator the posterior probability of an edge to be present in the graph and decide to keep edges such that the posterior probability is greater than 95%.

3.2.2 Estimating the 0-1 conditional independence graph

The presence of edges in the graph is now defined through zero and first-order conditionnal distributions. Precisely, for each pair of nodes (a, b), let $R_{ab/\emptyset}$ be the correlation between the variables X_a and X_b, and for each $c \in \Gamma^{-\{a,b\}}$, let $R_{ab/c}$ be the correlation between X_a and X_b conditionnally to X_c. There exists an edge between nodes (a, b) if $R_{ab/\emptyset} \neq 0$ and $R_{ab/c} \neq 0$ for all $c \in \Gamma^{-\{a,b\}}$, or equivalently if

$$\phi_{a,b} = \min \left\{ |R_{ab/c}|, c \in \Gamma^{-\{a,b\}} \cup \emptyset \right\} \tag{3.1}$$

is non zero. Therefore, detecting edges in the graph remains to testing $p(p-1)/2$ statistical hypotheses: For each (a,b), $1 \leq a < b \leq 1$, there exists an edge between nodes (a,b) if the hypothesis "$\phi_{ab} = 0''$ is rejected. Wille and Bühlmann propose the following testing procedure: For each (a,b) and $c \in \Gamma^{-\{a,b\}} \cup \emptyset$ the likelihood ratio test statistic of the hypothesis "$R_{ab/c} = 0$" is calculated as well as the corresponding p-value denoted $P(a,b/c)$. Then the hypothesis "$\phi_{ab} = 0''$ is rejected at level α if

$$P_{\max}(a,b) = \max\Big\{P(a,b/c), c \in \Gamma^{-\{a,b\}} \cup \emptyset\Big\} \leq \alpha.$$

It remains to calculate the adjusted p-values to take into account the multiplicity of hypotheses to test, considering for example the Bonferroni procedure or the Benjamini-Hochberg one's.

Considering 0-1 conditionnal independence in place of full conditionnal independence has several advantages. The test statistics are very easy to calculate. For each hypothesis to test, "$R_{ab/c} = 0$", one considers the marginal distribution of the 3-random Gaussian vector $(X_a, X_b, X_c)^T$. Therefore, provided that n is large enough, the distribution of the likelihood ratio test statistic of the hypothesis "$R_{ab/c} = 0$" is well approximated by the distribution of a χ^2 with 1 degree of freedom. Note that it not necessary to assume that p is small. It follows, that, for each (a,b), the probability to detect an edge between a and b when it does not exist is smaller than α, if n is large (see Proposition 3 in [12]). Moreover it can be shown that if p increases with n in such a way that $\log(p)/n$ tends to 0 when n tends to infinity, then the estimators of the $R_{ab/c}$'s are uniformly convergent for all $a, b \in \Gamma$ and $c \in \Gamma^{-\{a,b\}} \cup \emptyset$.

The counterpart of this method is that 0-1 conditionnal independence graphs do not generally coincide with concentrations graphs, though the links between both graphs can be established in some cases.

Castelo and Roverato [1] and Malouche and Sevestre [5] proposed a similar approach for estimating "up to q"-order conditionnal independence graphs where the presence/absence of edges is associated to all marginal distributions up to the order $q+2$. We present simulation results only for the Wille and Bühlmann procedure.

3.2.3 Estimating the neighbors

For each node a, the set of neighbors of a, denoted $ne(a)$, is defined as the set of nodes in $\Gamma^{-\{a\}}$ that are connected with a. Thus detecting the neighbors of all nodes leads to detecting the edges in the graph. Because of the Gaussian assumption on the distribution of $\mathbf{X}$, for each variable X_a, a conditional regression model can be defined as follows:

$$X_a = \sum_{b \in \Gamma^{-\{a\}}} \theta_b^a X_b + \varepsilon_a$$

where the parameters θ_b^a are equal to $-K_{a,b}/K_{a,a}$. The variable ε_a is distributed as a centered Gaussian variable and is independent from the X_b's for all $b \in \Gamma^{-\{a\}}$. Meinshausen and Bühlmann [6] proposed to detect the non zero coefficients of the regression of X_a on the variables X_b for $b \in \Gamma^{-\{a\}}$ on the basis of the n-sample $(\mathcal{X}_1, \ldots, \mathcal{X}_n)$, using the LASSO method as a model selection procedure. Precisely, for a given smoothing parameter λ, the

estimators of $\{\theta_b^a, b \in \Gamma^{-\{a\}}\}$ minimize the sum of squares penalized by the ℓ_1-norm of the parameters vector:

$$\sum_{i=1}^{n} \left(X_{ia} - \sum_{b \in \Gamma^{-\{a\}}} \theta_b^a X_{ib} \right)^2 + \lambda \sum_{b \in \Gamma^{-\{a\}}} |\theta_b^a| \,. \tag{3.2}$$

For a given smoothing parameter λ, the solution to this minimization problem is given by a set of $(\widehat{\theta}_b^a, b \in \Gamma^{-\{a\}})$ that are either equal to zero or not. The set of nodes $b \in \Gamma^{-\{a\}}$ such that $\widehat{\theta}_b^a$ is non zero constitutes $\widehat{ne(a)}$, the estimated set of neighbors of the node a. Two estimated graphs may be deduced from all these $\widehat{ne(a)}$ for $a = 1, \ldots, p$, depending on whether we decide to put an edge between nodes a and b if both $\widehat{\theta}_b^a$ and $\widehat{\theta}_a^b$ are non zero or if one of these is non zero.

Meinshausen and Bühlmann proved that, under some conditions the method is consistent, namely that the probability for $\widehat{ne(a)}$ to be exactly equal to $ne(a)$ tends to one. The asymptotic framework considers sparse graphs with a large number of variables: the maximum number of neighbors growes slower than n while the number of nodes p may grow like any power of n. Finally the smoothing parameter λ is assumed to decrease to zero at a rate smaller than $n^{-1/2}$.

For the sake of application they propose a choice of λ such that the probability to connect two distinct connectivity components of the graph is bounded by some α. This choice is based on the Bonferroni inequality and it is assumed that the variance of the variables X_a for $a = 1, \ldots, p$ are all equal to one.

3.3 Simulations

3.3.1 Methods of simulation

3.3.1.1 Simulating a graph

Graphs were simulated according to two different approaches.

The first approach is based on the Erdös-Rényi model, noted ER model, which assumes that edges are independent and occur with the same probability. Practically, we fix the number of nodes p and the percentages of edges η then we draw the number of edges according to a binomial distribution with parameters $p(p-1)/2$, η. Next we choose uniformly and independently the positions of the edges.

The second approach was proposed by Daudin et al. [2] to take into account the topological features of biological networks such as connectivity degree or clustering coefficient. Their model called Erdos-Rényi Mixtures for Graphs, noted $ERMG$, supposes that nodes are spread into Q clusters with probabilities $\{p_1, \ldots, p_Q\}$, and that the connection probabilities of each cluster and between clusters are heterogenous. These connection probabilities constitute the connectivity matrix C. The parameters available from Daudin et al. [2] study, correspond to a graph with 199 nodes. As we wanted to study the influence of p, we adapted those parameters to our simulations. However, we kept the same graph structure by taking a large weakly connected cluster, a small highly connected cluster and the same group connection structure. Thus we used the following parameter values

$$Q = 4, \qquad (p_1, \ldots, p_Q) = \begin{pmatrix} 0.07 & 0.1 & 0.18 & 0.65 \end{pmatrix} \tag{3.3}$$

$$C = \begin{pmatrix} 0.999 & 10^{-6} & 10^{-6} & 0.005 \\ 10^{-6} & 0.4 & 0.014 & 0.003 \\ 10^{-6} & 0.014 & 0.2065 & 0.011 \\ 0.005 & 0.003 & 0.011 & 0.013 \end{pmatrix} \tag{3.4}$$

that lead to a mean percentage of edges η equals to 2.5%.

Whatever the approach, we finally obtain a matrix composed of 0 and 1, the values 1 indicating the edge positions in the corresponding graph. This matrix is denoted the incidence matrix.

3.3.1.2 Simulating the data

From the incidence matrix of a given graph, we simulated n observations as follows: first we generate a partial correlation matrix Π by remplacing the values 1 indicating the edge positions in the incidence matrix, by values drawn from the uniform distribution between -1 and 1. Then we compute columm-wise sums of the absolute values and set the corresponding diagonal element equal to this sum plus a small constant. This ensures that the resulting matrix is diagonally dominant and thus positive definite. Next we standardize the matrix so that each diagonal entry equals to 1. Finally, we generate n independent samples from the multivariate normal distribution with mean zero, unit variance, and correlation structure associated to the partial correlation matrix Π.

3.3.2 Simulation setup

We simulated graphs and data for different values of p, η, n, and we estimated graphs from these data using three different approaches:

- the $\widehat{\Pi}^{\text{bagged}}$ approach, proposed by Schäfer and Strimmer [10, 9] with the decision rule based on posterior probabilities,

- the 0-1 conditionnal independence graph approach, proposed by Wille and Bühlmann [12] with the decision rule based on the adjusted p-values following the Benjamini-Hochberg procedure,

- the Lasso approach, with the two variants *and* and *or* proposed by Meinshausen and Bühlmann [6].

In the continuation of this document we will respectively denote these methods as *bagging*, WB, $MB.and$ and $MB.or$.

To assess the performance of the investigated methods we compared each simulated graph with the estimated graph by counting true positives TP (correctly identified edges), false positive FP (wrongly detected edges), true negatives TN (correctly identified zero-edges), and false negatives FN (not recognized edges). From those quantities we estimated the power and the false discovery rate FDR, which are defined by:

$$power = E\left(\frac{TP}{TP + FN}\right)$$

$$FDR = E\left(\frac{FP}{TP+FP}|(TP+FP)>0\right)$$

The power and FDR values presented in this work, are the means over 2000 simulations (according to our preliminary results which showed that the stability of the FDR estimation was reached with 2000 simulations).

The performance of the methods were evaluated for several combinations of the parameters p, η and n, in regards to the problematic we wanted to investigate. Moreover, the parameter values are chosen in order to both make the computer time reasonable and extrapolate the results to biological fields.

The first problematic we have focused on, is the influence of the sample size. To this aim, we simulated random graphs fixing the number of nodes p equal to 30, η equal to 2.5% and varying the number of observations n in $\{15, 22, 30, 60\}$. Secondly, we investigated the sparsity assumption common to all methods taking η in $\{0\%, 2.5\%, 4\%, 5\%, 10\%\}$. Third, we were interested in the influence of the node number p. So, we increased p and chose n in order to keep the p/n rates similar to the p/n values used in the first considered point.

For all of these three studies, graphs were simulated with the ER method.

The forth problematic we investigated concerns the influence of the graph structure. In this goal, we also simulated graphs with the $ERMG$ method fixing p equal to 30 and varying n in $\{15, 22, 30, 60\}$.

Finally, we focused on the method proposed by Wille and Bühlmann to evaluate the consequence of estimating the 0-1 graph instead of the concentration graph. For this purpose we fixed $p = 30$ and varied η from 0.025 to 0.2 and n from 60 to 1200.

3.3.3 The results

3.3.3.1 Comparing the methods

As shown in Figure 3.1a), the FDR values never exceed 0.06 except with the *bagging* method for $n = 15$. The FDR values obtained with $MB.or$ remain steady around 0.055 whereas the FDR values obtained with $MB.and$ never exceed 0.01. FDR from *bagging* reaches 0.18 when $n = 15$ then deeply declines below 0.03. The power represented in Figure 3.1b) gradually increases with the number of observations n except with the *bagging* method which shows a drop for $n = p$. This drop mentionned by the authors is not yet explained. Let us notice that $MB.and$ and $MB.or$ do not work when $n = 15$. Indeed, for each variable a, the LASSO estimators of the parameters θ_b^a for $b \in \Gamma^{-\{a\}}$ are calculated in two steps: first the LARS algorithm is used for ranking the variables $\mathbf{X}^{-\{a\}}$ according to the covariance structure of the matrix $\mathbf{X}$, then the estimators of $(\theta_b^a, b \in \Gamma^{-\{a\}})$ that minimize Equation (3.2) are calculated for the chosen value of λ. It can be shown that if the data matrix is centered and scaled before the analysis such that $\sum_{i=1}^n X_{ib} = 0$ and $\sum_{i=1}^n X_{ib}^2 = 1$ for all b, and if $\lambda \geq 2$, then these estimators are all equal to zero. Therefore, if λ is chosen as proposed by Meinshausen and Bühlmann, namely

$$\lambda = \frac{2}{\sqrt{n}}\Phi^{-1}\left(1-\frac{\alpha}{2p^2}\right),$$

where Φ is the distribution function of a centered gaussian variable with variance 1, then the parameters will be estimated by zero whatever the data if

$$n \leq \left\{\Phi^{-1}(1-\alpha/2p^2)\right\}^2. \tag{3.5}$$

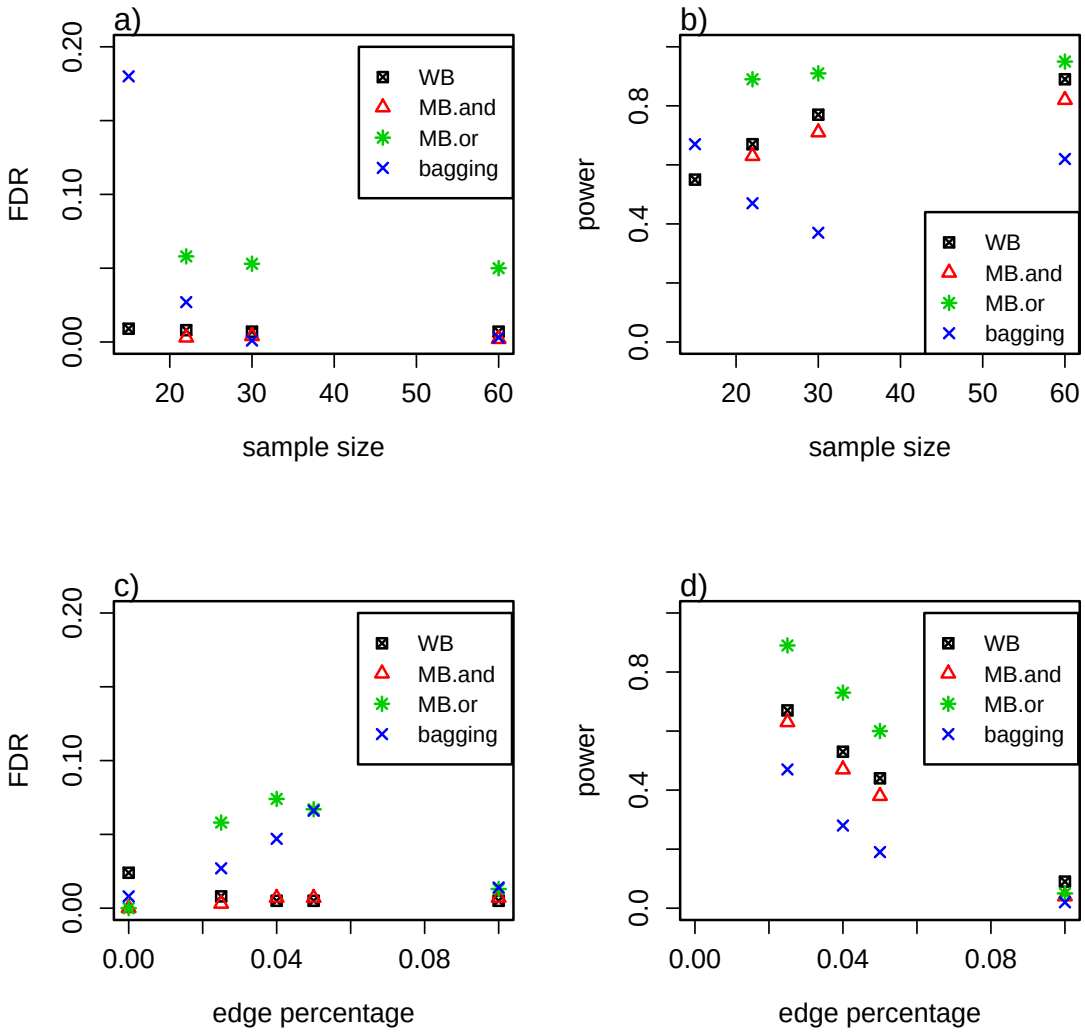

Figure 3.1: FDR and power of the different methods tested, in function of the sample size (a) and b), respectively) and the edge percentage (c) and d), respectively). Graphics a) and b) were obtained with $\eta = 0.025$. Graphics c) and d) were obtained with $n = 22$.

The influence of the edge percentages η is shown in Figure 3.1c) and 3.1d), for $n = 22$. The FDR values, shown in Figure 3.1c), stay very low with WB and $MB.and$ methods, but exceed 0.05 with *bagging* and $MB.or$ when η in $\{2.5\%, 4\%, 5\%\}$. Indeed, for all methods the power dramatically falls as η increases. When η equals 0.1, the power is close to 0 whatever the method used. Similar graphics were obtained for $n = 30$ and $n = 60$.

Considering the reliability feature (low FDR), the results presented in Figure 3.1 reveal that the $MB.and$ and WB methods perform quite well in all cases. Referring to the power, the $MB.or$ method outperfoms the others but we should pay attention to the high FDR values that it produces. The $MB.and$ and WB methods are slightly less powerfull than the $MB.or$ method with the advantage of producing small FDR values. The *bagging* appears as the less competitive method in terms of power. All methods similarly show a strong decrease of the power when η increases. Note that these methods were based on the sparsity assumption.

3.3.3.2 Influence of the number of nodes

In the preceding paragraph, we showed that all methods lose in power when η increases. So we now investigate the influence of the number of nodes, p, on that loss of power. Results (Figure 3.2) are shown for $MB.and$ procedure, the behavior of the other procedures being similar. The graphic in Fig.3.2 represents the power in function of η, for different values of p, with the constraint $p = n$. One can see that whatever the value of p, the power decreases when η increases. However, the larger p, the faster the loss of power with η. Consequently, all methods are efficient for sparse graphs, and the edge percentage from which the methods fail depends on the number of nodes.

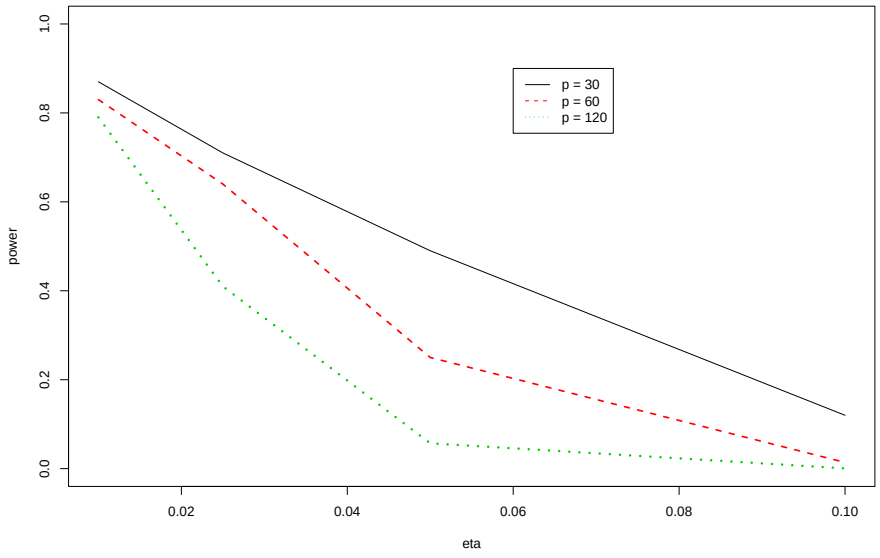

Figure 3.2: Power according to η, for different values of p with $n = p$. and for Meinshausen and Bühlmann method using its *and* variant.

3.3.3.3 Influence of the numbers of neighbors

In Section 3.3.3.1 we showed that if the graph is highly connected, the methods are not powerful anymore. In this section we aim at understanding why, and we show in particular the behavior of the methods according to the number of neighbors of the nodes. We focus on the procedure proposed by Meinshausen and Bühlmann, the more natural one to study the number of neighbors, and we consider the experiment simulation for $p = 30$, $\eta = 0.025$ and $n = 30$. For each node of the 2000 simulated graphs we count the number of neighbors and the number of correctly detected neighbors. In Table 3.1 we present for i in $\{1, \ldots, 5\}$, the number n_i of nodes with i neighbors and the percentage $p_{i,j}$ of nodes for which the method has correctly detected j neighbors, for j in $\{0, \ldots, i\}$.

i	1	2	3	4	5
n_i	21398	7684	1788	287	38

percentage $p_{i,j}$ obtained with $MB.and$

j \ i	1	2	3	4	5
0	0.182	0.078	0.062	0.066	0.105
1	0.818	0.643	0.465	0.380	0.263
2		0.279	0.414	0.411	0.500
3			0.059	0.139	0.132
4				0.004	0.000

percentage $p_{i,j}$ obtained with $MB.or$

j \ i	1	2	3	4	5
0	0.022	0.016	0.030	0.049	0.079
1	0.978	0.267	0.097	0.063	0.105
2		0.717	0.370	0.195	0.079
3			0.503	0.387	0.132
4				0.306	0.447
5					0.158

Table 3.1: n_i is the number of nodes with i neighbors and $p_{i,j}$ the percentage of nodes for which j neighbors have been correctly detected by the method of Meinshausen and Bühlmann with $n = 30$. Graphs are simulated according to the ER model with $p = 30$, $\eta = 0.025$.

The percentage $(p_{ii})_{i=1,\ldots,5}$ of nodes for which the whole set of neighbors is correctly detected decreases when the number of neighbors i increases. In other words when a node has several neighbors, it often happens that at least one neighbor is not detected. This may explain the loss of power previously observed (see Section 3.3.3.1) when η increases, because the average number of neighbors increases with η.

Let us now compare the results obtained with $MB.and$ and $MB.or$ procedures. In Section 3.3.3.1 we showed that $MB.or$ procedure is more powerful and we recover in Table 3.1 that the percentages of nodes for which the whole set of neighbors is correctly detected are significantly larger with $MB.or$ procedure than with $MB.and$ procedure. Consider for example, as illustrated in Figure 3.3, a node a with two neighbors b and c such that a is the only neighbor of b and of c. As it has been noticed just before, the procedure of Meinshausen and Bühlmann will detect more easily that a is the neighbor of b and c than b and c are both neighbors of a. This is the reason why $MB.or$ procedure is more powerful than $MB.and$ procedure.

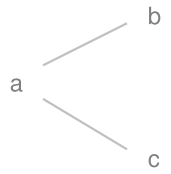

Figure 3.3: node a with two neighbors b and c such that a is the only neighbor of b and of c

3.3.3.4 Influence of the graph structure

In this section we present results of simulation when graphs are simulated according to the ERMG model described in Section 3.3.1.1. Our aim is to evaluate the influence of heteregeneous clusters in the graph. Results are shown in Figure 3.4 for $p = 30$ and for n taking the values 15, 22, 30 and 60. For the parameters given in Equations (3.3) and (3.4) the percentage η of edges equals 0.025 which makes the results comparable with those of Figure 3.1a) and 3.1b).

The methods behave similarly whatever the model used to simulate the graphs. As in Figure 3.1a) the FDR value obtained with *bagging* is high when $n = 15$ then deeply declines, and as in Figure 3.1b) we observe a drop of the power for $n = p$. Moreover we recover than the FDR values stay very low with WB and $MB.and$ procedure and are larger with $MB.or$ procedure. Referring to the power, as in Figure 3.1b) the $MB.or$ procedure outperforms the others.

The main difference when graphs are simulated according to the ERMG models is that the power remains under 0.8 even for large n (Figure 3.4b) whereas it achieves 0.95 when ER model is used (Figure 3.1b). So, the methods are less powerful when graphs are simulated according to the ERMG model than according to the ER model. The next section shed light on this loss of power.

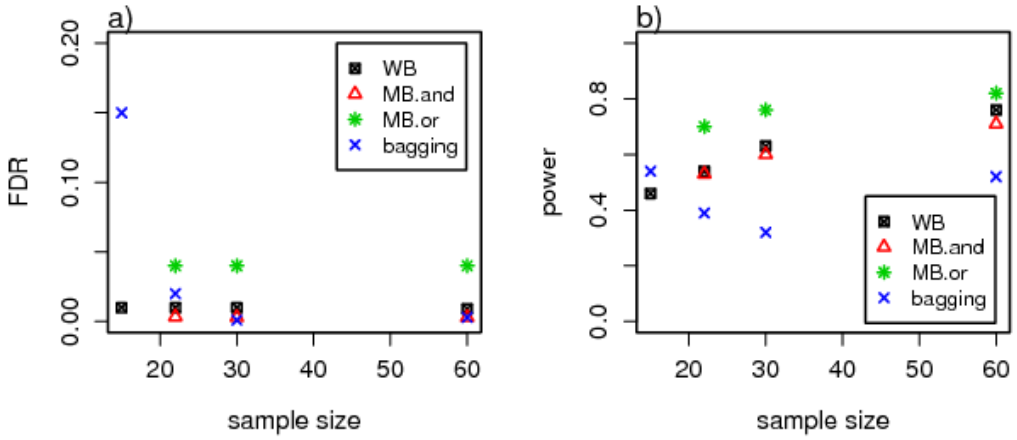

Figure 3.4: FDR (a)) and power (b)) obtained with the different methods tested, in function of the sample size. Graphs were simulated according to the ERMG model with $p = 30$

3.3.3.5 Influence of the neighborhood structure

In this section we study why the methods are less powerful when graphs are simulated according to the ERMG model than according to the ER model and we underline in particular the influence of the neighborhood structure.

We consider the same experiment study as in Section 3.3.3.3 except that the graphs are simulated according to the ERMG model. In Table 3.2, one can read for each i in $\{1,\dots,6\}$, the number n_i of nodes with i neighbors and the percentage $p_{i,j}$ of nodes for which the method has correctly detected j neighbors, for j in $\{0,\dots,i\}$. Results are obtained with the procedure $MB.or$.

i	1	2	3	4	5	6
n_i	16326	6720	2603	941	332	63

percentage $p_{i,j}$ obtained with $MB.or$

j \ i	1	2	3	4	5	6
0	0.042	0.129	0.289	0.530	0.654	0.841
1	0.958	0.305	0.206	0.148	0.133	0.048
2		0.566	0.256	0.128	0.075	0.048
3			0.249	0.121	0.057	0.032
4				0.073	0.054	0.000
5					0.027	0.032
6						0.000

Table 3.2: n_i is the number of nodes with i neighbors and $p_{i,j}$ the percentage of nodes for which j neighbors have been correctly detected by the $MB.or$ procedure with $n = 30$. Graphs are simulated according to the ERMG model with $p = 30$

Comparing Tables 3.1 and 3.2 shows that the number of nodes with 1 and 2 neighbors is smaller for the ERMG model than for the ER model, while the number of nodes with more than 3 neighbors is greater. Therefore, because the power of the procedure for detecting the neighbors of nodes decreases with the number of neighbors, it is more difficult to estimate graphs simulated according to the ERMG model than to the ER model.

Let us now compare the percentages p_{ij}. For nodes with one neighbor, the percentages p_{1j} are similar in both tables. But when the number of neighbors i is larger than one, the percentage of nodes p_{ii} for which the whole set of neighbors is correctly detected are smaller in Table 3.2 than in Table 3.1. Moreover the main difference between Table 3.1 and 3.2 concerns the percentage of nodes for which no neighbor is detected: these percentages p_{i0} are very small in Table 3.1, but large in Table 3.2 and increases with the number of neighbours. In other words, for graphs simulated according to the EMRG model, detecting no neighbor at all happens often, especially for nodes with a large number of neighbors.

This can be explained by the structure of the neighbors, which is more complex for graphs simulated according to the ERMG model. This is indeed illustrated above.

In the following we present the FDR and the power estimated into each cluster and between the clusters. We first simulate a graph $\mathcal{G}$ according to the ERMG model with the parameters defined in Section 3.3.1.1 in order to fix the number of nodes and the localisation of edges in each cluster and between clusters. We simulate a graph with $p = 120$ nodes to ensure that each cluster contains a minimal number of nodes. We denote by $(n_1, \ldots, n_Q)$ the number of nodes in each of the clusters and by N_{edges} the matrix which specifies the number of edges into each cluster and between two clusters. For the simulated graph $\mathcal{G}$, these parameters equal:

$$(n_1, \ldots, n_Q) = \begin{pmatrix} 7 & 11 & 23 & 79 \end{pmatrix}$$

and

$$N_{edges} = \begin{pmatrix} 21 & 0 & 0 & 3 \\ 0 & 21 & 6 & 3 \\ 0 & 6 & 38 & 19 \\ 3 & 3 & 19 & 35 \end{pmatrix}$$

We simulate 2000 data matrix as described in Section 3.3.1.2 from this graph $\mathcal{G}$ and we estimate the FDR and the power for detecting edges in and between clusters. The results obtained with the *MB.or* procedure with $n = p$ are presented in the matrices FDR and *power* given at Equations (3.6) and (3.7). The component (a, b) of the matrix FDR (respectively *power*) gives the estimated false discovery rate (respectively power) of edges between clusters a and b. When there is no edge between two clusters, estimating the power does not make any sense and we put Na. The elements on the diagonal give the estimated false discovery rate (respectively power) of edges in each cluster.

$$FDR = \begin{pmatrix} 0.000 & 0.001 & 0.016 & 0.008 \\ 0.001 & 0.005 & 0.004 & 0.012 \\ 0.016 & 0.004 & 0.006 & 0.014 \\ 0.008 & 0.012 & 0.014 & 0.021 \end{pmatrix} \tag{3.6}$$

$$power = \begin{pmatrix} 0.10 & \text{Na} & \text{Na} & 0.46 \\ \text{Na} & 0.26 & 0.22 & 0.61 \\ \text{Na} & 0.22 & 0.29 & 0.61 \\ 0.46 & 0.61 & 0.61 & 0.87 \end{pmatrix} \tag{3.7}$$

Let us notice that the FDR estimated in and between clusters are small. Moreover the estimated powers vary a lot according to the clusters and depend on the percentage of edges. Indeed, in the first cluster which contains 21 edges among the $n_1(n_1 - 1)/2 = 21$ possible edges, the power is very small whereas in the fourth cluster which contains 35 edges among the $n_4(n_4 - 1)/2 = 3081$ possible edges, the power is large. The neighbors of the neighbors also influence the power. Let us illustrate this by comparing the power for

detecting edges between the second and third clusters, $power[2,3] = 0.22$, with the power for detecting edges in the fourth cluster, $power[4,4] = 0.87$. It appears that it is more difficult to detect edges between clusters 2 and 3 than in cluster 4, while in both cases the percentage of edges to detect is approximately equal to 0.01. This comes from the fact that clusters 2 and 3 are both highly connected. Therefore these two clusters involve nodes for which the structure of the neighbors is complex.

Because of these highly connected parts, the power estimated on the whole graph $\mathcal{G}$ is smaller than if the edges were distributed uniformly in the graph. Indeed, the FDR and the power estimated on the whole graph $\mathcal{G}$ equal respectively 0.016 and 0.44. For graphs simulated according to the ER model with $p = 120$ and $\eta = 0.025$ the average FDR and power estimated over 2000 simulations with the $MB.or$ procedure and $n = 120$ equal respectively 0.009 and 0.50.

3.3.3.6 Inferring a concentration graph using a 0-1 conditionnal independence graph

If the gaussian distribution is *faithful* for the concentration graph $\mathcal{G}$ (see Proposition 1 in [12]), then all edges in $\mathcal{G}$ are edges in the 0-1 conditionnal independence graph denoted $\mathcal{G}_{\{0,1\}}$. A comparison between $\mathcal{G}$ and $\mathcal{G}_{\{0,1\}}$ is given at Table 3.3. For each concentration matrix whose values are simulated as described in Section 3.3.1.2, and for each pair (a, b), $1 \leq a < b \leq 1$, we calculated $\phi_{a,b}$ defined at Equation (3.1). It appears that, as it was already noticed by Wille and Bühlmann, the number of edges in $\mathcal{G}_{\{0,1\}}$ may be considerably larger than in $\mathcal{G}$. The power and the FDR for estimating the graph $\mathcal{G}$ are reported on

	$\mathcal{G} \cap \mathcal{G}_{\{0,1\}}$			$\mathcal{G}_{\{0,1\}} \setminus \mathcal{G}$		
η	Number	mean	range	Number	mean	range
0.025	11	0.72	$[10^{-4}, 0.99]$	0.3	0.09	$[10^{-4}, 0.33]$
0.05	22	0.57	$[10^{-5}, 0.99]$	17	0.05	$[10^{-6}, 0.46]$
0.1	43	0.35	$[10^{-7}, 0.99]$	217	0.02	$[10^{-9}, 0.41]$
0.15	65	0.24	$[10^{-9}, 0.99]$	322	0.012	$[10^{-9}, 0.30]$
0.2	87	0.18	$[10^{-8}, 0.99]$	337	0.01	$[10^{-9}, 0.21]$
0.3	131	0.11		304	0.009	

Table 3.3: Comparison of $\mathcal{G}_{\{0,1\}}$ and $\mathcal{G}$ for $p = 30$ and several values of η. The column $\mathcal{G} \cap \mathcal{G}_{\{0,1\}}$ gives the mean (over 2000 simulations) number of edges that are both in $\mathcal{G}$ and $\mathcal{G}_{\{0,1\}}$, followed by the mean and range of the $\phi_{a,b}$'s corresponding to these edges. The column $\mathcal{G}_{\{0,1\}} \setminus \mathcal{G}$ gives the sames results for edges that are in $\mathcal{G}_{\{0,1\}}$ and not in $\mathcal{G}$. In all simulations the edges of $\mathcal{G}$ are edges of $\mathcal{G}_{\{0,1\}}$.

Figure 3.5, graphs a) and b). It shows that the FDR increases with n and is maximum for $\eta = 10\%$. This behaviour can be easily explained by looking at graphs c) and d) where the FDR and the power for estimating $\mathcal{G}_{\{0,1\}}$ are reported. It shows that the FDR for estimating $\mathcal{G}_{\{0,1\}}$ stays small, and that, as expected, the power increases with n. Unfortunatly the edges detected as positive in $\mathcal{G}_{\{0,1\}}$ are not in $\mathcal{G}$, leading to increase the FDR for detecting edges in $\mathcal{G}$.

When η is small, say $\eta \leq 2.5\%$, the number of edges that are in $\mathcal{G}_{\{0,1\}}$ but not in $\mathcal{G}$ is very small, and then the FDR for detecting edges in $\mathcal{G}$ is not changed. But when η is large the FDR becomes very large, up to 20% for $\eta = 10\%$. Nevertheless when η is larger, the FDR decreases. This can be explained by the values of the $\phi_{a,b}$'s that are smaller when η increases as it is shown in table 3.3. Obviously the behaviour of the procedure proposed by Wille and Bühlmann shown in this simulation study may depend on the way we simulate the concentration matrix. Nevertheless, we have to keep in mind that if the graph is highly connected, or if a part of it is highly connected, then, inferring a concentration graph on the basis of its approximation by a 0-1 conditionnal independence graph, may lead to detect edges wrongly.

3.4 Application to biological data

In this section, we apply the different methods to the multivariate flow cytometry data produced by Sachs et al. [8]. These data concern a human T cell signaling pathway whose deregulation may lead to carcenogenesis. Therefore, this pathway was extensively studied in the literature and a network involving 11 proteins and 18 interactions was conventionally accepted [8]. This network is denoted $\mathcal{G}_{raf}$ and is represented in Figure 3.6. The data from Sachs et al.consist of quantitative amounts of these 11 proteins, simultaneously measured from single cells under several perturbation conditions. In the sequel, we focus on one general perturbation (anti-CD3/CD28 + ICAM-2) that overall stimulates the cellular signaling network. In this condition the quantities of the 11 proteins are measured in 902 cells. Let denote D this data set constituted of $p = 11$ variables and $n = 902$ observations. Contrary to most of postgenomic data, flow cytometry data provide a large sample of observations that allow us to measure the influence of the sample size on the power of the estimation methods. From this data set we first infer the network using the three presented methods and we refer to the literature to assess the inferred graphs. As such abundance of data is rarely available in postgenomic data, we secondly carry out a study to determine the influence of the number of observations on the methods.

We represent the estimated graphs in Figure 3.7. The graphs inferred with the methods of Schäfer and Strimmer and Wille and Bühlmann are identical. This graph is denoted $\mathcal{G}_1$ and involves 10 edges. The two variants of Meinshausen and Bühlmann procedure infer the same graph. This graph, denoted $\mathcal{G}_2$ involves 9 edges and is identical to $\mathcal{G}_1$ except for the edges between PKA and $Erk1/2$ which is missing.

To assess the quality of the methods, we refer to the conventionally acccepted network shown in Figure 3.6. This network involves 18 connections among which 16 connections are well-established. As the data set D is obtained by considering only one general perturbation condition we do not expect the methods to detect all the connections established in the literature. In fact, 10 connections are detected by two of the three methods, among which 9 are well-established or cited at least once in the literature. Moreover the three methods detect a supplementary connection between $p38$ and JNK. This connection is also detected by Sachs et al. using a bayesian approach. So we assume that this $p38$-JNK connection is not wrongly detected and that the graph $\mathcal{G}_1$ represents the conditional independance structure of the data set D.

We now investigate the influence of the number of observations n on the power of the methods for estimating the graph $\mathcal{G}_1$. We choose n equal to 15, 30, 100, 200, and 300.

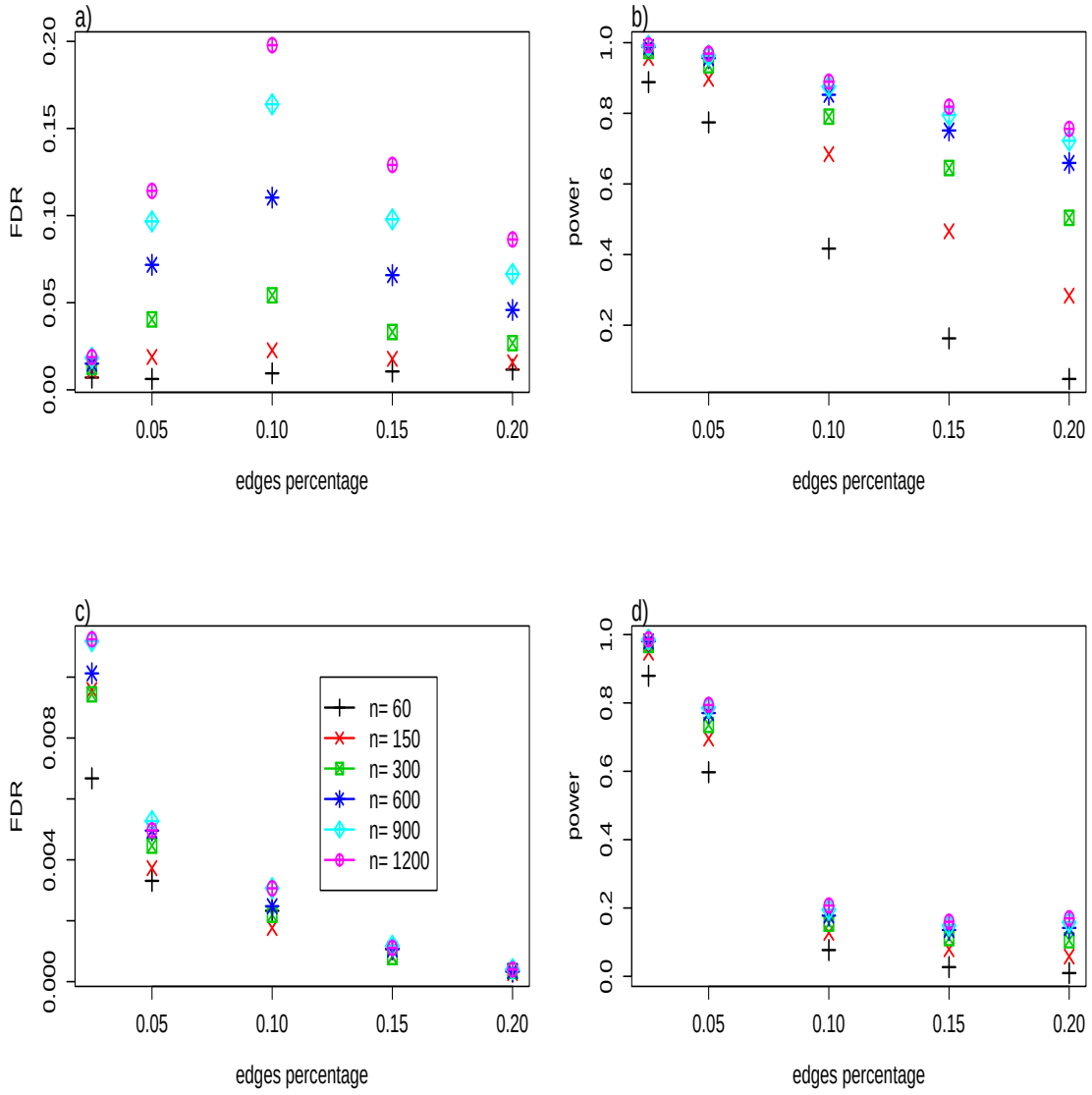

Figure 3.5: Wille and Bühlmann method. FDR and power for estimating $\mathcal{G}$, graphs a) and b), and $\mathcal{G}_{\{0,1\}}$, graphs c) and d), in function of η for $p = 30$ and different values of n.

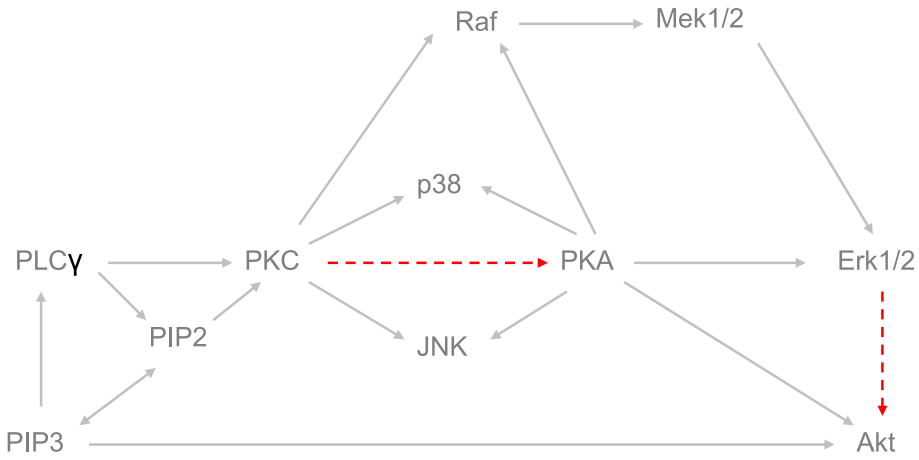

Figure 3.6: $\mathcal{G}_{raf}$. Classic signaling network of the human T cell pathway. The connections well-established in the literature are in grey and the connections cited at least once in the literature are represented by red dotted lines.

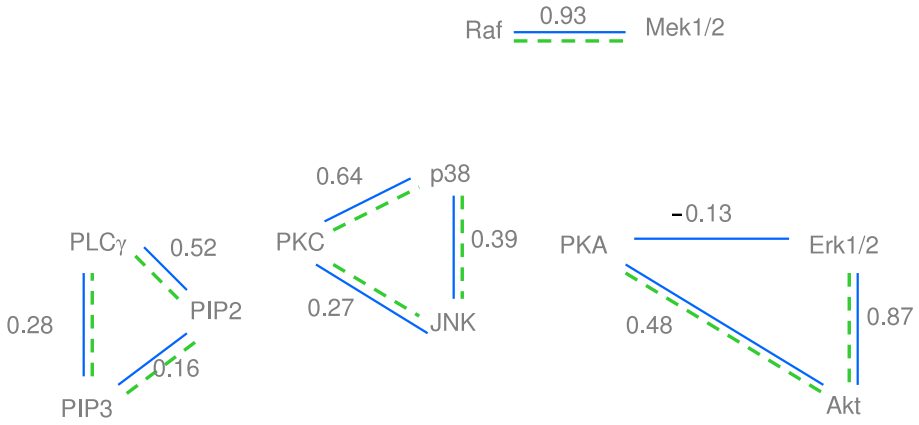

Figure 3.7: Inferred graphs. The graph $\mathcal{G}_1$ estimated with the *bagging* and WB procedures is represented in blue. The graph $\mathcal{G}_2$ estimated with the $MB.and$ and $MB.or$ procedures is in green dotted line. For each edge are reported the values of the partial correlation matrix associated to the data set D

For each value of n we take 2000 samples of size n from D without replacement. From each sample, we estimate graphs using the three methods and we compare each estimated graph with the graph $\mathcal{G}_1$. We compute the porportion of wrongly detected edges among the detected edges and the proportion of correctly identified edges among the edges of $\mathcal{G}_1$. The means of these quantities over the 2000 simulations are denoted FDR and power. Results are presented in Table 3.4. As expected, the power of all the methods increases with the number of observations n. However n has to be large to detect most of the edges. It comes from the fact that the graph $\mathcal{G}_1$ involves 11 proteins and 10 edges, which corresponds to a large percentage of edges equal to 18%. In this study, we notice that the edges $Raf - Mek1/2$ and $Erk1/2 - Akt$ are detected in most of the simulations even for small n and whatever the method; on the contrary the edge $PKA - Erk1/2$ is less detected. It is in accordance with the values of the partial correlation matrix for each edge, given in Figure 3.7: the largest values of the partial correlation matrix correpond indeed to the edges the most detected. Let us now compare the methods according to the number of observations at our disposal. When n is small equal to 15, the WB procedure is the most powerful and the FDR is small. When n is moderate equal to 30 or 100, we advise to use $MB.or$ procedure, because the FDR is small and the benefit in power is large. When n is very large, referring to the power all the methods are equivalent; the FDR obtained with the $MB.and$ procedure being null, this procedure is recommended.

FDR

n	*bagging*	WB	$MB.and$	$MB.or$
15	0.0227	0.0037	0.0007	0.0017
30	0.0159	0.0020	0.0011	0.0044
100	0.0117	0.0017	0.0001	0.0067
200	0.0098	0.0010	0.0000	0.0111
300	0.0056	0.0005	0.0000	0.0136

Power

n	*bagging*	WB	$MB.and$	$MB.or$
15	0.27	0.33	0.23	0.26
30	0.43	0.47	0.47	0.62
100	0.68	0.69	0.68	0.77
200	0.79	0.79	0.77	0.81
300	0.85	0.83	0.82	0.83

Table 3.4: FDR and power for estimating the graph $\mathcal{G}_1$. Results for the different methods and for different values of n.

3.5 Conclusion

When n is small relative to p, then all methods lack of power. Moreover, the two Meinshausen and Bühlmann procedures can not be used if n satisfies Equation (3.5), and the bagging method must be avoided because of its high false discovery rate. If the graph is known to be sparse such that the 0-1 conditionnal independence graph is nearly equal to the concentration graph, then the Wille and Bühlmann method can be used.

When n is large, the estimation methods perform better. Nevertheless, as it was already noticed by Schäffer and Strimmer, the bagging method lacks of power when $n = p$. Concerning the Wille and Bühlmann approach, if the 0-1 conditionnal independence graph contains many edges that are not in the concentration graph, then the power for detecting edges in the 0-1 graph increases with n, leading to false discovery edges in the concentration graph. The Lasso appproach involved in the MB procedure has good properties. If one can accept a false discovery rate of the order 5%, then we recommend to use the variant $MB.or$ which is more powerfull than the variant $MB.and$. This last one must be preferred when the false discovery rate has to be very small.

The structure of the graph should also be considered: if the edges are not uniformly distributed over the nodes as in the Erdös-Rényi model, then edges localized in highly connected parts of the graph or edges joining two highly connected parts may be difficult to detect.

So, methods inferring graphs do not behave equivalently faced to the graph and dataset structures. Consequently we have to pay attention to the validity domain of each method before using it. A way of making sure that an edge is correctly detected could be in implementing several methods and comparing their results.

References

[1] Castelo, R., and Roverato, A. A robust procedure for gaussian graphical model search from microarray data with p larger than n. *Journal of Machine Learning Research 7* (2006), 2621–2650.

[2] Daudin, J. J., Picard, F., and Robin, S. A mixture model for random graphs. Tech. Rep. RR-5840, INRIA, Rapport de Recherche, 2006.

[3] Drton, M., and Perlman, M. Multiple testing and error control in gaussian graphical model selection. *Statistical Sciences, To appear* (2007).

[4] Li, H., and Gui, J. Gradient directed regularization for sparse gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics 7*, 2 (2006), 302–317.

[5] Malouche, D., and Sevestre, S. Estimating high dimensional faithful gaussian graphical models : upc-algorithm. Tech. Rep. arXiv:0705.1613, Technical Report, 2007.

[6] Meinshausen, N., and Bühlmann, P. High dimensional graphs and variable selection with the Lasso. *Annals of Statistics 34*, 3 (2006), 1436–1462.

[7] Okamoto, S., Yamanishi, Y., Ehira, S., Kawashima, S., Tonomura, K., and Kanehisa, M. Prediction of nitrogen metabolism-related genes in anabaena by kernel-based network analysis. *Proteomics 7*, 6 (2007), 900–909.

[8] Sachs, K., Perez, O., D.Pe'er, Lauffenburger, D. A., and Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science 308* (2005), 523–529.

[9] Schäfer, J., and Strimmer, K. An empirical bayes approach to inferring large-scale gene association nerworks. *Bioinformatics 21*, 6 (2005), 754–764.

[10] Schäfer, J., and Strimmer, K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology 4* (2005), 1–32.

[11] Werhli, A., and Husmeier, D. Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge. *Statistical Applications in Genetics and Molecular Biology 6* (2007).

[12] Wille, A., and Bühlmann, P. Low-order conditional independence graphs for inferring genetic networks. *Statistical Applications in Genetics and Molecular Biology 5* (2006), 1–34.

[13] Yellaboina, S., Goyal, K., and Mande, S. Inferring genome-wide functional linkages in e-coli by combining improved genome context methods: Comparison with high-throughput experimental data. *Genome Research 17*, 4 (2007), 527–535.

Troisième partie

Test de validation de graphe

Chapter 4

Goodness-of-fit tests: Gaussian regression with random covariates

Ce chapitre présente un travail réalisé en collaboration avec Nicolas Verzelen

Abstract

Let $(Y, (X_i)_{i\in\mathcal{I}})$ be a zero mean Gaussian vector and V be a subset of $\mathcal{I}$. Suppose we are given n i.i.d. replications of the vector (Y, X). We propose a new test for testing that Y is independent of $(X_i)_{i\in\mathcal{I}\setminus V}$ conditionally to $(X_i)_{i\in V}$ against the general alternative that it is not. This procedure does not depend on any prior information on the covariance of X or the variance of Y and applies in a high dimensional setting. It is based on a model of Gaussian regression with random Gaussian covariates. We give non asymptotic properties of the test and we prove that our test is rate optimal (up to a possible $\log(n)$ factor) over various classes of alternatives under some additional assumptions. Besides, it allows us to derive non asymptotic minimax rates of testing in this setting. Finally, we carry out a simulation study in order to evaluate the performance of our procedure.

4.1 Introduction

We consider the following regression model

$$Y = \sum_{i\in\mathcal{I}} \theta_i X_i + \epsilon \tag{4.1}$$

where θ is an unknown vector of $\mathbb{R}^{\mathcal{I}}$. The vector X follows a zero mean Gaussian distribution with non singular covariance matrix Σ and ϵ is a zero mean Gaussian random variable independent of X. We note p the cardinal of $\mathcal{I}$ and $\mathrm{var}(Y)$ the variance of Y. Straightforwardly, the variance of ϵ corresponds to the conditional variance of Y given X, $\mathrm{var}(Y|X)$.

We are given n i.i.d. replications of the vector (Y, X). Let us respectively note $\mathbf{Y}$ and $\mathbf{X}_i$ the vectors of the n observations of Y and X_i for any $i \in \mathcal{I}$. Let V be a subset of $\mathcal{I}$, then X_V refers to the set $\{X_i, i \in V\}$ and θ_V stands for the sequence $(\theta_i)_{i\in V}$. The

first purpose of this paper is to propose a test of the null hypothesis "$\theta_{\mathcal{I}\setminus V} = 0$" against the general alternative "$\theta_{\mathcal{I}\setminus V} \neq 0$" under no prior knowledge of the covariance of X, the variance of ϵ, nor the variance of Y. Note that the property "$\theta_{\mathcal{I}\setminus V} = 0$" is equivalent to "$Y$ is independent of $X_{\mathcal{I}\setminus V}$ conditionally to X_V". Moreover, we want to be able to consider the difficult case of tests in a high dimensional setting: the number of covariates p is possibly much larger than the number of observations n. Such situations arise in many statistical applications like in genomics or biomedical imaging. Our issue of testing is the counterpart of the problem of estimation in a high dimensional setting considered for instance by Meinshausen and Bühlmann [6] or Candès and Tao [3]. Our second purpose is to use our procedure to derive non asymptotic minimax rates of testing over various alternatives.

4.1.1 Application to Gaussian Graphical Models (GGM)

Our work was originally motivated by the following question: let $(Z_j)_{j\in\mathcal{J}}$ be a random vector which follows a zero mean Gaussian distribution whose covariance matrix is non singular. We observe n i.i.d. replications of this vector Z and we are given a graph $\mathcal{G}$. How can we test that Z is a Gaussian graphical model (GGM) with respect to the graph $\mathcal{G}$? See Lauritzen [5] for definitions and main properties of GGM. This issue of testing in graphical modelling is extensively studied and it is beyond the scope of this paper to give an exhaustive review of it. See Drton and Perlman [4] for the description of various testing procedures. However, lots of them either cannot be applied in a high dimensional setting ($|\mathcal{J}| \geq n$) or lack of theoritical justifications in a non asymptotic way. In Chapter 5, we define a test based on the present work. It benefits the hability of handling high dimensional GGM and exhibits minimax properties. Besides we show numerical evidence of its efficiency; see Chapter 5 for more details. We only present here the idea underlying our approach.

For any $j \in \mathcal{J}$, we note $N(j)$ the set of neighbours of j in the graph $\mathcal{G}$. Testing that Z is a GGM with respect to $\mathcal{G}$ is equivalent to testing that the random variable Z_j conditionnally to $(Z_l)_{l\in N(j)}$ is independent of $(Z_l)_{l\in\mathcal{J}\setminus(N(j)\cup\{j\})}$ for any $j \in \mathcal{J}$. As Z follows a Gaussian distribution, the distribution of Z_j conditionally to the other variables decomposes as follows:

$$Z_j = \sum_{k\in\mathcal{J}\setminus\{j\}} \theta_k Z_k + \epsilon_j,$$

where ϵ_j is normal and independent of $(Z_k)_{k\in\mathcal{J}\setminus\{j\}}$. Then, the statement of conditional independency is equivalent to $\theta_{\mathcal{J}\setminus\{j\}\cup N(j)} = 0$. This approach based on conditional regression is also used for estimation by Meinshausen and Bühlmann [6].

4.1.2 Connection with tests in fixed design regression

Our work is directly inspired by the testing procedure of Baraud *et al.* [2] in fixed design regression framework. Contrary to model (4.1), the problem of hypothesis testing in fixed design regression has been extensively studied. That is why we will use the results in this framework as a benchmark for the theoritical bounds in our model (4.1). Let us define

this second regression model:

$$Y_i = f_i + \sigma\epsilon_i, \quad i \in \{1, \dots, N\}, \tag{4.2}$$

where f is an unknown vector of $\mathbb{R}^N$, σ some unknown positive number, and the ϵ_i's a sequence of i.i.d. standard Gaussian random variables. The problem at hand is testing that f belongs to a linear subspace of $\mathbb{R}^N$ against the alternative that it does not. We refer to Baraud *et al.* [2] for a short review of non parametric tests in this framework. Besides, we are interested in the performance of the procedures from a minimax perspective. To our knowledge, there has been no results in model (4.1). On the other hand, there are numerous papers on this issue in the fixed design regression model. First, we refer to the seminal work of Ingster (1993,a,b,c) which gives asymptotic minimax rates over non parametric alternatives. Our work is closely related to the results of Baraud [1] where he gives non asymptotic minimax rates of testing over ellipsoids or sparse signals. Throughout the paper, we highlight the link between the minimax rates in fixed and in random design.

4.1.3 Principle of our testing procedure

Let us briefly describe the idea underlying our testing procedure. Let m be a subset of $\mathcal{I} \setminus V$. We respectively define S_V and $S_{V\cup m}$ as the linear subspaces of $\mathbb{R}^{\mathcal{I}}$ such that $\theta_{\mathcal{I}\setminus V} = 0$, respectively $\theta_{\mathcal{I}\setminus(V\cup m)} = 0$. We note d and D_m for the cardinalities of V and m and N_m refers to $N_m = n - d - D_m$. If $N_m > 0$, we define the Fisher statistic ϕ_m by

$$\phi_m(\mathbf{Y}, \mathbf{X}) = \frac{N_m \|\Pi_{V\cup m}\mathbf{Y} - \Pi_V \mathbf{Y}\|_n^2}{D_m \|\mathbf{Y} - \Pi_{V\cup m}\mathbf{Y}\|_n^2}, \tag{4.3}$$

where Π_V refers to the orthogonal projection onto the space generated by the vectors $(\mathbf{X}_i)_{i\in V}$ and $\|.\|_n$ is the canonical norm in $\mathbb{R}^n$.

Let us consider a finite collection $\mathcal{M}$ of non empty subsets of $\mathcal{I} \setminus V$ such that for each $m \in \mathcal{M}$, $N_m > 0$. Our testing procedure consists of doing a Fisher test for each $m \in \mathcal{M}$. We define $\{\alpha_m, m \in \mathcal{M}\}$ a suitable collection of numbers in $]0, 1[$ (which possibly depends on $\mathbf{X}$). For each $m \in \mathcal{M}$, we do the Fisher test ϕ_m of level α_m of:

$$H_0 : \theta \in S_V \quad \text{against the alternative} \quad H_{1,m} : \theta \in S_{V\cup m} \setminus S_V$$

and we decide to reject the null hypothesis if one of those Fisher tests does.

The main advantage of our procedure is that it is very flexible in the choices of the model $m \in \mathcal{M}$ and in the choices of the weights $\{\alpha_m\}$. Consequently, if we choose a suitable collection $\mathcal{M}$, the test is powerful over a large class of alternatives as shown in Sections 4.3, 4.4, and 4.5.

Finally, let us mention that our procedure easily extends to the case where the expectation of the random vector (Y, X) is unknown. Let $\overline{\mathbf{X}}$ and $\overline{\mathbf{Y}}$ denote the projections of $\mathbf{X}$ and $\mathbf{Y}$ onto the unit vector $\mathbf{1}$. Then, one only has to apply the procedure to $(\mathbf{Y} - \overline{\mathbf{Y}}, \mathbf{X} - \overline{\mathbf{X}})$ and to replace d by $d+1$. The properties of the test remain unchanged and one can adapt all the proofs to the price of more technicalities.

4.1.4 Minimax rates of testing

In order to examine the quality of our tests, we will compare their performance with the minimax rates of testing. That is why we now define precisely what we mean by the (α, δ)-minimax rate of testing over a set Θ. We write $\mathbb{R}^{\mathcal{I}}$ for $\mathbb{R}^{\mathcal{I}}$ endowed with the euclidean norm

$$\|\theta\|^2 := \theta^t \Sigma \theta = \mathrm{var}\left(\sum_{i\in\mathcal{I}} \theta_i X_i\right) . \tag{4.4}$$

As ϵ and X are independent, we derive from the definition of $\|.\|^2$ that $\mathrm{var}(Y) = \|\theta\|^2 + \mathrm{var}(Y|X)$. Let us remark that $\mathrm{var}(Y|X)$ does not depend on X. If we make vary $\|\theta\|$, either the quantity $\mathrm{var}(Y)$ or $\mathrm{var}(Y|X)$ has to vary. In the sequel, we suppose that $\mathrm{var}(Y)$ is fixed. We briefly justify this choice in Section 4.4.2. Consequently, if $\|\theta\|^2$ is increasing, then $\mathrm{var}(Y|X)$ has to decrease so that the sum remains constant. Let α be a number in $]0;1[$ and let δ be a number in $]0;1-\alpha[$ (typically small). For a given vector θ, matrix Σ and $\mathrm{var}(Y)$, we denote $\mathbb{P}_\theta$ the joint distribution of $(\mathbf{Y}, \mathbf{X})$. For the sake of simplicity, we do not emphasize the dependence of $\mathbb{P}_\theta$ on $\mathrm{var}(Y)$ or Σ. Let ψ_α be a test of level α of the hypothesis "$\theta = 0$" against the hypothesis "$\theta \in \Theta \setminus 0$". In our framework, it is natural to measure the performance of ψ_α using the quantity $\rho\left(\psi_\alpha, \Theta, \delta, \mathrm{var}(Y), \Sigma\right)$ defined by:

$$\rho\left(\psi_\alpha, \Theta, \delta, \mathrm{var}(Y), \Sigma\right) =$$
$$\inf\left\{\rho > 0,\ \inf\left\{\mathbb{P}_\theta(\psi_\alpha = 1),\ \theta \in \Theta \text{ and } \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq \rho^2\right\} \geq 1-\delta\right\},$$

where the quantity

$$r_{s/n}(\theta) = \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \tag{4.5}$$

appears naturally as it corresponds to the ratio $\|\theta\|^2/\mathrm{var}(Y|X)$ which is the quantity of information brought by X (i.e. the signal) over the conditional variance of Y (i.e. the noise). We aim at describing the quantity

$$\inf_{\psi_\alpha} \rho\left(\psi_\alpha, \Theta, \delta, \mathrm{var}(Y), \Sigma\right) = \rho\left(\Theta, \alpha, \delta, \mathrm{var}(Y), \Sigma\right), \tag{4.6}$$

where the infimum is taken over all the level-α tests ψ_α. We call this quantity the (α, δ)-minimax rate of testing over Θ.

A dual notion of this ρ function is the function β_Σ. For any $\Theta \subset \mathbb{R}^{\mathcal{I}}$ and $\alpha \in]0,1[$, we denote $\beta_\Sigma(\Theta)$ the quantity

$$\beta_\Sigma\left(\Theta\right) = \inf_{\psi_\alpha} \sup_{\theta\in\Theta} \mathbb{P}_\theta\left[\psi_\alpha = 0\right],$$

where the infimum is taken over all level-α tests ψ_α and where we recall that Σ refers to the covariance matrix of X.

4.1.5 Organization of the Paper

We present the procedure in Section 4.2. In Section 4.3, we give a general theorem which characterizes a set of vectors θ over which the test is powerful in a non asymptotic setting. In Section 4.4 and 4.5, we apply our procedure to define tests and study their optimality for two different classes of alternatives. More precisely, in Section 4.4 we test 0 against the class of θ whose components equal 0, except at most k of them (k is supposed small). We define a test which under mild conditions achieves the minimax rate of testing. When the covariates are independent, it is interesting to note that the minimax rates exhibits the same ranges in our statistical model (4.1) and in fixed design regression model (4.2). In section 4.5, we define two procedures which achieve the simultaneous minimax rates of testing over large classes of ellipsoids (to sometimes the price of a $\log(p)$ factor). Besides, we show that the problem of adaptation over classes of ellipsoids is impossible without a loss in efficiency. This was previously pointed out by Spokoiny [7] in fixed design regression framework. The simulation studies are presented in Section 4.6. The proofs are presented in Annexe A.

Let us now introduce some notations that will be used throughout the paper. For $x, y \in \mathbb{R}$, we set

$$x \wedge y = \inf\{x, y\},\ x \vee y = \sup\{x, y\}.$$

For any $u \in \mathbb{R}$, $\bar{F}_{D,N}(u)$ denotes the probability for a Fisher variable with D and N degrees of freedom to be larger than u.

4.2 The Testing procedure

In this section, we adapt the testing procedure of Baraud *et al.* [2] in our random design model (4.1).

4.2.1 Description of the procedure

Let us first fix some level $\alpha \in]0,1[$. Throughout this paper, we suppose that $n \geq d+2$. Let us consider a finite collection $\mathcal{M}$ of non empty subsets of $\mathcal{I} \setminus V$ such that for all $m \in \mathcal{M}$, $1 \leq D_m \leq n-d-1$. Most of the notations used in this definition were defined in Section 4.1.3. We introduce the following test of level α. We reject H_0:"$\theta \in S_V$" when the statistic

$$T_\alpha = \sup_{m \in \mathcal{M}} \left\{ \phi_m(\mathbf{Y}, \mathbf{X}) - \bar{F}^{-1}_{D_m, N_m}(\alpha_m(\mathbf{X})) \right\} \tag{4.7}$$

is positive, where the collection of weights $\{\alpha_m(\mathbf{X}), m \in \mathcal{M}\}$ is chosen according to one of the two following procedures:

P_1 : The α_m 's do not depend on $\mathbf{X}$ and satisfy the equality :

$$\sum_{m \in \mathcal{M}} \alpha_m = \alpha\ . \tag{4.8}$$

P_2 : For all $m \in \mathcal{M}$, $\alpha_m(\mathbf{X}) = q_{\mathbf{X},\alpha}$, the α-quantile of the distribution of the random

variable

$$\inf_{m\in\mathcal{M}} \bar{F}_{D_m,N_m}\left(\frac{\|\Pi_{V\cup m}(\boldsymbol{\epsilon})-\Pi_V(\boldsymbol{\epsilon})\|_n^2/D_m}{\|\boldsymbol{\epsilon}-\Pi_{V\cup m}(\boldsymbol{\epsilon})\|_n^2/N_m}\right) \tag{4.9}$$

conditionally to $\mathbf{X}$.

Note that it is easy to compute the quantity $q_{\mathbf{X},\alpha}$. Let Z be a standard Gaussian random vector of size n independent of $\mathbf{X}$. As $\boldsymbol{\epsilon}$ is independent of $\mathbf{X}$, the distribution of (4.9) conditionally to $\mathbf{X}$ is the same as the distribution of

$$\inf_{m\in\mathcal{M}} \bar{F}_{D_m,N_m}\left(\frac{\|\Pi_{V\cup m}(Z)-\Pi_V(Z)\|^2/D_m}{\|Z-\Pi_{V\cup m}(Z)\|^2/N_m}\right)$$

conditionally to $\mathbf{X}$. Thus, we can easily work out its quantile using Monte-Carlo method.

4.2.2 Behavior of the test under the null hypothesis

The test associated with Procedure P_1 corresponds to a Bonferroni test and is of size less than α by arguing as follows: let θ be an element of S_V (defined in Section 4.1.3),

$$\mathbb{P}_\theta(T_\alpha>0) \quad\leq\quad \sum_{m\in\mathcal{M}} \mathbb{P}_\theta\left(\phi_m(\mathbf{X},\mathbf{Y})-\bar{F}^{-1}_{D_m,N_m}(\alpha_m)>0\right),$$

where $\phi_m(\mathbf{X},\mathbf{Y})$ is defined in (4.3). We now define the test statistic $\phi_{m,\alpha_m}(\mathbf{X},\mathbf{Y})$ as

$$\phi_{m,\alpha_m}(\mathbf{X},\mathbf{Y})=\phi_m(\mathbf{X},\mathbf{Y})-\bar{F}^{-1}_{D_m,N_m}(\alpha_m). \tag{4.10}$$

The test is rejected if $\phi_{m,\alpha_m}(\mathbf{X},\mathbf{Y})$ is positive. As θ belongs to S_V, $\Pi_{V\cup m}\mathbf{Y}-\Pi_V\mathbf{Y}=\Pi_{V\cup m}\boldsymbol{\epsilon}-\Pi_V\boldsymbol{\epsilon}$ and $\mathbf{Y}-\Pi_{V\cup m}\mathbf{Y}=\boldsymbol{\epsilon}-\Pi_{V\cup m}\boldsymbol{\epsilon}$. Then, the quantity $\phi_m(\mathbf{X},\mathbf{Y})$ is equal to

$$\phi_m(\mathbf{X},\mathbf{Y})=\frac{N_m\|\Pi_{V\cup m}\boldsymbol{\epsilon}-\Pi_V\boldsymbol{\epsilon}\|_n^2}{D_m\|\boldsymbol{\epsilon}-\Pi_{V\cup m}\boldsymbol{\epsilon}\|_n^2}.$$

Because $\boldsymbol{\epsilon}$ is independent of $\mathbf{X}$, the distribution of $\phi_m(\mathbf{X},\mathbf{Y})$ conditionally to $\mathbf{X}$ is a Fisher distribution with D_m and N_m degrees of freedom. As a consequence, $\phi_{m,\alpha_m}(\mathbf{X},\mathbf{Y})$ is a Fisher test with D_m and N_m degrees of freedom. It follows that:

$$\mathbb{P}_\theta(T_\alpha>0)\leq\sum_{m\in\mathcal{M}}\alpha_m\leq\alpha.$$

Procedure P_1 is therefore conservative.

The test associated with Procedure P_2 has the property to be of size exactly α. More precisely, for any $\theta\in S_V$, we have that

$$\mathbb{P}_\theta(T_\alpha>0|\mathbf{X})=\alpha \quad \mathbf{X}\text{ a.s. }.$$

The result follows from the fact that $q_{\mathbf{X},\alpha}$ satisfies

$$\mathbb{P}_\theta\left(\sup_{m\in\mathcal{M}}\left\{\frac{N_m\|\Pi_{V\cup m}(\boldsymbol{\epsilon})-\Pi_V(\boldsymbol{\epsilon})\|_n^2}{D_m\|\boldsymbol{\epsilon}-\Pi_{V\cup m}(\boldsymbol{\epsilon})\|_n^2}-\bar{F}^{-1}_{D_m,N_m}\left(q_{\mathbf{X},\alpha}\right)\right\}>0\middle|\,\mathbf{X}\right)=\alpha,$$

and that for any $\theta\in S_V$, $\Pi_{V\cup m}\mathbf{Y}-\Pi_V\mathbf{Y}=\Pi_{V\cup m}\boldsymbol{\epsilon}-\Pi_V\boldsymbol{\epsilon}$ and $\mathbf{Y}-\Pi_{V\cup m}\mathbf{Y}=\boldsymbol{\epsilon}-\Pi_{V\cup m}\boldsymbol{\epsilon}$.

4.2.3 Comparison of Procedures P_1 and P_2

We show in this section that the test (4.7) with Procedure P_2 is more powerful than the corresponding test defined with Procedure P_1 with weights $\alpha_m = \alpha/|\mathcal{M}|$. We respectively refer to T^1_α and T^2_α for these two tests associated with Procedure P_1 and P_2. More precisely, let us prove that

$$\forall \theta \notin S_V,\ \mathbb{P}_\theta\left(T^2_\alpha(\mathbf{X},\mathbf{Y}) > 0 \middle| \mathbf{X}\right) \geq \mathbb{P}_\theta\left(T^1_\alpha(\mathbf{X},\mathbf{Y}) > 0 \,|\, \mathbf{X}\right) \quad \mathbf{X}\ a.s.\ . \tag{4.11}$$

In fact this previous inequality is straighforward when considering the definitions of T^1_α and T^2_α:

$$\begin{aligned} T^1_\alpha(\mathbf{X},\mathbf{Y}) &= \sup_{m\in\mathcal{M}} \left\{\phi_m(\mathbf{X},\mathbf{Y}) - \bar{F}^{-1}_{D_m,N_m}(\alpha/|\mathcal{M}|)\right\} \\ T^2_\alpha(\mathbf{X},\mathbf{Y}) &= \sup_{m\in\mathcal{M}} \left\{\phi_m(\mathbf{X},\mathbf{Y}) - \bar{F}^{-1}_{D_m,N_m}(q_{\mathbf{X},\alpha})\right\} \end{aligned}$$

Conditionally on $\mathbf{X}$, the size of T^1_α is smaller than α, whereas the size T^2_α is exactly α. As a consequence $q_{\mathbf{X},\alpha} \geq \alpha/|\mathcal{M}|$ as the statistics T^1_α and T^2_α differ only through these quantities. Thus, $T^2_\alpha(\mathbf{X},\mathbf{Y}) \geq T^1_\alpha(\mathbf{X},\mathbf{Y})$, $(\mathbf{X},\mathbf{Y})$ almost surely and the result (4.11) follows.

The choice of Procedure P_1 allows to avoid the computation of the quantile $q_{\mathbf{X},\alpha}$ and possibly permits to give a Bayesian flavor to the choice of the weights. On the other hand, Procedure P_2 is more powerful than the corresponding test with Procedure P_1. This will be illustrated in Section 4.6. In the next three sections we study the power and rates of testing of T_α with Procedure P_1.

4.3 Power of the Test

In this section, we aim at describing a set of vectors θ in $\mathbb{R}^{\mathcal{I}}$ over which the test defined in Section 4.2 with Procedure P_1 is powerful. We note that since Procedure P_2 is more powerful than Procedure P_1 with $\alpha_m = \alpha/|\mathcal{M}|$, the test with Procedure P_2 will also be powerful on this set of θ.

Let α and δ be two numbers in $]0,1[$, and let $\{\alpha_m, m \in \mathcal{M}\}$ be weights such that $\sum_{m\in\mathcal{M}} \alpha_m \leq \alpha$. We introduce some quantities that depend on α_m, δ, D_m, and N_m. We set $L = \log\left(\frac{1}{\delta}\right)$ and for any $m \in \mathcal{M}$, we define $L_m = \log\left(\frac{1}{\alpha_m}\right)$, $k_m = 2\exp(4L_m/N_m)$, and $l_m = \left(1 + 2\sqrt{\frac{L}{N_m}} + 2\frac{L}{N_m}\right)$.

Under the following condition, k_m and l_m behave like constants:

$(H_{\mathcal{M}})$ For all $m \in \mathcal{M}$, $\alpha_m \geq \exp(-N_m/10)$ and $\delta \geq \exp(-N_m/21)$.

For typical choices of the collections $\mathcal{M}$ and $\{\alpha_m, m \in \mathcal{M}\}$, these conditions are fulfilled. In Sections 4.4 and 4.5, we discuss these assumptions for various settings. Let us now turn to the main result.

Theorem 3. *Let T_α be the test procedure defined by (4.7). We assume that $n > d+2$. Then $\mathbb{P}_\theta(T_\alpha > 0) \geq 1-\delta$ for all θ belonging to the set*

$$\mathcal{F}_{\mathcal{M}}(\delta) = \left\{ \theta \in \mathbb{R}^{\mathcal{I}}, \exists m \in \mathcal{M} : \frac{\mathrm{var}(Y|X_V) - \mathrm{var}(Y|X_{V\cup m})}{\mathrm{var}(Y|X_{V\cup m})} \geq \Delta(m) \right\},$$

where

$$\Delta(m) = \frac{4l_m\sqrt{D_m \log\left(\frac{1}{\alpha_m\delta}\right)}\left(1+\sqrt{\frac{D_m}{N_m}}\right) + \log\left(\frac{1}{\alpha_m\delta}\right)\left[8 \vee k_m l_m\left(1+2\frac{D_m}{N_m}\right)\right]}{(n-d)\left(1-1/2\vee 2\sqrt{\frac{2L}{N_m}}\right)}. \quad (4.12)$$

Under the hypothesis $H_{\mathcal{M}}$, for any $m \in \mathcal{M}$,

$$\Delta(m) \leq \frac{C_1\sqrt{D_m \log\left(\frac{1}{\alpha_m\delta}\right)}\left(1+\sqrt{\frac{D_m}{N_m}}\right) + C_2\left(1+2\frac{D_m}{N_m}\right)\log\left(\frac{1}{\alpha_m\delta}\right)}{n-d}, \quad (4.13)$$

where C_1 and C_2 are universal constants.

This result is similar to Theorem 1 in Baraud *et al.* [2] in fixed design regression framework and the same comment also holds: The test T_α under procedure P_1 has a power comparable to the best of the tests among the family $\{\phi_{m,\alpha}, m \in \mathcal{M}\}$. Indeed, let us assume for instance that $V = \{0\}$ and that the α_m are chosen to be equal to $\alpha/|\mathcal{M}|$. The test T_α defined by (4.7) is equivalent to doing several tests of $\theta = 0$ against $\theta \in S_m$ at level α_m for $m \in \mathcal{M}$ and it rejects the null hypothesis if one of those tests does. From Theorem 3, we know that under the hypothesis $H_{\mathcal{M}}$ this test has a power greater than $1-\delta$ over the set of vectors θ belonging to $\bigcup_{m\in\mathcal{M}} \mathcal{F}'_m(\delta,\alpha_m)$ where

$$\mathcal{F}'_m(\delta,\alpha_m) = \left\{ \theta \in \Theta, \frac{\mathrm{var}(Y) - \mathrm{var}(Y|X_m)}{\mathrm{var}(Y|X_m)} \geq \frac{C'_1(D_m,N_m)}{n}\left(\sqrt{D_m\log\left(\frac{1}{\alpha_m\delta}\right)} + \log\left(\frac{1}{\alpha_m\delta}\right)\right)\right\}. \quad (4.14)$$

Besides, $C'_1(D_m,N_m)$ behaves like a constant if the ratio D_m/N_m is bounded. Let us compare this result with the set of θ over which the Fisher test $\phi_{m,\alpha}$ at level α has a power greater than $1-\delta$. Applying Theorem 3, we know that it contains $\mathcal{F}'_m(\delta,\alpha)$. Moreover, the following Proposition shows that this set is not much larger than $\mathcal{F}'_m(\delta,\alpha)$:

Proposition 4. *Let $\delta \in]0, 1-\alpha[$ and*

$$t(\alpha,\delta) = \sqrt{\log\left(1+8\left(1-\alpha-\delta\right)^2\right)}\left[1 \wedge \sqrt{\log\left(1+8\left(1-\alpha-\delta\right)^2\right)/(2\log 2)}\right].$$

If

$$\frac{\mathrm{var}(Y) - \mathrm{var}(Y|X_m)}{\mathrm{var}(Y|X_m)} \leq t(\alpha,\delta)\frac{\sqrt{D_m}}{n},$$

then $\mathbb{P}_\theta(\phi_m > 0) \leq 1-\delta$.

The proof is postponed to Annexe A.2 and is based on a lower bound of the minimax rate of testing.

$\mathcal{F}'_m(\delta,\alpha)$ and $\mathcal{F}'_m(\delta,\alpha_m)$ defined in (4.14) differ from the fact that $\log(1/\alpha)$ is replaced by $\log(1/\alpha_m)$. For the main applications that we will study in section 4.4, 4.5, and 4.6, the difference $\log(1/\alpha_m) - \log(1/\alpha)$ is of order $k\log(ep/k)$ where k is a "small" integer of the order $\log(n)$ or $\log\log n$. Thus, for each $\delta \in]0, 1-\alpha[$, the test based on T_α has a power greater than $1-\delta$ over a class of vectors which is close to $\bigcup_{m\in\mathcal{M}} \mathcal{F}'_m(\delta,\alpha)$. It follows that for each $\theta \neq 0$ the power of this test under $\mathbb{P}_\theta$ is comparable to the best of the tests among the family $\{\phi_{m,\alpha}, m \in \mathcal{M}\}$.

In the next two sections, we use this theorem to establish rates of testing against different types of alternatives. First, we give an upper bound for the rate of testing $\theta = 0$ against a class of θ for which a lot of components are equal to 0. In Section 4.5, we study the rates of testing and simultaneous rates of testing $\theta = 0$ against classes of ellipsoids. For the sake of simplicity, we will only consider the case $V = \{0\}$. Nevertheless, the procedure T_α defined in (4.7) applies in the same way when one considers more complex null hypothesis and the rates of testing are unchanged except that we have to replace n by $n-d$ and $\mathrm{var}(Y)$ by $\mathrm{var}(Y|X_V)$.

4.4 Detecting non-zero coordinates

Let us fix a number k between 1 and p. In this section, we are interested in testing $\theta = 0$ against the class of θ with a most k non zero components. For each pair of integers (k,p) with $k \leq p$, let $\mathcal{M}(k,p)$ be the class of all subsets of $\mathcal{I} = \{1,\dots,p\}$ of cardinality k. The set $\Theta[k,p]$ stands for the subset of $\theta \in \mathbb{R}^{\mathcal{I}}$, such that at most k coordinates of θ are non-zero.

First, we define a test T_α of the form (4.7) with Procedure P_1, and we derive an upper bound for the rate of testing of T_α against the alternative $\theta \in \Theta[k,p]$. Then, we show that this procedure is rate optimal when all the covariates are independent. Finally, we study the optimality of the test when $k = 1$ for some examples of matrices Σ.

4.4.1 Rate of testing of T_α

Proposition 5. *We consider the set of models $\mathcal{M} = \mathcal{M}(k,p)$. We use the test T_α under Procedure P_1 and we take the weights α_m all equal to $\alpha/|\mathcal{M}|$. Let us suppose that n satisfies:*

$$n \geq k + \left[10\left[\log\left(\frac{1}{\alpha}\right) + k\log\left(\frac{ep}{k}\right)\right] \vee 21\log\left(1/\delta\right)\right].$$

Let us set the quantity

$$\rho'^2_{k,n,p} = \frac{C_3 k\log\left(\frac{ep}{k}\right) + C_4\left[\sqrt{k\log\left(\frac{1}{\alpha\delta}\right)} \vee \log\left(\frac{1}{\alpha\delta}\right)\right]}{n}, \tag{4.15}$$

where C_3 and C_4 are universal constants.
Then, for any θ in $\Theta[k,p]$, such that $\frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} \geq \rho'^2_{k,n,p}$, we have $\mathbb{P}_\theta(T_\alpha > 0) \geq 1-\delta$.

This Proposition follows easily from Theorem 3 and its proof is given in Annexe A.1. Let us note that this upper bound does not directly depend on the covariance matrix of the vector X. We will further discuss this result after deriving lower bounds for the minimax rate of testing in this setting.

4.4.2 Minimax lower bounds for independent covariates

In the statistical framework considered here, the problem of giving minimax rates of testing under no prior knowledge of the covariance of X and of $\mathrm{var}(Y)$ is open. That is why we shall only derive lower bounds when $\mathrm{var}(Y)$ and the covariance matrix of X are known. In this section, we give non asymptotic lower bounds for the (α, δ)-minimax rate of testing over the set $\Theta[k,p]$ when the covariance matrix of X is the identity matrix. As these bounds coincide with the upper bound obtained in Section 4.4.1, this will show that our test T_α is rate optimal.

In order to simplify the notations, we set $\eta = 2(1-\alpha-\delta)$ and $\mathcal{L}(\eta) = \frac{\log(1+2\eta^2)}{2}$. We first give a lower bound for the (α, δ)-minimax rate of detection of all p non-zero coordinates, as we will need it later.

Proposition 6. *Let us suppose that* $\mathrm{var}(Y)$ *is known. Let us set* $\rho^2_{p,n}$ *such that:*

$$\rho^2_{p,n} = \sqrt{2}\left[\sqrt{\mathcal{L}(\eta)} \wedge \frac{\mathcal{L}(\eta)}{\sqrt{\log(2)}}\right]\frac{\sqrt{p}}{n}. \tag{4.16}$$

Then for all $\rho < \rho_{p,n}$,

$$\beta_\Sigma\left(\left\{\theta \in \Theta[p,p], \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} = \rho^2\right\}\right) \geq \delta,$$

where we recall that Σ *is the covariance matrix of* X.

We now turn to the lower bound for the (α, δ)-minimax rate of testing against $\theta \in \Theta[k,p]$.

Theorem 7. *Let us set* $\rho^2_{k,p,n}$ *such that*

$$\rho^2_{k,p,n} = \frac{k(\mathcal{L}(\eta) \wedge 1)}{n}\log\left(1 + \frac{p}{k^2} + \sqrt{2\frac{p}{k^2}}\right). \tag{4.17}$$

Moreover, we suppose that the covariance of X *is the identity matrix* I. *Then, for all* $\rho < \rho_{k,n,p}$,

$$\beta_I\left(\left\{\theta \in \Theta[k,p], \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} = \rho^2\right\}\right) \geq \delta.$$

where the quantity $\mathrm{var}(Y)$ *is known.*

If $\alpha + \delta \leq 53\%$, *then one has*

$$\rho^2_{k,n,p} \geq \frac{k}{2n}\log\left(1 + \frac{p}{k^2} \vee \sqrt{\frac{p}{k^2}}\right).$$

This result implies the lower bound

$$\rho\left(\Theta[k,p],\alpha,\delta,\mathrm{var}(Y),I\right)\right) \geq \rho^2_{k,p,n}.$$

The proof is given in Annexe A.2. To the price of more technicity, it is possible to prove that the lower bound still holds if the variables (X_i) are assumed independent with known variances possibly different. This theorem recovers approximately the lower bounds for the minimax rates of testing in signal detection framework obtained by Baraud [1]. The main difference lies in the fact that we suppose $\mathrm{var}(Y)$ known which in the signal detection framework translates in the fact that we would know the quantity $\|f\|^2+\sigma^2$.

We are now in position to compare the results of Proposition 5 and Theorem 7. We distinguish between the values of k.

- When $k \leq p^\gamma$ for some $\gamma < 1/2$, if n is large enough to satisfy the assumption of Proposition 5, the quantities $\rho^2_{k,n,p}$ and $\rho'^2_{k,n,p}$ are both of the order $\frac{k\log(p)}{n}$ times a constant (which depends on γ, α, and δ). This shows that the lower bound given in Theorem 7 is sharp. Additionally, in this case, the procedure T_α defined in Proposition 5 follows approximately the minimax rate of testing. We recall that our procedure T_α does not depend on the knowledge of $\mathrm{var}(Y)$ and $\mathrm{corr}(X)$. In applications, this choice of a small k typically corresponds to testing a Gaussian graphical model with respect to a graph $\mathcal{G}$, when the number of nodes is large and the graph is supposed to be sparse.

- When $\sqrt{p} \leq k \leq p$, the lower bound and the upper bound do not coincide anymore. Nevertheless, if $n \geq (1+\gamma)p$ for some $\gamma > 0$, Theorem 3 shows that the test $\phi_{\mathcal{I},\alpha}$ defined in (4.10) has power greater than δ over the vectors θ which satisfy

$$\frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} \geq C(\gamma,\alpha,\delta)\frac{\sqrt{p}}{n}. \tag{4.18}$$

  This upper bound and the lower bound do not depend on k. Here again, the lower bound obtained in Theorem 7 is sharp and the test $\phi_{\mathcal{I},\alpha}$ defined previously is rate optimal. The fact that the rate of testing stabilizes around $\sqrt{p}/n$ for $k > \sqrt{p}$ also appears in signal detection and there is a discussion of this phenomenon in Baraud [1].

- When $k < \sqrt{p}$ and k is close to $\sqrt{p}$, the lower bound and the upper bound given by Proposition 5 differ from at most a $\log(p)$ factor. For instance, if k is of order $\sqrt{p}/\log p$, the lower bound in Theorem 7 is of order $\sqrt{p}\log\log p/\log p$ and the upper bound is of order $\sqrt{p}$. We do not know if any of this bound is sharp and if the minimax rates of testing coincide when $\mathrm{var}(Y)$ is fixed and when it is not fixed.

All in all, the minimax rates of testing exhibit the same range of rates in our framework as in signal detection (Baraud [1]) when the covariates are independent. Moreover, our result shows that the minimax rate of testing is slower when the $(X_i)_{i\in\mathcal{I}}$ are independent than for any form of dependence. Indeed, the upper bounds obtained in Proposition 5 and in (4.18) do not depend on the covariance of X. Then, a natural question arises: is the test statistic T_α rate optimal for other correlation of X? We will partially answer this question only when testing against the alternative $\theta \in \Theta[1,p]$.

4.4.3 Minimax rates for dependent covariates

In this section, we look for the minimax rate of testing $\theta = 0$ against $\theta \in \Theta[1,p]$ when the covariates X_i are no longer independent. We know that this rate is between the orders $\frac{1}{n}$, which is the minimax rate of testing when we know which coordinate is non-zero, and $\frac{\log(p)}{n}$, the minimax rate of testing for independent covariates.

Proposition 8. *Let us suppose that there exists a positive number c such that for any $i \neq j$,*

$$|corr(X_i, X_j)| \leq c.$$

We define $\rho^2_{1,p,n,c}$ as

$$\rho^2_{1,p,n,c} = \frac{1}{n}\left[\log(1+\eta^2 p) \wedge \frac{1}{c}\log\left(1+\eta^2\right)\right]. \tag{4.19}$$

Then for any $\rho < \rho_{1,p,n,c}$,

$$\beta_\Sigma\left(\left\{\theta \in \Theta[1,p], \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} = \rho^2\right\}\right) \geq \delta,$$

where Σ refers to the covariance matrix of X.

Remark: If the correlation between the covariates is smaller than $1/\log(p)$, then the minimax rate of testing is of the same order as in the independent case. If the correlation between the covariates is larger, we show in the following Proposition that under some additional assumption, the rate is faster.

Proposition 9. *Let us suppose that the correlation between X_i and X_j is exactly $c > 0$ for any $i \neq j$. Moreover n satisfies the following condition:*

$$n \geq \left[C_5\left(1+\log\left(\frac{2p}{\alpha}\right)\right)\right] \vee \left[C_6 \log\left(\frac{1}{\delta}\right)\right] \tag{4.20}$$

If $\alpha < 60\%$ and $\delta < 60\%$ the test T_α defined by

$$T_\alpha = \left[\sup_{2\leq i\leq p} \phi_{\{i\},\alpha/(2(p-1))}\right] \vee \phi_{\{1\},\alpha/2}$$

satisfies

$$\mathbb{P}_0\left(T_\alpha > 0\right) \leq \alpha \text{ and } \mathbb{P}_\theta\left(T_\alpha > 0\right) \geq 1-\delta,$$

for any θ in $\Theta[1,p]$ such that

$$\frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq \rho'^2_{1,n,p,c},$$

where

$$\rho'^2_{1,n,p,c} = \frac{C_7}{n}\left(\log\left(\frac{2p}{\alpha\delta}\right) \bigwedge \frac{1}{c}\log\left(\frac{2}{\alpha\delta}\right)\right), \tag{4.21}$$

and C_5, C_6, and C_7 are universal constants.

Consequently, when the correlation between X_i and X_j is a positive constant c, the minimax rate of testing is of order $\frac{\log(p)\wedge(1/c)}{n}$. When the correlation coefficient c is small, the minimax rate of testing coincides with the independent case, and when c is larger those rates differ. Therefore, the test T_α defined in Proposition 5 is not rate optimal when the correlation is known and is large. Indeed, when the correlation between the covariates is large, all the tests statistics ϕ_{m,α_m} defining T_α are highly correlated. The choice of the weights α_m in Procedure P_1 corresponds to a Bonferroni procedure. The loss due to a Bonferroni procedure is precisely large when the tests are positively correlated.

This example shows the limits of Procedure P_1. However, it is not very realistic to suppose that the covariates have a constant correlation, for instance when one considers a GGM. Indeed, we expect that the correlation between two covariates is large if they are neighbours in the graph and smaller if they are far (w.r.t. the graph distance). That is why we derive lower bounds of the rate of testing for other kind of correlation matrices often used to model stationary processes.

Proposition 10. *Let $X_1, \dots, X_p$ form a stationary process on the one dimensional torus. More precisely, the correlation between X_i and X_j is a function of $|i-j|_p$ where $|.|_p$ refers to the toroidal distance defined by:*

$$|i-j|_p = (|i-j|) \wedge (p - |i-j|)$$

$\Sigma_1(w)$ and $\Sigma_2(t)$ respectively refer to the correlation matrix of X such that

$$\begin{aligned} \mathrm{corr}(X_i, X_j) &= \exp(-w|i-j|_p) \text{ where } w > 0, \\ \mathrm{corr}(X_i, X_j) &= (1+|i-j|_p)^{-t} \text{ where } t > 0. \end{aligned}$$

Let us set $\rho^2_{1,p,n,\Sigma_1}(w)$ and $\rho^2_{1,p,n,\Sigma_2}(t)$ such that:

$$\begin{aligned} \rho^2_{1,p,n,\Sigma_1}(w) &= \frac{1}{n}\log\left(1 + 2p\eta^2\frac{1+e^{-w}}{1-e^{-w}}\right) \\ \rho^2_{1,p,n,\Sigma_2}(t) &= \begin{cases} \frac{1}{n}\log\left(1+\frac{p(t-1)\eta^2}{2}\right) & \text{if} \quad t > 1 \\ \frac{1}{n}\log\left(1+\frac{2p\eta^2}{1+2\log(p-1)}\right) & \text{if} \quad t = 1 \\ \frac{1}{n}\log\left(1+p^t 2^{-t}(1-t)\eta^2\right) & \text{if} \quad 0 < t < 1. \end{cases} \end{aligned}$$

Then, for any $\rho^2 \le \rho^2_{1,p,n,\Sigma_1}(w)$,

$$\beta_{\Sigma_1(w)}\left(\left\{\theta \in \Theta[1,p], \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} = \rho^2\right\}\right) \ge \delta,$$

and for any $\rho \le \rho^2_{1,p,n,\Sigma_2}(t)$,

$$\beta_{\Sigma_2(t)}\left(\left\{\theta \in \Theta[1,p], \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} = \rho^2\right\}\right) \ge \delta.$$

All in all, these lower bounds are of order $\frac{\log p}{n}$. As a consequence, for any of these correlation models the minimax rate of testing is of the same order as the minimax rate of testing for independent covariates. This means, that our test T_α defined in Proposition 5

is rate-optimal for these correlations matrices.

To conclude, when $k \leq p^\gamma$ (for $\gamma \leq 1/2$), the test T_α defined in Proposition 5 is approximately (α, δ)-minimax against the alternative $\theta \in \Theta[k,p]$, when neither $\mathrm{var}(Y)$ nor the covariance matrix of X is fixed. Indeed, the rate of testing of T_α coincide (up to a constant) with the following quantity:

$$\rho\left(\Theta[k,p], \alpha, \delta\right) = \sup_{\mathrm{var}(Y)>0, \Sigma>0} \rho\left(\Theta[k,p], \alpha, \delta, \mathrm{var}(Y), \Sigma\right),$$

where the supremum is taken over all positive $\mathrm{var}(Y)$ and every positive definite matrix Σ. When $k \geq \sqrt{p}$ and when $n \geq (1+\gamma)p$ (for $\gamma > 0$), the test defined in (4.18) has the same behavior.

However, our procedure does not adapt to Σ: for some correlation matrices (as for instance in Proposition 9), T_α with Procedure P_1 is not rate optimal. Nevertheless, we believe and this will be illustrated in Section 4.6 that Procedure P_2 slightly improves the power of the test in practice.

4.5 Rates of testing on "ellipsoids" and adaptation

In this section, we define tests T_α of the form (4.7) in order to test simultaneously $\theta = 0$ against θ belongs to some classes of ellipsoids. We will study their rates and show that they are optimal at sometimes the price of a $\log p$ factor. In this Section, $\mathcal{I}$ is supposed again to be $\{1, \ldots, p\}$. In the sequel for any non increasing sequence $(a_i)_{1 \leq i \leq p+1}$ such that $a_1 = 1$ and $a_{p+1} = 0$ and any $R > 0$, we define the ellipsoid $\mathcal{E}_a(R)$ by

$$\mathcal{E}_a(R) = \left\{ \theta \in \mathbb{R}^{\mathcal{I}}, \sum_{i=1}^{p} \frac{\mathrm{var}(Y|X_{m_{i-1}}) - \mathrm{var}(Y|X_{m_i})}{a_i^2} \leq R^2 \mathrm{var}(Y|X) \right\}, \tag{4.22}$$

where for any $1 \leq i \leq p$, m_i refers to the set $\{1, \ldots, i\}$ and $m_0 = \varnothing$.

Let us explain why we call this set an ellipsoid. For instance, let us suppose that the (X_i) are independent identically distributed with variance one. In this case, the difference $\mathrm{var}(Y|X_{m_{i-1}}) - \mathrm{var}(Y|X_{m_i})$ equals $|\theta_i|^2$ and the definition of $\mathcal{E}_a(R)$ translates in

$$\mathcal{E}_a(R) = \left\{ \theta \in \mathbb{R}^{\mathcal{I}}, \sum_{i=1}^{p} \frac{|\theta_i|^2}{a_i^2} \leq R^2 \mathrm{var}(Y|X) \right\}.$$

The main difference between this definition and the classical definition of an ellipsoid in the fixed design regression framework (as for instance in Baraud [1]) is the presence of the term $\mathrm{var}(Y|X)$. We added this quantity in order to be able to derive lower bounds of the minimax rate. If the X_i are not i.i.d. with unit variance, it is always possible to create a sequence X_i' of i.i.d. standard gaussian variables by orthogonalizing the X_i using the Gram-Schmidt process. If we call θ' the vector in $\mathbb{R}^{\mathcal{I}}$ such that $X\theta = X'\theta'$, it is straightforward to show that $\mathrm{var}(Y|X_{m_{i-1}}) - \mathrm{var}(Y|X_{m_i}) = |\theta_i'|^2$. We can then express $\mathcal{E}_a(R)$ using the coordinates of θ' as previously:

$$\mathcal{E}_a(R) = \left\{ \theta \in \mathbb{R}^{\mathcal{I}}, \sum_{i=1}^{p} \frac{|\theta_i'|^2}{a_i^2} \leq R^2 \mathrm{var}(Y|X) \right\}.$$

The main advantage of definition (4.22) is that it does not depend on the covariance of X.

In the sequel we also consider the special case of ellipsoids with polynomial decay,

$$\mathcal{E}'_s(R) = \left\{ \theta \in \mathbb{R}^{\mathcal{I}}, \sum_{i=1}^{p} \frac{\mathrm{var}(Y|X_{m_{i-1}}) - \mathrm{var}(Y|X_{m_i})}{i^{-2s}\mathrm{var}(Y|X)} \leq R^2 \right\}, \tag{4.23}$$

where $s > 0$ and $R > 0$. First, we define two tests procedures of the form (4.7) and evaluate their power respectively on the ellipsoids $\mathcal{E}_a(R)$ and on the ellipsoids $\mathcal{E}'_s(R)$. Then, we give some lower bounds for the (α, δ)-simultaneous minimax rates of testing. Extensions to more general l_p balls with $0 < p < 2$ are possible to the price of more technicalities by adapting the results of Section 4 in Baraud [1].

4.5.1 Simultaneous Rates of testing of T_α over classes of ellipsoids

First, we define a test of the form (4.7) in order to test $\theta = 0$ against θ belongs to any of the ellipsoids $\mathcal{E}_a(R)$. For any $x > 0$, $[x]$ denotes the integer part of x.

The class of models $\mathcal{M}$ and the weights α_m depend on n and p:

- If $n < 2p$, we take the set $\mathcal{M}$ to be $\cup_{1 \leq k \leq [n/2]} m_k$ and all the weights α_m are equal to $\alpha/|\mathcal{M}|$.
- If $n \geq 2p$, we take the set $\mathcal{M}$ to be $\cup_{1 \leq k \leq p} m_k$. α_{m_p} equals $\alpha/2$ and for any k between 1 and $p-1$, α_{m_k} is chosen to be $\alpha/(2(p-1))$.

Proposition 11. *Let us assume that*

$$n \geq 42 \left(\log\left(\frac{40}{\alpha}\right) \vee \log\left(\frac{1}{\delta}\right) \right) \tag{4.24}$$

For any ellipsoid $\mathcal{E}_a(R)$, the test T_α defined by (4.7) with Procedure P_1 and with the class of models given just above satisfies

$$\mathbb{P}_0\left(T_\alpha \leq 0\right) \geq 1 - \alpha,$$

and $\mathbb{P}_\theta\left(T_\alpha > 0\right) \geq 1 - \delta$ for all $\theta \in \mathcal{E}_a(R)$ such that

$$\frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq C_8 \left(\inf_{1 \leq i \leq [n/2]} \left[a_{i+1}^2 R^2 + \frac{\sqrt{i \log\left(\frac{n/2}{\alpha\delta}\right)}}{n} \right] + \frac{1}{n} \log\left(\frac{n/2}{\alpha\delta}\right) \right) \tag{4.25}$$

if $n < 2p$, or

$$\frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq C_8 \left[\inf_{1 \leq i \leq p-1} \left[a_{i+1}^2 R^2 + \frac{\sqrt{i \log\left(\frac{2(p-1)}{\alpha\delta}\right)}}{n} \right] + \frac{\log\left(\frac{2(p-1)}{\alpha\delta}\right)}{n} \right] \bigwedge \left[\frac{\sqrt{p \log\left(\frac{2}{\alpha\delta}\right)} + \log\left(\frac{2}{\alpha\delta}\right)}{n} \right] \tag{4.26}$$

if $n \geq 2p$.

All in all, for large values of n, the rate of testing is of the order

$$\sup_{1\leq i\leq p}\left[a_i^2R^2\wedge\frac{\sqrt{i\log(p)}}{n}\right].$$

We will show in next subsection that the minimax rate of testing for an ellipsoid is of order:

$$\sup_{1\leq i\leq p}\left[a_i^2R^2\wedge\frac{\sqrt{i}}{n}\right].$$

Besides, we will show in Proposition 16 that a loss in $\sqrt{\log\log p}$ is unavoidable if one considers the simultaneous minimax rates of testing over a family of nested ellipsoids. We do not know if the term $\sqrt{\log(p)}$ is optimal for testing simultaneously against all the ellipsoids $\mathcal{E}_a(R)$ for all sequences (a_i) and all $R>0$. When n is smaller than $2p$, we obtain comparable results except that we are unable to consider alternatives in large dimensions.

We now turn to define a procedure of the form (4.7) in order to test simultaneously that $\theta=0$ against θ belongs to any of the $\mathcal{E}'_s(R)$. For this, we introduce the following collection of models $\mathcal{M}$ and weights α_m:

- If $n<2p$, we take the set $\mathcal{M}$ to be $\cup m_k$ where k belongs to $\left\{2^j,j\geq 0\right\}\cap\{1,\dots,[n/2]\}$ and all the weights α_m are chosen to be $\alpha/|\mathcal{M}|$.
- If $n\geq 2p$, we take the set $\mathcal{M}$ to be $\cup m_k$ where k belongs to $\left(\left\{2^j,j\geq 0\right\}\cap\{1,\dots,p\}\right)\cup\{p\}$, α_{m_p} equals $\alpha/2$ and for any k in the model between 1 and $p-1$, α_{m_k} is chosen to be $\alpha/(2(|\mathcal{M}|-1))$.

Proposition 12. *Let us assume that*

$$n\geq 42\left(\log\left(\frac{40}{\alpha}\right)\vee\log\left(\frac{1}{\delta}\right)\right) \tag{4.27}$$

and that $R^2\geq\sqrt{\log\log n}/n$. *For any* $s>0$, *the test procedure* T_α *defined by (4.7) with Procedure* P_1 *and with a class of models given just above satisfies:*

$$\mathbb{P}_0\left(T_\alpha>0\right)\leq 1-\alpha,$$

and $\mathbb{P}_\theta\left(T_\alpha>0\right)\geq 1-\delta$ *for any* $\theta\in\mathcal{E}'_s(R)$ *such that*

$$\frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2}\geq C_9\left(\alpha,\delta\right)\left[R^{2/(1+4s)}\left(\frac{\sqrt{\log\log n}}{n}\right)^{4s/(1+4s)}+R^2\left(n/2\right)^{-2s}+\frac{\log\log n}{n}\right] \tag{4.28}$$

if $n<2p$ *or*

$$\frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2}\geq C_9\left(\alpha,\delta\right)\left(\left[R^{2/(1+4s)}\left(\frac{\sqrt{\log\log p}}{n}\right)^{4s/(1+4s)}+\frac{\log\log p}{n}\right]\bigwedge\frac{\sqrt{p}}{n}\right) \tag{4.29}$$

if $n\geq 2p$. $C_9(\alpha,\delta)$ *is a constant which only depends on* α *and* δ.

Again, we retrieve similar results to those of Corollary 2 in Baraud et al. [2] in the fixed design regression framework. For $s > 1/4$ and $n < 2p$, the rate of testing is of order $\left(\frac{\sqrt{\log\log n}}{n}\right)^{4s/(1+4s)}$. We show in the next subsection that this logarithmic factor is due to the adaptative property of the test. If $s \leq 1/4$, the rate is of order n^{-2s}. When $n \geq 2p$, the rate is of order $\left(\frac{\sqrt{\log\log p}}{n}\right)^{4s/(1+4s)} \wedge \left(\frac{\sqrt{p}}{n}\right)$, and we mention at the end of the next subsection that it is optimal.

Here again, it is possible to define these tests with Procedure P_2 in order to improve the power of the test (see Section 4.6 for numerical results).

4.5.2 Minimax lower bounds

We first establish the (α, δ)-minimax rate of testing over an ellipsoid when the variance of Y and the covariance matrix of X are known.

Proposition 13. *Let us set the sequence $(a_i)_{1\leq i\leq p+1}$ and the positive number R. We introduce*

$$\rho^2_{a,n}(R) = \sup_{1\leq i\leq p} \left[\rho^2_{i,n} \wedge a_i^2 R^2\right], \tag{4.30}$$

where $\rho^2_{i,n}$ is defined by (4.16), then for any non singular covariance matrix Σ we have

$$\beta_\Sigma\left(\left\{\theta \in \mathcal{E}_a(R), \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq \rho^2_{a,n}(R)\right\}\right) \geq \delta,$$

where the quantity var(Y) is fixed. If $\alpha + \delta \leq 47\%$ then

$$\rho^2_{a,n}(R) \geq \sup_{1\leq i\leq p}\left[\frac{\sqrt{i}}{n} \wedge a_i^2 R^2\right].$$

This lower bound is once more analogous to the one in the fixed design regression framework. Contrary to the lower bounds in previous section, this bound does not depend on the covariance of the covariates. We now look for an upper bound of the minimax rate of testing over a given ellipsoid. First, we need to define the quantity D^* as:

$$D^* = \inf\left\{1 \leq i \leq p, a_i^2 R^2 \leq \frac{\sqrt{i}}{n}\right\}$$

with the convention that $\inf \varnothing = p$.

We get the corresponding upper bound only if D^* is not too large compared to n, as shown by the following proposition.

Proposition 14. *Let us assume that $n \geq 20\log\left(\frac{1}{\alpha}\right) \vee 41\log\left(\frac{2}{\delta}\right)$. If $R^2 > \frac{1}{n}$ and $D^* \leq n/2$, the test $\phi_{m_{D^*},\alpha}$ defined by (4.10) satisfies*

$$\mathbb{P}_0\left[\phi_{m_{D^*},\alpha} = 1\right] \leq \alpha \text{ and } \mathbb{P}_\theta\left[\phi_{m_{D^*},\alpha} = 0\right] \leq \delta$$

for all $\theta \in \mathcal{E}_a(R)$ *such that*

$$\frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq C_{10}(\alpha, \delta) \sup_{1 \leq i \leq d} \left[\frac{\sqrt{i}}{n} \wedge a_i^2 R \right],$$

where $C_{10}(\alpha, \delta)$ *is a constant which only depends on* α *and* δ.

If $n \geq 2D^*$, the rates of testing on an ellipsoid are analogous to the rates on an ellipsoid in fixed design regression framework (see for instance Baraud [1]). If D^* is large and n is small, the bounds in Proposition 13 and 14 do not coincide. In this case, we do not know if this comes from the fact that the test in Proposition 14 does not depend on the knowledge of $\mathrm{var}(Y)$ or if one of the bounds in Proposition 13 and 14 is not sharp.

We are now interested in computing lower bounds of rates of testing simultaneously over a family of ellipsoids, in order to compare them with rates obtained in Section 4.5.1. First, we need a lower bound for the minimax simultaneous rates of testing over nested linear spaces. We recall that for any $D \in \{1, \ldots, p\}$, S_{m_D} stands for the linear spaces of vectors θ such that only their D first coordinates are possibly non zero.

Proposition 15. *For* $D \geq 2$, *let us set*

$$\bar{\rho}^2_{D,n} = \frac{1}{2\sqrt{\log(2)}} \left(1 \wedge \log(1 + 2\eta^2)\right) \frac{\sqrt{\log\log D + 1}\sqrt{D}}{n}. \tag{4.31}$$

Then, the following lower bound holds

$$\beta_I \left(\bigcup_{1 \leq D \leq p} \left\{ \theta \in S_{m_D}, \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} = r_D^2 \right\} \right) \geq \delta,$$

if for all D *between* 1 *and* p, $r_D \leq \bar{\rho}_{D,n}$

Using this Proposition, it is possible to get a lower bound for the simultaneous rate of testing over a family of nested ellipsoids.

Proposition 16. *We fix a sequence* $(a_i)_{1 \leq i \leq p+1}$. *For each* $R > 0$, *let us set*

$$\bar{\rho}^2_{a,R,n} = \sup_{1 \leq D \leq p} \left[\bar{\rho}^2_{D,n} \wedge (R^2 a_D^2) \right]. \tag{4.32}$$

where $\bar{\rho}_{D,n}$ *is given by (4.31). Then, for any non singular covariance matrix* Σ *of the vector* X,

$$\beta_\Sigma \left(\bigcup_{R>0} \left\{ \theta \in \mathcal{E}_a(R), \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \leq \bar{\rho}^2_{a,R,n} \right\} \right) \geq \delta.$$

This Proposition shows that the problem of adaptation is impossible in this setting: it is impossible to define a test which is simultaneously minimax over a class of nested ellipsoids (for $R > 0$). This is also the case in fixed design as proved by Spokoiny [7] for the case of Besov bodies. The loss of a term of the order $\sqrt{\log\log p}/n$ is unavoidable.

As a special case of Proposition 16, it is possible to compute a lower bound for the simultaneous minimax rate over $\mathcal{E}_s(R)$ where R describes the positive numbers. After calculation, we find that the lower bound is of the order:

$$\left(\frac{\log\log p}{n}\right)^{\frac{4s}{1+4s}} \bigwedge \frac{\sqrt{p\log\log p}}{n}.$$

This shows that the power of the test T_α obtained in (4.29) for $n \geq 2p$ is optimal when $R^2 \geq \sqrt{\log\log n}/n$. However, when $n < 2p$ and $s \leq 1/4$, we do not know if the rate n^{-2s} is optimal or not.

To conclude, when $n \geq 2p$ the test T_α defined in Proposition 12 is rate optimal over the classes of ellipsoids $\mathcal{E}'_s(R)$. On the other hand, the test T_α defined in Proposition 11 is not rate optimal simultaneously over all the ellipsoids $\mathcal{E}_a(R)$ and suffers a loss of a $\sqrt{\log p}$ factor even when $n \geq 2p$.

4.6 Simulations studies

The purpose of this simulation study is threefold. First, we illustrate the theoretical results established in previous sections. Second, we show that our procedure is easy to implement for different choices of collections $\mathcal{M}$. Our third purpose is to compare the efficiency of Procedures P_1 and P_2. Indeed, for a given collection $\mathcal{M}$, we know from Section 4.2.3 that the test (4.7) based on Procedure P_2 is more powerful than the corresponding test based on P_1. However, the computation of the quantity $q_{\mathbf{X},\alpha}$ is possibly time consuming and we therefore want to know if the benefit in power is worth the computation time.

To our knowledge, when the number of covariates p is larger than the number of observations n there is no test with which we can compare our procedure.

4.6.1 Simulation experiments

We consider the regression model (4.1) with $\mathcal{I} = \{1, \ldots, p\}$ and test the null hypothesis $''\theta = 0''$, which is equivalent to "Y is independent of X", at level $\alpha = 5\%$. Let $(X_i)_{1\leqslant i\leqslant p}$ be a collection of p Gaussian variables with unit variance. The random variable is defined as follows: $Y = \sum_{i=1}^p \theta_i X_i + \varepsilon$ where ε is a zero mean gaussian variable with variance $1 - \|\theta\|^2$ independent of X.

We consider two simulation experiments described below.

1. First simulation experiment: The correlation between X_i and X_j is a constant c for any $i \neq j$. Besides, in this experiment the parameter θ is chosen such that only one of its components is possibly non zero. This corresponds to the situation considered in Section 4.4. First, the number of covariates p is fixed equal to 30 and the number of observations n is taken equal to 10 and 15. We choose for c three different values 0, 0.1, and 0.8, allowing thus to compare the procedure for independent, weakly and highly correlated covariates. We estimate the level of the test by taking $\theta_1 = 0$ and the power by taking for θ_1 the values 0.8 and 0.9. Theses choices of θ lead to a small and a large signal/noise ratio $r_{s/n}$ defined in (4.5) and equal in this experiment to $\theta_1^2/(1-\theta_1^2)$. Second, we examine the behavior of the tests when p increases and when the covariates are highly correlated: p equals 100 and 500, n equals 10 and 15, θ_1 is set to 0 and 0.8, and c is chosen to be 0.8.

2. Second simulation experiment: The covariates $(X_i)_{1\leqslant i\leqslant p}$ are independent. The number of covariates p equals 500 and the number of observations n equals 50 and 100. We set for any $i \in \{1,\dots,p\}$, $\theta_i = Ri^{-s}$. We estimate the level of the test by taking $R = 0$ and the power by taking for (R,s) the value $(0.2, 0.5)$, which corresponds to a slow decrease of the $(\theta_i)_{1\leqslant i\leqslant p}$. It was pointed out in the beginning of Section 4.5 that $|\theta_i|^2$ equals $\mathrm{var}(Y|X_{m_{i-1}}) - \mathrm{var}(Y|X_{m_i})$. Thus, $|\theta_i|^2$ represents the benefit in term of conditional variance brought by the variable X_i.

We use our testing procedure defined in (4.7) with different collections $\mathcal{M}$ and different choices for the weights $\{\alpha_m, m \in \mathcal{M}\}$.

The collections $\mathcal{M}$: we define three classes. Let us set $J_{n,p} = p \wedge [\frac{n}{2}]$, where $[x]$ denotes the integer part of x and let us define:

$$\begin{aligned}\mathcal{M}^1 &= \{i, 1 \leqslant i \leqslant p\}\\ \mathcal{M}^2 &= \{m_k = (1,2,\dots,k), 1 \leqslant k \leqslant J_{n,p}\}\}\\ \mathcal{M}^3 &= \{m_k = (1,2,\dots,k), k \in \{2^j, j \geq 0\} \cap \{1,\dots,J_{n,p}\}\}\end{aligned}$$

We evaluate the performance of our testing procedure with $\mathcal{M} = \mathcal{M}^1$ in the first simulation experiment, and $\mathcal{M} = \mathcal{M}^2$ and $\mathcal{M}^3$ in the second simulation experiment.

The collections $\{\alpha_m, m \in \mathcal{M}\}$: We consider Procedures P_1 and P_2 defined in section 4.2. When we are using the procedure P_1, the α_m's equal $\alpha/|\mathcal{M}|$ where $|\mathcal{M}|$ denotes the cardinality of the collection $\mathcal{M}$. The quantity $q_{\mathbf{X},\alpha}$ that occurs in the procedure P_2 is computed by simulation. We use 1000 simulations for the estimation of $q_{\mathbf{X},\alpha}$. In the sequel we note $T_{\mathcal{M}^i,P_j}$ the test (4.7) with collection $\mathcal{M}^i$ and Procedure P_j.

In the first experiment, when p is large we also consider two other tests

1. The test $\phi_{\{1\},\alpha}$ (defined in Equation 4.10) of the hypothesis $\theta_1 = 0$ against the alternative $\theta_1 \neq 0$. This test corresponds to the single test when we know which coordinate is non zero.

2. The test $\phi_{\{2\},\alpha}$ of $\theta_2 = 0$ against $\theta_2 \neq 0$. This test corresponds to a single test where the model under the alternative is wrong. Adapting the proof of Proposition 9, we know that this test is approximately minimax on $\Theta[1,p]$ if the correlation between the covariates is constant and large. There is nothing special about the number 2, we could use any i between 2 and p.

Contrary to our procedures, these two tests are based on a deep knowledge of θ or $\mathrm{var}(X)$. We only use them as a benchmark to evaluate the performance of our procedure. We aim at showing that our test with Procedure P_2 is more powerful than $\phi_{\{2\},\alpha}$ and is close to the test $\phi_{\{1\},\alpha}$.

We estimate the level and the power of the testing procedures with 1000 simulations. For each simulation, we simulate the gaussian vector $(X_1,\dots,X_p)$ and then simulate the variable Y as described in the two simulation experiments.

Null hypothesis is true, $\theta_1 = 0$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.043	0.045
15	0.044	0.049

Null hypothesis is false

$\theta_1 = 0.8$, $r_{s/n} = 1.78$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.48	0.48
15	0.81	0.81

$\theta_1 = 0.9$, $r_{s/n} = 4.26$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.86	0.86
15	0.99	0.99

Table 4.1: First simulation study, independent case: $p = 30$, $c = 0$. Percentages of rejection and value of the signal/noise ratio $r_{s/n}$.

4.6.2 Results of the simulation

The results of the first simulation experiment for $c = 0$ are given in Table 4.1. As expected, the power of the tests increases with the number of observations n and with the signal/noise ratio $r_{s/n}$. If the signal/noise ratio is large enough, we obtain powerful tests even if the number of covariates p is larger than the number of obervations.

In Table 4.2 we present results of the first simulation experiment for $\theta_1 = 0.8$ when c varies.
Let us first compare the results for independent, weakly and highly correlated covariates when using Procedure P_1. The level and the power of the test for weakly correlated covariates are similar to the level and the power obtained in the independent case. Hence, we recover the remark following Proposition 8: when the correlation coefficient between the covariates is small, the minimax rate is of the same order as in the independent case. The test for highly correlated covariates is more powerful than the test for independent covariates, recovering thus the remark following Theorem 7: the worst case from a minimax rate perspective is the case where the covariates are independent. Let us now compare Procedures P_1 and P_2. In the case of independent or weakly correlated covariates, they give similar results. For highly correlated covariates, the power of $T_{\mathcal{M}^1,P_2}$ is much larger than the one of $T_{\mathcal{M}^1,P_1}$.

In Table 4.3 we present results of the multiple testing procedure and of the two tests $\phi_{\{1\},\alpha}$ and $\phi_{\{2\},\alpha}$ when $c = 0.8$ and the number of covariates p is large. As expected, because the collection of models $\mathcal{M}^1$ depends on p, Procedure P_1 is too conservative when p increases. For $p = 100$, the power of the test based on Procedure P_1 is similar to the power of the test $\phi_{\{2\},\alpha}$, while when p is larger, $T_{\mathcal{M}^1,P_1}$ is less powerful than $\phi_{\{2\},\alpha}$. Procedure P_2 is therefore recommanded in case of a large number of highly correlated covariates. The test based on Procedure P_2 is indeed more powerful than $\phi_{\{2\},\alpha}$, and its power is close to the one of $\phi_{\{1\},\alpha}$. We recall that this last test is based on the knowledge of the non-zero component of θ contrary to ours. In practice, we advise to use Procedure P_2 if the number of covariates p is large, as Procedure P_1 becomes too conservative, especially if the covariates are correlated.

Null hypothesis is true, $\theta_1 = 0$

$c = 0$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.043	0.045
15	0.044	0.049

$c = 0.1$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.042	0.04
15	0.058	0.06

$c = 0.8$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.018	0.045
15	0.019	0.052

Null hypothesis is false, $\theta_1 = 0.8$

$c = 0$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.48	0.48
15	0.81	0.81

$c = 0.1$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.49	0.49
15	0.81	0.82

$c = 0.8$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.64	0.77
15	0.89	0.94

Table 4.2: First simulation study, independent and dependent case. $p = 30$ Percentages of rejection.

Null hypothesis is true, $\theta_1 = 0$

$p = 100$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$	$\phi_{\{1\},\alpha}$	$\phi_{\{2\},\alpha}$
10	0.01	0.056	0.051	0.050
15	0.016	0.053	0.047	0.050

$p = 500$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$	$\phi_{\{1\},\alpha}$	$\phi_{\{2\},\alpha}$
10	0.009	0.044	0.040	0.043
15	0.011	0.040	0.042	0.039

Null hypothesis is false, $\theta_1 = 0.8$

$p = 100$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$	$\phi_{\{1\},\alpha}$	$\phi_{\{2\},\alpha}$
10	0.60	0.77	0.91	0.62
15	0.85	0.92	0.99	0.82

$p = 500$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$	$\phi_{\{1\},\alpha}$	$\phi_{\{2\},\alpha}$
10	0.52	0.76	0.91	0.63
15	0.77	0.94	0.99	0.83

Table 4.3: First simulation study, dependent case: $c = 0.8$. Percentages of rejection.

Null hypothesis is true, $R = 0$

n	$T_{\mathcal{M}^2,P_1}$	$T_{\mathcal{M}^2,P_2}$	$T_{\mathcal{M}^3,P_1}$	$T_{\mathcal{M}^3,P_2}$
50	0.013	0.052	0.036	0.059
100	0.009	0.059	0.042	0.059

Null hypothesis is false, $R = 0.2, s = 0.5$

n	$T_{\mathcal{M}^2,P_1}$	$T_{\mathcal{M}^2,P_2}$	$T_{\mathcal{M}^3,P_1}$	$T_{\mathcal{M}^3,P_2}$
50	0.17	0.33	0.31	0.38
100	0.42	0.66	0.62	0.69

Table 4.4: Second simulation study. Percentages of rejection.

The results of the second simulation experiment are given in Table 4.4. As expected, Procedure P_2 improves the power of the test and the test $T_{\mathcal{M}^3,P_2}$ has the greatest power. In this setting, one should prefer the collection $\mathcal{M}^3$ to $\mathcal{M}^2$. This was previously pointed out in Section 4.5 from a theoretical point of view. Although $T_{\mathcal{M}^3,P_1}$ is conservative, it is a good compromise for practical issues: it is very easy and fast to implement and its performances are good.

References

[1] Baraud, Y. Non-asymptotic rates of testing in signal detection. *Bernoulli 8*, 5 (2002), 577–606.

[2] Baraud, Y., Huet, S., and Laurent, B. Adaptative tests of linear hypotheses by model selection. *Annals of Statistics 31*, 1 (2003), 225–251.

[3] Candès, E., and Tao, T. The dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics (in press)* (2007).

[4] Drton, M., and Perlman, M. Mutiple testing and error control in Gaussian graphical model selection. *arXiv.math/0508267* (2007).

[5] Lauritzen, S. L. *Graphical Models*. Oxford University Press, New York, 1996.

[6] Meinshausen, N., and Bühlmann, P. High dimensional graphs and variable selection with the Lasso. *Annals of Statistics 34*, 3 (2006), 1436–1462.

[7] Spokoiny, V. G. Adaptative hypothesis testing using wavelets. Annals of Statistics *24* (1996), 2477–2498.

Chapter 5

Tests for Gaussian graphical models

Ce chapitre présente un travail réalisé en collaboration avec Nicolas Verzelen

Abstract

In this chapter, we construct procedures for testing the graph of a Gaussian graphical model. Our approach is based on the connection between local Markov property and conditional regression of a Gaussian random variable. Thus, we adapt the testing procedures defined in Chapter 4 to this framework. These new tests then share the appealing theoretical properties proved in Chapter 4. Besides, they are able to handle the important issue of graph testing in a high dimensional setting: the number of observations may be much smaller than the number of nodes. Finally, a large part of this study is deserved to illustrate and discuss the application of our procedures to simulated data and to biological data.

5.1 Introduction

Consider $X = (X_1, \ldots, X_p)^t$ a random vector distributed as a multivariate Gaussian $\mathcal{N}(0, \Sigma)$. Throughout this paper, we assume that the matrix Σ is non-singular. The conditional independence structure of this distribution can be represented by an undirected graph $\mathcal{G} = (\Gamma, E)$ where $\Gamma = \{1, \ldots, p\}$ is the set of nodes and E the set of edges. There is an edge between nodes a and b if and only if the random variables X_a and X_b are conditionally dependent given all remaining variables $X_{-\{a,b\}} = \{X_i, i \in \Gamma \setminus \{a, b\}\}$. The random vector X is then said to be a Gaussian graphical model with respect to the graph $\mathcal{G}$. Given a node $a \in \Gamma$, we define its neighborhood $ne(a)$ as the set of nodes $b \in \Gamma \setminus \{a\}$ such that $(a, b) \in E$. We say that X follows the local Markov property at node a with respect to the graph $\mathcal{G}$ if X_a is independent from $\{X_i, i \in \Gamma \setminus (ne(a) \cup \{a\})\}$ given $\{X_i, i \in ne(a)\}$. Lauritzen [9] shows that X is a Gaussian graphical model with respect to $\mathcal{G}$ if and only if it follows the local Markov property at each node $a \in \Gamma$.

Graphical models are widely used in spatial statistics (e.g. Cressie [2], Chs. 6 and 7), image analysis Geman and Geman [8], probabilistic expert systems Cowell *et al.* [1].

Recently, they have been applied to the analysis of biological data. A large literature is now devoted to the study of gene expression data from microarray experiment. Indeed, biologists aim at inferring the network regulating the expression of the genes under study. In the graphical context, we assume that these expression data follow the distribution of a Gaussian graphical model with respect to the genetic network. Given a n sample of these gene expression quantities, one aims at estimating the underlying graph. However, these data particularly suffer from the curse of dimensionality: whereas the number of genes greatly increases, the number of samples stays relatively small. There is an active research in building and studying algorithms to perform the graph selection in such a high dimensional setting (see for instance Meinshausen and Bühlmann [10] ; Wille and Bühlmann [13]; Schäfer and Strimmer [12]).

In many applications, the biologists have at least a partial knowledge of the genetic network and they want to assess the quality of their model thanks to gene expression data. That is the reason why we are interested in the problem of testing the graph of a Gaussian graphical model. More precisely, suppose we are given a n-sample of the vector X and an undirected graph $\mathcal{G} = (\Gamma, E)$. In the present paper, we construct testing procedures of the hypothesis "X follows the local Markov property at the node a with respect to the graph $\mathcal{G}$" against the hypothesis that it does not. In the following, we refer to such tests as *test of neighborhood*. We deduce from it testing procedures of the hypothesis "X is a Gaussian graphical model with respect to the graph $\mathcal{G}$" against the hypothesis that it is not. We call these tests *tests of graph*. Note that we aim at detecting if there are missing neighbours not if there are false neighbours. If the sample size n is larger than the number of nodes p, the issue of testing is extensively studied and there exists a variety of procedures based on asymptotic analysis; see for instance Schäfer and Strimmer [12] and Drton and Perlman [4]. We refer to Drton and Perlman [5] for a detailed review. However, most of these procedures mainly aim at performing graph estimation and not assessing a given graph. Besides, most of them do not apply when p is much larger than n. On the contrary, our procedures apply in a high dimensional setting whenever the graph $\mathcal{G}$ is sparse.

In Section 5.2.1.1 we highlight the connection between test of neighborhood and test in Gaussian linear regression in random Gaussian design. That is why our procedures are based on tests of linear hypothesis in this regression framework introduced in Chapter 4. These procedures are feasible in a high dimensional setting and we control exactly their Family-wise error rate. Besides, we are able to exhibit non asymptotic results on their power. Finally, we apply our procedures to simulated data in Section 5.3 and to real data sets in Section 5.4.

In the sequel, we denote $\overline{ne}(a) := ne(a) \cup \{a\}$ for any node $a \in \Gamma$.

5.2 Description of the testing procedures

5.2.1 Test of neighborhood

5.2.1.1 Connection with conditional Gaussian regression

In this part, we highlight the connection between the local Markov property and conditional regression of a Gaussian random variable. We define precisely the testing procedure in the

next part, following the approach introduced in Chapter 4.

Let $\mathcal{G} = (\Gamma, E)$ be an undirected graph and $a \in \Gamma$ be a node of this graph. We want to test the hypothesis "X_a is independent from $X_{\Gamma \setminus \overline{ne}(a)}$ conditionally to $X_{ne(a)}$" against the general alternative that it is not. This hypothesis corresponds to the local Markov property defined in Lauritzen [9] of X at the node a. In order to perform this test, we use a different caracterisation of conditional independence.

Let us consider the conditional distribution of X_a given all remaining variables $X_{-a} = \{X_b, b \in \Gamma \setminus \{a\}\}$. Using standard Gaussian properties (see for instance Lauritzen [9] appendix C), we know that this conditional distribution is a Gaussian distribution whose mean is a linear combination of elements in X_{-a} and whose variance does not depend on X_{-a}. Hence, we can decompose X_a as:

$$X_a = \sum_{b \in \Gamma \setminus a} \theta_b^a X_b + \epsilon_a, \tag{5.1}$$

where θ^a is a vector of coefficients in $\mathbb{R}^{p-1}$ and ϵ_a is a zero mean Gaussian random variable independant from X_{-a} whose variance equals the conditional variance of X_a given X_{-a}, $\mathrm{var}(X_a|X_{-a})$. The vector θ^a is determined by the inverse covariance matrix K of X (see Edwards [6]). More precisely, $\theta_b^a = -K[a,b]/K[a,a]$ for any $b \neq a$ and $\mathrm{var}(X_a|X_{-a}) = 1/K[a,a]$. As a consequence, the set of non-zero coefficients of θ^a corresponds to the non zero-components of the a-th row of K. Equivalently, there is an edge between the nodes a and b in the graph if the quantity $K[a,b]$ is not zero. For any set $V \subset \Gamma \setminus \{a\}$, θ_V^a denotes the sequence $(\theta_b^a)_{b \in V}$.

Testing the null-hypothesis "X_a is independant from $X_{\Gamma \setminus \overline{ne}(a)}$ conditionally to $X_{ne(a)}$" against the general alternative is therefore equivalent to testing the null-hypothesis $H_{0,a}$: "$\theta^a_{\Gamma \setminus \overline{ne}(a)} = 0$" against the general alternative $H_{1,a}$: "$\theta^a_{\Gamma \setminus \overline{ne}(a)} \neq 0$". Consequently, the test of neighborhood amounts to goodness-of-fit tests for Gaussian regression with random Gaussian covariates as considered in Chapter 4.

5.2.1.2 Description of the procedure

In this part, we adapt the test introduced in Chapter 4 to our statistical context. We are given n observations of the vector $X = (X_1, \ldots, X_p)^t$. For any $a \in \Gamma$, let us note $\mathbf{X}_a$ the n-vector of observations of X_a and $\mathbf{X}_{-a}$ the set of vectors $\mathbf{X}_b$ where b belongs to $\Gamma \setminus \{a\}$. The joint distribution of (X_a, X_{-a}) is uniquely defined by the vector θ^a, the covariance matrix of X_{-a} denoted Σ_{-a}, and $\mathrm{var}(X_a|X_{-a})$ the conditional variance of X_a. In the sequel, $\mathbb{P}_{\theta^a}$ refers to the joint distribution of $(\mathbf{X}_a, \mathbf{X}_{-a})$. For the sake of simplicity, we do not emphasize the dependence of $\mathbb{P}_{\theta^a}$ on Σ_{-a} and $\mathrm{var}(X_a|X_{-a})$.

Let us first fix some level $\alpha \in]0,1[$ and let m be a subset of $\Gamma \setminus \overline{ne}(a)$. In the sequel d_a and D_m denote the cardinalities of $ne(a)$ and m, and we define N_m as $n - d_a - D_m$. We assume that $n \geq d_a + 2$.

We define the Fisher statistic ϕ_m by

$$\phi_m(\mathbf{X}_a, \mathbf{X}_{-a}) = \frac{N_m \|\Pi_{ne(a) \cup m} \mathbf{X}_a - \Pi_{ne(a)} \mathbf{X}_a\|_n^2}{D_m \|\mathbf{X}_a - \Pi_{ne(a) \cup m} \mathbf{X}_a\|_n^2}, \tag{5.2}$$

where $\|.\|_n$ is the canonical norm in $\mathbb{R}^n$, and $\Pi_{ne(a)}$ and $\Pi_{ne(a) \cup m}$ respectively refer to the orthogonal projection onto the space generated by the vectors $(\mathbf{X}_b)_{b \in ne(a)}$ and to the

orthogonal projection onto the space generated by the vectors $(\mathbf{X}_b)_{b\in ne(a)\cup m}$. Then, ϕ_m corresponds to the statistic of the Fisher test of the null hypothesis

$$H_{0,a} : \theta_{\Gamma\setminus\overline{ne}(a)} = 0 \text{ against the alternative } H_{1,a,m} : \theta_{\Gamma\setminus\overline{ne}(a)} \neq 0 \text{ and } \theta_{\Gamma\setminus(\overline{ne}(a)\cup m)} = 0. \tag{5.3}$$

In the sequel, $\Pi_{ne(a)^\perp}$ stands for the orthogonal projection along the space generated by $(\mathbf{X}_b)$ with b belonging to $ne(a)$. Let us consider a finite collection $\mathcal{M}_a$ of non empty subsets of $\Gamma \setminus \overline{ne}(a)$. For all $m \in \mathcal{M}_a$, the cardinality D_m must be smaller than $n - d_a$. We define $\{\alpha_m, m \in \mathcal{M}_a\}$ a suitable collection of numbers in $]0,1[$ (which possibly depend on $\mathbf{X}_{-a}$). Our testing procedure consists in doing for each $m \in \mathcal{M}_a$ the Fisher test based on the statistic ϕ_m defined in Equation (5.2) at level α_m and rejecting the null hypothesis $H_{0,a}$ if one of those tests does. More precisely, we define the test T_α as

$$T_\alpha = \sup_{m\in\mathcal{M}_a} \left\{ \phi_m(\mathbf{X}_a, \mathbf{X}_{-a}) - \bar{F}^{-1}_{D_m,N_m}(\alpha_m(\mathbf{X}_{-a})) \right\}, \tag{5.4}$$

where for any $u \in \mathbb{R}$, $\bar{F}_{D,N}(u)$ denotes the probability for a Fisher variable with D and N degrees of freedom to be larger than u. We therefore reject the null hypothesis when T_α is positive. The main difference between this procedure and the one defined in Chapter 4 lies in the fact that we now deal with possibly random collection of models.

In order to ensure that the level T_α is less than α, the collection of weights $\{\alpha_m(\mathbf{X}_{-a}), m \in \mathcal{M}_a\}$ in $]0,1[$ must satisfy the property: for all $\theta \in \mathbb{R}^{p-1}$ such that $\theta_{\Gamma\setminus\overline{ne}(a)} = 0$, then $\mathbb{P}_\theta(T_\alpha > 0) \leq \alpha$.

We choose the collection $\{\alpha_m(\mathbf{X}_{-a}), m \in \mathcal{M}_a\}$ in accordance with one of the two following procedures :

- P_1 : The α_m 's do not depend on $\mathbf{X}_{-a}$ and satisfy the equality :

$$\sum_{m\in\mathcal{M}_a} \alpha_m = \alpha \tag{5.5}$$

- P_2 : For all $m \in \mathcal{M}_a$, $\alpha_m(\mathbf{X}_{-a}) = q_{\mathbf{X}_{-a},\alpha}$, where $q_{\mathbf{X}_{-a},\alpha}$ is defined conditionally to $\mathbf{X}_{-a}$ as the α-quantile of the distribution of the random variable

$$\inf_{m\in\mathcal{M}_a} \bar{F}_{D_m,N_m}\left(\phi_m(\boldsymbol{\epsilon}_a, \mathbf{X}_{-a})\right) \tag{5.6}$$

  Note that this last distribution does not depend on the variance of $\boldsymbol{\epsilon}_a$ and thus we can work out $q_{\mathbf{X}_{-a},\alpha}$ using Monte-Carlo method.

5.2.1.3 Comparison of Procedures P_1 and P_2

If the collection of models is not random, one can either use Procedure P_1 or P_2. In Chapter 4, Section 2.2, we show that the test T_α with Procedure P_1 has a size less than α, whereas the size of T_α with Procedure P_2 is exactly α. We deduce from this fact that the test T_α with procedure P_2 is more powerful than the corresponding test defined with Procedure P_1 with weights $\alpha_m = \alpha/|\mathcal{M}_a|$ (see Chapter 4, Section 2.3).

On the one hand the choice of Procedure P_1 allows to avoid the computation of the quantile $q_{\mathbf{X}_{-a},\alpha}$ and possibly permits to give a Bayesian flavor to the choice of the weights.

On the other hand, Procedure P_1 becomes too conservative when the collection of models $\mathcal{M}_a$ is large. This is often the case when the number p of nodes in the graph is large. That is why we advise to use Procedure P_2 when considering large graphs. We compare both Procedures in practice in Chapter 4 Section 6 and in Section 5.3.

5.2.1.4 Collection of models $\mathcal{M}_a$

The main advantage of our procedure is that it is very flexible in the choices of the models $m \in \mathcal{M}_a$. If we choose suitable collections $\mathcal{M}_a$, the test is powerful over a large class of alternatives as shown in Chapter 4 for non random collections. In this part, we propose two relevant classes of models $\mathcal{M}_a^1$ and $\mathcal{M}_a^2$ for our issue of test of neighborhood.

The collection $\mathcal{M}_a^1$ is defined as $\mathcal{M}_a^1 = \{\{b\}, b \in \Gamma \setminus \overline{ne}(a)\}$ and consists in taking each node in $\Gamma \setminus \overline{ne}(a)$ in turn. In Section 5.2.2, we present theoretical results of the power of T_α with collection $\mathcal{M}_a^1$ and Procedure P_1. This collection presents the advantage to be relatively small compared to other possible collections and the obtained procedure is consequently computationally attractive.

We have shown in Chapter 4, and this will be illustrated again in Section 5.3, that if there are several non-zero coefficients in $\theta^a_{\Gamma\setminus\overline{ne}(a)}$, considering models of larger dimensions can improve the performance of the test. For instance, if we are given an order on the nodes and if the vector θ^a belongs to an ellipsoid relative to this order, one should choose the collection of nested models defined by this order (see Chapter 4, Section 5). There is not such an order in our context as we do not know in principle which node are more relevant to test. That is why we propose to use the LARS (least angle regression) algorithm introduced by Efron *et al.* [7]. This model selection algorithm provides an order of relevance of the covariates in linear regression. Besides, one of its main advantage lies in its computationally attractiveness. The collection of models $\mathcal{M}_a^2$ is built as follows. We first choose an integer J which corresponds to the maximal size of the models we want to consider. We advise to take J smaller than $n/2$. Then, we apply the LARS algorithm to the response $\Pi_{ne(a)^\perp}\mathbf{X}_a$ with the set of covariates $\Pi_{ne(a)^\perp}\mathbf{X}_b$ where $b \in \Gamma \setminus \overline{ne}(a)$ and we obtain the sequence $s_{LARS} = (j_1, \ldots, j_J)$. Finally we define the collection $\mathcal{M}_a^2$ as:

$$\mathcal{M}_a^2 = \{\{j_1, \ldots, j_k\}, 1 \leq k \leq J\}$$

As the collection of models $\mathcal{M}_a^2$ given by the LARS algorithm now depends on the data, we need do define a new procedure to handle random collections.

Suppose we are given a random collection of models $\mathcal{M}_a$ which only depends on

$$\Psi(\mathbf{X}_a, \mathbf{X}_{-a}) := \left(\frac{\Pi_{ne(a)^\perp}\mathbf{X}_a}{\|\Pi_{ne(a)^\perp}\mathbf{X}_a\|_n}, \mathbf{X}_{-a} \right) ,$$

then we shall use the test statistic (5.4) with weights given by the procedure P_3 defined as follows:

- P_3: For all $m \in \mathcal{M}_a\left[\Psi\left(\mathbf{X}_a, \mathbf{X}_{-a}\right)\right]$, $\alpha_m(\mathbf{X}_{-a}) = q'_{\mathbf{X}_{-a},\alpha}$, the α-quantile of the distribution of the random variable

$$\inf_{m \in \mathcal{M}_a[\Psi(\boldsymbol{\epsilon}_a, \mathbf{X}_{-a})]} \bar{F}_{D_m, N_m}\left(\phi_m(\boldsymbol{\epsilon}_a, \mathbf{X}_{-a})\right) \tag{5.7}$$

conditionally to $\mathbf{X}_{-a}$. As for the procedure P_2, the distribution of (5.7) does not depend on the variance of $\boldsymbol{\epsilon}_a$ and thus we are able to compute $q'_{\mathbf{X}_{-a},\alpha}$ using Monte-Carlo method.

Clearly, if the collection of models is not random, Procedures P_2 and P_3 lead to the same weights. As with Procedure P_2, the size of T_α with Procedure P_3 is exactly α. More Precisely, for any $\theta^a \in \mathbb{R}^{p-1}$ such that $\theta^a_{\Gamma \setminus \overline{ne}(a)} = 0$, we have that

$$\mathbb{P}_{\theta^a}\left(T_\alpha | \mathbf{X}_{-a}\right) = \alpha \ \ \mathbf{X}_{-a} \text{ a.s. } .$$

The result follows from the fact that $q'_{\mathbf{X}_{-a},\alpha}$ satisfies

$$\mathbb{P}_{\theta^a}\left(\sup_{m \in \mathcal{M}_a[\Psi(\boldsymbol{\epsilon}_a, \mathbf{X}_{-a})]} \left\{\phi_m(\boldsymbol{\epsilon}_a, \mathbf{X}_{-a}) - \bar{F}^{-1}_{D_m, N_m}\left(q'_{\mathbf{X}_{-a},\alpha}\right)\right\} > 0 \middle| \mathbf{X}_{-a}\right) = \alpha,$$

and for any $\theta^a \in \mathbb{R}^{p-1}$ such that $\theta_{\Gamma \setminus \overline{ne}(a)} = 0$,

$$\begin{aligned} \Pi_{ne(a)\cup m}\mathbf{X}_a - \Pi_{ne(a)}\mathbf{X}_a &= \Pi_{ne(a)\cup m}\boldsymbol{\epsilon}_a - \Pi_{ne(a)}\boldsymbol{\epsilon}_a \\ \text{and } \mathbf{X}_a - \Pi_{ne(a)\cup m}\mathbf{X}_a &= \boldsymbol{\epsilon}_a - \Pi_{ne(a)\cup m}\boldsymbol{\epsilon}_a \ . \end{aligned}$$

As the sequence of relevant variables given by the LARS algorithm does not depend on the norm of the reponse, the collection $\mathcal{M}^2_a$ only depends on $\Psi(\mathbf{X}_a, \mathbf{X}_{-a})$ and thus we are able to apply Procedure P_3.

5.2.2 Properties of the test of neighborhood with collection $\mathcal{M}^1_a$

For the convenience of the reader, we recall in this part some of the theoretical results established in Chapter 4. First, we give a proposition which caracterizes the set of vectors θ^a over which the test T_α with the collection $\mathcal{M}^1_a$ and weights $\alpha_m = \alpha/|\mathcal{M}^1_a|$ is powerful. We shall then discuss the optimality of this test.

Proposition 17. *Let us assume that n satisfies:*

$$n - d_a - 1 \geq \left[10 \log\left(\frac{p - d_a - 1}{\alpha}\right) \vee 21 \log\left(1/\delta\right)\right].$$

Let us set the quantity

$$\rho^2_{n-d_a, p-d_a} = \frac{C_1}{n - d_a} \log\left(\frac{p - d_a - 1}{\alpha\delta}\right), \tag{5.8}$$

where C_1 is a universal constant. For any θ^a in $\mathbb{R}^{\Gamma \setminus \{a\}}$, $\mathbb{P}_\theta\left(T_\alpha > 0\right) \geq 1 - \delta$ if there exists $b \in \Gamma \setminus \overline{ne}(a)$ such that

$$\frac{\mathrm{var}_{\theta^a}(X_a | X_{ne(a)}) - \mathrm{var}_{\theta^a}(X_a | X_{ne(a)\cup\{b\}})}{\mathrm{var}_{\theta^a}(X_a | X_{ne(a)\cup\{b\}})} \geq \rho^2_{n-d_a, p-d_a}. \tag{5.9}$$

This proposition is a straightforward corollary of Theorem 3 in Chapter 4. One interprets the quantity appearing in (5.9) as follows: the quotient of conditional variances measures the ratio of the quantity of information brought by X_i for the prediction of X_a to the part of X_a not explained by $X_{ne(a)\cup\{i\}}$. In other words, the test T_α has a power larger than δ for vectors θ^a such that there exists a node $i \in \Gamma \setminus \overline{ne}(a)$ which improves enough the prediction of X_a.

This test is optimal in the minimax sense if we test against the alternative "$\theta^a_{\Gamma\setminus\overline{ne}(a)}$ has only one non-zero component" and if the covariates are independent (see Chapter 4, Section 4.2). The condition of independence for covariates is unrealistic in a Gaussian graphical context, but it is nevertheless relevant as the independent case is an important benchmark from the minimax point of view (see Chapter 4, Section 4.2 for more details). When the covariates are correlated we know from a simulation study (Chapter 4, Section 6) that using Procedure P_2 slightly improves the power of the test T_α.

5.2.3 Test of graph

From the test of neighborhood we define a procedure to test a graph. More precisely, we test the null hypothesis H_0 that "X is a Gaussian graphical model with respect to $\mathcal{G}$" against the alternative that it is not. Let $\{\alpha_a, a \in \Gamma\}$ be a collection of numbers in $]0,1[$. For each node $a \in \Gamma$, we test at level α_a the neighborhood of the node a with one of the procedures explained in Section 5.2.1.2. We decide to reject the null hypothesis H_0 as soon as one of the test $T^a_{\alpha_a}$ is rejected. We obtain a test of level α of the graph $\mathcal{G}$ if we take $\{\alpha_a, a \in \Gamma\}$ such that $\sum_{a\in\Gamma} \alpha_a = \alpha$. In the sequel we choose $\alpha_a = \alpha/p$ for each $a \in \Gamma$.

This procedure corresponds to a Bonferroni choice of the weights. As a consequence, if the number p of nodes is very large, our test may suffer a loss of its size. This restricts ourselves to consider tests of graph only for relatively small graphs, or for subgraphs of a large graph. Let us recall that when we apply the test of neighborhood to one node, the number p of nodes can be arbitrary large without any loss in the size of the test, provided that we use Procedure P_2 or P_3.

5.3 Simulations

In this section we present two simulation studies. First we study the test of graph when the number of nodes is small. On the one hand we compare the efficiency of Procedures P_1 and P_2 and on the other hand we show the influence of the percentage of edges in the graph on the power of the test. Second, we study the test of neighborhood when p is large, illustrating thus the power of our procedure in a high dimensional setting. Besides, we compare the efficiency of the tests based on the collections of models $\mathcal{M}^1_a$ and $\mathcal{M}^2_a$ defined in Section 5.2.1.4.

5.3.1 Simulation of a GGM

5.3.1.1 Simulation of a graph

In our simulations we use two different methods to generate random graphs. The first one allows to control the number of nodes p and the percentages of edges η in the graph. It consists in choosing uniformly and independently the positions of the $\eta \times p(p-1)/2$ edges.

We use this method in the simulation experiment on the test of graph, with different values of η to measure the influence of the percentage of edges on the test.

However, the vertices of real-world networks are often structured in clusters, i.e groups of proteins functionally related, with different connectivity properties. That is why Daudin *et al.* [3] proposed a model called ERMG for Erdös-Rényi Mixtures for Graphs, which describes the way edges connect nodes, accounting for some groups of nodes, and some preferential connections between the groups. The ERMG model assumes that the nodes are spread into Q clusters with probabilities $\{p_1, \ldots, p_Q\}$. We are given a connectivity matrix C of size $Q \times Q$ which specifies the probability of connection between two nodes according to the clusters they belong to. More precisely, the probability that two nodes belonging to the clusters i and j share an edge equals $C[i, j]$. We use this method to generate a graph in the simulation experiment on the test of neighborhood, with the following parameters provided by Daudin *et al.* [3]: $p = 199$ nodes, $Q = 7$ clusters, the probabilities $(p_1, \ldots, p_Q)$ and the connectivity matrix C equal:

$$(p_1, \ldots, p_Q) = \begin{pmatrix} 0.038 & 0.052 & 0.060 & 0.082 & 0.083 & 0.125 & 0.560 \end{pmatrix} \tag{5.10}$$

$$C = \begin{pmatrix} 0.999 & 0.319 & 1e-06 & 0.116 & 1e-06 & 1e-06 & 0.007 \\ 0.319 & 0.869 & 1e-06 & 1e-06 & 0.140 & 0.004 & 0.002 \\ 1e-06 & 1e-06 & 0.467 & 0.0155 & 0.005 & 0.014 & 0.004 \\ 0.116 & 1e-06 & 0.016 & 0.216 & 1e-06 & 0.017 & 0.005 \\ 1e-06 & 0.140 & 0.005 & 1e-06 & 0.229 & 1e-06 & 0.004 \\ 1e-06 & 0.004 & 0.014 & 0.017 & 1e-06 & 0.239 & 0.013 \\ 0.007 & 0.002 & 0.004 & 0.005 & 0.0041 & 0.0129 & 0.0163 \end{pmatrix} \tag{5.11}$$

Using these parameters, the percentage of edges η in the graph equals 2.5%.

5.3.1.2 Simulation of the data

Given a graph we generate random vectors whose conditional independence structure is represented by the graph.

First, we generate the partial correlation matrix Π as follows : to a graph with p nodes we associate a symmetric $p \times p$ matrix U such that for any $(i, j) \in \{1, \ldots, p\}^2$, $U[i, j]$ is drawn from the uniform distribution between -1 and 1 if there is an edge between the nodes i and j and $U[i, j]$ is set to 0 in the other case. We then compute column-wise sums of the absolute values of the matrix U entries, and set the corresponding diagonal element equal to this sum plus a small constant. This ensures that the resulting matrix is diagonally dominant and thus positive definite. Finally, we standardize the matrix so that the diagonal entries all equal 1 to obtain the simulated partial correlation matrix Π.

Second, we simulate data of the sample size n. We generate n independent samples from the multivariate normal distribution with mean zero, unit variance, and correlation structure associated to the partial correlation matrix Π. In the sequel, we note $\mathbf{X}$ the $n \times p$ associated data matrix.

5.3.2 Simulation setup

5.3.2.1 Simulation study on the test of graph

We evaluate the performance of the test of graph, first with simulations on randomly generated graphs, and secondly on a network coming from the data base KEGG.

1. First simulation experiment: We estimate the level and the power of the test of graph with 1000 simulations. For fixed parameters (p, η, n), we generate 1000 graphs by using the first method described in Section 5.3.1.1 and 1000 data matrices as described in Section 5.3.1.2. Let $\mathcal{G}^s$ and $\mathbf{X}^s$ for $s = 1, \ldots, 1000$ denote the graphs and the data matrices for the 1000 simulations. For each simulation s, we test the null hypothesis "$\mathbf{X}^s$ is a Gaussian graphical model with respect to the graph $\mathcal{G}^s$". We thus estimate the level of the test by dividing the number of simulations for which we reject the null hypothesis by 1000. Let q be a number in $]0, 1[$. For each simulation s, let $\mathcal{G}^s_{-q}$ be the graph built from the graph $\mathcal{G}^s$ in which we delete randomly $q\frac{p(p-1)}{2}\eta$ edges. For each simulation s, we test the null hypothesis "$\mathbf{X}^s$ is a Gaussian graphical model with respect to the graph $\mathcal{G}^s_{-q}$". We estimate the power of the test by dividing the number of simulations for which we reject the null hypotheses by 1000.

   The number of variables p is set to 15, whereas the number of observations n is taken equal to 10, 15 and 30 to study the effect of the sample size. We examine the influence of the percentage of edges in the graph, by taking $\eta = 0.1$ and 0.15. Besides, we show the effect of the percentage q of missing edges on the power, by presenting the results for q equal to 10%, 40% and 100%.

2. Second simulation experiment: This simulation is based on the cell cycle of yeast (Saccharomyces cerevisiae). This experiment aims at showing the performance of our procedure with simulations on a real biological network. The graph correponding to the cell cycle of yeast is available in the data base KEGG from the following website: `http://www.genome.jp/kegg/pathway/sce/sce04111.html`. We focus on a part of this pathway involving 16 proteins and 18 interactions. The graph, denoted in the sequel $\mathcal{G}_{cellcycle}$ is shown in Figure 5.1. We estimate the level and the power of the test by simulating 1000 data matrix $(\mathbf{X}^s)_{s=1,\ldots,1000}$ from the graph $\mathcal{G}_{cellcycle}$ as described in Section 5.3.1.2. We first estimate the level of the test by testing for each simulation s, the null hypothesis "$\mathbf{X}^s$ is a Gaussian graphical model with respect to the graph $\mathcal{G}_{cellcycle}$". Then, we delete the three edges involving the protein complex $SCF\ Cdc4$ in $\mathcal{G}_{cellcycle}$ in order to define the graph $\mathcal{G}^{-Cdc4}_{cellcycle}$. This protein complex $SCF\ Cdc4$ participates in cell death. We estimate the power of the test by testing for each simulation s the null hypothesis "$\mathbf{X}^s$ is a Gaussian graphical model with respect to the graph $\mathcal{G}^{-Cdc4}_{cellcycle}$. In other words we evaluate the ability of our procedure to detect the link of the protein complex $SCF\ Cdc4$ with the cell cycle.

5.3.2.2 Simulation study for the test of neighborhood

We first simulate a graph $\mathcal{G}$ according to the ERMG model described in Section 5.3.1.1 with $p = 199$ nodes, $Q = 7$ clusters, and the parameters $(p_1, \ldots, p_Q)$ and the matrix C defined in Equations (5.10) and (5.11). We then focus on a node a of this graph, chosen

such that it has several neighbours. In our simulation this node has 6 neighbours. Let us denote $ne(a)$ its neighborhood given by the graph $\mathcal{G}$. We simulate 1000 data matrix as described in Section 5.3.1.2 from the graph $\mathcal{G}$ and estimate the level of the test by testing the null hypothesis that the node a has no other neighbour that the set $ne(a)$, and the power by testing the null hypothesis that the node a has no neighbour. We present results for the sample size n equal to 50, 100 and 200.

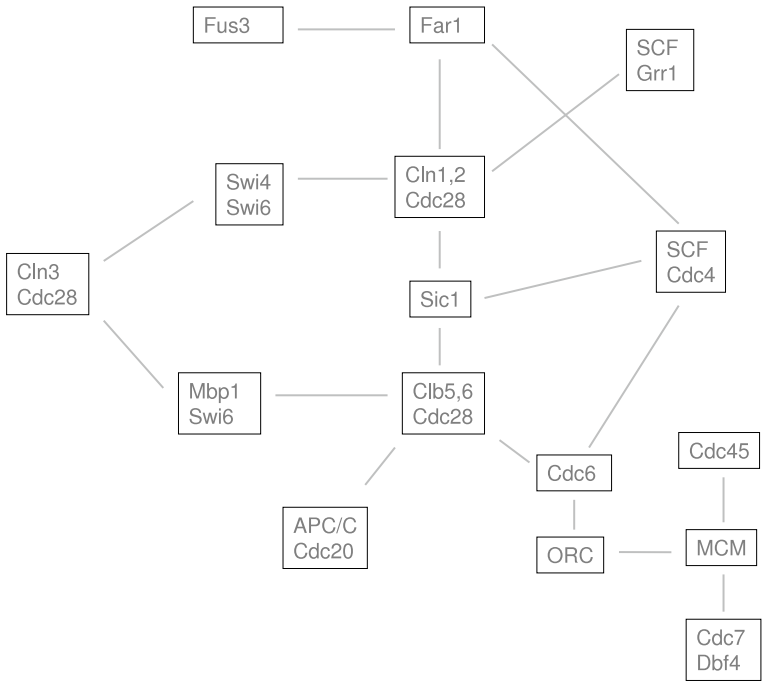

Figure 5.1: $\mathcal{G}_{cellcycle}$

5.3.2.3 Collections of models $\mathcal{M}_a$ and collections $\{\alpha_m, m \in \mathcal{M}_a\}$:

For each node a, we use the testing procedure defined in (5.4) with different collections $\mathcal{M}_a$ and different choices for the weights $\{\alpha_m, m \in \mathcal{M}_a\}$. Let us recall that $ne(a)$ denotes the neighborhood of the node a under the null hypothesis and α_a the level of the test of neighborhood for the node a. For the test of graph we choose $\alpha_a = \alpha/p$ and for the test of neighborhood α_a equals α.

The collections $\mathcal{M}_a$: we consider the two collections defined in Section 5.2.1.4.

$$\mathcal{M}_a^1 = \{\{b\}, b \in \Gamma \setminus \overline{ne}(a)\}.$$

$$\mathcal{M}_a^2 = \{\{j_1, \ldots, j_k\}, 1 \leq k \leq J\}$$

where $S_{Lars}\left[\Psi\left(\mathbf{X}_a, \mathbf{X}_{-a}\right)\right] = \{j_1, j_2, \ldots, j_J\}$ is the sequence given by the LARS algorithm for the prediction of $\Pi_{ne(a)^\perp}\mathbf{X}_a$ with the set of covariates $\Pi_{ne(a)^\perp}\mathbf{X}_b$ where $b \in$

$\Gamma \setminus \overline{ne}(a)$. The maximum number of steps J is taken equal to 10. We evaluate the performance of our testing procedure with $\mathcal{M}_a^1$ in the simulation experiment on the test of graph, and we compare collections $\mathcal{M}_a^1$ and $\mathcal{M}_a^2$ in the simulation experiment on the test of neighborhood. Indeed, in the second simulation experiment p and consequently the collection $\mathcal{M}_a^1$ is large. It is therefore interesting to compare the two collections from a computing-time point of view.

The collection $\{\alpha_m, m \in \mathcal{M}_a\}$ *:* When we consider the collection of models $\mathcal{M}_a^1$ we use either Procedure P_1 or Procedure P_2 defined in Section 5.2.1.2. For Procedure P_1 the α_m's are taken equal to $\alpha_a/|\mathcal{M}_a|$. The quantity $q_{\mathbf{X}_{-a},\alpha_a}$ occurring in Procedure P_2 is evaluated by simulation. Let Z be a standard Gaussian random vector of size n independent from $\mathbf{X}_{-a}$. As $\boldsymbol{\epsilon}_a$ is independant from $\mathbf{X}_{-a}$, the distribution of (5.6) conditionally to $\mathbf{X}_{-a}$ is the same as the distribution of

$$\inf_{m \subset \mathcal{M}_a} \bar{F}_{D_m,N_m} \frac{\|\Pi_{ne(a)\cup m}(Z) - \Pi_{ne(a)}(Z)\|^2/D_m}{\|Z - \Pi_{ne(a)\cup m}(Z)\|^2/N_m}$$

conditionally to $\mathbf{X}_{-a}$. Consequently, we estimate the quantile $q_{\mathbf{X}_{-a},\alpha_a}$ by a Monte-Carlo method with 1000 samples. When we use the collection of models $\mathcal{M}_a^2$ we apply Procedure P_3. The quantile $q'_{\mathbf{X}_{-a},\alpha_a}$ is again computed by a Monte-Carlo method with 1000 simulations.. The difference with the simulation of $q_{\mathbf{X}_{-a},\alpha_a}$ lies in the fact that the collection $\mathcal{M}_a^2$ is random and depends on ϵ_a. For each simulation, let Z be a standard Gaussian random vector of size n independant from $\mathbf{X}_{-a}$. We apply the LARS algorithm for the prediction of $\Pi_{ne(a)^\perp} Z$ with the set of covariates $\Pi_{ne(a)^\perp} \mathbf{X}_b$ where $b \in \Gamma_{-a} \setminus ne(a)$. We obtain the sequence $S_{Lars}\left[\Psi\left(Z, \mathbf{X}_{-a}\right)\right]$ which leads to the collection of models $\mathcal{M}_a^2\left[\Psi\left(Z, \mathbf{X}_{-a}\right)\right]$. As $\boldsymbol{\epsilon}_a$ is independant from $\mathbf{X}_{-a}$, the distribution of (5.7) conditionally to $\mathbf{X}_{-a}$ is the same as the distribution of

$$\inf_{m \in \mathcal{M}_a[\Psi(Z,\mathbf{X}_{-a})]} \bar{F}_{D_m,N_m} \left(\frac{\|\Pi_{ne(a)\cup m} Z - \Pi_{ne(a)} Z\|_n^2/D_m}{\|Z - \Pi_{ne(a)\cup m} Z\|_n^2/N_m} \right)$$

conditionally to $\mathbf{X}_{-a}$ and we therefore estimate the quantile $q'_{\mathbf{X}_{-a},\alpha_a}$. In the sequel, we note $T_{\mathcal{M}_a^i,P_j}$ the test (5.4) with collection $\mathcal{M}_a^i$ and Procedure P_j.

5.3.3 The results

In Table 5.1 and 5.2 we present results of the first simulation experiment on the test of graph respectively for $\eta = 0.1$ and $\eta = 0.15$. As expected, the power of the tests increases with the number of observations n. Besides, the power of the tests increases also with the percentage of missing edges q, the tests being indeed more powerful when the graphs under the null and the alternative hypotheses are more different. As expected the tests based on Procedure P_2 are more powerful than the corresponding tests based on Procedure P_1. However because p is small, the difference between the two procedures is not really significant. Nevertheless, Procedure P_1 may become too conservative when p is large. As expected, its implementation is faster : for $p = 15$ and $n = 10$ a single simulation using Procedure P_1 is 60 times faster than a single simulation using Procedure P_2. For p small, Procedure P_1 is therefore a good compromise in practice, Procedure P_2 being

rather recommended when considering large graphs. Let us now compare the influence of η on the power of the test. When the percentage of edges η in the graph increases, the tests are less powerful. It is especially significant for $q = 10\%$. In fact, when η increases the average number of neighbours for each node increases as well. In practice, the test of neighborhood is less powerful for a node which already has several neighbours under the null hypothesis. Consequently, the issue of testing the graph is more difficult when η is large.

Estimated levels

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.028	0.046
15	0.035	0.061
30	0.033	0.054

Estimated powers

$q = 10\%$			$q = 40\%$			$q = 100\%$		
n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$	n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$	n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.73	0.75	10	0.94	0.94	10	0.99	0.99
15	0.83	0.84	15	0.97	0.98	15	1	1
30	0.95	0.95	30	1	1	30	1	1

Table 5.1: Test of graph, first simulation. $\eta = 0.1$. Estimated levels and powers. The nominal level is $\alpha = 5\%$.

Estimated levels

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.031	0.050
15	0.044	0.053
30	0.041	0.058

Estimated powers

$q = 10\%$			$q = 40\%$			$q = 100\%$		
n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$	n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$	n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.28	0.32	10	0.70	0.72	10	0.90	0.91
15	0.44	0.46	15	0.87	0.88	15	0.99	0.99
30	0.73	0.75	30	0.99	0.99	30	1	1

Table 5.2: Test of graph, first simulation. $\eta = 0.15$. Estimated levels and powers. The nominal level is $\alpha = 5\%$.

In Table 5.3 we give the results of the second simulation experiment for the test of graph. The percentage of edges in the graph $\mathcal{G}_{cellcycle}$ equals 15%, whereas the ratio of missing edges is $q = 1/6$ as we delete 3 edges among 18 in $\mathcal{G}_{cellcycle}$. In fact, as q is between 10% and 40% the power of the tests in this setting are comparable to the results in Table 5.2. For $n = 20$ observations, the test is powerful and detects the relation between the protein complex $SCF\ Cdc4$ and the cell cycle with large probability. Even when n is smaller than p, the test detects the relation with a moderate probability.

Estimated levels

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.040	0.055
20	0.046	0.063
30	0.040	0.058

Estimated powers

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.43	0.46
20	0.76	0.79
30	0.89	0.90

Table 5.3: Test of graph, second simulation experiment. Estimated levels and powers. The nominal level is $\alpha = 5\%$.

In Table 5.4 we give the results of the simulation experiment on the test of neighborhood. For $n = 50$ and 100 the test is more powerful when using the collection of models $\mathcal{M}_a^1$ whereas when n is larger both procedures exhibit a comparable power. This comes from the fact that the test with collection $\mathcal{M}_a^2$ is performed in two steps: first, the selection of the relevant covariates using LARS and second, the test (5.4) itself. When n is small, LARS makes mistakes and possibly selects irrelevant covariates. In this case, the collection of models is bad and the test seldom rejects. When n is large, LARS often selects the relevant variables and the test $T_{\mathcal{M}^2,P_3}$ therefore takes advantage of exploiting models of several dimensions. However, its performances are not much better than the ones of $T_{\mathcal{M}^1,P_2}$ even when n is large. Let us now compare the two collections from a computing-time point of view. For $p = 200$ and $n = 100$ a single simulation using collection $\mathcal{M}_a^1$ is almost three times longer than using collection $\mathcal{M}_a^2$. In conclusion it seems natural to exploit model of several dimensions especially when we consider the test of neighborhood for a node which has several missing neighbours. On the one hand the LARS algorithm does not really improve the performance of the test. On the other hand, using collection $\mathcal{M}_a^2$ is computationally more attractive that using collection $\mathcal{M}_a^1$.

Estimated levels

n	$T_{\mathcal{M}^1,P_2}$	$T_{\mathcal{M}^2,P_3}$
50	0.056	0.052
100	0.044	0.054
200	0.041	0.043

Estimated powers

n	$T_{\mathcal{M}^1,P_2}$	$T_{\mathcal{M}^2,P_3}$
50	0.19	0.15
100	0.47	0.41
200	0.85	0.86

Table 5.4: Test of neighborhood for the simulation experiment described in Section 5.3.2.2. Estimated levels and powers. The nominal level is $\alpha = 5\%$.

5.4 Application to biological data

In this section, we apply the test of graph to the multivariate flow cytometry data produced by Sachs *et al.* [11]. These data concern a human T cell signaling pathway whose deregulation may lead to carcenogenesis. Therefore, this pathway was extensively studied in the literature and a network involving 11 proteins and 16 interactions was conventionally accepted (Sachs *et al.* [11]). See Figure 5.2 for a representation of this network. The data from Sachs consist of quantitative amounts of these 11 proteins, simultaneously measured from single cells under perturbation conditions. In the sequel, we focus on one general perturbation (anti-CD3/CD28 + ICAM-2) that overall stimulates the cellular signaling network. In this condition the quantities of the 11 proteins are measured in 902 cells. Let denote D this data set constituted of $p = 11$ variables and $n = 902$ observations. Contrary to most of postgenomic data, flow cytometry data provide a large sample of observations that allow us to measure the influence of the sample size on the power. From this data set we infer the network using three methods and we apply our test of graph as a tool to validate these estimations. As such abondance of data is rarely available in postgenomic data, we secondly carry out a simulation study to determine the influence of the number of observations on the test. From the empirical covariance matrix obtained with the whole data set D, we generate data of different sample sizes and we evaluate the performance of the test with respect to the sample size.

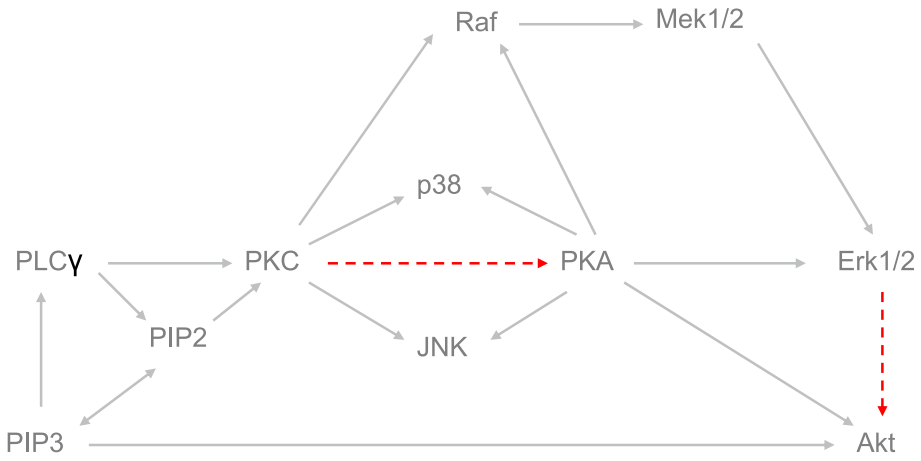

Figure 5.2: Classic signaling network of the human T cell pathway. The connections well-established in the literature are in grey and the connections cited at least once in the literature are represented by red dotted lines.

We use the methods proposed by Drton and Perlman [4], Wille and Bühlmann [13], and Meinshausen and Bühlmann [10] to infer the network. Let us briefly describe them. The SINful approach introduced by Drton and Perlman is a model selection algorithm based on multiple testing. For any couple of nodes they perform a test of existence of an edge between these two nodes and select the graph by computing the simultaneous

p-values of these tests. This method assumes that the number of observations n is larger than the number of variables p. The two other methods have been recently proposed to deal with the usual fact in genomics of p large and n small. Wille and Bühlmann [13] estimate a lower-order conditional independence graph instead of the concentration graph, while Meinshausen and Bühlmann [10] estimate the neighborhood of any node with the Lasso method. We represent the three estimated graphs in Figure 5.3.

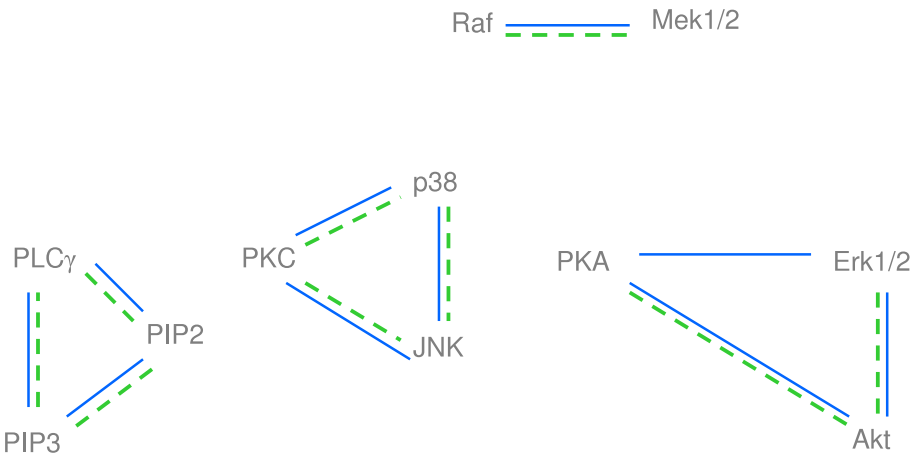

Figure 5.3: Inferred graphs. The graphs estimated with the methods of Drton and Perlman and Wille and Bühlmann are identical and represented in blue. The graph estimated with the method of Meinshausen and Bühlmann is in green dotted line

Let us define the graph $\mathcal{G}_\cap$ as the intersection of the graph estimated by these three methods and of the graph with the connections well-established in the literature. This graph $\mathcal{G}_\cap$ is represented in Figure 5.4. We test with our procedure the null hypothesis $H_{\mathcal{G}_\cap}$:"the data set D follows the distribution of a Gaussian graphical model with respect to the graph $\mathcal{G}_\cap$". We use for each node a of the graph the collection of models $\mathcal{M}_a^1$ defined in Section 5.2.1.4 and the procedure P_1. As p is small, the benefit of Procedure P_2 on P_1 is indeed not significant and the implementation of P_1 is faster. If we apply our procedure at level $\alpha = 5\%$, we reject the null hypothesis $H_{\mathcal{G}_\cap}$. As our procedure consists in testing the neighborhood of each node, it is interesting to look for the nodes for which the test of neighborhood is rejected, and for any rejected neighborhood which alternative leads to this rejection. In Table 5.5 we enumerate the nodes for which the test of neighborhood is rejected and the alternatives which lead to this decision.

As the connection $PKA - Erk1/2$ is well-established and the connection $Erk1/2 - Akt$ is cited at least once in the literature, we decide to add those two edges in the graph $\mathcal{G}_\cap$, defining thus a new graph $\mathcal{G}_2$ shown in Figure 5.5. The test of the null hypothesis $H_{\mathcal{G}_2}$ at level $\alpha = 5\%$: "the data set D follows the distribution of a Gaussian graphical model with respect to the graph $\mathcal{G}_2$" is rejected. The reason is that the tests concerning respectively nodes $p38$ and JNK are rejected when we consider in the alternative respectively nodes JNK and $p38$.

Let us therefore define a new graph $\mathcal{G}_T$ by adding the connection $p38 - JNK$, even if this connection is not well-established in the literature. Let us note that the graph $\mathcal{G}_T$

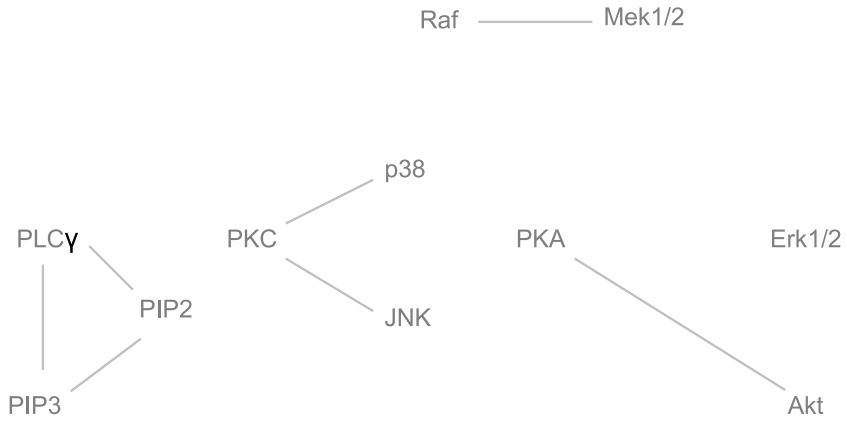

Figure 5.4: Graph $\mathcal{G}_\cap$

Rejection of the neighborhood of

node	because of node(s)
Erk1/2	Akt, PKA
Akt	Erk1/2
PKA	Erk1/2
p38	JNK
JNK	p38

Table 5.5: Rejection of $H_{\mathcal{G}_\cap}$

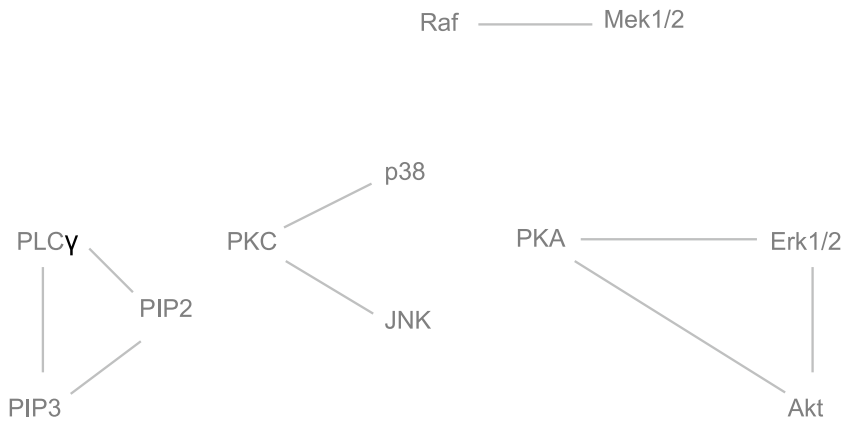

Figure 5.5: Graph $\mathcal{G}_2$

is the same as the network inferred by Sachs *et al.* [11] with approximatively the same data set by using a Bayesian approach. We apply our test of graph and we accept the hypothesis that the data set D is a Gaussian graphical model with respect to the graph $\mathcal{G}_T$ at the level $\alpha = 5\%$. As n is large we can use the result of the test with confidence and assume that the graph $\mathcal{G}_T$ (Figure 5.6) represents the conditional independence structure of the data set D.

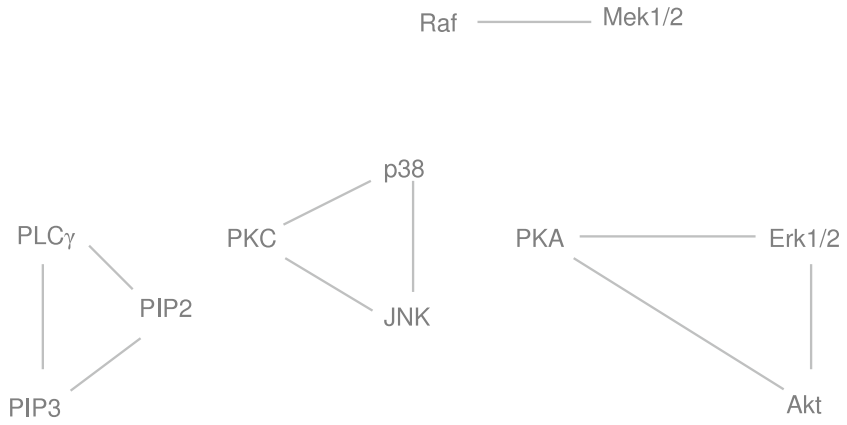

Figure 5.6: Graph $\mathcal{G}_T$

We now carry out a simulation study from this data set to determine the influence of the number of observations n on the power of our procedure. From the empirical covariance matrix obtained with the data set D, we generate 1000 simulated data $(\mathbf{X}^s)_{s=1,\ldots,1000}$ of different sample sizes n whose conditional independence structure is represented by the graph $\mathcal{G}_T$. First, we estimate the level of the test for different values of n by testing for each simulation that $\mathbf{X}^s$ is a Gaussian graphical model with respect to the graph $\mathcal{G}_T$. Second, we delete the two edges involving protein PKC in $\mathcal{G}_T$ in order to define $\mathcal{G}_T^-$. We estimate the power of the test for different values of n by testing for each simulation that $\mathbf{X}^s$ is a Gaussian graphical model with respect to the graph $\mathcal{G}_T^-$.

Estimated levels

n	$T_{\mathcal{M}^1,P_1}$
10	0.032
15	0.036
20	0.033

Estimated powers

n	$T_{\mathcal{M}^1,P_1}$
10	0.49
15	0.86
20	0.97

Table 5.6: Sachs data. Estimated levels and powers

The results of the simulation study from the selected Sachs'data are presented in Table 5.6. We recall that the graph involves $p = 11$ proteins and we take for the sample size n the values 10, 15, and 20. As expected, the power of the test increases with the number

of observations n. However, the number of observations do not have to be very large to obtain a powerful test. For $n = 15$ observations the test is able to recover that the protein PKC is not independent from the proteins $p38$ and JNK with large probability.

References

[1] Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. *Probabilistic Networks and Expert Systems*. Statistics for Engineering and Information Science. Springer-Verlag, New-York, 1999.

[2] Cressie, N. *Statistics for Spatial Data*, revised ed. Wiley, New York, 1993.

[3] Daudin, J. J., Picard, F., and Robin, S. A mixture model for random graphs. Tech. Rep. RR-5840, INRIA, Rapport de Recherche, 2006.

[4] Drton, M., and Perlman, M. A SINful approach to Gaussian graphical model selection. Tech. Rep. 457, Dept. of Statistics, University of Washington, Seattle, 2004.

[5] Drton, M., and Perlman, M. Mutiple testing and error control in Gaussian graphical model selection. *arXiv.math/0508267* (2007).

[6] Edwards, D. M. *Introduction to Graphical Modelling*, second ed. Sringer-Verlag, New-York, 2000.

[7] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *Annals of Statistics 32*, 2 (2004), 407–499.

[8] Geman, S., and Geman, D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6* (1984), 721–741.

[9] Lauritzen, S. L. *Graphical Models*. Oxford University Press, New York, 1996.

[10] Meinshausen, N., and Bühlmann, P. High dimensional graphs and variable selection with the Lasso. *Annals of Statistics 34*, 3 (2006), 1436–1462.

[11] Sachs, K., Perez, O., D.Pe'er, Lauffenburger, D. A., and Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science 308* (2005), 523–529.

[12] Schäfer, J., and Strimmer, K. An empirical bayes approach to inferring large-scale gene association nerworks. *Bioinformatics 21* (2005), 754–764.

[13] Wille, A., and Bühlmann, P. Low-order conditional independence graphs for inferring genetic networks. *Stat. Appl. Genet. Mol. Biol. 5* (2006).

Annexe

Annexe A

Preuves du chapitre 4

Ces preuves ont été établies par Nicolas Verzelen.

A.1 Proofs of Theorem 3, Proposition 5, 9, 11, 12, and 14

A.1.1 Proof of Theorem 3

This proof follows the same approach as the proof of Theorem 1 in [2]. The main differences and difficulties come from the fact that the design is now random.

Using the definition of T_α, we notice that $\mathbb{P}_\theta(T_\alpha \leq 0) \leq \inf_{m\in\mathcal{M}} P_\theta(m)$ where

$$P_\theta(m) = \mathbb{P}_\theta\left(\frac{N_m\|\Pi_{V\cup m}\mathbf{Y} - \Pi_V\mathbf{Y}\|_n^2}{D_m\|\mathbf{Y} - \Pi_{V\cup m}\mathbf{Y}\|_n^2} \leq \overline{F}^{-1}_{D_m,N_m}(\alpha_m)\right). \tag{A.1}$$

First, we derive the distribution of the test statistic $\phi_m(\mathbf{X},\mathbf{Y})$ under $\mathbb{P}_\theta$, then we give an upper bound for $P_\theta(m)$ and finally we shall gather the results in order to find a subset of $\mathbb{R}^{\mathcal{I}}$ over which the power of T_α is larger than δ.

The distribution of Y conditionally to the set of variables $(X_{V\cup m})$ is of the form

$$Y = \sum_{i\in V\cup m} \theta_i^{V\cup m} X_i + \epsilon^{V\cup m}, \tag{A.2}$$

where the vector $\theta^{V\cup m}$ is a constant and $\epsilon^{V\cup m}$ is a zero mean gaussian variable independent of $X_{V\cup m}$, whose variance is $\mathrm{var}(Y|X_{V\cup m})$. As a consequence, $\|\mathbf{Y} - \Pi_{V\cup m}\mathbf{Y}\|_n^2$ is exactly $\|\Pi_{(V\cup m)^\perp}\boldsymbol{\epsilon}^{V\cup m}\|_n^2$, where $\Pi_{(V\cup m)^\perp}$ denotes the orthogonal projection along the space generated by $(\mathbf{X}_i)_{i\in V\cup m}$.
Using the same decomposition of $\mathbf{Y}$ one simplifies the numerator of $\phi_m(\mathbf{X},\mathbf{Y})$:

$$\|\Pi_{V\cup m}\mathbf{Y} - \Pi_V\mathbf{Y}\|_n^2 = \left\|\sum_{i\in V\cup m} \theta_i^{V\cup m}(\mathbf{X}_i - \Pi_V\mathbf{X}_i) + \Pi_{V^\perp\cap(V\cup m)}\boldsymbol{\epsilon}^{V\cup m}\right\|_n^2,$$

where $\Pi_{V^\perp\cap(V\cup m)}$ is the orthogonal projection onto the intersection between the space generated by $(\mathbf{X}_i)_{i\in V\cup m}$ and the orthogonal of the space generated by $(\mathbf{X}_i)_{i\in V}$.

For any $i \in m$, let us consider the conditional distribution of X_i with respect to $\mathbf{X}_V$,

$$\mathbf{X}_i = \sum_{j \in V} \theta_j^{V,i} \mathbf{X}_j + \boldsymbol{\epsilon}_i^V . \tag{A.3}$$

where $\theta_j^{V,i}$ are constants and $\boldsymbol{\epsilon}_i^V$ is a zero-mean normal gaussian random variable whose variance is $\mathrm{var}(X_i|X_V)$ and which is independent of $\mathbf{X}_V$. This enables us to express

$$\mathbf{X}_i - \Pi_V \mathbf{X}_i = \Pi_{V^\perp \cap (V \cup m)} \boldsymbol{\epsilon}_i^V, \quad \text{for all } i \in m \ .$$

Therefore, we decompose $\phi_m(\mathbf{X}, \mathbf{Y})$ in

$$\phi_m(\mathbf{X}, \mathbf{Y}) = \frac{N_m \| \Pi_{V^\perp \cap (V \cup m)} \left(\sum_{i \in m} \theta_i^{V \cup m} \boldsymbol{\epsilon}_i^V + \boldsymbol{\epsilon}^{V \cup m} \right) \|_n^2}{D_m \| \Pi_{(V \cup m)^\perp} \boldsymbol{\epsilon}^{V \cup m} \|_n^2}. \tag{A.4}$$

Let us define the random variable $Z_m^{(1)}$ and $Z_m^{(2)}$ where $Z_m^{(1)}$ refers to the numerator of (A.4) divided by N_m and $Z_m^{(2)}$ to the denominator divided by D_m. We now prove that $Z_m^{(1)}$ and $Z_m^{(2)}$ are independent.

The variables $(\boldsymbol{\epsilon}_j^V)_{j \in m}$ are $\sigma(\mathbf{X}_{V \cup m})$-measurable as linear combinations of elements in $\mathbf{X}_{V \cup m}$. Moreover, $\boldsymbol{\epsilon}^{V \cup m}$ follows a zero mean normal distribution with covariance matrix $\mathrm{var}(Y|X_{V \cup m}) I_n$ and is independent of $\mathbf{X}_{V \cup m}$. As a consequence, conditionally to $\mathbf{X}_{V \cup m}$, $Z_m^{(1)}$ and $Z_m^{(2)}$ are independent by Cochran's Theorem as they correspond to projections onto two sets orthogonal from each other. Additionally, $Z_m^{(2)}$ is independent of $\mathbf{X}_{V \cup m}$. Indeed, almost surely conditionally to $\mathbf{X}_{V \cup m}$, $Z_m^{(2)} / \mathrm{var}(Y|X_{V \cup m})$ follows a χ^2 distribution with N_m degrees of freedom. This distribution does not depend on $\mathbf{X}_{V \cup m}$. As $Z_m^{(1)}$ and $Z_m^{(2)}$ are independent conditionally to $\mathbf{X}_{V \cup m}$ and as $Z_m^{(2)}$ is independent of $X_{V \cup m}$, $Z_m^{(1)}$ and $Z_m^{(2)}$ are independent.

As $\boldsymbol{\epsilon}_j^V$ is a linear combination of the columns of $\mathbf{X}_{V \cup m}$, $Z_m^{(1)}$ follows a non-central χ^2 distribution conditionally to $\mathbf{X}_{V \cup m}$:

$$(Z_m^{(1)}|X_{V \cup m}) \sim \mathrm{var}(Y|X_{V \cup m}) \chi^2 \left(\frac{\left\| \sum_{j \in m} \theta_j^{V \cup m} \Pi_{(V \cup m) \cap V^\perp} \boldsymbol{\epsilon}_j^V \right\|_n^2}{\mathrm{var}(Y|X_{V \cup m})}, D_m \right).$$

Let us derive the distribution of the non-central parameter. First, we simplify the projection term as $\boldsymbol{\epsilon}_j^V$ is a linear combinations of elements of $\mathbf{X}_{V \cup m}$.

$$\Pi_{(V \cup m) \cap V^\perp} \boldsymbol{\epsilon}_j^V = \Pi_{V \cup m} \boldsymbol{\epsilon}_j^V - \Pi_V \boldsymbol{\epsilon}_j^V = \Pi_{V^\perp} \boldsymbol{\epsilon}_j^V .$$

Let us define κ_m^2 as

$$\kappa_m^2 := \frac{\mathrm{var}\left(\sum_{j \in m} \theta_j^{V \cup m} \epsilon_j^V \right)}{\mathrm{var}(Y|X_{V \cup m})}.$$

As the variable $\sum_{j\in m}\theta_j^{V\cup m}\boldsymbol{\epsilon}_j^V$ is independent of $\mathbf{X}_V$, and as almost surely the dimension of the vector space generated by $\mathbf{X}_V$ is d, we are able to derive the distribution of the non-central parameter:

$$\frac{\left\|\sum_{j\in m}\theta_j^{V\cup m}\Pi_{V^\perp}\boldsymbol{\epsilon}_j^V\right\|_n^2}{\mathrm{var}(Y|X_{V\cup m})}\sim\kappa_m^2\chi^2(n-d).$$

To sum up, let us express simply the distribution of $Z_m^{(1)}$. Let U,V and W be three independent random variables which respectively follow a χ^2 distribution with $n-d$ degrees of freedom, a standard normal distribution and a χ^2 distribution with D_m-1 degrees of freedom. Then,

$$Z_m^{(1)}\sim\mathrm{var}(Y|X_{V\cup m})\left[\left(\kappa_m\sqrt{U}+V\right)^2+W\right]. \tag{A.5}$$

In fact, κ_m^2 easily simplifies in a quotient of conditional variances. Let us first express $\mathrm{var}(Y|X_V)$ in term of $\mathrm{var}(Y|X_{m\cup V})$ using the decomposition (A.2) of Y.

$$\begin{aligned}\mathrm{var}(Y|X_V) &= \mathrm{var}\left(\sum_{j\in V\cup m}\theta_j^{V\cup m}X_j+\epsilon^{V\cup m}\,|X_V\right)\\ &= \mathrm{var}\left(\sum_{j\in V\cup m}\theta_j^{V\cup m}X_j\,|X_V\right)+\mathrm{var}\left(\epsilon^{V\cup m}\,|X_V\right)\\ &= \mathrm{var}\left(\sum_{j\in V\cup m}\theta_j^{V\cup m}X_j\,|X_V\right)+\mathrm{var}\left(Y\,|X_{V\cup m}\right),\end{aligned} \tag{A.6}$$

as $\epsilon^{V\cup m}$ is independent of $X_{V\cup m}$. Now using the definition of ϵ_j^V in (A.3), it turns out that

$$\begin{aligned}\mathrm{var}\left(\sum_{j\in V\cup m}\theta_j^{V\cup m}X_j|X_V\right) &= \mathrm{var}\left(\sum_{j\in m}\theta_j^{V\cup m}X_j|X_V\right)\\ &= \mathrm{var}\left(\sum_{j\in m}\theta_j^{V\cup m}\epsilon_j^V\,|X_V\right)\\ &= \mathrm{var}\left(\sum_{j\in m}\theta_j^{V\cup m}\epsilon_j^V\right),\end{aligned} \tag{A.7}$$

as the $(\epsilon_j^V)_{j\in m}$ are independent of X_V. Gathering formulae (A.6) and (A.7), we get

$$\kappa_m^2=\frac{\mathrm{var}(Y|X_V)-\mathrm{var}(Y|X_{V\cup m})}{\mathrm{var}(Y|X_{V\cup m})}. \tag{A.8}$$

As we know the distribution of $\phi_m(\mathbf{X}, \mathbf{Y})$ under the distribution $\mathbb{P}_\theta$, we are now in position to work out precise upper bounds for $P_\theta(m)$.

$$\begin{aligned} P_\theta(m) &= \mathbb{P}_\theta\left(\frac{N_m}{D_m} Z_m^{(1)} \leq \overline{F}^{-1}_{D_m,N_m}(\alpha_m) Z_m^{(2)}\right) \\ &= \mathbb{P}_\theta\left(\frac{1}{\mathrm{var}(Y|X_{V\cup m})}\left(\frac{D_m}{N_m}\overline{F}^{-1}_{D_m,N_m}(\alpha_m) Z_m^{(2)} - Z_m^{(1)}\right) \geq 0\right). \end{aligned} \tag{A.9}$$

Let us call Z_m the random variable defined by

$$Z_m := \frac{1}{\mathrm{var}(Y|X_{V\cup m})}\left(\frac{D_m}{N_m}\overline{F}^{-1}_{D_m,N_m}(\alpha_m) Z_m^{(2)} - Z_m^{(1)}\right).$$

It is possible to control the quantity $P_\theta(m)$ by bounding the deviations of Z_m.

Lemma 18. *For any $x > 0$, the random variable Z_m defined above satisfies the inequality:*

$$\mathbb{P}_\theta(Z_m - \mathbb{E}(Z_m) \geq c_m x + 2\sqrt{v_m x}) \leq \exp(-x), \tag{A.10}$$

where c_m and v_m refer to:

$$\begin{aligned} c_m &:= \frac{2D_m}{N_m}\overline{F}^{-1}_{D_m,N_m}(\alpha_m), \\ v_m &:= D_m + (n-d)\left(2\kappa_m^2 + \kappa_m^4\right) + \frac{D_m^2}{N_m}\left(\overline{F}^{-1}_{D_m,N_m}(\alpha_m)\right)^2. \end{aligned}$$

We now apply this lemma choosing $x = L$,

$$\mathbb{P}_\theta\left(Z_m \geq \mathbb{E}(Z_m) + c_m L + 2\sqrt{v_m L}\right) \leq \delta.$$

Therefore, $\mathbb{P}_\theta(T_\alpha \leq 0) \leq \delta$ if for some $m \in \mathcal{M}$,

$$\mathbb{E}(Z_m) + c_m L + 2\sqrt{v_m L} \leq 0. \tag{A.11}$$

It is straightforward to compute the expectation of Z_m:

$$\mathbb{E}(Z_m) = -\kappa_m^2(n-d) - D_m + D_m\overline{F}^{-1}_{D_m,N_m}(\alpha_m).$$

Using this last equality, condition (A.11) is equivalent to the following inequality:

$$\kappa_m^2(n-d) \geq D_m\left(\overline{F}^{-1}_{D_m,N_m}(\alpha_m) - 1\right) + c_m L + 2\sqrt{L v_m}. \tag{A.12}$$

Thanks to the definition of v_m in Lemma 18, we now bound the term $2\sqrt{Lv_m}$. If $\kappa_m^2 \geq 2$, then

$$2\sqrt{Lv_m} \leq 2\sqrt{LD_m} + 2\kappa_m^2\sqrt{2L(n-d)} + 2\overline{F}^{-1}_{D_m,N_m}(\alpha_m) D_m\sqrt{\frac{L}{N_m}}.$$

On the other hand, if $\kappa_m^2 \leq 2$, we obtain an alternative upper bound using the inequality $2uv \leq 4u^2 + v^2/4$,

$$2\sqrt{Lv_m} \quad \leq \quad 2\sqrt{LD_m} + (n-d)\kappa_m^2/2 + 8L + 2\overline{F}^{-1}_{D_m,N_m}(\alpha_m) D_m \sqrt{\frac{L}{N_m}}.$$

Gathering these two inequalities, whatever the value of κ_m^2,

$$2\sqrt{Lv_m} \leq 2\sqrt{LD_m} + (n-d)\kappa_m^2 \left(1/2 \vee 2\sqrt{\frac{2L}{n-d}}\right) + 2\overline{F}^{-1}_{D_m,N_m}(\alpha_m) D_m \sqrt{\frac{L}{N_m}} + 8L. \quad \text{(A.13)}$$

Combining the upper bound (A.13) with condition (A.12) enables to give a condition in term of κ_m^2. Indeed, $P_\theta(m) \leq \delta$ if

$$\kappa_m^2 \geq \frac{\frac{D_m}{N_m}\overline{F}^{-1}_{D_m,N_m}(\alpha_m)\left[N_m + 2\sqrt{N_m L} + 2L\right] + 2\sqrt{D_m L} - D_m + 8L}{(n-d)\left(1 - \left(\frac{1}{2} \vee 2\sqrt{\frac{2L}{n-d}}\right)\right)}. \quad \text{(A.14)}$$

To bound $\overline{F}^{-1}_{D_m,N_m}(\alpha_m)$, we use Lemma 1 in [2]:

Lemma 19. *Let $u \in]0,1[$ and $\overline{F}^{-1}_{D,N}(u)$ be the $1-u$ quantile of a Fisher random variable with D and N degrees of freedom. Then we have*

$$\begin{aligned} D\overline{F}^{-1}_{D,N}(u) \quad \leq \quad & D + 2\sqrt{D\left(1 + \frac{D}{N}\right)\log\left(\frac{1}{u}\right)} \\ & + \left(1 + 2\frac{D}{N}\right)\frac{N}{2}\left[\exp\left(\frac{4}{N}\log\left(\frac{1}{u}\right)\right) - 1\right]. \end{aligned} \quad \text{(A.15)}$$

Gathering inequalities (A.14) and (A.15), $P_\theta(m) \leq \delta$ if

$$\kappa_m^2 \geq \frac{A + B}{(n-d)\left(1 - \left(\frac{1}{2} \vee 2\sqrt{\frac{2L}{n-d}}\right)\right)}, \quad \text{(A.16)}$$

where

$$\begin{aligned} A \quad := \quad & 2\sqrt{D_m\left(1 + \frac{D_m}{N_m}\right)\log\left(\frac{1}{\alpha_m}\right)}\left[1 + 2\sqrt{\frac{L}{N_m}} + 2\frac{L}{N_m}\right] + 2D_m\left[\sqrt{\frac{L}{N_m}} + \frac{L}{N_m}\right] \\ & + 2\sqrt{D_m L}, \\ B \quad := \quad & \left(1 + 2\frac{D_m}{N_m}\right)\frac{N_m}{2}\left[\exp\left(\frac{4}{N}\log\left(\frac{1}{\alpha_m}\right)\right) - 1\right]\left(1 + 2\sqrt{\frac{L}{N_m}} + 2\frac{L}{N_m}\right) + 8L. \end{aligned}$$

By factorizing and bounding the last two terms of A, we get

$$\begin{aligned} 2D_m\left[\sqrt{\frac{L}{N_m}} + \frac{L}{N_m}\right] + 2\sqrt{D_m L} \quad &= \quad 2\sqrt{D_m L}\left(1 + \sqrt{\frac{D_m}{N_m}} + \sqrt{\frac{D_m}{N_m}}\sqrt{\frac{L}{N_m}}\right) \\ &\leq \quad 2\sqrt{D_m L}\left[1 + \sqrt{\frac{D_m}{N_m}}\right]\left[1 + 2\sqrt{\frac{L}{N_m}} + 2\frac{L}{N_m}\right]. \end{aligned}$$

It follows that

$$\begin{aligned} A &\leq 2\sqrt{D_m}\left(1+\sqrt{\frac{D_m}{N_m}}\right)\left(1+2\sqrt{\frac{L}{N_m}}+2\frac{L}{N_m}\right)\left[\sqrt{L}+\sqrt{L_m}\right] \\ &\leq 4\sqrt{D_m}l_m\left(1+\sqrt{\frac{D_m}{N_m}}\right)\left[\sqrt{\log\left(\frac{1}{\alpha_m\delta}\right)}\right]. \end{aligned} \tag{A.17}$$

Using the inequality $\exp(u)-1 \leq u\exp(u)$ which holds for all $u>0$, we derive that

$$\begin{aligned} B &\leq 2\left(1+2\frac{D_m}{N_m}\right)\log\left(\frac{1}{\alpha_m}\right)\exp\left[\frac{4}{N_m}\log\left(\frac{1}{\alpha_m}\right)\right]\left[1+2\sqrt{\frac{L}{N_m}}+2\frac{L}{N_m}\right]+8L \\ &\leq \log\left(\frac{1}{\alpha_m\delta}\right)\left(8\vee k_m l_m\left(1+\frac{2D_m}{N_m}\right)\right). \end{aligned} \tag{A.18}$$

Combining inequalities (A.16), (A.17), and (A.18) we obtain the condition (4.12). Under assumption $H_{\mathcal{M}}$, $L_m \leq N_m/10$ for all $m \in \mathcal{M}$ and $L \leq N_m/21$. The terms L/N_m, L_m/N_m, k_m, and l_m are bounded by a constant and the second part of the theorem follows easily.

A.1.2 Proof of Lemma 18

We prove this deviation inequality thanks to Laplace method. First of all, one has to upper bound the Laplace transform of the variable

$$Z_m \sim \frac{D_m}{N_m}\overline{F}^{-1}_{D_m,N_m}(\alpha_m)T - \left((\kappa_m\sqrt{U}+V)^2+W\right),$$

where we recall that T, U, V, and W are independent random variables which follow respectively a χ^2 distribution with N_m degrees of freedom, a χ^2 distribution with $n-d$ degrees of freedom, a standard normal distribution and a χ^2 distribution with D_m-1 degree of freedom. To keep the formulae as short as possible, λ_m will refer to $\frac{D_m}{N_m}\overline{F}^{-1}_{D_m,N_m}(\alpha_m)$.

$$\begin{aligned} &\mathbb{E}\left[\exp\left(-t\left(\kappa_m\sqrt{U}+V\right)^2\right)\right] \\ &= \int \exp\left(-t\left(\kappa_m\|x\|_{n-d}+y\right)^2\right)\frac{1}{(2\pi)^{(n-d+1)/2}}\exp\left(-\frac{\|x\|^2_{n-d}-y^2}{2}\right)dxdy \\ &= \frac{1}{\sqrt{1+2t}}\left[\frac{1+2t}{1+2t[\kappa_m^2+1]}\right]^{(n-d)/2}, \end{aligned}$$

by standard Gaussian computation. After multiplication by the Laplace transform of $\frac{D_m}{N_m}\overline{F}^{-1}_{D_m,N_m}(\alpha_m)T$ and W, we get:

$$\mathbb{E}\left[\exp\left(tZ_m\right)\right] = \frac{(1+2t)^{N_m/2}}{(1+2t[\kappa_m^2+1])^{(n-d)/2}\left(1-2t\lambda_m\right)^{N_m/2}}.$$

Clearly, the expectation of Z_m is

$$\mathbb{E}(Z_m) = \lambda_m N_m - (\kappa_m^2(n-d)+D_m).$$

One then obtains $\Psi_m(t)$ the log-Laplace transform of $Z_m - \mathbb{E}(Z_m)$:

$$\begin{aligned}\Psi_m(t) &= \frac{N_m}{2}\log\left(\frac{1+2t}{1-2t\lambda_m}\right) - \frac{n-d}{2}\log\left(1+2t[\kappa_m^2+1]\right) - t\mathbb{E}(Z_m)\\ &= -\frac{D_m}{2}\log(1+2t) - \frac{n-d}{2}\log\left(1+\frac{2t\kappa_m^2}{1+2t}\right) - \frac{N_m}{2}\log(1-2t\lambda_m) - t\mathbb{E}(Z_m).\end{aligned}$$

Using the inequality $\log(1+u) \geq u - u^2/2$ which holds for all $u > 0$, we derive that for any $t \geq 0$,

$$\begin{aligned}\Psi_m(t) &\leq D_m t^2 + (n-d)\left[-\frac{t\kappa_m^2}{1+2t} + t\kappa_m^2 + \frac{t^2\kappa_m^4}{(1+2t)^2}\right] - \frac{N_m}{2}\log(1-2t\lambda_m) - t\lambda_m N_m\\ &\leq D_m t^2 + (n-d)\left[\frac{2t^2\kappa_m^2}{1+2t} + \frac{t^2\kappa_m^4}{(1+2t)^2}\right] - \frac{N_m}{2}\log(1-2t\lambda_m) - t\lambda_m N_m\\ &\leq t^2\left[D_m + (n-d)(2\kappa_m^2+\kappa_m^4)\right] - \frac{N_m}{2}\log(1-2t\lambda_m) - t\lambda_m N_m.\end{aligned}$$

For any $0 \leq u \leq 1/2$, it holds that $-u - 1/2\log(1-2u) \leq \frac{u^2}{1-2u}$ (compare the power series). As a consequence, for any $0 \leq t \leq \frac{\lambda_m}{2}$,

$$\begin{aligned}\Psi_m(t) &\leq t^2\left[D_m + (n-d)(2\kappa_m^2+\kappa_m^4)\right] + N_m\frac{\lambda_m^2 t^2}{1-2t\lambda_m}\\ &\leq \frac{t^2}{1-2\lambda_m t}\left(D_m + (n-d)(2\kappa_m^2+\kappa_m^4) + N_m\lambda_m^2\right). \qquad \text{(A.19)}\end{aligned}$$

We now refer to [3], where it is proved that if

$$\log\left(\mathbb{E}\left[e^{tZ}\right]\right) \leq \frac{vt^2}{2(1-ct)},$$

then for any positive x,

$$\mathbb{P}\left(Z \geq cx + \sqrt{2vx}\right) \leq e^{-x}.$$

Applying this property to the upper bound (A.19) and replacing λ_m by its value enable to prove (A.10).

A.1.3 Proof of Proposition 5

We first recall the classical upper bound for the binomial coefficient (see for instance (2.9) in [4]).

$$\log|\mathcal{M}(k,p)| = \log\binom{p}{k} \leq k\log\left(\frac{ep}{k}\right).$$

As a consequence, $\log(1/\alpha_m) \leq \log(1/\alpha) + k\log\left(\frac{ep}{k}\right)$. The assumption on n in Proposition 5 therefore implies hypothesis $H_{\mathcal{M}}$ applied to this class of models. Thus, we are in position to apply the second result of Theorem 3. Moreover, the assumption on n implies that

$n \geq 11k$ and D_m/N_m is thus smaller than $1/10$ for any model m in $\mathcal{M}(k,p)$. Formula (4.13) in Theorem 3 then translates into

$$\triangle(m) \leq \frac{(1+\sqrt{0.1})C_1\left(\sqrt{k^2\log\left(\frac{ep}{k}\right)}+\sqrt{k\log\left(\frac{1}{\alpha\delta}\right)}\right)+1.2C_2\left(k\log\left(\frac{ep}{k}\right)+\log\left(\frac{1}{\alpha\delta}\right)\right)}{n},$$

and it follows that Proposition 5 holds.

A.1.4 Proof of Proposition 9

We fix the constant C_5 to be $10 \vee 2C_4'$ where C_4' is defined below and C_6 to be 21. This choice of constants allows the procedure T_α to satisfy Hypothesis $H_{\mathcal{M}}$. An argument similar to the proof of Proposition 5 allows to show easily that there exists a universal constant C_3' such that if we set

$$\rho_1'^2 \quad = \quad \frac{C_3'\left(\log(p)+\log\left(\frac{2}{\alpha\delta}\right)\right)}{n} = \frac{C_3'}{n}\log\left(\frac{2p}{\alpha\delta}\right), \tag{A.20}$$

then $\frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} \geq \rho_1'^2$ implies that $\mathbb{P}_\theta\left(T_\alpha > 0\right) \geq 1-\delta$. Here, the factor 2 in the logarithm comes from the fact that some weights α_m equal $\alpha/(2p)$.

Let ρ^2 and λ^2 be two positive numbers such that $\frac{\lambda^2}{\mathrm{var}(Y)-\lambda^2} = \rho^2$ and let $\theta \in \Theta[1,p]$ such that $\|\theta\|^2 = \lambda^2$. As $\mathrm{corr}(X_1, X_i) = c$ for any i in $\{2 \dots p\}$,

$$\frac{\mathrm{var}(Y) - \mathrm{var}(Y|X_1)}{\mathrm{var}(Y|X_1)} \geq \frac{c\lambda^2}{\mathrm{var}(Y) - c\lambda^2}.$$

We now apply Theorem 3 to $\phi_{\{1\},\alpha/2}$ under $H_{\mathcal{M}}$. There exists a universal constant C_4' such that $\mathbb{P}_\theta\left(\phi_{\{1\},\alpha/2} > 0\right) \geq 1-\delta$ if

$$\frac{c\lambda^2}{\mathrm{var}(Y) - c\lambda^2} \geq \frac{C_4'}{n}\log\left(\frac{2}{\alpha\delta}\right).$$

This last condition is equivalent to

$$\frac{\lambda^2}{\mathrm{var}(Y)} \geq \frac{C_4'}{nc + cC_4'\log\left(\frac{2}{\alpha\delta}\right)}\log\left(\frac{2}{\alpha\delta}\right). \tag{A.21}$$

Let us assume that $c \geq \log\left(\frac{2}{\alpha\delta}\right)/\log\left(\frac{2p}{\alpha\delta}\right)$. As $n \geq 2C_4'\log\left(\frac{2p}{\alpha\delta}\right)$ (hypothesis (4.20) and definition of C_5), $nc \geq 2C_4'\log\left(\frac{2}{\alpha\delta}\right)$. As a consequence, condition (A.21) is implied by:

$$\rho^2 \geq \frac{2C_4'}{nc}\log\left(\frac{2}{\alpha\delta}\right). \tag{A.22}$$

Let us define C_7 as the supremum of C_3' and $2C_4'$. Combining (A.20) and (A.22) allows to conclude that $\mathbb{P}_\theta\left(T_\alpha > 0\right) \geq 1-\delta$ if

$$\rho^2 \geq \frac{C_7}{n}\left(\log\left(\frac{2p}{\alpha\delta}\right) \bigwedge \frac{1}{c}\log\left(\frac{2}{\alpha\delta}\right)\right).$$

If c is smaller than $\log\left(\frac{2}{\alpha\delta}\right)/\log\left(\frac{2p}{\alpha\delta}\right)$, this last result also holds by (A.20).

A.1.5 Proof of Proposition 11

First, we have to check that the test T_α satisfies condition $H_{\mathcal{M}}$. As the dimension of each model is smaller than $n/2$, for any model m in $\mathcal{M}$, N_m is larger than $n/2$. Moreover, for any model m in $\mathcal{M}$, α_m is larger than $\alpha/(2|\mathcal{M}|)$ and $|\mathcal{M}|$ is smaller than $n/2$. As a consequence, the first condition of $H_{\mathcal{M}}$ is implied by the inequality:

$$n \geq 20 \log\left(\frac{n}{\alpha}\right). \tag{A.23}$$

Hypothesis (4.24) implies that $n/2 \geq 20\log\left(\frac{40}{\alpha}\right)$. Besides, for any $n > 0$ it holds that $n/2 \geq 20\log\left(\frac{n}{40}\right)$. Combining these two lower bounds enables to obtain (A.23). The second condition of $H_{\mathcal{M}}$ holds if $n \geq 42\log\left(\frac{1}{\delta}\right)$ which is a consequence of hypothesis (4.24).

Let us first consider the case $n < 2p$. Let us apply Theorem 3 under hypothesis $H_{\mathcal{M}}$ to T_α. $\mathbb{P}_\theta(T_\alpha > 0) \geq 1-\delta$ for all $\theta \in \Theta$ such that

$$\exists i \in \{1,\dots,[n/2]\}, \frac{\mathrm{var}(Y) - \mathrm{var}\left(Y|X_{m_i}\right)}{\mathrm{var}\left(Y|X_{m_i}\right)} \geq C_5' \frac{\sqrt{i\log\left(\frac{[n/2]}{\alpha\delta}\right)} + \log\left(\frac{[n/2]}{\alpha\delta}\right)}{n}, \tag{A.24}$$

where C_5' is universal constant (equals $2C_1 \vee 4C_2$).

Let θ be an element of $\mathcal{E}_a(R)$ which satisfies

$$\|\theta\|^2 \geq (1+C_5')\left(\mathrm{var}(Y|X_{m_i}) - \mathrm{var}(Y|X)\right) + (1+C_5')\mathrm{var}(Y|X)\frac{\sqrt{i\log\left(\frac{[n/2]}{\alpha\delta}\right)} + \log\left(\frac{[n/2]}{\alpha\delta}\right)}{n}.$$

Using hypothesis (4.24), we show that, for any i between 1 and $[n/2]$, $\frac{\sqrt{i\log\left(\frac{[n/2]}{\alpha\delta}\right)}+\log\left(\frac{[n/2]}{\alpha_\delta}\right)}{n} \leq 1$. It is then straighforward to check that θ satisfies (A.24).
As θ belongs to the set $\mathcal{E}_a(R)$,

$$\begin{aligned}\mathrm{var}(Y|X_{m_i}) - \mathrm{var}(Y|X) &= a_{i+1}^2 \mathrm{var}(Y|X) \sum_{j=i+1}^{p} \frac{\mathrm{var}(Y|X_{m_{j-1}}) - \mathrm{var}(Y|X_{m_j})}{a_{i+1}^2 \mathrm{var}(Y|X)} \\ &\leq a_{i+1}^2 \mathrm{var}(Y|X) R^2.\end{aligned}$$

As a consequence if θ belongs to $\mathcal{E}_a(R)$ and satisfies

$$\|\theta\|^2 \geq (1+C_5')\mathrm{var}(Y|X)\left[\left(a_{i+1}^2 R^2 + \frac{\sqrt{i\log\left(\frac{[n/2]}{\alpha\delta}\right)}}{n}\right) + \frac{1}{n}\log\left(\frac{[n/2]}{\alpha\delta}\right)\right],$$

then $\mathbb{P}_\theta(T_\alpha \leq 0) \leq \delta$. Gathering this condition for any i between 1 and $[n/2]$ allows to conclude that if θ satifies

$$\frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq (1+C_5')\left[\inf_{1\leq i\leq [n/2]}\left(a_{i+1}^2 R^2 + \frac{\sqrt{i\log\left(\frac{n/2}{\alpha\delta}\right)}}{n}\right) + \frac{1}{n}\log\left(\frac{n/2}{\alpha\delta}\right)\right],$$

then $\mathbb{P}_\theta(T_\alpha \leq 0) \leq \delta$.

Let us now turn to the case $n \geq 2p$. Let us consider T_α as the supremum of $p-1$ tests of level $\alpha/2(p-1)$ and one test of level $\alpha/2$. By considering the $p-1$ firsts tests, we obtain as in the previous case that $\mathbb{P}_\theta(T_\alpha \leq 0) \leq \delta$ if

$$\frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq (1 + C_5') \left[\inf_{1 \leq i \leq (p-1)} \left(a_{i+1}^2 R^2 + \frac{\sqrt{i \log\left(\frac{(p-1)/2}{\alpha\delta}\right)}}{n} \right) + \frac{1}{n} \log\left(\frac{(p-1)/2}{\alpha\delta}\right) \right].$$

On the other hand, using the last test statistic $\phi_{\mathcal{I},\alpha/2}$, $\mathbb{P}_\theta(T_\alpha \leq 0) \leq \delta$ if

$$\frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq C_5' \frac{\sqrt{p \log\left(\frac{2}{\alpha\delta}\right)} + \log\left(\frac{2}{\alpha\delta}\right)}{n}.$$

Gathering these two conditions allows to prove (4.26).

A.1.6 Proof of Proposition 12

The approach behind this proof is similar to the one for Proposition 11. First, we check that our class of models $\mathcal{M}$ and weights α_m satisfy hypothesis $H_{\mathcal{M}}$ as in the previous proof.

Let us give a sharper upper bound on $|\mathcal{M}|$:

$$|\mathcal{M}| \leq 1 + \log(n/2 \wedge p)/\log(2) \leq \log(n \wedge 2p)/\log(2). \tag{A.25}$$

We deduce from (A.25) that there exists a constant $C(\alpha, \delta)$ only depending on α and δ such that for all $m \in \mathcal{M}$,

$$\log\left(\frac{1}{\alpha_m \delta}\right) \leq C(\alpha, \delta) \log\log(n \wedge p).$$

First, let us consider the case $n < 2p$. We apply Theorem 3 under the assumption $H_{\mathcal{M}}$. As in the proof of Proposition 11, we obtain that $\mathbb{P}_\theta(T_\alpha > 0) \geq 1 - \delta$ if

$$\frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq C'(\alpha, \delta) \left[\inf_{i \in \{2^j, j \geq 0\} \cap \{1, \ldots, [n/2]\}} \left(R^2 (i+1)^{-2s} + \frac{\sqrt{i \log\log n}}{n} \right) + \frac{\log\log n}{n} \right],$$

where $C'(\alpha, \delta)$ is a constant which only depends on α and δ. It is worth noting that $R^2 i^{-2s} \leq \frac{\sqrt{i \log\log n}}{n}$ if and only if

$$i \geq i^* = \left(\frac{R^2 n}{\sqrt{\log\log n}} \right)^{2/(1+4s)}$$

Under the assumption on R, i^* is larger than one. Let us distinguish between two cases. If there exists i' in $\{2^j, j \geq 0\} \cap \{1, \ldots, [n/2]\}$ such that $i^* \leq i'$, one can take $i' \leq 2i^*$ and then

$$\begin{aligned} \inf_{i \in \{2^j, j \geq 0\} \cap \{1, \ldots, [n/2]\}} \left(R^2 i^{-2s} + \frac{\sqrt{i \log\log n}}{n} \right) &\leq 2\frac{\sqrt{i' \log\log n}}{n} \\ &\leq 2\sqrt{2} R^{2/(1+4s)} \left(\frac{\sqrt{\log\log n}}{n} \right)^{4s/(1+4s)} \end{aligned} \quad \text{(A.26)}$$

Else, we take $i' \in \{2^j, j \geq 0\} \cap \{1, \ldots, [n/2]\}$ such that $n/4 \leq i' \leq n/2$. Since $i' \leq (i^* \wedge n/2)$ we obtain that

$$\inf_{i \in \{2^j, j \geq 0\} \cap \{1, \ldots, [n/2]\}} \left(R^2 i^{-2s} + \frac{\sqrt{i \log\log n}}{n} \right) \leq 2R^2 i'^{-2s} \leq 2R^2 \left(\frac{n}{2} \right)^{-2s}. \quad \text{(A.27)}$$

Gathering inequalities (A.26) and (A.27) allows to prove (4.28).

We now turn to the case $n \geq 2p$. As in the proof of Proposition 11, we divide the proof into two parts: first we give an upper bound of the power for the $|\mathcal{M}| - 1$ first tests which define T_α and then we give an upper bound for the last test $\phi_{\mathcal{I},\alpha/2}$. Combining these two inequalities allows us to prove (4.29).

A.1.7 Proof of Proposition 14

We first note that the assumption about R^2 implies that $D^* \geq 2$. As N_m is larger than $n/2$, it is straightforward to show that this test satisfies condition $H_{\mathcal{M}}$. As a consequence, we can apply the second part of Theorem 3. $\mathbb{P}_\theta(T_\alpha^* \leq 0) \leq \delta$ for any θ such that

$$\frac{\mathrm{var}(Y) - \mathrm{var}(Y|X_{m_{D^*}})}{\mathrm{var}(Y|X_{m_{D^*}})} \geq C_2'(\alpha, \delta) \frac{\sqrt{D^*}}{n}, \quad \text{(A.28)}$$

where $C_2'(\alpha, \delta)$ only depends on α and δ. Now, we use the same sketch as in the proof of Proposition 11. For any $\theta \in \mathcal{E}_a(R)$, condition (A.28) is equivalent to:

$$\|\theta\|^2 \geq \big(\mathrm{var}(Y|X_{m_{D^*}}) - \mathrm{var}(Y|X)\big) \left(1 + C_2'(\alpha, \delta) \frac{\sqrt{D^*}}{n} \right) + \mathrm{var}(Y|X) C_2'(\alpha, \delta) \frac{\sqrt{D^*}}{n} \quad \text{(A.29)}$$

Moreover, as θ belongs to $\mathcal{E}_a(R)$,

$$\mathrm{var}(Y|X_{m_{D^*}}) - \mathrm{var}(Y|X) \leq a_{D^*+1}^2 R^2 \mathrm{var}(Y|X) \leq a_{D^*}^2 \mathrm{var}(Y|X) R^2.$$

As $\sqrt{D^*}/n$ is smaller than one, condition (A.29) is implied by

$$\frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq (1 + C_2'(\alpha, \delta)) \left(a_{D^*}^2 R^2 + \frac{\sqrt{D^*}}{n} \right).$$

As $a_{D^*}^2 R^2$ is smaller than $\frac{\sqrt{D^*}}{n}$ which is smaller $\sup_{1 \leq i \leq p} \left[\frac{\sqrt{i}}{n} \wedge a_i^2 R^2 \right]$, it turns out that $\mathbb{P}_\theta(T_\alpha^* = 0) \leq \delta$ for any θ belonging to $\mathcal{E}_a(R)$ such that

$$\frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq 2(1 + C_2'(\alpha, \delta)) \sup_{1 \leq i \leq p} \left[\frac{\sqrt{i}}{n} \wedge a_i^2 R^2 \right].$$

A.2 Proofs of Theorem 7, Proposition 4, 6, 8, 10, 13, 15, and 16

A.2.1 Proof of Theorem 7

This proof follows the general method for obtaining lower bounds described in Section 7.1 in [1]. We first remind the reader of the main arguments of the approach applied to our model. Let ρ be some positive number and μ_ρ be some probability measure on

$$\Theta[k,p,\rho] := \left\{ \theta \in \Theta[k,p], \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} = \rho \right\}.$$

We define $\mathbb{P}_{\mu_\rho} = \int \mathbb{P}_\theta \mathrm{d}\mu_\rho(\theta)$ and Φ_α the set of level-α tests of the hypothesis "$\theta = 0$". Then,

$$\begin{aligned} \beta_I(\Theta[k,p,\rho]) &\geq \inf_{\phi_\alpha \in \Phi_\alpha} \mathbb{P}_{\mu_\rho}[\phi_\alpha = 0] \\ &\geq 1 - \alpha - \sup_{A,\ \mathbb{P}_0(A) \leq \alpha} |\mathbb{P}_{\mu_\rho}(A) - \mathbb{P}_0(A)| \\ &\geq 1 - \alpha - \frac{1}{2}\|\mathbb{P}_{\mu_\rho} - \mathbb{P}_0\|_{TV}, \end{aligned} \tag{A.30}$$

where $\|\mathbb{P}_{\mu_\rho} - \mathbb{P}_0\|_{TV}$ denotes the total variation norm between the probabilities $\mathbb{P}_{\mu_\rho}$ and $\mathbb{P}_0$. If we suppose that $\mathbb{P}_{\mu_\rho}$ is absolutely continuous with respect to $\mathbb{P}_0$, we can upper bound the norm in total variation between these two probabilities as follows. We define

$$L_{\mu_\rho}(\mathbf{Y}, \mathbf{X}) := \frac{\mathrm{d}\mathbb{P}_{\mu_\rho}}{\mathrm{d}\mathbb{P}_0}(\mathbf{Y}, \mathbf{X}).$$

Then, we get the upper bound

$$\begin{aligned} \|\mathbb{P}_{\mu_\rho} - \mathbb{P}_0\|_{TV} &= \int |L_{\mu_\rho}(\mathbf{Y}, \mathbf{X}) - 1| \mathrm{d}\mathbb{P}_0(\mathbf{Y}, \mathbf{X}) \\ &\leq \left(\mathbb{E}_0 \left[L^2_{\mu_\rho}(\mathbf{Y}, \mathbf{X}) \right] - 1 \right)^{1/2}. \end{aligned}$$

Thus, we deduce from (A.30) that

$$\beta_I(\Theta[k,p,\rho]) \geq 1 - \alpha - \frac{1}{2}\left(\mathbb{E}_0 \left[L^2_{\mu_\rho}(\mathbf{Y}, \mathbf{X}) \right] - 1 \right)^{1/2}.$$

If we find a number $\rho^* = \rho^*(\eta)$ such that

$$\log\left(\mathbb{E}_0 \left[L^2_{\mu_{\rho^*}}(\mathbf{Y}, \mathbf{X}) \right] \right) \leq \mathcal{L}(\eta), \tag{A.31}$$

then for any $\rho \leq \rho^*$,

$$\beta_I(\Theta[k,p,\rho]) \geq 1 - \alpha - \frac{\eta}{2} = \delta.$$

To apply this method, we first have to define a suitable prior μ_ρ on $\Theta[k,p,\rho]$. Let $\widehat{m}$ be some random variable uniformly distributed over $\mathcal{M}(k,p)$ and for each $m \in \mathcal{M}(k,p)$, let

$\epsilon^m = (\epsilon_j^m)_{j\in m}$ be a sequence of independent Rademacher random variables. We assume that for all $m \in \mathcal{M}(k,p)$, ϵ^m and $\widehat{m}$ are independent. Let ρ be given and μ_ρ be the distribution of the random variable $\widehat{\theta} = \sum_{j\in\widehat{m}} \lambda \epsilon_j^{\widehat{m}} e_j$ where

$$\lambda^2 := \frac{\mathrm{var}(Y)\rho^2}{k(1+\rho^2)},$$

and where $(e_j)_{j\in\mathcal{I}}$ is the orthonormal family of vectors of $\mathbb{R}^{\mathcal{I}}$ defined by

$$(e_j)_i = 1 \text{ if } i = j \text{ and } (e_i)_j = 0 \text{ otherwise.}$$

Straightforwardly, μ_ρ is supported by $\Theta[k,p,\rho]$. For any m in $\mathcal{M}(k,p)$ and any vector $(\zeta_j^m)_{j\in m}$ with values in $\{-1;1\}$, let $\mu_{m,\zeta^m,\rho}$ be the dirac measure on $\sum_{j\in m}\lambda\zeta_j^m e_j$. For any m in $\mathcal{M}(k,p)$, $\mu_{m,\rho}$ denotes the distribution of the random variable $\sum_{j\in m}\lambda\zeta_j^m e_j$ where (ζ_j^m) is a sequence of independent Rademacher random variables. These definitions easily imply

$$L_{\mu_\rho}(\mathbf{Y},\mathbf{X}) = \frac{1}{\binom{p}{k}} \sum_{m\in\mathcal{M}(k,p)} L_{\mu_{m,\rho}}(\mathbf{Y},\mathbf{X}) = \frac{1}{2^k\binom{p}{k}} \sum_{m\in\mathcal{M}(k,p)} \sum_{\zeta^m\in\{-1,1\}^k} L_{\mu_{m,\zeta^m\rho}}(\mathbf{Y},\mathbf{X}).$$

We aim at bounding the quantity $\mathbb{E}_0(L_{\mu_\rho}^2)$ and obtaining an inequality of the form (A.31). First, we work out $L_{\mu_{m,\zeta^m,\rho}}$:

$$\begin{aligned} L_{\mu_{m,\zeta^m,\rho}}(\mathbf{Y},\mathbf{X}) &= \left[\left(\frac{1}{1-\frac{\lambda^2 k}{\mathrm{var}(Y)}}\right)^{n/2} \exp\left(-\frac{\|\mathbf{Y}\|_n^2}{2}\frac{\lambda^2 k}{\mathrm{var}(Y)(\mathrm{var}(Y)-\lambda^2 k)}\right.\right. \\ &+ \left.\left. \lambda\sum_{j\in m}\zeta_j^m \frac{<\mathbf{Y},\mathbf{X}_j>_n}{\mathrm{var}(Y)-\lambda^2 k} - \lambda^2\sum_{j,j'\in m}\zeta_j^m\zeta_{j'}^m\frac{<\mathbf{X}_j,\mathbf{X}_{j'}>_n}{2(\mathrm{var}(Y)-\lambda^2 k)}\right)\right] \quad \text{(A.32)} \end{aligned}$$

where $<.>_n$ refers to the canonical inner product in $\mathbb{R}^n$.

Let us fix m_1 and m_2 in $\mathcal{M}(k,p)$ and two vectors ζ^1 and ζ^2 respectively associated to m_1 and m_2. We aim at computing the quantity $\mathbb{E}_0\left(L_{\mu_{m_1,\zeta^1,\rho}}(\mathbf{Y},\mathbf{X})L_{\mu_{m_2,\zeta^2,\rho}}(\mathbf{Y},\mathbf{X})\right)$. First, we decompose the set $m_1\cup m_2$ into four sets (which possibly are empty): $m_1\setminus m_2$, $m_2\setminus m_1$, m_3, and m_4, where m_3 and m_4 are defined by:

$$\begin{aligned} m_3 &:= \left\{j\in m_1\cap m_2 | \zeta_j^1 = \zeta_j^2\right\} \\ m_4 &:= \left\{j\in m_1\cap m_2 | \zeta_j^1 = -\zeta_j^2\right\}. \end{aligned}$$

For the sake of simplicity, we reorder the elements of $m_1\cup m_2$ from 1 to $|m_1\cup m_2|$ such that the first elements belong to $m_1\setminus m_2$, then to $m_2\setminus m_1$ and so on. Moreover, we define the vector $\zeta\in\mathbb{R}^{|m_1\cup m_2|}$ such that $\zeta_j = \zeta_j^1$ if $j\in m_1$ and $\zeta_j = \zeta_j^2$ if $j\in m_2\setminus m_1$. Using these notations, we compute the expectation of $L_{m_1,\zeta^1,\rho}(\mathbf{Y},\mathbf{X})L_{m_2,\zeta^2,\rho}(\mathbf{Y},\mathbf{X})$.

$$\mathbb{E}_0\left(L_{\mu_{m_1,\zeta^1,\rho}}(\mathbf{Y},\mathbf{X})L_{\mu_{m_2,\zeta^2,\rho}}(\mathbf{Y},\mathbf{X})\right) = \left(\frac{1}{\mathrm{var}(Y)(1-\frac{\lambda^2 k}{\mathrm{var}(Y)})^2}\right)^{n/2} |A|^{-n/2}\ , \quad \text{(A.33)}$$

where $|.|$ refers to the determinant and A is a symmetric square matrix of size $|m_1 \cup m_2|+1$ such that:

$$A[1,j] := \begin{cases} \frac{\text{var}(Y)+\lambda^2 k}{\text{var}(Y)(\text{var}(Y)-\lambda^2 k)} & \text{if} \quad j=1 \\ -\frac{\lambda\zeta_{j-1}}{\text{var}(Y)-\lambda^2 k} & \text{if} \quad (j-1)\in m_1 \triangle m_2 \\ -2\frac{\lambda\zeta_{j-1}}{\text{var}(Y)-\lambda^2 k} & \text{if} \quad (j-1)\in m_3 \\ 0 & \text{if} \quad (j-1)\in m_4\ , \end{cases}$$

where $m_1 \triangle m_2$ refers to $(m_1 \cup m_2) \setminus (m_1 \cap m_2)$. For any $i>1$ and $j>1$, A satisfies

$$A[i,j] := \begin{cases} \lambda^2\frac{\zeta_{i-1}\zeta_{j-1}}{\text{var}(Y)-\lambda^2 k}+\delta_{i,j} & \text{if} \quad (i-1,j-1)\in (m_1\setminus m_2)\times m_1 \\ \lambda^2\frac{\zeta_{i-1}\zeta_{j-1}}{\text{var}(Y)-\lambda^2 k}+\delta_{i,j} & \text{if} \quad (i-1,j-1)\in (m_2\setminus m_1)\times (m_2\setminus m_1\cup m_3) \\ -\lambda^2\frac{\zeta_{i-1}\zeta_{j-1}}{\text{var}(Y)-\lambda^2 k} & \text{if} \quad (i-1,j-1)\in (m_2\setminus m_1)\times m_4 \\ 2\lambda^2\frac{\zeta_{i-1}\zeta_{j-1}}{\text{var}(Y)-\lambda^2 k}+\delta_{i,j} & \text{if} \quad (i-1,j-1)\in [m_3\times m_3]\cup[m_4\times m_4] \\ 0 & \text{else,} \end{cases},$$

where $\delta_{i,j}$ is the indicator function of $i=j$.

After some linear transformation on the lines of the matrix A, it is possible to express its determinant into

$$|A| = \frac{\text{var}(Y)+\lambda^2 k}{\text{var}(Y)(\text{var}(Y)-\lambda^2 k)}\left|I_{|m_1\cup m_2|}+C\right|,$$

where $I_{|m_1\cup m_2|}$ is the identity matrix of size $|m_1 \cup m_2|$. C is a symmetric matrix of size $|m_1 \cup m_2|$ such that for any (i,j),

$$C[i,j] = \zeta_i\zeta_j D[i,j]$$

and D is a block symmetric matrix defined by

$$D := \begin{bmatrix} \frac{\lambda^4 k}{\text{var}^2(Y)-\lambda^4k^2} & \frac{-\lambda^2\text{var}(Y)}{\text{var}^2(Y)-\lambda^4k^2} & \frac{-\lambda^2}{\text{var}(Y)+\lambda^2k} & \frac{\lambda^2}{\text{var}(Y)-\lambda^2k} \\ \frac{-\lambda^2\text{var}(Y)}{\text{var}^2(Y)-\lambda^4k^2} & \frac{\lambda^4 k}{\text{var}^2(Y)-\lambda^4k^2} & \frac{-\lambda^2}{\text{var}(Y)+\lambda^2k} & \frac{-\lambda^2}{\text{var}(Y)-\lambda^2k} \\ \frac{-\lambda^2}{\text{var}(Y)+\lambda^2k} & \frac{-\lambda^2}{\text{var}(Y)+\lambda^2k} & \frac{-2\lambda^2}{\text{var}(Y)+\lambda^2k} & 0 \\ \frac{\lambda^2}{\text{var}(Y)-\lambda^2k} & \frac{-\lambda^2}{\text{var}(Y)-\lambda^2k} & 0 & \frac{2\lambda^2}{\text{var}(Y)-\lambda^2k} \end{bmatrix}.$$

Each block corresponds to one of the four previously defined subsets of $m_1 \cup m_2$ (i.e. $m_1 \setminus m_2$, $m_2 \setminus m_1$, m_3, and m_4). The matrix D is of rank at most four. By computing its non-zero eigenvalues, it is then straightforward to derive the determinant of A

$$|A| = \frac{\left[\text{var}(Y)-\lambda^2(2|m_3|-|m_1\cap m_2|)\right]^2}{\text{var}(Y)(\text{var}(Y)-\lambda^2k)^2}.$$

Gathering this equality with (A.33) yields

$$\mathbb{E}_0\left(L_{\mu_{m_1,\zeta^1,\rho}}(\mathbf{Y},\mathbf{X})L_{\mu_{m_2,\zeta^2,\rho}}(\mathbf{Y},\mathbf{X})\right) = \left[\frac{1}{1-\frac{\lambda^2(2|m_3|-|m_1\cap m_2|)}{\text{var}(Y)}}\right]^n. \tag{A.34}$$

Then, we take the expectation with respect to ζ^1, ζ^2, m_1 and m_2. When m_1 and m_2 are fixed the expression (A.34) depends on ζ^1 and ζ^2 only towards the cardinality of m_3. As ζ^1 and ζ^2 correspond to independent Rademacher variables, the random variable $2|m_3|-|m_1\cap m_2|$ follows the distribution of Z, a sum of $|m_1\cap m_2|$ independent rademacher variables and

$$\mathbb{E}_0(L_{\mu_{m_1},\rho}(\mathbf{Y},\mathbf{X})L_{\mu_{m_2},\rho}(\mathbf{Y},\mathbf{X})) \;=\; \mathbb{E}_0\left[\frac{1}{1-\frac{\lambda^2 Z}{\mathrm{var}(Y)}}\right]^n. \tag{A.35}$$

When Z is non-positive, this expression is smaller than one. On the other hand, when Z is non-negative:

$$\begin{aligned}\left[\frac{1}{1-\frac{\lambda^2 Z}{\mathrm{var}(Y)}}\right]^n &= \exp\left(n\log\left(\frac{1}{1-\frac{\lambda^2 Z}{\mathrm{var}(Y)}}\right)\right)\\ &\leq \exp\left[n\frac{\frac{\lambda^2 Z}{\mathrm{var}(Y)}}{1-\frac{\lambda^2 Z}{\mathrm{var}(Y)}}\right]\\ &\leq \exp\left[n\frac{\frac{\lambda^2 Z}{\mathrm{var}(Y)}}{1-\frac{\lambda^2 k}{\mathrm{var}(Y)}}\right],\end{aligned}$$

as $\log(1+x)\leq x$ and as Z is smaller than k. We define an event $\mathbb{A}$ such that $\{Z>0\}\subset\mathbb{A}\subset\{Z\geq 0\}$ and $\mathbb{P}(\mathbb{A})=\frac{1}{2}$. This is always possible as the random variable Z is symmetric. As a consequence, on the event $\mathbb{A}^c$, the quantity (A.35) is smaller or equal to one. All in all, we bound (A.35) by:

$$\mathbb{E}_0(L_{\mu_{m_1},\rho}(\mathbf{Y},\mathbf{X})L_{\mu_{m_2},\rho}(\mathbf{Y},\mathbf{X}))\leq\frac{1}{2}+\mathbb{E}_0\left[\mathbf{1}_{\mathbb{A}}\exp\left[n\frac{\frac{\lambda^2 Z}{\mathrm{var}(Y)}}{1-\frac{\lambda^2 k}{\mathrm{var}(Y)}}\right]\right], \tag{A.36}$$

where $\mathbf{1}_{\mathbb{A}}$ is the indicator function of the event $\mathbb{A}$. We now apply Hölder's inequality with a parameter $v\in]0;1]$, which will be fixed later.

$$\begin{aligned}\mathbb{E}_0\left[\mathbf{1}_{\mathbb{A}}\exp\left[n\frac{\frac{\lambda^2 Z}{\mathrm{var}(Y)}}{1-\frac{\lambda^2 k}{\mathrm{var}(Y)}}\right]\right] &\leq \mathbb{P}(\mathbb{A})^{1-v}\left[\mathbb{E}_0\exp\left(\frac{n}{v}\frac{\frac{\lambda^2 Z}{\mathrm{var}(Y)}}{1-\frac{\lambda^2 k}{\mathrm{var}(Y)}}\right)\right]^v\\ &\leq \left(\frac{1}{2}\right)^{1-v}\left[\cosh\left(\frac{n\lambda^2}{v(\mathrm{var}(Y)-\lambda^2 k)}\right)\right]^{|m_1\cap m_2|v}\end{aligned} \tag{A.37}$$

Gathering inequalities (A.36) and (A.37) yields

$$\mathbb{E}_0\left[L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})\right]\leq\frac{1}{2}+\left(\frac{1}{2}\right)^{1-v}\frac{1}{\binom{p}{k}^2}\sum_{m_1,m_2\in\mathcal{M}(k,p)}\cosh\left(\frac{n\lambda^2}{v(\mathrm{var}(Y)-\lambda^2 k)}\right)^{|m_1\cap m_2|v}.$$

Following the approach of [1] in Section 7.2, we note that if m_1 and m_2 are taken uniformly and indepently in $\mathcal{M}(k,p)$, then $|m_1\cap m_2|$ is distributed as a Hypergeometric

distribution with parameters p, k, and k/p. Thus, we derive that

$$\mathbb{E}_0\left[L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})\right] \leq \frac{1}{2} + \left(\frac{1}{2}\right)^{1-v} \mathbb{E}\left(\cosh\left(\frac{n\lambda^2}{v(\mathrm{var}(Y)-\lambda^2 k)}\right)^{vT}\right) \tag{A.38}$$

where T is a random variable distributed according to a Hypergeometric distribution with parameters p, k and k/p. We know from Aldous (1985, p.173) that T has the same distribution as the random variable $\mathbb{E}(W|\mathcal{B}_p)$ where W is binomial random variable of parameters k, k/p and $\mathcal{B}_p$ some suitable σ-algebra. By a convexity argument, we then upper bound (A.38).

$$\begin{aligned}
\mathbb{E}_0\left[L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})\right] &\leq \frac{1}{2} + \left(\frac{1}{2}\right)^{1-v} \mathbb{E}\left(\cosh\left(\frac{n\lambda^2}{v(\mathrm{var}(Y)-\lambda^2 k)}\right)^{vW}\right) \\
&= \frac{1}{2} + \left(\frac{1}{2}\right)^{1-v} \left(1 + \frac{k}{p}\left(\cosh\left(\frac{n\lambda^2}{v(\mathrm{var}(Y)-\lambda^2 k)}\right)^{v} - 1\right)\right)^{k} \\
&= \frac{1}{2} + \left(\frac{1}{2}\right)^{1-v} \exp\left[k\log\left(1 + \frac{k}{p}\left(\cosh\left(\frac{n\lambda^2}{v(\mathrm{var}(Y)-\lambda^2 k)}\right)^{v} - 1\right)\right)\right].
\end{aligned}$$

To get the upper bound on the total variation distance appearing in (A.30), we aim at constraining this last expression to be smaller than $1+\eta^2$. This is equivalent to the following inequality:

$$2^v \exp\left[k\log\left(1 + \frac{k}{p}\left(\cosh\left(\frac{n\lambda^2 k}{vk(\mathrm{var}(Y)-\lambda^2 k)}\right)^{v} - 1\right)\right)\right] \leq 1 + 2\eta^2 \ . \tag{A.39}$$

We now choose $v = \frac{\mathcal{L}(\eta)}{\log(2)} \wedge 1$. If v is strictly smaller than one, then (A.39) is equivalent to:

$$k\log\left[1 + \frac{k}{p}\left(\cosh\left(\frac{n\lambda^2 k}{vk(\mathrm{var}(Y)-\lambda^2 k)}\right)^{v} - 1\right)\right] \leq \frac{\log(1+2\eta^2)}{2}. \tag{A.40}$$

It is straightforward to show that this last inequality also implies (A.39) if v equals one. We now suppose that

$$\frac{n\lambda^2}{v(\mathrm{var}(Y)-\lambda^2 k)} \leq \log\left((1+u)^{\frac{1}{v}} + \sqrt{(1+u)^{\frac{2}{v}} - 1}\right), \tag{A.41}$$

where $u = \frac{p\mathcal{L}(\eta)}{k^2}$. Using the classical equality $\cosh\left[\log(1+x+\sqrt{2x+x^2})\right] = 1+x$ with $x = (1+u)^{\frac{1}{v}} - 1$, we deduce that inequality (A.41) implies (A.40) because

$$\begin{aligned}
k\log\left(1 + \frac{k}{p}\left(\cosh\left(\frac{n\lambda^2 k}{vk(\mathrm{var}(Y)-\lambda^2 k)}\right)^{v} - 1\right)\right) &\leq k\log\left(1 + \frac{k}{p}u\right) \\
&\leq \frac{k^2}{p}u \leq \mathcal{L}(\eta) \ .
\end{aligned}$$

For any $\beta \geq 1$ and any $x > 0$, it holds that $(1+x)^\beta \geq 1 + \beta x$. As $\frac{1}{v} \geq 1$, condition (A.41) is implied by:

$$\frac{\lambda^2 k}{\mathrm{var}(Y) - \lambda^2 k} \leq \frac{kv}{n} \log\left(1 + \frac{u}{v} + \sqrt{\frac{2u}{v}}\right).$$

One then combines the previous inequality with the definitions of u and v to obtain the upper bound

$$\frac{\lambda^2 k}{\mathrm{var}(Y) - \lambda^2 k} \leq \frac{k}{n}\left(\frac{\mathcal{L}(\eta)}{\log(2)} \wedge 1\right) \log\left(1 + \frac{p(\log(2) \vee \mathcal{L}(\eta))}{k^2} + \sqrt{\frac{2p(\log(2) \vee \mathcal{L}(\eta))}{k^2}}\right) .$$

For any x positive and any u between 0 and 1, $\log(1+ux) \geq u\log(1+x)$. As a consequence, the previous inequality is implied by:

$$\begin{aligned}\frac{\lambda^2 k}{\mathrm{var}(Y) - \lambda^2 k} &\leq \frac{k}{n}\left(\frac{\mathcal{L}(\eta)}{\log(2)} \wedge 1\right) ([\mathcal{L}(\eta) \vee \log(2)] \wedge 1) \log\left(1 + \frac{p}{k^2} + \sqrt{\frac{2p}{k^2}}\right) \\ &= \frac{k}{n}(\mathcal{L}(\eta) \wedge 1) \log\left(1 + \frac{p}{k^2} + \sqrt{\frac{2p}{k^2}}\right) .\end{aligned}$$

To resume, if we take ρ^2 smaller than this last quantity, then

$$\beta_I(\Theta[k, p, \rho]) \geq \delta .$$

To prove the second part of the theorem, one has to observe that $\alpha + \delta \leq 53\%$ implies that $\mathcal{L}(\eta) \geq \frac{1}{2}$.

A.2.2 Proof of Proposition 6

Let us first assume that the covariance matrix of X is the identity. We argue as in the proof of Theorem 7 taking $k = p$. The sketch of the proof remains unchanged except that we slightly modify the last part. Inequality (A.40) becomes

$$pv \log\left(\cosh\left(\frac{n\lambda^2 p}{vp(\mathrm{var}(Y) - \lambda^2 p)}\right)\right) \leq \mathcal{L}(\eta),$$

where we recall that $v = \frac{\mathcal{L}(\eta)}{\log 2} \wedge 1$. For all $x \in \mathbb{R}$, $\cosh(x) \leq \exp(x^2/2)$. Consequently, the previous inequality is implied by

$$\frac{\lambda^2 p}{\mathrm{var}(Y) - \lambda^2 p} \leq \sqrt{2v\mathcal{L}(\eta)}\frac{\sqrt{p}}{n},$$

and the result follows easily.

If we no longer assume that the covariance matrix Σ is the identity, we orthogonalize the sequence X_i thanks to Gram-Schmidt process. Applying the previous argument to this new sequence of covariates allows to conclude.

A.2.3 Proof of Proposition 4

Let us apply proposition 6. For any $\rho \leq s(\alpha,\delta)\frac{\sqrt{D_m}}{n}$ there exists some $\theta \in S_m$ such that $\frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} = \rho^2$ and $\mathbb{P}_\theta(T_m \leq 0) \geq \delta$. Here, $s(\alpha,\delta)$ refers to some function only depending on α and δ. In the proof of Theorem 3, we have shown in (A.4) and following equalities that the distribution of the test statistic ϕ_m only depends on the quantity $\kappa_m^2 = \frac{\mathrm{var}(Y)-\mathrm{var}(Y|X_m)}{\mathrm{var}(Y|X_m)}$. Let θ' be an element of S_m such that $\kappa_m^2 = \rho^2$. The distribution of ϕ_m under $\mathbb{P}_{\theta'}$ is the same as its distribution under $\mathbb{P}_\theta$, and therefore

$$\mathbb{P}_{\theta'}\left(T_m \leq 0\right) \geq \delta.$$

A.2.4 Proof of Proposition 8

This lower bound for dependent gaussian covariates is proved through the same approach as Theorem 7. We define the measure μ_ρ as in that proof. Under the hypothesis H_0, Y is independent of X. We note Σ the covariance matrix of X and $\mathbb{E}_{0,\Sigma}$ stands for the distribution of $(\mathbf{Y},\mathbf{X})$ under H_0 in order to emphasize the dependence on Σ.

First, one has to upper bound the quantity $\mathbb{E}_{0,\Sigma}\left[L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})\right]$. For the sake of simplicity, we make the hypothesis that every covariate X_j has variance 1. If this is not the case, we only have to rescale these variables. The quantity $\mathrm{corr}(i,j)$ refers to the correlation between X_i and X_j. As we only consider the case $k=1$, the set of models m in $\mathcal{M}(1,p)$ is in correspondance with the set $\{1,\ldots,p\}$.

$$\mathbb{E}_{0,\Sigma}\left(L_{\mu_{i,\zeta^1,\rho}}(\mathbf{Y},\mathbf{X})L_{\mu_{j,\zeta^2,\rho}}(\mathbf{Y},\mathbf{X})\right) = \left(\frac{\mathrm{var}(Y)}{\mathrm{var}(Y)-\mathrm{corr}(i,j)\lambda^2\zeta^1\zeta^2}\right)^n.$$

When i and j are fixed, we upper bound the expectation of this quantity with respect to ζ^1 and ζ^2 by

$$\mathbb{E}_{0,\Sigma}\left(L_{\mu_{i,\rho}}(\mathbf{Y},\mathbf{X})L_{\mu_{j,\rho}}(\mathbf{Y},\mathbf{X})\right) \leq \frac{1}{2} + \frac{1}{2}\left(\frac{\mathrm{var}(Y)}{\mathrm{var}(Y)-|\mathrm{corr}(i,j)|\lambda^2}\right)^n. \tag{A.42}$$

If $i \neq j$, $|\mathrm{corr}(i,j)|$ is smaller than c and if $i=j$, $\mathrm{corr}(i,j)$ is exactly one. As a consequence, taking the expectation of (A.42) with respect to i and j yields the upper bound

$$\mathbb{E}_{0,\Sigma}\left(L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})\right) \leq \frac{1}{2} + \frac{1}{2}\left(\frac{1}{p}\left(\frac{\mathrm{var}(Y)}{\mathrm{var}(Y)-\lambda^2}\right)^n + \frac{p-1}{p}\left(\frac{\mathrm{var}(Y)}{\mathrm{var}(Y)-c\lambda^2}\right)^n\right). \tag{A.43}$$

Recall that we want to constrain this quantity (A.43) to be smaller than $1+\eta^2$. In particular, this holds if the two following inequalities hold:

$$\frac{1}{p}\left(\frac{\mathrm{var}(Y)}{\mathrm{var}(Y)-\lambda^2}\right)^n \leq \frac{1}{p} + \eta^2 \tag{A.44}$$

$$\frac{p-1}{p}\left(\frac{\mathrm{var}(Y)}{\mathrm{var}(Y)-c\lambda^2}\right)^n \leq \frac{p-1}{p} + \eta^2. \tag{A.45}$$

One then uses the inequality $\log(\frac{1}{1-x}) \leq \frac{x}{1-x}$ which holds for any positive x smaller than one. Condition (A.44) holds if

$$\frac{\lambda^2}{\mathrm{var}(Y)-\lambda^2} \leq \frac{1}{n}\log(1+p\eta^2), \tag{A.46}$$

whereas condition (A.45) is implied by

$$\frac{c\lambda^2}{\mathrm{var}(Y) - c\lambda^2} \leq \frac{1}{n} \log\left(1 + \frac{p}{p-1}\eta^2\right) .$$

As c is smaller than one and $\frac{p}{p-1}$ is larger than 1, this last inequality holds if

$$\frac{\lambda^2}{\mathrm{var}(Y) - \lambda^2} \leq \frac{1}{nc} \log(1+\eta^2) . \tag{A.47}$$

Gathering conditions (A.46) and (A.47) allows to conclude and to obtain the desired lower bound (4.19).

A.2.5 Proof of Proposition 10

The sketch of the proof and the notations are analogous to the one in Proposition 8. The upper bound (A.42) still holds:

$$\mathbb{E}_{0,\Sigma}\left(L_{\mu_{i,\rho}}(\mathbf{Y},\mathbf{X}) L_{\mu_{j,\rho}}(\mathbf{Y},\mathbf{X})\right) \leq \frac{1}{2} + \frac{1}{2}\left(\frac{\mathrm{var}(Y)}{\mathrm{var}(Y) - |\mathrm{corr}(i,j)|\lambda^2}\right)^n .$$

Using the stationarity of the covariance function, we derive from (A.42) the following upper bound:

$$\mathbb{E}_{0,\Sigma}\left(L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})\right) \leq \frac{1}{2} + \frac{1}{2p}\sum_{i=0}^{p-1}\left(\frac{\mathrm{var}(Y)}{\mathrm{var}(Y) - \lambda^2|\mathrm{corr}(0,i)|}\right)^n ,$$

where $\mathrm{corr}(0,i)$ equals $\mathrm{corr}(X_1, X_{i+1})$. As previously, we want to constrain this quantity to be smaller than $1+\eta^2$. In particular, this is implied if for any i between 0 and $p-1$:

$$\left(\frac{\mathrm{var}(Y)}{\mathrm{var}(Y) - \lambda^2|\mathrm{corr}(i,0)|}\right)^n \leq 1 + \frac{2p\eta^2|\mathrm{corr}(i,0)|}{\sum_{i=0}^{p-1}|\mathrm{corr}(i,0)|}.$$

Using the inequality $\log(1+u) \leq u$, it is straighforward to show that this previous inequality holds if

$$\frac{\lambda^2}{\mathrm{var}(Y) - \lambda^2|\mathrm{corr}(i,0)|} \leq \frac{1}{n|\mathrm{corr}(i,0)|}\log\left(1 + \frac{2p\eta^2|\mathrm{corr}(i,0)|}{\sum_{i=0}^{p-1}|\mathrm{corr}(i,0)|}\right).$$

As $|\mathrm{corr}(i,0)|$ is smaller than one for any i between 0 and $p-1$, it follows that $\mathbb{E}_{0,\Sigma}\left(L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})\right)$ is smaller than $1+\eta^2$ if

$$\rho^2 \leq \bigwedge_{i=0}^{p-1} \frac{1}{n|\mathrm{corr}(i,0)|}\log\left(1 + \frac{2p\eta^2|\mathrm{corr}(0,i)|}{\sum_{i=0}^{p-1}|\mathrm{corr}(i,0)|}\right).$$

We now apply the convexity inequality $\log(1+ux) \geq u\log(1+x)$ which holds for any positive x and any u between 0 and 1 to obtain the condition

$$\rho^2 \leq \frac{1}{n}\log\left(1 + \frac{2p\eta^2}{\sum_{i=0}^{p-1}|\mathrm{corr}(i,0)|}\right). \tag{A.48}$$

It turns out we only have to upper bound the sum of $|\mathrm{corr}(i,0)|$ for the different types of correlation:

1. For $\mathrm{corr}(i,j) = \exp(-w|i-j|_p)$, the sum is clearly bounded by $1 + 2\frac{e^{-w}}{1-e^{-w}}$ and condition (A.48) simplifies as
$$\rho^2 \leq \frac{1}{n} \log\left(1 + 2p\eta^2 \frac{1-e^{-w}}{1+e^{-w}}\right).$$

2. if $\mathrm{corr}(i,j) = (1+|i-j|_p)^{-t}$ for t strictly larger than one, then $\sum_{i=0}^{p-1} |\mathrm{corr}(i,0)| \leq 1 + \frac{2}{t-1}$ and condition (A.48) simplifies as
$$\rho^2 \leq \frac{1}{n} \log\left(1 + \frac{2p(t-1)\eta^2}{t+1}\right).$$

3. if $\mathrm{corr}(i,j) = (1+|i-j|_p)^{-1}$ then $\sum_{i=0}^{p-1} |\mathrm{corr}(i,0)| \leq 1 + 2\log(p-1)$ and condition (A.48) simplifies as
$$\rho^2 \leq \frac{1}{n} \log\left(1 + \frac{2p\eta^2}{1+2\log(p-1)}\right).$$

4. if $\mathrm{corr}(i,j) = (1+|i-j|_p)^{-t}$ for $0 < t < 1$, then
$$\sum_{i=0}^{p-1} |\mathrm{corr}(i,0)| \leq 1 + \frac{2}{1-t}\left[\left(\frac{p}{2}\right)^{1-t} - 1\right] \leq \frac{2}{1-t}\left(\frac{p}{2}\right)^{1-t}$$
and condition (A.48) simplifies as
$$\rho^2 \leq \frac{1}{n} \log\left(1 + p^t 2^{1-t}(1-t)\eta^2\right).$$

A.2.6 Proof of Proposition 13

For each dimension D between 1 and p, we define $r_D^2 = \rho_{D,n}^2 \wedge a_D^2 R^2$. Let us fix some $D \in \{1, \ldots, p\}$. Since $r_D^2 \leq a_D^2$ and since the a_j's are non increasing,
$$\sum_{j=1}^{D} \frac{\mathrm{var}(Y|X_{m_{j-1}}) - \mathrm{var}(Y|X_{m_j})}{a_j^2} \leq \mathrm{var}(Y|X)R^2,$$
for all $\theta \in S_{m_D}$ such that $\frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} = r_D^2$. Indeed, $\|\theta\|^2 = \sum_{j=1}^{D} \mathrm{var}(Y|X_{m_{j-1}}) - \mathrm{var}(Y|X_{m_j})$ and $\mathrm{var}(Y) - \|\theta\|^2 = \mathrm{var}(Y|X)$. As a consequence,
$$\left\{\theta \in S_{m_D}, \frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} = r_D^2\right\} \subset \left\{\theta \in \mathcal{E}_a(R), \frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} \geq r_D^2\right\}.$$
Since $r_D \leq \rho_{D,n}$, we deduce from Proposition 6 that
$$\beta_\Sigma\left(\left\{\theta \in \mathcal{E}_a(R), \frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} \geq r_D^2\right\}\right) \geq \delta.$$
The first result of Proposition 13 follows by gathering these lower bounds for all D between 1 and p.

Moreover, $\rho_{i,n}^2$ is defined in Proposition 6 as $\rho_{i,n}^2 = \sqrt{2}\left[\sqrt{\mathcal{L}(\eta)} \wedge \frac{\mathcal{L}(\eta)}{\sqrt{\log 2}}\right]\frac{\sqrt{i}}{n}$. If $\alpha + \delta \leq 47\%$, it is straighforward to show that $\rho_{i,n}^2 \geq \frac{\sqrt{i}}{n}$.

A.2.7 Proof of Proposition 15

We first need the following Lemma.

Lemma 20. *We consider $(I_j)_{j\in\mathcal{J}}$ a partition of $\mathcal{I}$. For each $j \in \mathcal{J}$ let $p(j) = |I_j|$. For any $j \in \mathcal{J}$, we define Θ_j as the set of $\theta \in \mathbb{R}^{\mathcal{I}}$ such that their support is included in I_j. For any sequence of positive weights k_j such that*

$$\sum_{j\in\mathcal{J}} k_j = 1,$$

it holds that

$$\beta_I\left(\bigcup_{j\in\mathcal{J}}\left\{\theta\in\Theta_j, \frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} = r_j^2\right\}\right) \geq \delta \ ,$$

if for all $j \in \mathcal{J}$, $r_j \leq \rho_{p(j),n}(\eta/\sqrt{k_j})$, where the function $\rho_{p(j),n}$ is defined by (4.16).

For all $j \geq 0$ such that $2^{j+1}-1 \in \mathcal{I}$ (i.e. for all $j \leq J$ where $J = \log(p+1)/\log(2)-1$), let $\bar{S}_j$ be the linear span of the e_k's for $k \in \{2^j, \ldots, 2^{j+1}-1\}$. Then, $\dim(\bar{S}_j) = 2^j$ and $\bar{S}_j \subset S_{m_D}$ for $D = D(j) = 2^{j+1}-1$. It is straighforward to show that

$$\bigcup_{j=0}^{J} \bar{S}_j[r_{D(j)}] \subset \bigcup_{j=0}^{J} S_{m_{D(j)}}[r_{D(j)}] \subset \bigcup_{D=1}^{p} S_{m_D}[r_D] \ ,$$

where $\bar{S}_j[r_D(j)] = \left\{\theta \in \bar{S}_j, \frac{\|\theta\|^2}{var(Y)-\|\theta\|^2} = r_{D(j)}^2\right\}$ and $S_{m_D}[r_D] = \left\{\theta \in S_{m_D}, \frac{\|\theta\|^2}{var(Y)-\|\theta\|^2} = r_D^2\right\}$.

We now apply Lemma 20 with $k_j := \left[(j+1)^2 R(p)\right]^{-1}$ where $R(p) = \sum_{k=0}^{J} 1/(k+1)^2$ to show that

$$\beta_I\left(\bigcup_{D=1}^{p}\left\{\theta\in S_{m_D}, \frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} = r_D^2\right\}\right) \geq \delta \ ,$$

if for all those $D = D(j)$

$$r_D^2 \leq \sqrt{\log(1+2\eta^2/k_j)}\left(1 \wedge \frac{\sqrt{\log(1+2\eta^2/k_j)}}{\sqrt{2\log 2}}\right)\frac{\sqrt{D}}{n}.$$

For $D = D(j)$, this last quantity equals

$$\begin{aligned}&\sqrt{\log(1+2\eta^2/k_j)}\left(1 \wedge \frac{\sqrt{\log(1+2\eta^2/k_j)}}{\sqrt{2\log 2}}\right)\frac{\sqrt{D}}{n} \\ &\quad\geq \sqrt{\log(1+2\eta^2(j+1)^2R(p))}\left(1 \wedge \frac{\sqrt{\log(1+2\eta^2)}}{\sqrt{2\log 2}}\right)\frac{2^{j/2}}{n} \ .\end{aligned} \tag{A.49}$$

It remains to check that (A.49) is larger than $\bar{\rho}_{D(j),n}$. Using $j+1 = \log(D+1)/\log(2) \geq \log(D+1)$, we get $2^{j/2} \geq \sqrt{D/2}$. Thanks to the convexity inequality $\log(1+ux) \geq$

$u \log(1+x)$, which holds for any $x > 0$ and any $u \in]0,1]$, we obtain

$$\begin{aligned}\sqrt{\log(1+2\eta^2(j+1)^2R(p))}2^{j/2} &\geq \sqrt{D/2}\left(\eta\sqrt{2R(p)} \wedge 1\right)\sqrt{\log\left[1+\log^2(D+1)\right]} \\ &\geq \left((\eta\sqrt{2}) \wedge 1\right)\sqrt{\log\log^2(D+1)}\sqrt{D/2}, \\ &\geq \left(1 \wedge \sqrt{\log(1+2\eta^2)}\right)\sqrt{\log\log(D+1)}\sqrt{D}\ ,\end{aligned}$$

as $R(p)$ is larger than one for any $p \geq 1$. All in all, we get the lower bound

$$\begin{aligned}&\sqrt{\log(1+2\eta^2(j+1)^2R(p))}\left(1 \wedge \frac{\sqrt{\log(1+2\eta^2)}}{\sqrt{2\log 2}}\right)2^{j/2} \\ &\geq \frac{1}{2\sqrt{\log(2)}}\left(1 \wedge \log(1+2\eta^2)\right)\sqrt{\log\log(D+1)}\sqrt{D} \\ &= \bar{\rho}^2_{D,n}\ .\end{aligned}$$

Thus, if for all $1 \leq D \leq p$, r_D^2 is smaller than $\bar{\rho}^2_{D,n}$, it holds that

$$\beta_I\left(\bigcup_{D=1}^{p}\left\{\theta \in S_{m_D}, \frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} = r_D^2\right\}\right) \geq \delta\ .$$

A.2.8 Proof of Lemma 20

Using a similar approach to the proof of Theorem 7, we know that for each $r_j \leq \tilde{\rho}_j(\eta/\sqrt{k_j})$ there exists some measure μ_j over

$$\Theta_j[r_j] := \left\{\theta \in \Theta_j, \frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} = r_j^2\right\}$$

such that

$$\mathbb{E}_0\left[L^2_{\mu_j}(Y,X)\right] \leq 1+\eta^2/k_j\ . \tag{A.50}$$

We now define a probability measure $\mu = \sum_{j\in\mathcal{J}} k_j\mu_j$ over $\bigcup_{j\in\mathcal{J}}\Theta_j[r_j]$. L_{μ_j} refers to the density of $\mathbb{P}_{\mu_j}$ with respect to $\mathbb{P}_0$. Thus,

$$L_\mu(Y) = \frac{\mathrm{d}\mathbb{P}_\mu}{\mathrm{d}\mathbb{P}_0}(\mathbf{Y},\mathbf{X}) = \sum_{j\in\mathcal{J}} k_j L_{\mu_j}(\mathbf{Y},\mathbf{X})\ ,$$

and

$$\mathbb{E}_0\left[L^2_\mu(\mathbf{Y},\mathbf{X})\right] = \sum_{j,j'\in\mathcal{J}} k_j k_{j'}\mathbb{E}_0\left[L_{\mu_j}(\mathbf{Y},\mathbf{X})L_{\mu_{j'}}(\mathbf{Y},\mathbf{X})\right]\ .$$

Using expression (A.34), it is straightforward to show that if $j \neq j'$, then

$$\mathbb{E}_0\left[L_{\mu_j}(\mathbf{Y},\mathbf{X})L_{\mu_{j'}}(\mathbf{Y},\mathbf{X})\right] = 1.$$

This follows from the fact that the sets Θ_j and $\Theta_{j'}$ are orthogonal with respect to the inner product (4.4). Thus,

$$\mathbb{E}_0\left[L_\mu(\mathbf{Y},\mathbf{X})\right] = 1 + \sum_{j\in\mathcal{J}} k_j^2 \left(\mathbb{E}_0\left[L_{\mu_j}^2(\mathbf{Y},\mathbf{X})\right] - 1\right) \leq 1 + \eta^2$$

thanks to (A.50). Using the argument (A.31) as in the proof of Theorem 7 allows to conclude.

A.2.9 Proof of Proposition 16

First of all, we only have to consider the case where the covariance matrix of X is the identity. If this is not the case, one only has to apply Gram-Schmidt process to X and thus obtain a vector X' and a new basis for Θ which is orthonormal. We refer to the beginning of Section 4.5 for more details.

Like the previous bounds for ellipsoids, we adapt the approach of Section 6 in [1]. We use the same notations as in proof of Proposition 13. Let $D^*(R) \in \{1,\dots,p\}$ an integer which achieves the supremum of $\bar{\rho}_D^2 \wedge (R^2 a_D^2) = \bar{r}_D^2$. As in proof of Proposition 13, for any $R > 0$,

$$\left\{\theta \in S_{m_{D^*(R)}}, \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} = r_{D^*(R)}^2\right\} \subset \left\{\theta \in \mathcal{E}_a(R), \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq r_{D^*(R)}^2\right\} .$$

When R varies, $D^*(R)$ describes $\{1,\dots,p\}$. Thus, we obtain

$$\begin{aligned}\bigcup_{1\leq D\leq p} \left\{\theta \in S_{m_D}, \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} = r_D^2\right\} &= \bigcup_{R>0} \left\{\theta \in S_{m_{D^*(R)}}, \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} = r_{D^*(R)}^2\right\} \\ &\subset \bigcup_{R>0} \left\{\theta \in \mathcal{E}_a(R) \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq r_{D^*(R)}^2\right\} ,\end{aligned}$$

and the result follows from proposition 15.

References

[1] Baraud, Y. Non-asymptotic rates of testing in signal detection. *Bernoulli 8*, 5 (2002), 577–606.

[2] Baraud, Y., Huet, S., and Laurent, B. Adaptative tests of linear hypotheses by model selection. *Annals of Statistics 31*, 1 (2003), 225–251.

[3] Birgé, L., and Massart, P. Minimum contrast estimators on sieves: exponentntial bounds and rates of convergence. Bernoulli *4*, 3 (1998), 329–375.

[4] Massart, P. Concentration Inequalities and Model Selection*, École d'été de probabilités de Saint Flour XXXIII*, vol. 1896 of *Lecture Notes in Mathematics*. Springer-Verlag, 2007.