ANR ClasSel
Livrable 2.3
Sélection de modèle : Aspects non asymptotiques

Aurélie Boisbunon, Stéphane Canu, Dominique Fourdrinier, Ali Righi

13 septembre 2010

Résumé

Ce livrable présente les aspects non asymptotiques de la sélection de modèles du point de vue décisionnel. Il regroupe deux documents. Le premier est une application directe de l'état de l'art sur la sélection de modèles développé dans le livrable 2.1. Il considère l'estimateur du Lasso, très utilisé dans de nombreux domaines tels la biologie, à la place de l'estimateur des moindres carrés.

Le deuxième document considère quant à lui une approche bayésienne. Il a été soumis à Journal of Multivariate Analysis.

Chapitre 1

Sélection de variables dans le modèle linéaire

Sélection de variables dans le modèle linéaire

Aurélie Boisbunon, Dominique Fourdrinier, Stéphane Canu
Université de Rouen et INSA de Rouen, LITIS EA 4108
Avenue de l'Université, BP 12
76801 Saint-Etienne du Rouvray, France

13 septembre 2010

Résumé

Nous présentons dans ce rapport une nouvelle procédure de sélection de variables, basée sur l'estimation de coût. Cette procédure possède l'avantage d'être applicable dans un cadre distributionnel plus général que le gaussien : le cadre sphérique. Nous présentons également une application à l'estimateur du *lasso*, avec des résultats de simulations pour différents exemples de distributions. Nous comparons nos résultats à l'AIC et au BIC.

AMS 2010 subject classifications : 62C20, 62F07, 62F10, 62H12, 62J05, 62J07.

Mots-clés : Sélection de variables, modèle linéaire, estimation de coût, lois à symétrie sphérique.

Table des matières

1 Introduction **3**

2 Contexte et notations **6**

3 Estimation de coût **10**

4 Simulations **16**

5 Conclusion **22**

A Lois à symétrie sphérique **23**

1 Introduction

Dans le contexte de la régression, le modèle linéaire, bien que le plus simple des modèles, continue d'intéresser nombre de chercheurs de tous domaines. En effet, pour certains types de problèmes, comme les problèmes à grandes dimensions notamment, on préfère souvent la simplicité à l'ajustement le plus proche de la réalité.

Nous considérons donc ici le modèle linéaire suivant :

$$Y = X\beta + \varepsilon \tag{1.1}$$

où Y est un vecteur aléatoire dans $\mathbb{R}^n$, X est une matrice fixée et connue dans l'espace $\mathcal{M}_{n,p}$ des matrices de dimension $n \times p$, β est le vecteur inconnu des coefficients de la régression, et ε est le vecteur des erreurs dans $\mathbb{R}^n$. On supposera par ailleurs que $p < n$.

L'estimateur classique utilisé pour la régression linéaire est l'estimateur des moindres carrés :

$$\hat{\beta}^{MC} = \arg\min_{\beta \in \mathbb{R}^p} \| Y - X\beta \|^2 = (X^T X)^{-1} X^T Y \tag{1.2}$$

$\hat{\beta}^{MC}$ est un estimateur sans biais de β. Il est intéressant dans le sens où c'est l'estimateur de plus faible variance parmi la classe des estimateurs sans biais de β, et il est minimax. Cependant, il possède l'inconvénient d'être instable et difficilement interprétable, particulièrement dans le cas où p est grand. Par ailleurs, Stein [11] a montré que, pour $p \geq 3$, cet estimateur est inadmissible, autrement dit il existe des estimateurs biaisés de plus faible risque. La recherche s'est donc naturellement tournée vers une amélioration de l'estimation de β par rapport aux moindres carrés.

Dans le cadre des problèmes à grandes dimensions, on fait une hypothèse supplémentaire permettant de simplifier le problème et d'en améliorer l'interprétabilité : parmi les variables X^i, $i = 1, \ldots, p$, dont on dispose, seule une partie d'entre elles a une influence importante sur la variable étudiée. Se pose alors la question du nombre et de la sélection des variables explicatives. Si on appelle I l'ensemble des variables d'influence, $I \subset \{1, \ldots, p\}$, le modèle (1.1) se réduit au modèle suivant :

$$Y = X_I \beta_I + \varepsilon \tag{1.3}$$

où X_I correspond à la matrice X réduite au sous-ensemble I. Le problème est alors la recherche du meilleur sous-ensemble I.

Plusieurs méthodes ont été proposées jusque-là dans ce contexte de sélection de variables. On peut citer entre autres la *subset selection*, ou *hard thresholding*, qui utilise l'estimation des moindres carrés sur un sous-ensemble des variables (cf. Hastie et al [9]) ; le *lasso* (Least Absolute Shrinkage and Selection Operator), ou *soft thresholding*, introduit par Donoho et Johnstone en 1994 [2] et Tibshirani en 1996 [14], qui translate et tronque la solution des moindres carrés et qui propose un chemin de régularisation ; la *bridge regression* [7], qui généralise les deux premières ; ou encore le *SCAD* (Smoothly Clipped Absolute Deviation) [3], qui combine le *hard* et le *soft thresholding*.

Toutes ces solutions sont basées sur la minimisation d'un critère d'attache aux données, ici la somme des erreurs quadratique, pénalisé par des contraintes prenant en compte le nombre effectif de paramètres :

$$\hat{\beta}^{subset} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \| Y - X\beta \|^2 + \lambda \| \beta \|_0 \right\} \tag{1.4}$$

$$\hat{\beta}^{lasso} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \| Y - X\beta \|^2 + \lambda \| \beta \|_1 \right\} \tag{1.5}$$

$$\hat{\beta}^{bridge} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \| Y - X\beta \|^2 + \lambda \| \beta \|_q \right\}, \quad q \geq 0, \tag{1.6}$$

où λ est une constante positive.

Ces méthodes peuvent se résumer sous la forme suivante :

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \{L(Y, X\beta) + \lambda J(\beta)\}, \quad \lambda \geq 0 \tag{1.7}$$

où $L(Y, X\beta)$ est le terme d'attache aux données et $J(\beta)$ représente le terme de pénalité.

Comme le montrent les équations (1.4), (1.5), (1.6), et (1.7), la pénalisation est réglée par un hyperparamètre, appelé ici λ, dont la valeur est plus ou moins forte selon le niveau de sélection recherché. Tout le problème se situe donc dans l'estimation de cet hyperparamètre pour obtenir un nombre optimal de variables.

La validation croisée [13] permet d'obtenir sur une partie des données une estimation du paramètre que l'on cherche à évaluer, et d'en calculer le coût quadratique empirique sur les données restantes, en réitérant cette opération plusieurs fois sur des partitions différentes. La moyenne des coûts empiriques obtenus amène à choisir l'hyperparamètre optimal. Cette méthode a fait ses preuves dans de nombreuses applications, mais sa complexité de calculs la rend difficilement utilisable dans les problèmes à grandes dimensions. La validation croisée généralisée ([8]), quant à elle, utilise une estimation du risque et de ce fait réduit la complexité par rapport à la précédente, mais nécessite encore beaucoup de calculs.

D'autres critères, comme l'Akaike Information Criterion (AIC) [1] ou le critère d'information bayésien (BIC) [10], ont été initialement conçus pour l'estimateur des moindres carrés avec bruit gaussien, puis développés dans le cadre bayésien de l'estimation de la log-vraisemblance. Ces critères ont fait leurs preuves lorsque la distribution des erreurs est définie, mais il est plus difficile de les estimer dès qu'on relâche cette hypothèse. Par ailleurs, l'AIC a tendance à sélectionner des modèles trop complexes et le BIC des modèles trop simples.

Nous proposons d'utiliser une nouvelle procédure, basée sur l'estimation du coût pour un estimateur $\hat{\beta}$ spécifié. Le principe est le suivant : on évalue habituellement la qualité de $\hat{\beta}$ par le biais d'une fonction de coût $L(\beta, \hat{\beta})$. Cette fonction dépend de l'exemple précis sur lequel elle est calculée et peut donc varier fortement d'un exemple à l'autre. Ceci a amené certains chercheurs à considérer plutôt son risque $R(\hat{\beta}) = \mathbb{E}[L(\beta, \hat{\beta})]$. Cependant, le risque, lui, est totalement indépendant des observations, alors que celles-ci peuvent apporter une information. On souhaiterait donc un compromis entre le risque et le coût. De plus, aussi bien le risque comme le coût dépendent du paramètre inconnu β et ne sont donc pas accessibles directement.

Nous proposons donc d'estimer le coût $L(\beta, \hat{\beta})$ à partir des observations et indépendemment de β, pour un estimateur $\hat{\beta}$ fixé. En faisant varier l'hyperparamètre λ, on a une nouvelle valeur de $\hat{\beta}$, et donc une nouvelle estimation de son coût. Nous pouvons ainsi sélectionner la valeur de λ minimisant l'estimation du coût et ainsi obtenir le sous-ensemble I correspondant.

Notre procédure est semblable à l'*estimation sans biais du risque*, introduite par Stein en 1981 [12], dans le cadre de l'estimation du paramètre de position de la loi normale. L'estimation de coût comme sélecteur de variables a été développée récemment par Fourdrinier et Wells [6]. Dans leur article, les auteurs appliquent la procédure à la version réduite de l'estimateur des moindres carrés $\hat{\beta}_I^{MC} = (X_I^T X_I)^{-1} X_I^T Y$. L'inconvénient de cet estimateur est qu'il est nécessaire de calculer l'estimation de son coût pour les $(2^p - 1)$ sous-ensembles possibles afin d'en déterminer le minimum. Nous proposons donc de considérer plutôt l'estimateur du *lasso*. Cet estimateur proposant un chemin de régularisation, on a donc au maximum p sous-ensembles possibles à comparer. Il est donc beaucoup plus efficace dans la recherche du meilleur sous-ensemble I. On peut alors voir l'estimation de coût comme un critère d'arrêt pour le *lasso*.

Un autre intérêt de notre méthode est qu'elle peut s'appliquer à un type de lois plus général que le cas gaussien : les lois à symétrie sphérique. Ceci lui confère une propriété de robustesse distributionnelle.

Dans la partie 2, nous présentons le problème considéré et sa résolution par l'estimateur du *lasso*. Dans la partie 3, nous présentons le principe de l'estimation de coût ainsi que quelques résultats théoriques permettant la sélection de l'hyperparamètre optimal. Enfin, la partie 4 traite quelques exemples simulés et les résultats pratiques mis en comparaison avec les critères AIC et BIC.

2 Contexte et notations

On considère le modèle linéaire

$$Y = X\beta + \varepsilon \tag{2.1}$$

où Y est un vecteur aléatoire dans $\mathbb{R}^n$, X est une matrice fixée connue dans $\mathcal{M}_{n,p}$, avec $p < n$, β est le vecteur des coefficients de la régression, et ε est un vecteur aléatoire dans $\mathbb{R}^n$ représentant le bruit du modèle.

Notre hypothèse distributionnelle réside en ce que ε suit une loi à symétrie sphérique autour de 0, c'est-à-dire que cette loi est invariante par transformation orthogonale. Dans le cas où elle admet une densité, celle-ci est de la forme $t \mapsto f(\| t \|^2)$ pour une certaine fonction f de $\mathbb{R}_+$ dans $\mathbb{R}_+$. Lorsque f est spécifié comme étant $u \mapsto f(u) = 1/(2\pi)^{n/2} exp(-u/2)$, notre modèle correspond au modèle gaussien usuel. Une caractéristique de notre travail est que nous n'avons pas besoin de préciser la forme de f et, dans ce sens, nos résultats présentent une propriété de robustesse distributionnelle.

Notons que la loi de Y est translatée de la loi de ε par la translation $X\beta$: Y suit une loi à symétrie sphérique autour du paramètre de position $X\beta$ (voir annexe pour plus de développement sur les lois à symétrie sphérique). Nous noterons $\varepsilon \sim s.s.(0)$ et $Y \sim s.s.(X\beta)$.

Le problème que l'on souhaite résoudre ici est la sélection des variables les plus influentes sur la variable d'étude Y. Ceci équivaut à la recherche du sous-ensemble I inclus dans $\{1, \ldots, p\}$ tel que l'on puisse réduire le modèle (2.1) à :

$$Y = X_I \beta_I + \varepsilon \tag{2.2}$$

où X_I correspond à la matrice X réduite aux variables de I, autrement dit $X_I = (X^{i_1}, \ldots, X^{i_k})$, avec $I = \{i_1, \ldots, i_k\}$.

Une fois que l'on a défini un estimateur de β, par exemple l'estimateur du *lasso*, on souhaite en évaluer la qualité. Pour cela, on utilise une fonction de coût $L(\hat{\beta}, \beta)$ définie au préalable. La fonction la plus utilisée est le coût quadratique :

$$L(\hat{\beta}, \beta) = \| \hat{\beta} - \beta \|^2, \tag{2.3}$$

où $\hat{\beta}$ est une notation simplifiée pour $\hat{\beta}_{\hat{I}}$, $\hat{I}$ étant l'estimation de I.

Nous choisissons pour $\hat{\beta}$ l'estimateur du *lasso*, dont nous exposons quelques détails théoriques ainsi que les raisons de ce choix dans le paragraphe suivant.

2.1 Estimateur du *lasso*

Pour le modèle (2.2), on cherche à estimer β par la méthode du *lasso* (Least Absolute Shrinkage et Selection Operator). Cette méthode a été développée initialement par Tibshirani en 1996 [14], et généralise les travaux de Donoho et Johnstone [2] dans lesquels la matrice X était supposée orthogonale.

L'estimateur du *lasso* vérifie l'équation suivante :

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left(\| Y - X\beta \|^2 + \lambda \sum_{i=1}^{p} |\beta_i| \right), \quad \lambda > 0 \tag{2.4}$$

La pénalisation par la norme L_1 permet d'obtenir un certain nombre de coefficients nuls, donnant au *lasso* une propriété de "sparsité". En effet, une faible valeur de λ pénalise peu l'erreur quadratique moyenne et correspond donc à la solution des moindres carrés, tandis qu'une forte valeur aura au contraire pour effet de faire tendre la solution vers 0. De plus, cette

pénalisation est convexe, la solution est donc unique pour λ fixé. On peut donc obtenir un chemin de régularisation en faisant varier λ, depuis l'ensemble vide jusqu'à la sélection de la totalité des variables.

Lorsque la matrice X est orthogonale (*i.e.* $\hat{\beta}^{MC} = X^T Y$), l'estimateur du *lasso* de β, appelé *soft thresholding* par Donoho et Johnstone, prend la forme suivante, pour $\lambda \geq 0$ fixé :

$$\forall i = 1, \ldots, p \qquad \hat{\beta}_i^{lasso} = \Big(\hat{\beta}_i^{MC} - \lambda \operatorname{sgn}(\hat{\beta}_i^{MC})\Big)\mathbf{1}_{\{|\hat{\beta}_i^{MC}| > \lambda\}} \tag{2.5}$$

Le *lasso* translate donc les coefficients de l'estimateur des moindres carrés par un facteur constant, en les tronquant à 0. Cette forme du *lasso* est intéressante car elle permet une approche simplifiée du problème. Dans la suite de ce rapport, nous restreignons notre étude à ce cas orthogonal.

2.2 Forme canonique du modèle linéaire généralisé

Dans un souci de simplification des notations et calculs, nous utiliserons la forme canonique du modèle linéaire, définie dans le livre à paraître de Fourdrinier et Strawderman [5], qui consiste en une transformation orthogonale.

Soient :
- G_1 une matrice de dimensions $p \times n$ telle que les p lignes de G_1 couvrent l'espace des colonnes de X,
- et G_2 une matrice de dimensions $n-p \times n$, telle que $G_2^T G_1 = 0$, $G_1^T G_2 = 0$, et $G_2 G_2^T = I_{n-p}$.

La matrice $G = (G_1, G_2)^T$ est une matrice orthogonale, par laquelle on transforme le vecteur Y :

$$Y \quad \xrightarrow{G} \quad \begin{pmatrix} G_1 Y \\ G_2 Y \end{pmatrix} = \begin{pmatrix} G_1 X\beta \\ G_2 X\beta \end{pmatrix} + G\varepsilon$$

En posant $Z = G_1 Y$, $U = G_2 Y$, et $\theta = G_1 X\beta$, on obtient le modèle équivalent suivant (du fait de l'orthogonalité entre G_2 et l'espace des colonnes de X) :

$$\begin{pmatrix} Z \\ U \end{pmatrix} = \begin{pmatrix} \theta \\ 0 \end{pmatrix} + G\varepsilon \tag{2.6}$$

Si $Y \sim \mathcal{N}(X\beta, \sigma^2)$, alors le vecteur aléatoire $(Z, U)^T \sim \mathcal{N}((\theta, 0)^T, \sigma^2)$. Nous verrons dans l'annexe A que cette propriété se vérifie également pour les lois à symétrie sphérique : $(Z, U)^T \sim s.s.(\theta, 0)^T$ si $Y \sim s.s.(X\beta)$.

Comme on l'a défini précédemment, G_1^T et X appartiennent au même espace. Il existe donc une matrice A non singulière de dimensions $p \times p$ telle que $X = G_1^T A$. Par conséquent, on a que $\theta = A\beta$ et il est équivalent d'estimer β et θ.

En pratique, on peut utiliser la décomposition QR de X et poser :

$$G = Q^T \qquad \text{et} \qquad A = R.$$

De la même manière que l'on a défini une fonction de coût $L(\hat{\beta}, \beta)$ pour évaluer la qualité de l'estimateur $\hat{\beta}$, on définit la même fonction de coût $L(\varphi, \theta)$ pour un estimateur $\hat{\theta} = \varphi$ de θ. Le coût quadratique de φ est lié au coût quadratique de $\hat{\beta}$ de la manière suivante :

$$\begin{aligned} L(\varphi, \theta) &= \| \varphi - \theta \|^2 \\ &= \| A\hat{\beta} - A\beta \|^2 \\ &= (\hat{\beta} - \beta)^T A^T A (\hat{\beta} - \beta) \end{aligned}$$

Lorsque X est orthogonale, A est une matrice diagonale avec des 1 et des -1 sur la diagonale. Elle est donc également orthogonale, et on a égalité entre $L(\varphi, \theta)$ et $L(\hat{\beta}, \beta)$.

Sous la forme canonique, l'estimateur des moindres carrés devient :

$$\hat{\theta}^{MC} = \varphi_0(Z) = Z \tag{2.7}$$

En effet, on a :

$$\begin{array}{lll}
\hat{\theta}^{MC} & = A\hat{\beta}^{MC} & \\
& = A(X^TX)^{-1}X^TY & \\
& = A[(G_1^TA)^T(G_1^TA)]^{-1}(G_1^TA)^TY, & \text{en remplaçant } X \text{ par } G_1^TA \\
& = A(A^TG_1G_1^TA)^{-1}A^TG_1Y, & \text{or } G_1G_1^T = I_p \text{ et } G_1Y = Z \\
& = A(A^TA)^{-1}A^TZ, & \text{et } A \text{ étant carrée} \\
& = A(A)^{-1}(A^T)^{-1}A^TZ & \\
& = Z &
\end{array}$$

□

Quant à l'estimateur du *lasso* dans le cas orthogonal, il prend la forme :

$$\forall i = 1, \ldots, p \qquad \varphi_i^{lasso}(Z) = (Z_i - \lambda \text{sgn}(Z_i))\, \mathbf{1}_{\{|Z_i|>\lambda\}} \tag{2.8}$$

En effet, on a :

$$\begin{array}{rcl}
\varphi_i^{lasso} & = & A_{(i,i)}\hat{\beta}_i^{lasso} \\
& = & A_{(i,i)}\Big(\hat{\beta}_i^{MC} - \lambda\, \text{sgn}(\hat{\beta}_i^{MC})\Big)\mathbf{1}_{\{|\hat{\beta}_i^{MC}|>\lambda\}} \\
& = & \Big(A_{(i,i)}\hat{\beta}_i^{MC} - \lambda\, \text{sgn}(A_{(i,i)}\hat{\beta}_i^{MC})\Big)\mathbf{1}_{\{|\hat{\beta}_i^{MC}|>\lambda\}} \\
& = & \Big(\hat{\theta}_i^{MC} - \lambda\, \text{sgn}(\hat{\theta}_i^{MC})\Big)\mathbf{1}_{\{|\hat{\beta}_i^{MC}|>\lambda\}} \\
& = & \Big(Z_i - \lambda\, \text{sgn}(Z_i)\Big)\mathbf{1}_{\{|Z_i|>\lambda\}},
\end{array}$$

car $|\, \hat{\theta}_i \,| = |\, Z_i \,| = |\, A_{(i,i)}\hat{\beta}_i^{MC} \,| = |\, A_{(i,i)} \,||\, \hat{\beta}_i^{MC} \,| = |\, \hat{\beta}_i^{MC} \,|$.

On peut également écrire l'estimateur du *lasso* dans le cas orthogonal sous la forme :

$$\varphi^{lasso}(Z) = Z + g(Z) \tag{2.9}$$

avec $g_i(Z) = -\lambda\, \text{sgn}(Z_i)\mathbf{1}_{\{|Z_i|>\lambda\}} - Z_i\mathbf{1}_{\{|Z_i|\leq\lambda\}} \qquad \forall i = 1, \ldots, p$

2.3 Choix de l'hyperparamètre

Comme on l'a vu précédemment, l'estimateur du *lasso* fournit un chemin de régularisation. En effet, une valeur très grande de λ correspond à l'ensemble vide, où aucune variable n'est sélectionnée, et la diminution de λ fait entrer une à une les variables dans l'ensemble de sélection, jusqu'à obtenir l'ensemble complet $\{1, \ldots, p\}$ correspondant à la solution des moindres carrés. La taille de l'ensemble sélectionné varie donc de 0 à p par pas de 1. Il apparaît clairement que le choix de λ est crucial dans la détermination du sous-ensemble I.

Tibshirani [14] propose trois méthodes pour l'estimer : validation croisée, validation croisée généralisée, et estimation analytique du risque, également appelée estimation de coût. On peut y ajouter les critères d'information de type AIC et BIC.

Les deux premières sont très coûteuses en temps de calcul, et ne sont donc pas appropriées dans les problèmes à grandes dimensions.

Les deux dernières sont fiables lorsque les erreurs sont gaussiennes, ou tout du moins lorsque l'on fait une hypothèse stricte sur la forme de leur distribution, mais leurs performances sont moins bonnes lorsque l'on élargit cette hypothèse distributionnelle.

Nous allons voir la troisième méthode en détails dans la partie 3.

3 Estimation de coût

L'évaluation de l'estimateur φ^{lasso} d'un paramètre θ se fait habituellement au travers d'une fonction de coût $L(\varphi^{lasso}, \theta)$ définie au préalable, la valeur optimale étant celle minimisant le coût L. Cette fonction dépend des observations, et peut donc varier considérablement si l'on considère un nouveau jeu de données. Son risque lui est donc souvent préféré, perdant ainsi toute information apportée par les observations en en prenant l'espérance. Par ailleurs, aucune de ces deux fonctions ne sont accessibles directement car elles dépendent du paramètre inconnu θ. Généralement, elles sont estimées par validation croisée, mais cette technique se révèle trop coûteuse dans le cas de problèmes à grandes dimensions.

Nous proposons donc d'estimer le coût à partir des observations uniquement. On cherche ainsi un compromis entre entre la variabilité de l'évaluation face à l'aléa et l'indépendance totale aux données, le tout indépendamment du paramètre θ.

Soit $\delta(Z,U)$ un estimateur du coût $L(\varphi^{lasso}, \theta)$. $\delta(Z,U)$ dépend directement de φ^{lasso} et donc de l'hyperparamètre λ. En faisant varier λ, et si le coût L est convexe, on peut déterminer le minimum unique de $\delta(Z,U)$, amenant au choix correspondant pour λ. Lorsque l'on applique l'estimation de coût au *lasso*, le minimum obtenu peut être vu comme un critère d'arrêt.

Nous allons voir des exemples d'estimateurs de coûts dans les paragraphes suivants. Nous rappelons que nous utilisons ici le coût quadratique $L(\varphi,\theta) = \| \varphi - \theta \|^2$, mais que cette méthode est applicable à d'autres fonctions de coût.

Rappelons également l'estimateur du *lasso* dans le cas orthogonal :

$$\varphi^{lasso}(Z) = Z + g(Z) \tag{3.1}$$

où $g_i(Z) = -\lambda \operatorname{sgn}(Z_i)\mathbf{1}_{\{|Z_i|>\lambda\}} - Z_i \mathbf{1}_{\{|Z_i|\leq\lambda\}} \qquad \forall i = 1,\dots,p.$

3.1 Estimateur sans biais

Définition 1 *Soit $\delta_0(Z,U)$ un estimateur du coût $L(\varphi,\theta)$. $\delta_0(Z,U)$ est dit* sans biais *s'il vérifie :*

$$\mathbb{E}_\theta[\delta_0(Z,U)] = \mathbb{E}_\theta[L(\varphi,\theta)] = R(\varphi), \tag{3.2}$$

où $\mathbb{E}_\theta$ représente l'espérance par rapport à θ et $R(\varphi)$ représente le risque de φ.

Dans notre contexte, voyons ce que vaut le risque de φ^{lasso} :

$$\begin{aligned}
\mathbb{E}_\theta\big[\| \varphi^{lasso} - \theta \|^2 \big] &= \mathbb{E}_\theta\big[\| Z + g(Z) - \theta \|^2 \big] \\
&= \mathbb{E}_\theta\big[\| Z - \theta \|^2 + \| g(Z) \|^2 + 2(Z-\theta)^T g(Z) \big]
\end{aligned}$$

Or, la conséquence de la propriété 2 de l'annexe nous donne que :

$$\begin{aligned}
\mathbb{E}_\theta[\| Z - \theta \|^2 + \| U \|^2] &= \frac{\mathbb{E}_\theta[R^2]}{n} I_n \\
\Rightarrow \quad \mathbb{E}_\theta[\| Z - \theta \|^2] &= \frac{p}{n-p} \mathbb{E}_\theta[\| U \|^2].
\end{aligned}$$

Par ailleurs, on a :

$$\| g(Z) \|^2 = \sum_{i=1}^{p} \big[\lambda^2 \mathbf{1}_{\{|Z_i|>\lambda\}} + Z_i^2 \mathbf{1}_{\{|Z_i|\leq\lambda\}}\big] = \sum_{i=1}^{p} \lambda^2 \wedge Z_i^2.$$

La notation $a \wedge b$ correspond ici au minimum entre a et b.

Pour finir, le troisième terme du membre de droite donne, par application de l'égalité de Stein (pour plus de détails, voir le rapport technique de Fourdrinier [4]) :

$$\begin{aligned}\mathbb{E}_\theta[(Z-\theta)^T g(Z)] &= \frac{1}{n-p}\mathbb{E}_\theta[\| U \|^2 \text{ div } g(Z)] \\ &= \frac{1}{n-p}\mathbb{E}_\theta[\| U \|^2 (k-p)],\end{aligned}$$

où $k = \text{Card}\{i \setminus |Z_i| > \lambda\}$ est le nombre d'éléments sélectionnés. On obtient alors :

$$\mathbb{E}_\theta[\| \varphi^{lasso} - \theta \|^2] = \mathbb{E}_\theta \left[\frac{2k-p}{n-p} \| U \|^2 + \sum_{i=1}^{p} Z_i^2 \wedge \lambda^2 \right]. \tag{3.3}$$

Un estimateur sans biais de $R(\varphi^{lasso})$ sera donc :

$$\delta_0(Z,U) = \frac{2k-p}{n-p} \| U \|^2 + \sum_{i=1}^{p} Z_i^2 \wedge \lambda^2. \tag{3.4}$$

On voit ici apparaître deux éléments importants : d'un côté la différence entre le nombre d'éléments sélectionnés k et le nombre d'éléments non sélectionnés $p-k$ ($2k-p = k-(p-k)$) ; et d'un autre côté, on constate les valeurs mêmes des éléments non sélectionnés au travers du terme $\sum_{i=1}^{p} Z_i^2 \wedge \lambda^2$. On fait donc l'hypothèse que ces éléments mis de côté dans un premier temps apportent tout de même une information à prendre en compte pour valider la sélection effectuée. Le terme $\| U \|^2 /(n-p)$ est constant et correspond à une estimation de la variance du modèle.

Si le modèle compte des variables de bruit ou de faible importance dans la matrice X, les éléments de Z correspondant seront petits en valeur absolue. Si λ est tel que l'on ne sélectionne aucune de ces variables peu influentes, le terme $\sum_{i=1}^{p} Z_i^2 \wedge \lambda^2$ est faible. Si on augmente suffisamment λ, on enlève certaines variables importantes de la sélection, et alors ce même terme augmente fortement. Le premier terme, quant à lui, est diminué par un facteur de $\| U \|^2 /(n-p)$, ce qui ne permet pas de compenser le saut du deuxième terme.

3.2 Qualité des estimateurs de coût

Comme pour évaluer la qualité de l'estimateur φ de θ, nous souhaitons évaluer la qualité de l'estimateur δ de $L(\varphi, \theta)$. De la même manière, on définit une fonction de coût pour δ par $\mathcal{L}(\delta)$. Le "meilleur" estimateur du coût L sera celui minimisant le risque $\mathcal{R}(\delta) = \mathbb{E}_\theta[\mathcal{L}(\delta)]$.

Là encore, le choix le plus courant est le coût quadratique :

$$\mathcal{L}(\delta) = \left(\delta(Z,U) - \| \varphi - \theta \|^2\right)^2 \tag{3.5}$$

A noter que L et $\mathcal{L}$ ne sont pas nécessairement identiques.

$\mathcal{L}(\delta)$ et $\mathcal{R}(\delta)$ dépendent eux aussi du paramètre inconnu θ. Cependant, ils servent essentiellement au cadre théorique lorsque l'on cherche les conditions d'amélioration d'un estimateur du coût, comme nous le verrons dans le prochain paragraphe, ainsi qu'aux simulations pour vérifier ces conditions. Notre but *in fine* est de proposer un seul "bon" estimateur de coût dans le contexte étudié.

3.3 Estimateurs compétitifs

Selon Fourdrinier [4], l'estimateur sans biais peut être amélioré étant donné qu'il peut prendre des valeurs négatives, alors qu'il estime une quantité positive. Il propose deux estimateurs compétitifs, que nous allons voir ci-après.

3.3.1 Estimateur positif

Le premier estimateur proposé est la partie positive de l'estimateur sans biais δ_0 :

$$\delta_0^+(Z,U) = \max(\delta_0(Z,U), 0) \tag{3.6}$$

$\delta_0^+(Z,U)$ est toujours meilleur que $\delta_0(Z,U)$ au sens du risque quadratique. Pour la preuve, voir [4].

3.3.2 Estimateur avec fonction de correction

Le second type d'estimateurs proposé fait intervenir une fonction de correction $\gamma : \mathbb{R}^p \to \mathbb{R}$ deux fois faiblement différentiable :

$$\delta(Z,U) = \delta_0(Z,U) - \parallel U \parallel^4 \gamma(Z) \tag{3.7}$$

Cet estimateur tient son origine de deux constats : le premier vient de ce que l'estimateur sans biais du coût a tendance à surestimer le coût réel, et on cherche donc à le corriger par un terme négatif ; le deuxième est que la présence d'une puissance de $\parallel U \parallel$ ne nécessite aucune hypothèse distributionnelle et apporte ainsi une robustesse à l'estimateur.

$\delta(Z,U)$ est meilleur que $\delta_0(Z,U)$ au sens des moindres carrés si on a l'inégalité différentielle suivante, pour $p \geq 5$:

$$d_\theta = \mathbb{E}_\theta\left[\left(\delta - \parallel \varphi^{lasso} - \theta \parallel^2\right)^2 - \left(\delta_0 - \parallel \varphi^{lasso} - \theta \parallel^2\right)^2\right] \leq 0 \tag{3.8}$$

Or, d'après (3.7), on a :

$$\begin{aligned} d_\theta &= \mathbb{E}_\theta\left[\left(\delta_0 - \parallel U \parallel^4 \gamma(Z) - \parallel \varphi^{lasso} - \theta \parallel^2\right)^2 - \left(\delta_0 - \parallel \varphi^{lasso} - \theta \parallel^2\right)^2\right] \\ &= \mathbb{E}_\theta\left[\parallel U \parallel^8 \gamma^2(Z) - 2 \parallel U \parallel^4 \gamma(Z)\left(\delta_0 - \parallel \varphi^{lasso} - \theta \parallel^2\right)\right] \end{aligned}$$

Fourdrinier [4] développe le terme de droite et montre que la domination de $\delta(Z,U)$ sur $\delta_0(Z,U)$ a lieu si

$$\begin{aligned} d_\theta &= \mathbb{E}_\theta\left[\left\{\frac{2}{(n-P+4)(n-P+6)}\Delta\gamma(Z) + \gamma^2(Z)\right\} \parallel U \parallel^8\right. \\ &+ \left.\frac{4}{n-P+4}\left\{2\frac{P-2p}{n-P}\gamma(Z) + \nabla\gamma(Z)\cdot g(Z)\right\} \parallel U \parallel^6\right] \leq 0 \end{aligned} \tag{3.9}$$

Une condition suffisante pour que $\delta(Z,U)$ domine $\delta_0(Z,U)$ est alors que γ satisfasse les deux inégalités suivantes :

$$\frac{2}{(n-p+4)(n-p+6)}\Delta\gamma(Z) + \gamma^2(Z) \leq 0 \tag{3.10}$$

$$2\frac{p-2k}{n-p}\gamma(Z) + \nabla\gamma(Z)\cdot g(Z) \leq 0 \tag{3.11}$$

Nous proposons d'étudier les deux fonctions de correction (3.12) et (3.13).

$$\gamma_1(Z) = \frac{a_1}{\parallel Z \parallel^2}, \quad a_1 > 0 \tag{3.12}$$

$$\gamma_2(Z) = \frac{a_2}{\left[(k+1)Z^2_{(k+1)} + \sum_{i=k+2}^{p} Z^2_{(i)}\right]} \tag{3.13}$$

Ici, la notation $Z_{(i)}, i = k+1, \ldots, p$, correspond aux éléments inférieurs à λ en valeur absolue, et sont présentés comme si on les avait ordonnés : $|Z_{(1)}| > \cdots > |Z_{(p)}|$. $Z_{(k+1)}$ correspond donc au plus grand élément en valeur absolue immédiatement inférieur à λ. Il y a égalité entre les deux fonctions de correction lorsque λ est très grand. On voit que la première fonction de correction fait intervenir l'ensemble des éléments de Z, alors que la seconde ne considère que ceux qui n'ont pas été sélectionnés par l'estimateur du *lasso*.

Si l'on considère la correction positive $\gamma_1(Z) = a_1 / \parallel Z \parallel^2$, $a_1 > 0$, son gradient et son laplacien valent :

$$\nabla\gamma_1(Z) = -2a_1 Z / \parallel Z \parallel^4 \quad \text{et} \quad \Delta\gamma_1(Z) = -2a_1(p-4) / \parallel Z \parallel^4 .$$

L'inégalité (3.10) devient alors :

$$\frac{-4a_1(p-4)}{(n-p+4)(n-p+6) \parallel Z \parallel^4} + \frac{a_1^2}{\parallel Z \parallel^4} \leq 0$$

Ceci implique que l'on ait $0 \leq a_1 \leq 4(p-4)/[(n-p+4)(n-p+6)]$.

Comme $\gamma_1(Z) > 0$, $\forall Z \in \mathbb{R}^p$, l'inégalité (3.11) donne :

$$\begin{aligned} k &\geq \frac{p}{2} + \frac{n-p}{4}\left(\frac{\nabla\gamma_1(Z)\cdot g(Z)}{\gamma_1(Z)}\right) \\ &\geq \frac{p}{2} + \frac{n-p}{2}\left(\frac{\sum_{i=1}^{p}|Z_i|(\lambda \wedge |Z_i|)}{\parallel Z \parallel^2}\right) \\ &> \frac{p}{2} \end{aligned}$$

Cette inégalité implique que l'estimateur $\delta(\gamma_1)$ améliore l'estimateur sans biais δ_0 dès que $k > p/2$. Cependant, bien que l'on puisse obtenir une amélioration sur l'estimation, on remarque que la correction γ_1 est constante pour tout λ. Ceci signifie que le minimum sera le même pour les deux estimateurs $\delta(\gamma_1)$ et δ_0, et donc aucune amélioration n'est possible en terme de sélection de variables.

On considère maintenant la correction $\gamma_2(Z) = \dfrac{a_2}{(k+1)Z^2_{(k+1)} + \sum_{i=k+2}^{p} Z^2_{(i)}}$, où les $p-k$ variables considérées sont celles inférieures à λ en valeur absolue.

Son gradient et son laplacien valent :

$$\begin{aligned} \nabla\gamma_2(Z) &= \frac{-2a_2\left(0, \ldots, 0, (k+1)Z_{(k+1)}, Z_{(k+2)}, \ldots, Z_{(p)}\right)^T}{\left[(k+1)Z^2_{(k+1)} + \sum_{i=k+1}^{p} Z^2_{(i)}\right]^2} \\ \Delta\gamma_2(Z) &= \frac{-2a_2(p-2)}{\left[(k+1)Z^2_{(k+1)} + \sum_{i=k+2}^{p} Z^2_{(i)}\right]^2} + \frac{4a_2 k(k+1)Z^2_{(k+1)}}{\left[(k+1)Z^2_{(k+1)} + \sum_{i=k+2}^{p} Z^2_{(i)}\right]^3} \end{aligned}$$

L'inégalité (3.10) donne donc :

$$-\frac{4a_2}{(n-p+4)(n-p+6)}\left(p - 2 - \frac{2k(k+1)Z^2_{(k+1)}}{(k+1)Z^2_{(k+1)} + \sum_{i=k+2}^{p} Z^2_{(i)}}\right) + a_2^2 \leq 0$$

ce qui implique :

$$\frac{-4p}{(n-p+4)(n-p+6)} \leq a_2 \leq \frac{4(p-2)}{(n-p+4)(n-p+6)}.$$

Quant à l'inégalité (3.11), elle donne :

$$\begin{aligned}
&\frac{4}{n-p+4}\left(2\frac{p-2k}{n-p}\cdot\frac{a_2}{(k+1)Z_{(k+1)}^2+\sum_{i=k+2}^{p}Z_{(i)}^2}+\frac{2a_2\left[(k+1)Z_{(k+1)}^2+\sum_{i=k+2}^{p}Z_{(i)}^2\right]}{\left[(k+1)Z_{(k+1)}^2+\sum_{i=k+1}^{p}Z_{(i)}^2\right]^2}\right)\\
=\quad&\frac{8a_2}{(n-p+4)\left((k+1)Z_{(k+1)}^2+\sum_{i=k+2}^{p}Z_{(i)}^2\right)}\left(\frac{p-2k}{n-p}+1\right)\leq 0
\end{aligned}$$

On peut donc obtenir une amélioration en terme de risque quadratique par rapport à l'estimateur sans biais à partir de $k \geq n/2$ lorsque $a_2 > 0$ ou pour $k \leq n/2$ lorsque $a_2 < 0$. Il semble ici plus avantageux de choisir une valeur négative pour a_2 lorsque le rapport p/n est petit.

Cette fonction de correction est intéressante de par sa nature : c'est l'inverse du carré des éléments non sélectionnés par λ. Si λ est élevé, il rejette des variables importantes, qui donneront une valeur faible pour γ_2 et améliore donc peu l'estimateur sans biais δ_0. A l'inverse, lorsqu'on a sélectionné toutes les variables importantes, celles qui restent sont d'amplitude moindre, et donc γ_2 aura une valeur élevée. Il apparaît donc qu'il y a un grand avantage à tirer de cette fonction de correction dans le cadre de la sélection de variables. On voit aussi avec cette analyse que, si a_2 est positif, l'estimateur $\delta(\gamma_2)$ s'éloigne de δ_0 vers le bas à mesure que λ augmente, ce qui signifie qu'il est possible d'obtenir un estimateur décroissant en fonction de λ et donc un mauvais sélecteur de variables. Nous avons donc un double avantage à choisir a_2 positif. Le choix de a_2 doit être fait avec précaution. Une valeur trop faible l'éloigne peu de l'estimateur sans biais ce qui donne des minima très proches entre les deux fonctions et on peut donc se contenter de l'estimation de δ_0. Il faut donc choisir une valeur de a_2 suffisamment grande pour obtenir une différence notable entre les deux estimateurs.

4 Simulations

Nous nous proposons dans cette partie de tester les qualités des estimateurs δ_0 et $\delta(\gamma_2)$ dans un exemple simulé. Nous ne tiendrons pas compte de l'estimateur δ_0^+ car dans toutes nos simulations δ_0 est positif, ni de l'estimateur $\delta(\gamma_1)$ pour les raisons invoquées dans la partie précédente sur l'amélioration en terme de sélection.

4.1 Protocole

Pour toutes les simulations, on pose $n = 1000$, et $p = 250$. La matrice X est obtenue à partir de p vecteurs gaussiens de taille n, que l'on factorise par la méthode QR. Pour que X soit orthogonale, on pose $X = (Q^1, \dots, Q^p)$, les p premières colonnes de Q. Par la suite, la matrice X est gardée fixe.

Le vecteur β est fixé dans un premier temps à $(0, \dots, 0{,}10, \dots, 10{,}0, \dots, 0{,}10, \dots, 10{,}0, \dots, 0)^T$, où chaque série de même valeur a une longueur de $p/5 = 50$. Cet exemple est inspiré de l'exemple 4 dans [14]. Dans un deuxième temps, on génère β à partir de k coefficients non nuls suivant une loi normale $\mathcal{N}(0, 10)$ et $(n - k)$ coefficients nuls, où k est lui-même généré selon une loi uniforme discrète $\mathcal{U}_{\{1,\dots,p\}}$. Ceci nous permet d'étudier le comportement des estimateurs pour des exemples plus ou moins difficiles.

Enfin, on génère les erreurs ε comme produit d'une variable uniforme sur la sphère unité et d'un rayon, le rayon étant généré selon une loi à support $[0, +\infty[$ dont des exemples sont présentés dans le paragraphe suivant. Pour chaque exemple de distribution et chaque exemple de coefficient β, nous générons un nombre $r = 200$ de répliques des erreurs, toutes différentes, nous permettant d'étudier le comportement des estimateurs face à l'aléa.

Le vecteur d'observations de la variable d'étude Y est ainsi obtenu à partir de la matrice X, du vecteur fixé des coefficients β, et de chaque vecteur de bruit généré ε. Pour le modèle équivalent sous forme canonique, nous avons utilisé une nouvelle décomposition QR de X en posant $G = Q^T$, nous donnant les observations de Z et de U, ainsi que la valeur réelle du paramètre recherché θ.

On pose ensuite $\boldsymbol{\lambda}$ comme vecteur de taille m linéairement espacé entre 0 et $\max |Z_i|$. Enfin, pour chaque valeur de $\boldsymbol{\lambda}$ et pour chaque vecteur d'observations y, nous calculons l'estimation du *lasso* et les estimateurs de coût δ_0 et $\delta(\gamma_2)$, que nous comparons au coût quadratique réel de φ^{lasso}, ainsi qu'aux critères AIC et BIC.

4.2 Exemples de lois à symétrie sphérique

Nous présentons dans ce paragraphe quelques exemples de lois à symétrie sphérique, que nous utilisons pour les simulations.

Le tableau 1 présente les distributions du rayon d'une variable à symétrie sphérique, avec la distribution à symétrie sphérique associée quand celle-ci est connue. C représente la constante de normalisation, et la figure 1 en affiche les visualisations lorsque l'on fixe les paramètres de chacune des lois.

4.3 Choix de la constante

Comme nous l'avons précisé précédemment, le choix de a_2 est important pour obtenir un bon estimateur.

Les simulations montrent que le risque quadratique de $\delta(\gamma_2)$ est inférieur à celui de δ_0 lorsque a_2 est positif, et plus particulièrement pour $0 < a_2 \leq 4(p-2)/(n-p+4)(n-p+6)$. Cependant, pour ces valeurs de a_2, le minimum de $\delta(\gamma_2)$ est souvent obtenu pour les mêmes valeurs de λ

Nom	Distribution du rayon R	Distribution à symétrie sphérique associée
Chi $\chi(m)$	$h(r) = Cr^{m+n-1}\exp(-r^2), \quad m > 0$	Normale
Fisher $F(n,m)$	$h(r) = Cr^{n-1}(1+nr^2/m)^{-\frac{n+m}{2}}, \quad m \geq 4$	Student
Exponentielle $\mathcal{E}(\lambda)$	$h(r) = C\exp(-\lambda r)$	
Weibull $Wei(a,b)$	$h(r) = Cr^{n-1}\exp(-(r^2/b)^a), \quad a,b > 0$	

TAB. 1 – Exemples de distributions du rayon.

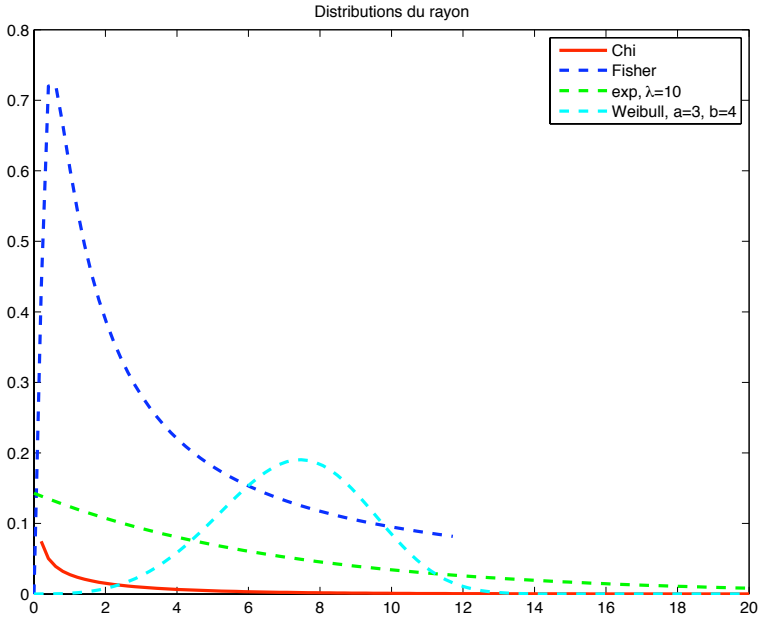

FIG. 1 – Forme générale des distributions des rayons

qu'avec δ_0, et parfois même il estime moins bien le nombre réel de coefficients non nuls. Les résultats des simulations nous montrent également qu'une amélioration est notable en matière de sélection de variables par rapport à l'estimateur sans biais lorsque la constante a_2 est négative, bien que son risque soit supérieur à celui de l'estimateur sans biais. On peut donc en conclure que le risque quadratique n'est pas un bon critère d'évaluation pour les estimateurs de coût lorsque l'on s'intéresse à la sélection des variables.

Par la suite, nous choisirons $a_2 = -0.07$, valeur qui a donné de bons résultats dans nos simulations. Le choix de a_2 sera étudié plus en détail dans nos futurs travaux.

4.4 Mesures de qualité des estimateurs

Dans ce paragraphe, nous présentons deux quantités nous permettant de mesurer la qualité de la sélection de variables effectuées par nos estimateurs.

Soient P et R deux mesures définies par :

$$P = \frac{\text{Card}\{j \setminus \hat{\beta}_j = \beta_j \neq 0\}}{\text{Card}\{j \setminus \hat{\beta}_j \neq 0\}}$$

$$R = \frac{\text{Card}\{j \setminus \hat{\beta}_j = \beta_j \neq 0\}}{\text{Card}\{j \setminus \beta_j \neq 0\}}$$

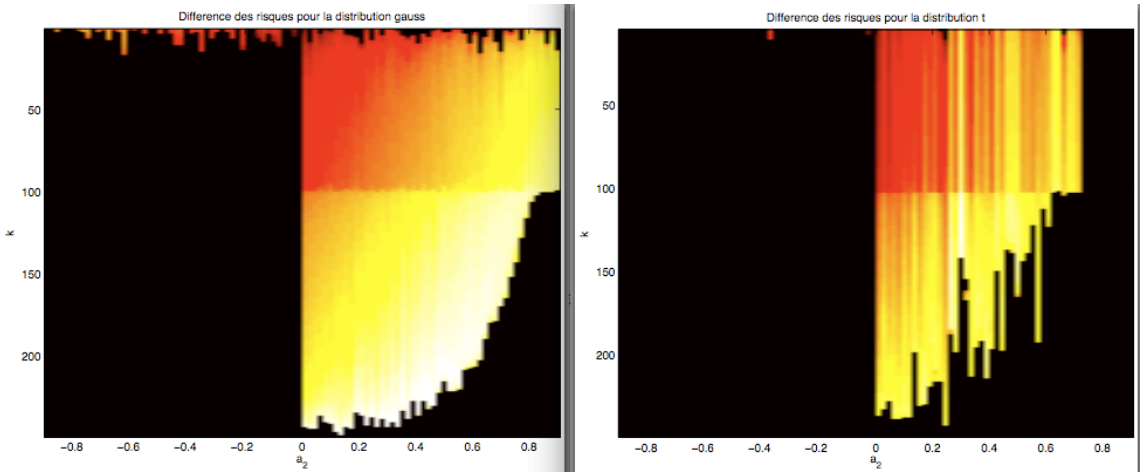

FIG. 2 – Évolution de ED_1 en fonction de a_2 et de k pour les distributions normale (gauche) et student (droite).

P est appelé *précision* et R *rappel*. On définit le F-score de la manière suivante :

$$\text{F-score} = 2\frac{P \cdot R}{P + R}. \tag{4.1}$$

Le F-score est une mesure de précision. Il vaut 1 si on trouve exactement tous les β_j non nuls et pas un de plus, et tend vers 0 si on est loin du vrai β.

Une autre mesure de la qualité des estimateurs est la probabilité empirique de sélectionner le bon sous-modèle :

$$\hat{\mathbb{P}}_{egalite}(\hat{\beta}) = \sum_{i=1}^{r} \mathbf{1}_{\{\hat{I}=I\}} \tag{4.2}$$

$$\hat{\mathbb{P}}_{inclusion}(\hat{\beta}) = \sum_{i=1}^{r} \mathbf{1}_{\{\hat{I}\supset I\}} \tag{4.3}$$

$$\hat{\mathbb{P}}_{inclusion}(\hat{\beta}, \alpha) = \sum_{i=1}^{r} \mathbf{1}_{\{\hat{I}\supset I\}\cap\{Card(\hat{I})=(1+\alpha)\times Card(I)\}} \tag{4.4}$$

4.5 Exemple avec β déterministe

On considère ici le cas $\beta = (0,\dots,0,10,\dots,10,0,\dots,0,10,\dots,10,0,\dots,0)^T$.

Les graphes de la figure 3 présentent l'évolution des estimateurs de coûts δ_0 et $\delta(\gamma_2)$ moyennés sur les r répliques en fonction du nombre de variables sélectionnées par le *lasso*. δ_0 y est représenté en noir, et $\delta(\gamma_2)$ en vert. On peut voir également le coût réel de l'estimateur du *lasso* en pointillés rouges, ainsi que l'AIC moyenné en magenta et le BIC moyenné en bleu. Pour chaque point d'estimation, on peut également voir l'écart-type des 100 répliques par les traits verticaux. Les quatre cadres correspondent aux quatre distributions de rayon citées dans le tableau 1. Enfin, la ligne verticale rouge est la droite d'équation $x = 100$, le nombre exacte de coefficients non nuls, et les croix et lignes de même couleur que les estimateurs indiquent la position de leur minimum.

On peut voir sur ces quatre figures que le nombre de variables sélectionné par l'estimateur δ_0 est en moyenne plus loin de la réalité que celui sélectionné par $\delta(\gamma_2)$, bien que sa courbe soit plus

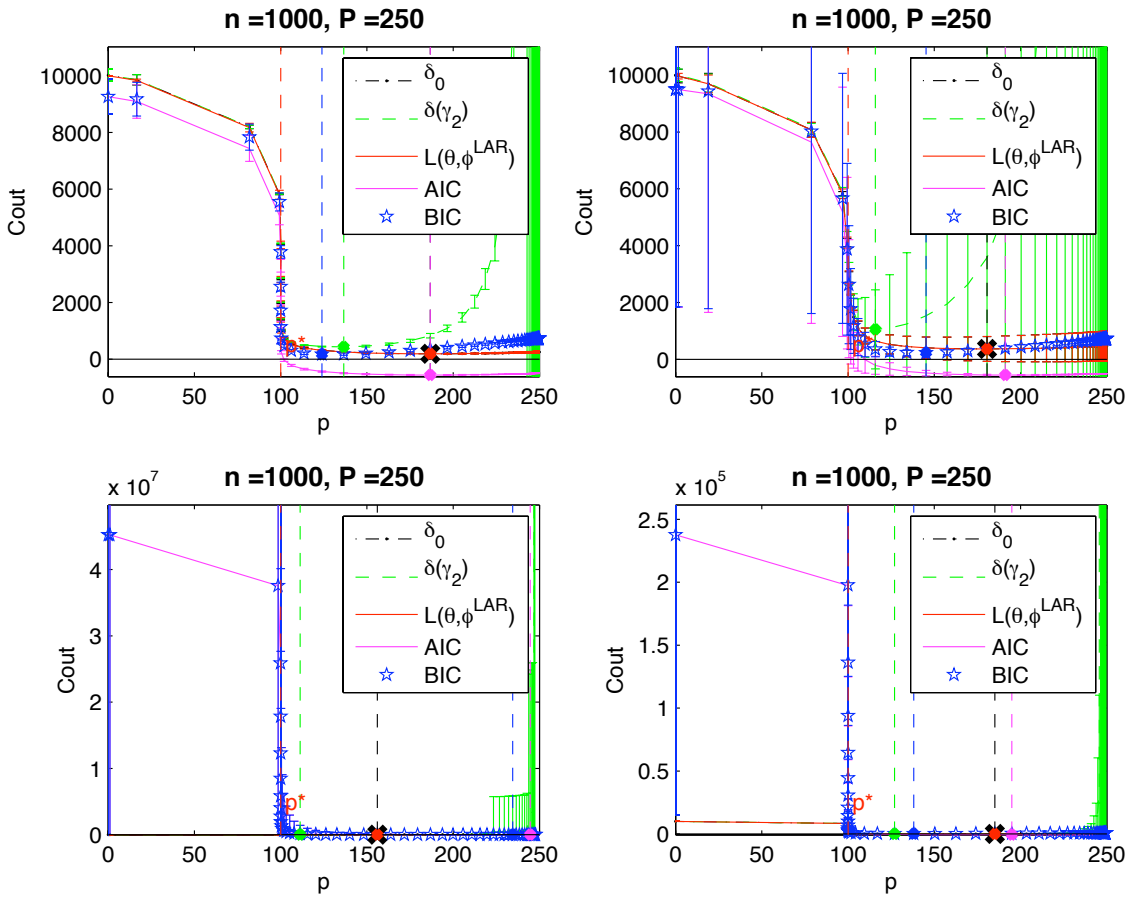

FIG. 3 – Évolution des estimateurs de coûts en fonction de k.

proche du vrai coût $L(\varphi^{lasso}, \theta)$. On remarque d'ailleurs ici que le vrai coût de φ^{lasso} continue de décroître bien après avoir sélectionné plus de variables que le nombre réel et serait lui-même un mauvais sélecteur de variables sur cet exemple si on pouvait y accéder directement. Une des raisons de ce phénomène vient de l'estimation biaisée du *lasso*. En effet, cet estimateur est connu pour être un bon sélecteur de variable, lorsque l'on fait un bon choix pour λ, mais il sous-estime la valeur des coefficients. A mesure que λ diminue, il se rapproche des vrais coefficients, mais dans le même temps il fait entrer des variables de bruit dans la sélection.

Par ailleurs, on constate que l'AIC sélectionne en général plus de variables que nos deux estimateurs. Le BIC, quant à lui, estime le mieux dans le cas gaussien, se situe entre nos deux estimateurs dans le cas d'un rayon Fisher et Weibull, et estime très mal dans le cas d'un rayon exponentiel. Pour le cas gaussien, ces remarques rejoignent ce que l'on sait de l'AIC qui a tendance a sélectionner des modèles complexes et du BIC qui au contraire sélectionne des modèles simples. Dans le cas exponentiel cependant, notre méthode estime beaucoup mieux que l'AIC et le BIC.

Si l'on regarde maintenant les mesures de qualité des modèles sélectionnés, présentées en figures 4 et 5, on remarque que l'ordre des estimateurs du meilleur au moins bon est à peu près conservé par rapport à la capacité à sélectionner le bon nombre de variables, hormis pour $\delta(\gamma_2)$ qui est plutôt mauvais. Ceci vient de ce que l'on met à 0 la fonction de correction γ_2 lorsqu'elle n'est pas calculable (i.e. lorsque $k = 250$), ce qui fausse l'évolution de l'estimation et donne souvent un minimum pour $k = 250$. Lorsque l'on enlève ces valeurs, l'estimateur est souvent presque aussi bon que le BIC, comme on peut le voir sur les figures 6 et 7.

Un autre point remarquable sur ces figures est que les mesures de qualité du BIC sont toujours très bonnes même dans le cas exponentiel. Ceci peut s'expliquer par la grande variance de l'estimateur dans ce cas, bien qu'il estime correctement le minimum pour un bon nombre de répliques. On peut faire la même remarque pour l'AIC sur le même exemple.

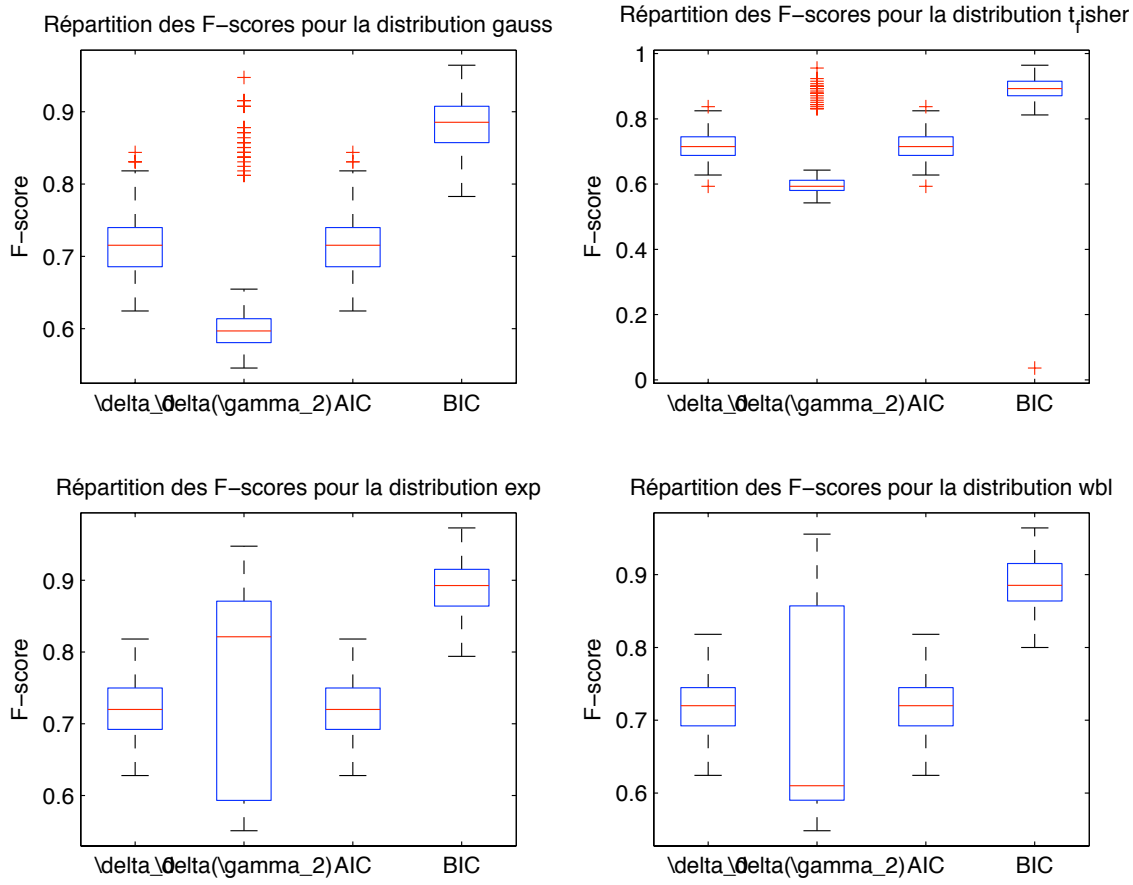

FIG. 4 – F-scores.

Pour résumer ces résultats, l'estimateur sans biais estime aussi mal que l'AIC sur cet exemple, et l'estimateur corrigé est un peu moins que le BIC mais s'en approche beaucoup.

Voyons si ce comportement se vérifie sur d'autres exemple.

4.6 Exemple avec β aléatoire

La figure 8 présente les F-scores moyennés sur les 200 répliques et pour 30 exemples différents : le F-score de l'estimateur sans biais δ_0 est en noir, celui de l'estimateur corrigé $\delta(\gamma_2)$ est en vert, celui de l'AIC en magenta et celui du BIC en bleu. L'estimateur sans biais est peu visible il a exactement les mêmes valeurs que l'AIC.

On peut voir sur cette figure que, pour la plupart des exemples, $\delta(\gamma_2)$ est meilleur que δ_0 en matière de sélection de variables. Le BIC est souvent meilleur encore que $\delta(\gamma_2)$, mais les deux estimateurs présentent des scores proches. L'AIC donne exactement les mêmes résultats que δ_0, les deux estimateurs sur-estiment souvent le bon nombre de variables et sont particulièrement mauvais lorsque ce nombre est faible (entre 5 et 20 sur 250). Les quatre estimateurs donnent d'excellents résultats pour les exemples 16, 17 et 19, où le nombre de variables explicatives varie de 197 à 212.

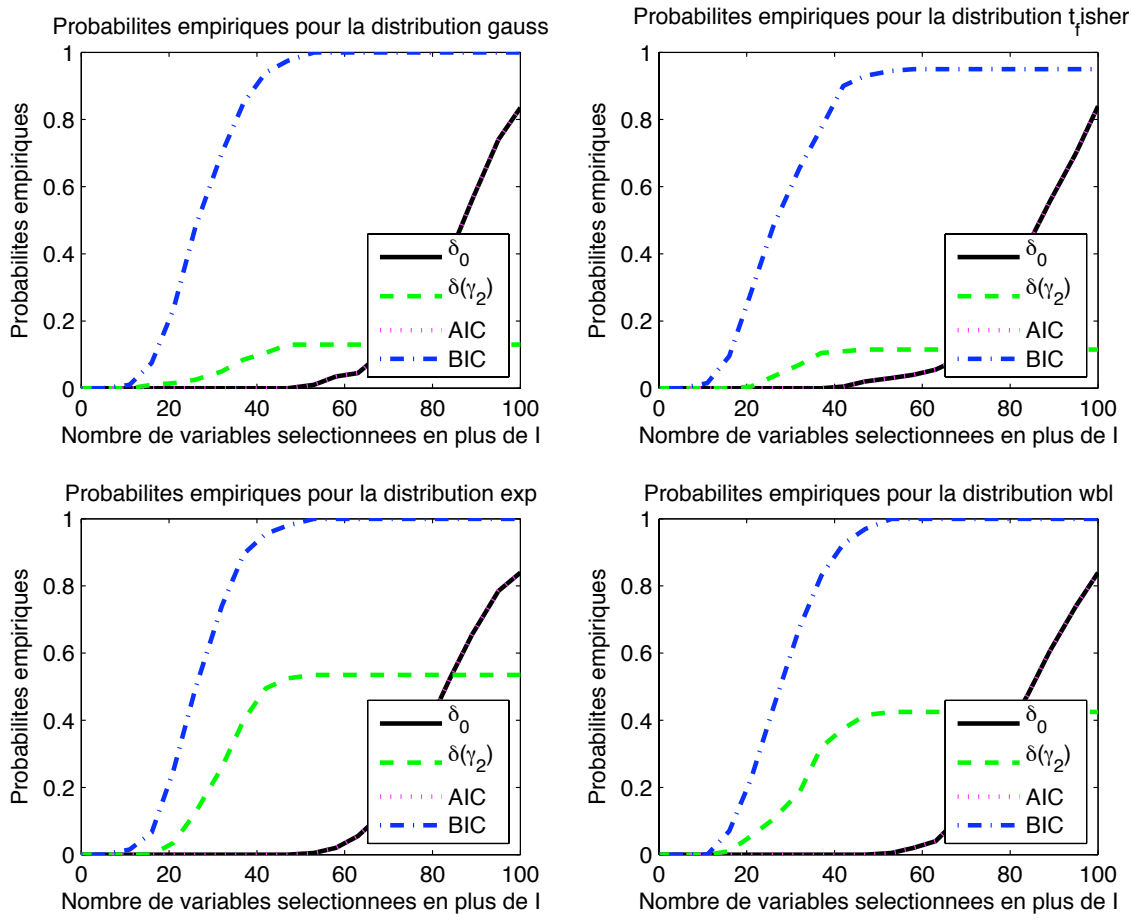

FIG. 5 – Probabilités empiriques.

5 Conclusion

Nous avons proposé dans ce rapport une nouvelle procédure pour la sélection de variables dans le modèle linéaire. Cette procédure possède l'avantage de pouvoir s'appliquer à un cadre distributionnelle large : les lois à symétrie sphérique. Nous avons testé cette méthode sur des exemples de ces lois, avec une contrainte d'orthogonalité pour la matrice de design. Les résultats obtenus jusqu'ici sont encourageants. Nous obtenons en effet de meilleurs résultats que l'AIC, et des résultats équivalents au BIC. Cependant, nos résultats pâtissent du biais de l'estimateur du *lasso*, et pourraient être améliorés en le remplaçant par l'estimateur des moindres carrés sur la sélection de variables obtenue par la *lasso*. Nous sommes déjà en cours d'étude pour ce nouvel estimateur des coefficients de la régression.

Dans la suite de nos travaux, nous chercherons à généraliser ces résultats en relâchant la contrainte d'orthogonalité de la matrice de design, nous permettant ainsi de tester la méthode avec des données réelles.

Par ailleurs, nous avons indiqué dans ce rapport que la méthode était applicable à d'autres fonctions de coût que le coût quadratique, ainsi qu'à d'autres estimateurs. En prenant le coût *hinge* par exemple, nous pourrions effectuer des tâches de classification au lieu de la régression.

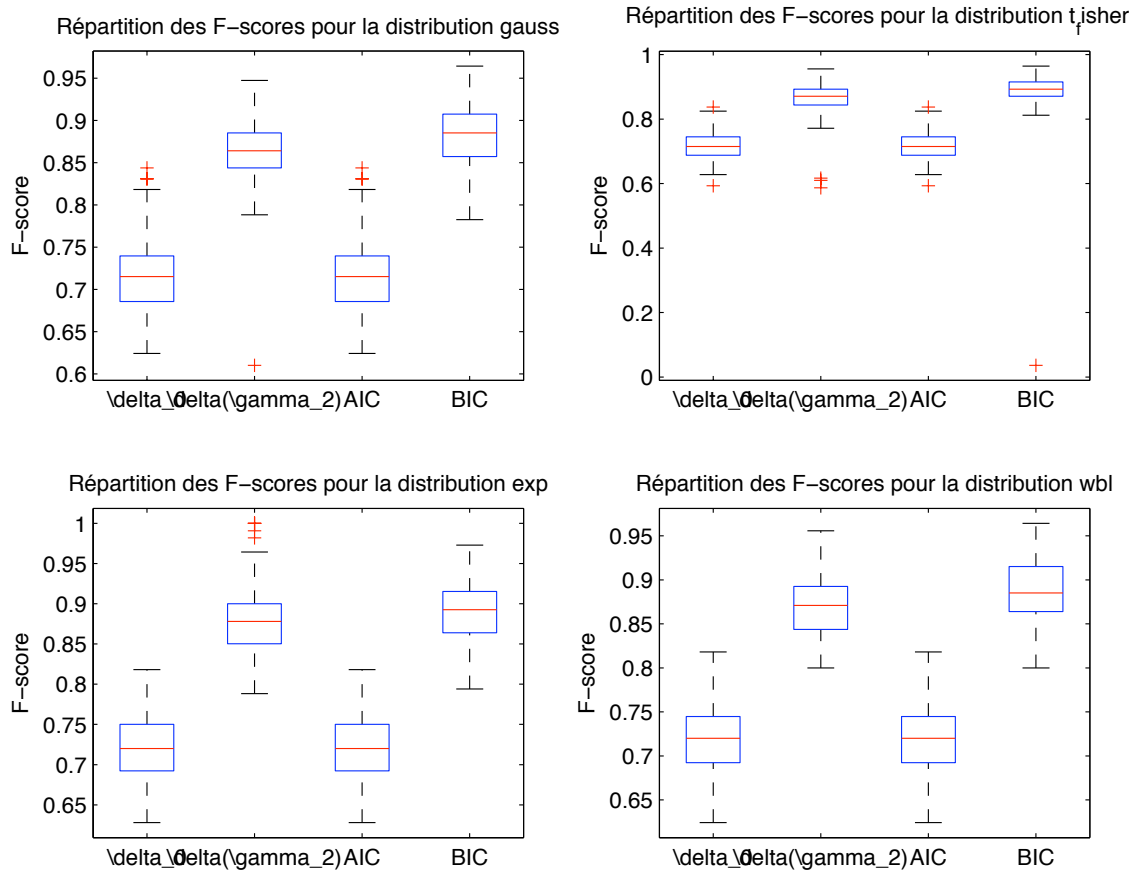

FIG. 6 – F-scores avec $\delta(\gamma 2)$ tronqué.

A Lois à symétrie sphérique

Dans cette annexe, nous présentons quelques aspects théoriques sur les lois à symétrie sphérique. Le modèle sphérique est une extension du modèle gaussien usuel. On garde de la loi normale la symétrie autour d'un paramètre de localisation, la moyenne. Mais la fonction de distribution n'a pas une forme définie comme la fonction gaussienne, ce qui permet de regrouper sous son nom d'autres lois telles la loi de Student et les mélanges de gaussiennes.

Les définitions et propriétés qui suivent sont tirées du livre à paraître de Fourdrinier et Strawderman [5]. Les preuves y sont disponibles.

A.1 Définitions et propriétés

Définition 2 (Vecteur à symétrie sphérique) *Un vecteur aléatoire $Y \in \mathbb{R}^n$ est dit à symétrie sphérique autour de $\mu \in \mathbb{R}^n$ si $Y - \mu$ est orthogonalement invariant, autrement dit si, pour toute transformation orthogonale H, le vecteur $V = H(Y - \mu)$ suit la même distribution que $Y - \mu$.*

On note $Y \sim s.s.(\mu)$.

Propriété 1 *Si $Y \sim s.s.(\mu)$, alors $HY \sim s.s.(H\mu)$ pour toute transformation orthogonale H.*

Propriété 2 (Représentation stochastique) *Si $Y \sim s.s.(\mu)$, alors Y peut s'écrire :*

$$Y = \| Y - \mu \| \cdot \frac{Y - \mu}{\| Y - \mu \|} + \mu = RU + \mu,$$

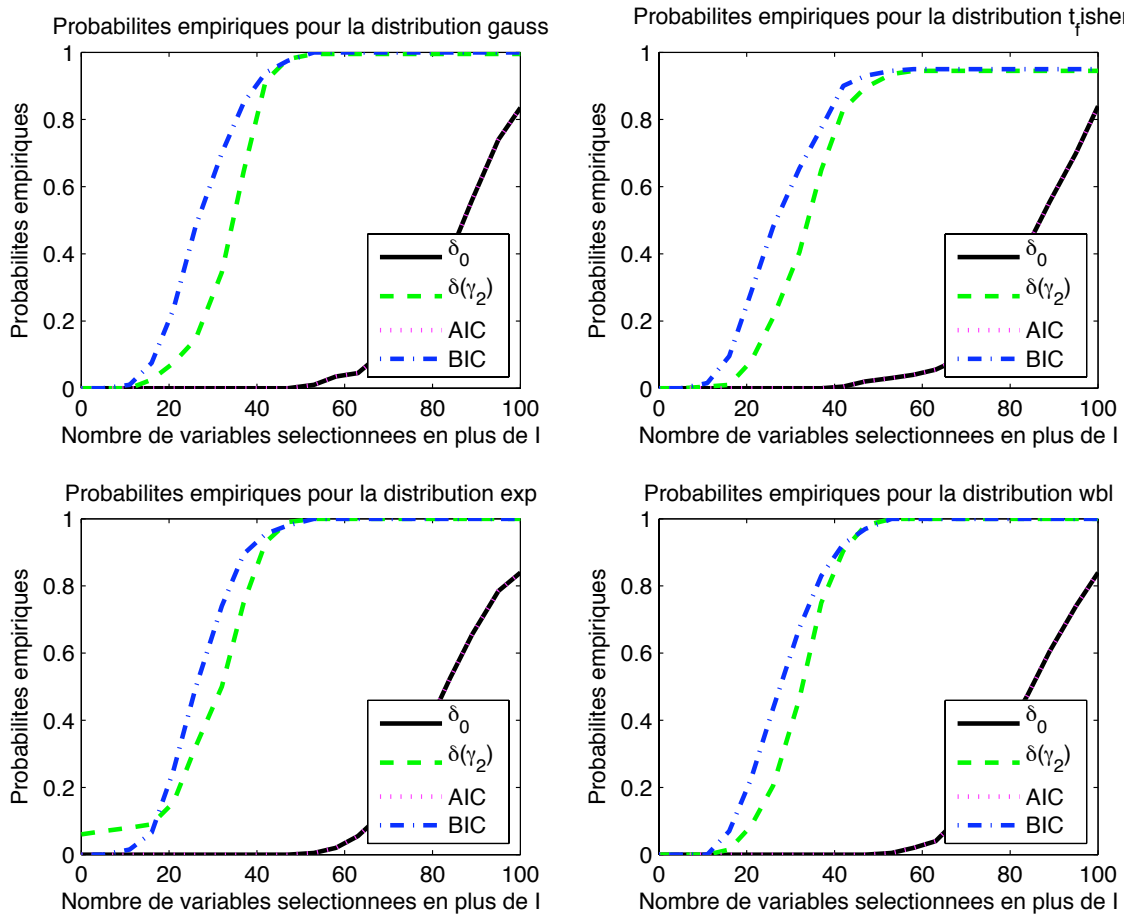

FIG. 7 – Probabilités empiriques avec $\delta(\gamma 2)$ tronqué.

où le rayon $R \in \mathbb{R}_+$ suit une distribution $h(R)$, et la direction U suit la distribution uniforme $\mathcal{U}_1$ sur la sphère unité et est telle que

$$\begin{array}{ll} \mathbb{E}[U] = 0 & \mathbb{E}[U_i^2] = 1/n \quad \forall i = 1, \ldots, n \\ \sum_{i=1}^n U_i^2 = 1 & \mathbb{E}[U_i U_j] = 0 \quad \forall j \neq i \end{array}$$

De plus, R et U sont indépendants.

Conséquence :

$$\begin{array}{lll} \mathbb{E}[Y] < \infty & \Leftrightarrow & \mathbb{E}[R] < \infty \quad (\mathbb{E}[Y] = \mu), \\ \mathrm{Cov}[Y] < \infty & \Leftrightarrow & \mathbb{E}[R^2] < \infty \quad (\mathrm{Cov}[Y] = \mathbb{E}[R^2] I_n/n). \end{array}$$

Définition 3 (Distribution uniforme) *La distribution uniforme sur la sphère $S_{R,\mu}$ de rayon R et de centre μ, $S_{R,\mu} = \{x \in \mathbb{R}^n \setminus \| x - \mu \| = R\}$, est définie de la manière suivante :*

$$\mathcal{U}_{R,\mu}(\Omega) = \frac{\sigma_{R,\mu}(\Omega)}{\sigma_{R,\mu}(S_{R,\mu})} = \frac{\sigma_{R,\mu}(\Omega)}{\sigma_{R,\mu}(S_{1,\mu}) R^{n-1}} = \mathcal{U}_1\Big(\frac{\Omega - \mu}{R}\Big) \tag{A.1}$$

pour tout ensemble de Borel Ω.

$\sigma_R(\Omega)$ est la mesure uniforme par rapport à la mesure de Lebesgue λ sur $\mathbb{R}^n$:

$$\sigma_R(\Omega) = \frac{n}{R} \lambda(\{ru \in \mathbb{R}^n \setminus 0 < r < R, u \in \Omega\}). \tag{A.2}$$

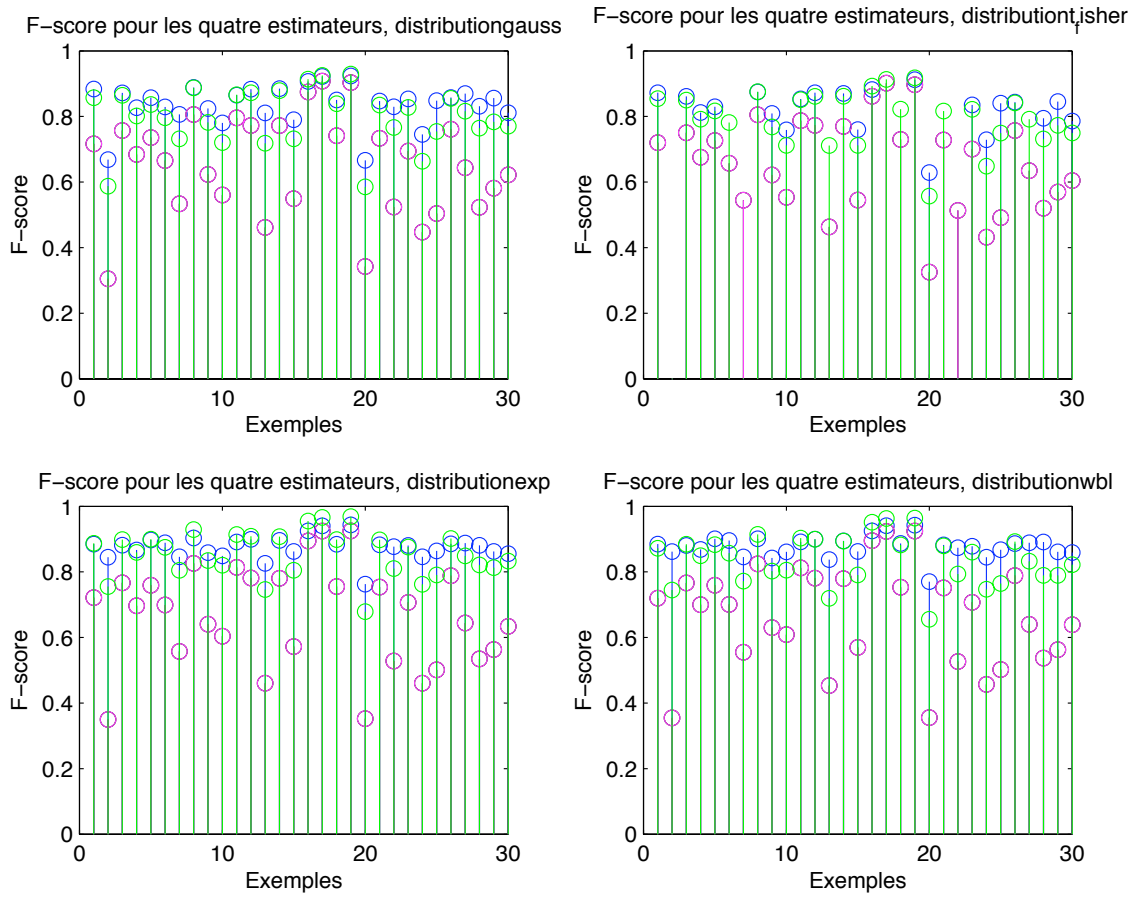

FIG. 8 – F-scores sur 30 exemples

De la propriété 2 et de la définition 3 se dégage la propriété suivante concernant l'intégration.

Propriété 3 (Intégration) *Pour toute fonction h intégrable, on a :*

$$\int_{\mathbb{R}^n} h(x)dx = \int_0^\infty \int_{S_R} h(x)d\sigma_R(x)dR \tag{A.3}$$

Le théorème suivant permet de définir la loi du rayon à partir de la loi de Y et inversement.

Théorème 1 *Soit $Y \in \mathbb{R}^n$ admettant une distribution à symétrie sphérique autour de $\mu \in \mathbb{R}^n$. Les propositions suivantes sont équivalentes :*

1. *Y admet une densité f par rapport à la mesure de Lebesgue dans $\mathbb{R}^n$.*
2. *$\| Y - \mu \|$ admet une densité h par rapport à la mesure de Lebesgue dans $\mathbb{R}_+$.*

De plus, si 1. et 2. sont vérifiées, il existe une fonction g de $\mathbb{R}_+$ dans $\mathbb{R}_+$ telle que :

$$f(y) = g(\| y - \mu \|^2) \qquad p.p. \tag{A.4}$$

$$h(r) = \frac{2\pi^{n/2}}{\Gamma(n/2)} r^{n-1} g(r^2) \qquad p.p. \tag{A.5}$$

Exemple : soit $Y \sim \mathcal{N}(\mu, 1)$, on a :

$$f(y) = \frac{1}{\sqrt{2\pi}} e^{-\|y-\mu\|^2/2} \quad \Rightarrow \quad h(r) = \frac{\sqrt{2}\pi^{(n-1)/2}}{\Gamma(n/2)} r^{n-1} e^{-r^2/2},$$

ce qui donne pour le rayon, en effectuant le changement de variable $t = r^2$, une loi $\chi^2(n)$.

A.2 Loi uniforme et coordonnées sphériques

$U \sim \mathcal{U}_R$ peut être transformé en coordonnées sphériques par le système d'équations φ_R : $(\rho_1, \ldots, \rho_{n-1}) \in V =]0, \pi[^{n-2} \times]0, 2\pi[\mapsto (u_1, \ldots, u_n)$ suivant :

$$\begin{aligned} u_1 &= R \sin\rho_1 \sin\rho_2 \ldots \sin\rho_{n-2} \sin\rho_{n-1} \\ u_2 &= R \sin\rho_1 \sin\rho_2 \ldots \sin\rho_{n-2} \cos\rho_{n-1} \\ u_3 &= R \sin\rho_1 \sin\rho_2 \ldots \cos\rho_{n-2} \\ &\vdots \\ u_{n-1} &= R \cos\rho_1 \end{aligned}$$

où $\rho_i \sim p_{n-i-1}(\rho_i)$, $i = 1, \ldots, n-2$, et $\rho_{n-1} \sim \mathcal{U}_{]0,2\pi[}$, avec

$$p_j(\rho) = \frac{j(j-2)(j-4)\ldots 1}{(j-1)(j-3)\ldots 1 \times 2^{j[2]}\pi^{1-j[2]}} \sin^j \rho.$$

Preuve :

Par changement de variables en coordonnées sphériques, l'intégrale de la distribution de Y devient :

$$\begin{aligned} 1 = \int_{\mathbb{R}^n} f(x)dx &= \int_0^\infty h(R)dR \int_{S_R} d\mathcal{U}_R(x) \\ &= \int_0^\infty h(R)dR \int_{S_R} \frac{1}{\sigma_1(S_1)R^{n-1}} d\sigma_R(x) \\ &= \int_0^\infty \sigma_1(S_1)R^{n-1}g(R^2)dR \times \\ &\quad \int_V \frac{1}{\sigma_1(S_1)R^{n-1}} \sin^{n-2}\rho_1 \ldots \sin\rho_{n-2} d\rho_1 \ldots d\rho_{n-1} \end{aligned}$$

Les ρ_j étant indépendants entre eux et avec le rayon, on peut séparer l'intégrale en :

$$\begin{aligned} \int_{\mathbb{R}^n} f(x)dx &= \int_0^\infty \sigma_1(S_1)R^{n-1}g(R^2)dR \times \\ &\quad \frac{1}{R^{n-1}} \int_0^\pi C_1 \sin^{n-2}\rho_1 d\rho_1 \ldots \int_0^\pi C_{n-2} \sin\rho_{n-2} d\rho_{n-2} \int_0^{2\pi} C_{n-1} d\rho_{n-1} \end{aligned}$$

où les C_j sont des constantes de normalisation telles que $\prod_{j=1}^{n-1} C_j = 1/\sigma_1(S_1)$.

Or,

$$\begin{aligned} I_n &= \int \sin^n(x)dx = -\frac{1}{n}\sin^{n-1}(x)\cos(x) + \frac{n-1}{n} I_{n-2} \\ &= -\frac{1}{n}\sin^{n-1}(x)\cos(x) + \frac{n-1}{n}\left(-\frac{1}{n-2}\sin^{n-3}(x)\cos(x) + \frac{n-3}{n-2}\right) \end{aligned}$$

$$\begin{aligned} &\ldots \\ &= -\cos(x)\sum_{i=1}^{[n/2]} K_i \sin^{n-2i}(x) + \frac{(n-1)(n-3)\ldots 1}{n(n-2)\ldots 1}\left(-\cos(x)\mathbf{1}_{\{n=2k\}} + x\mathbf{1}_{\{n=2k+1\}}\right) \end{aligned}$$

$$\Rightarrow \qquad I_n(]0,\pi[) = \frac{(n-1)(n-3)\ldots 1}{n(n-2)\ldots 1}\left(2 \times \mathbf{1}_{\{n=2k\}} + \pi \times \mathbf{1}_{\{n=2k+1\}}\right)$$

On obtient alors pour $j = 1, \dots, n-2$:

$$\int_0^\pi p_j(\rho)d\rho = \int_0^\pi C_j \sin^j(\rho)d\rho = 1 \qquad \Leftrightarrow \qquad C_j = \frac{j(j-2)(j-4)\dots 1}{(j-1)(j-3)\dots 1 \times 2^{j[2]}\pi^{1-j[2]}}$$
$$\int_{-\pi}^\pi p_{n-1}(\rho)d\rho = \int_{-\pi}^\pi C_{n-1}d\rho = 1 \qquad \Leftrightarrow \qquad C_{n-1} = \frac{1}{2\pi}$$

□

Le changement de variable en coordonnées sphériques nous permet de générer des lois à symétrie sphérique simplement, en générant d'un côté une direction U uniforme à partir des $n-1$ directions sur chaque dimension, puis en générant le rayon R selon la loi que l'on cherche à obtenir.

Références

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. *Budapest, Hungary*, pages 267–281, 1973.

[2] D.L. Donoho and J.M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3) :425, 1994.

[3] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456) :1348–1360, 2001.

[4] D. Fourdrinier. Loss estimation ot the lar estimator. Technical report, Université de Rouen, January 2010.

[5] D. Fourdrinier, W.E. Strawderman, and M.T. Wells. Improved estimation for spherically symmetric distributions. 2009.

[6] D. Fourdrinier and MT Wells. Comparaisons de procedures de selection d'un modele de regression : une approche decisionnelle. *Comptes rendus de l'Académie des sciences. Série 1, Mathématique*, 319(8) :865–870, 1994.

[7] I.E. Frank and J.H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2) :109–135, 1993.

[8] G.H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2) :215–223, 1979.

[9] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. *The elements of statistical learning : data mining, inference and prediction*, volume 27. Springer, 2005.

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

[11] C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 197–206, 1955.

[12] C.M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6) :1135–1151, 1981.

[13] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 111–147, 1974.

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

Chapitre 2

Improved generalized Bayes estimators of loss

Improved generalized Bayes estimators of loss

Dominique FOURDRINIER *
LITIS, EA 4108, Université de Rouen, France
and
Ali RIGHI †
LMRS, UMR 6085, Université de Rouen, France

August 28, 2010

Abstract

Let X be a random p-dimensional vector having a normal distribution with unknown mean θ and identity covariance matrix. As an estimator of θ, the observable X is itself a reference estimator in so far as it is the MLE, it is UMVUE and minimax under quadratic loss. In this paper, we consider estimators $\delta(X)$ of its quadratic loss $\|x-\theta\|^2$ under the new quadratic loss $\left(\delta(x)-\|x-\theta\|^2\right)^2$. More specifically, we envisage a Bayesian approach of this loss estimation problem and give sufficient conditions on the prior π and the marginal m under which the generalized Bayes estimator δ_π dominates the standard estimator $\delta_0(X)=p$, when $p \geq 5$. Examples illustrate the theory.

AMS 2000 subject classifications. Primary 62C15, 62C20, 62F10, 62H12.

Keywords and phrases: loss estimation, risk function, quadratic loss, generalized Bayes estimators, pseudo-Bayes estimators, spherically symmetric priors.

*Avenue de l'Université, BP 12, 76801 Saint-Étienne-du-Rouvray, France. The support of the ANR grant 08-EMER-002 is gratefully acknowledged.

†Avenue de l'Université, BP 12, 76801 Saint-Étienne-du-Rouvray, France

1 Introduction

Let $X \sim \mathcal{N}_p(\theta, I_p)$, that is, a p-dimensional Gaussian vector with unknown mean θ and identity covariance matrix I_p. As an estimator of θ, it is well known since Stein's findings [Ste56] that, under quadratic loss, X is inadmissible under quadratic loss and can be improved by alternative estimators when $p \geq 3$ (see [Ste81] for a general method which yields a wide class of improved estimators). However, it is still of interest to use X for simplicity reasons; note that it is the MLE, it is UMVU and it is minimax.

Thus Johnstone [Joh88], using X as an estimator of θ, considered the problem of estimating, for any observation x from X, its quadratic loss $\|x-\theta\|^2$. To this end, he envisaged to use real valued statistics δ as estimators of $\|x-\theta\|^2$ and evaluated them under the new quadratic loss

$$\left(\delta(X) - \|x-\theta\|^2\right)^2 . \tag{1.1}$$

He showed that the unbiased and constant estimator δ_0, that is

$$\delta_0(X) = p\,, \tag{1.2}$$

is admissible under the quadratic risk $E_\theta[(\delta(X) - \|X-\theta\|^2)^2]$ associated to (1.1) for $p \leq 4$. When $p \geq 5$, he gave as a general sufficient condition for a competitive estimator δ, written as

$$\delta(X) = p - \gamma(X) \tag{1.3}$$

for a certain function γ, to dominate δ_0, the differential inequality

$$2\Delta\gamma(x) + \gamma^2(x) < 0 \tag{1.4}$$

for all $x \in \mathbb{R}^p$, where $\Delta\gamma(x)$ denotes the Laplacian of γ at x. As a basic example, he proposed the function γ defined, for any $x \neq 0$, by $\gamma(x) = c/\|x\|^2$ where c is a positive constant and showed that, for $0 < c < 4(p-4)$, Inequality (1.4) is satisfied.

It is clear that, for regularity reason, Johnstone's estimators δ of the form $\delta(X) = p - c/\|X\|^2$ are not (generalized) Bayes estimators. As a Bayes estimate is a minimizer of the posterior expected loss, it is hence of interest to search if there exist priors for which the corresponding Bayes estimator improve on δ_0. Note also that such findings would be a first step in the quest of loss estimators which would be both dominating δ_0 and admissible.

Here, our goal is to find priors π on θ for which domination, under (1.1), of the corresponding Bayesian estimator δ_π over δ_0 is guaranteed. Note that such a Bayes estimator is expressed through the marginal m associated to π since it is of the form (1.3) with $\gamma(X) = -\Delta m(X)/m(X)$. Indeed, as, for any $x \in \mathbb{R}^p$, we have

$$m(x) = \int_{\mathbb{R}^p} \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{\|x-\theta\|^2}{2}\right) d\pi(\theta)\,,$$

it follows that the gradient of $m(x)$ is

$$\nabla m(x) = \int_{\mathbb{R}^p} -(x-\theta)\frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{\|x-\theta\|^2}{2}\right) d\pi(\theta).$$

and then that its Laplacian equals

$$\Delta m(x) = \int_{\mathbb{R}^p} (\|x-\theta\|^2 - p)\frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{\|x-\theta\|^2}{2}\right) d\pi(\theta).$$

Therefore it is clear that the Bayesian estimator δ_π expressed, for any $x \in \mathbb{R}^p$, as

$$\delta_\pi(x) = \frac{1}{m(x)} \int_{\mathbb{R}^p} \|x-\theta\|^2 \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{\|x-\theta\|^2}{2}\right) d\pi(\theta)$$

depends on the prior π through its marginal m and equals

$$\delta_\pi(x) = \delta_0(x) + \frac{\Delta m(x)}{m(x)} = p + \frac{\Delta m(x)}{m(x)}\,. \tag{1.5}$$

With a light abuse of notation, we will also use the notation δ_m instead of δ_π.

From the Bayesian perspective adopted here, it could be argued that, if the prior information is given by π, it could be more coherent to exploit it to estimate θ as well (what was done in Fourdriner and Strawderman [FS03]). Here X follows from a preliminary choice (as mentioned above), and hence, it is used without reference to π.

In Section 2, we give an expression of the unbiased estimator of the risk difference between δ_m and δ_0. We first study the case where the marginal m is, in fact, a pseudo marginal, that is a function m which may or may not be a marginal corresponding to some prior π. This approach allows us to guess what suitable priors should be. Thus, for certain constants $a \geq 0$ and $b > 0$, the improvement on δ_0 of estimators corresponding to the pseudo-marginals $m(x) = (\|x\|^2/2 + a)^{-b}$ (considered in [FS03]) allows to guess that domination over δ_0 of generalized Bayes estimators is thinkable thanks to priors of this form. In fact, we first give a direct proof of such an improvement through the prior $\pi(\theta) = 1/\|\theta\|^2$, but for large dimension (p has to be greater than or equal to 20). Next, in our main result, we are able to give general conditions on the prior π for which the corresponding generalized Bayes estimator dominates δ_0. As an interesting fact, these conditions are satisfied by the prior $\pi(\theta) = 1/\|\theta\|^2$ for smaller dimension; p just needs to be less than or equal to 12. More generally, considerations on the value of the constant b for priors of the form $\pi(\theta) = \|\theta\|^{-2b}$ allow to reach lower dimensions for p where the domination over δ_0 occurs.

Section 3 is devoted to examples which illustrate the theory. In Section 4, we give some concluding remarks and some perspectives. Finally, Section 5 is an appendix which contains technical results used in the proofs of our findings.

2 A class of improved generalized Bayes estimators

As Johnstone proved that, for $p \leq 4$, δ_0 is an admissible loss estimator, in the following, we assume that $p \geq 5$.

2.1 Pseudo Bayes estimators

In this section, we provide an unbiased estimator of the risk difference between an estimator δ_m of the form (1.5) and δ_0. Note that, following the Bock's approach [Boc88], the function m in (1.5) does not need to be a real marginal corresponding to a prior, which allows to enlarge the class of Bayes estimators to the pseudo-Bayes estimators. However, it should be noticed that, unlike the Bayes context where the marginal m is a smooth function (as a multiple of the Laplace transform of a function), a regularity condition is needed on m; following Johnstone [Joh88] we will assume that the correction factor of δ_0 in (1.5), that is $\Delta m/m$, is twice weakly differentiable (and hence, we will assume that m is four times weakly differentiable).

Lemma 2.1 *For any (pseudo) Bayes estimator δ_m of the form (1.5) such that $\Delta m/m$ is four times weakly differentiable, an unbiased estimator of the risk difference between δ_m and δ_0 is*

$$\zeta(X) = 3\left(\frac{\Delta m(X)}{m(X)}\right)^2 - 2\,\frac{\Delta^{(2)} m(X)}{m(X)} + 4\,\frac{\nabla m(X)}{m(X)} \cdot \nabla\left[\frac{\Delta m(X)}{m(X)}\right] \tag{2.1}$$

provided that

$$E_\theta\left[\left(\frac{\Delta m(X)}{m(X)}\right)^2\right] < \infty\,. \tag{2.2}$$

(Here the notation $\cdot$ holds for the inner product in $\mathbb{R}^p$).

PROOF For any twice weakly differentiable function γ from $\mathbb{R}^p$ onto $\mathbb{R}$, it is proved in [Joh88] that an unbiased estimator of the risk difference between $p - \gamma(X)$ and p is $2\,\Delta\gamma(X) + \gamma^2(X)$, provided that the finiteness risk condition $E_\theta[\gamma^2] < \infty$ for $p - \gamma(X)$ is satisfied. In other words, we have

$$E_\theta[(p - \gamma(X) - ||X-\theta||^2)^2] - E_\theta\left[(p - ||X-\theta||^2)^2\right] = E_\theta[2\,\Delta\gamma(X) + \gamma^2(X)]\,.$$

It follows that Expression (2.1) is

$$2\,\Delta\gamma(X) + \gamma^2(X) \quad \text{with} \quad \gamma(X) = -\frac{\Delta m(X)}{m(X)}\,. \tag{2.3}$$

Hence it will be derived through, for any $x \in \mathbb{R}^p$, the expression of the Laplacian

$$\Delta\left(\frac{\Delta m(x)}{m(x)}\right) = \frac{\Delta^{(2)} m(x)}{m(x)} + \Delta m(x)\,\Delta\left(\frac{1}{m(x)}\right) + 2\,\nabla(\Delta m(x)) \cdot \nabla\left(\frac{1}{m(x)}\right). \quad (2.4)$$

Now routine calculations yield

$$\begin{aligned} \Delta m(x)\,\Delta\left(\frac{1}{m(x)}\right) &= \Delta m(x)\,\mathrm{div}\left(\frac{-\nabla m(x)}{m^2(x)}\right) \\ &= -\left(\frac{\Delta m(x)}{m(x)}\right)^2 + 2\,\frac{\Delta m(x)}{m(x)}\left\|\frac{\nabla m(x)}{m(x)}\right\|^2 \end{aligned} \quad (2.5)$$

and

$$\begin{aligned} \nabla(\Delta m(x))\,.\,\nabla\left(\frac{1}{m(x)}\right) &= -\nabla\left(\frac{\Delta m(x)}{m(x)}\,m(x)\right).\nabla\left(\frac{-\nabla m(x)}{m^2(x)}\right) \\ &= -\nabla\left(\frac{\Delta m(x)}{m(x)}\right).\frac{\nabla m(x)}{m(x)} - \frac{\Delta m(x)}{m(x)}\left\|\frac{\nabla m(x)}{m(x)}\right\|^2. \end{aligned} \quad (2.6)$$

Therefore, setting the expressions (2.5) and (2.6) in (2.4), we obtain

$$\Delta\left(\frac{\Delta m(x)}{m(x)}\right) = \frac{\Delta^{(2)} m(x)}{m(x)} - \left(\frac{\Delta m(x)}{m(x)}\right)^2 - 2\,\nabla\left(\frac{\Delta m(x)}{m(x)}\right).\frac{\nabla m(x)}{m(x)}. \quad (2.7)$$

Finally, it is clear that (2.7) and (2.3) give rise to the expression (2.1). □

Although we are mainly interested in (generalized) Bayes estimators, the complexity of Formula (2.1) leads us to first illustrate it through pseudo-Bayes estimators, that is, to estimators which do not necessarily correspond to a prior (possibly improper). Our first example is the class, considered by Fourdrinier and Strawderman [FS03], of estimators $\delta_{a,b}$ corresponding to pseudo-marginals $m(x)$ of the form

$$m(x) = \left(\frac{1}{\frac{\|x\|^2}{2} + a}\right)^b, \quad (2.8)$$

where a and b are non negative constants.

Setting $y = \|x\|^2/2$ and $m(x) = f(y) = (1/(y+a))^b$, straightforward but tedious calculations (checked with MAPLE®) allow to express the unbiased estimator of the risk difference in (2.1) as

$$\eta(y) = \frac{-b}{(y+a)^2}\left\{A\left(\frac{y}{y+a}\right)^2 + B\left(\frac{y}{y+a}\right) + C\right\} \quad (2.9)$$

where

$$A = -4\,(b-3)\,(b+1)\,(b+4), \quad B = 4\left[p\,b^2 - (p+8)\,b - 4\,(p+2)\right]$$

and

$$C = p\left[-(p-4)\,b + 2\,(p+2)\right]\ .$$

It is worth noting that, as

$$E_\theta\left[\left(\frac{\Delta m(X)}{m(X)}\right)^2\right] \propto E_\theta\left[\frac{1}{||X||^4}\right]\ ,$$

which shows that the moment condition in (2.2) corresponds to the dimension condition $p \geq 5$.

We give hereafter conditions which guarantee that $\eta(y) \leq 0$ for all y and hence that δ_m dominates δ_0 (the proof is postponed to the appendix). In the case where $a = 0$, a necessary and sufficient condition for $\eta(y)$ to be non positive is, either $b \leq (p-2)/2$ when $5 \leq p \leq 15$, or, when $p \geq 16$, $b \leq (p-2-\sqrt{\Delta})/4$ or $(p-2+\sqrt{\Delta})/4 \leq b \leq (p-2)/2$, where $\Delta = p^2 - 20\,p + 68$. In the case where $a > 0$, a simple sufficient condition for $\eta(y)$ to be non positive is $b \leq 1$. Note that, with $m(x) \propto 1/\|x\|^2$, we recover Johnstone's estimator $p - 2\,(p-4)/\|X\|^2$ (see [Joh88]).

2.2 Improved generalized Bayes estimators of loss

In this section, we turn our attention to generalized Bayes estimators of loss. We first give, in Proposition 2.1 below a preliminary example of a generalized Bayes estimator δ_m corresponding to an improper prior $\pi(\theta)$ which shows that domination of δ_m over δ_0 can be easily demonstrate for priors of the form of the pseudo-marginal evoked in Subsection 2.1, but at the cost of a large dimension p. Next, we yield general conditions on the prior $\pi(\theta)$ to guarantee such a domination.

Proposition 2.1 *For the prior $\pi(\theta) = 1/\|\theta\|^2$, the corresponding generalized Bayes estimator δ_m dominates δ_0 provided that $p \geq 20$.*

PROOF We will need the following differential expressions related to the prior $\pi(\theta) = 1/\|\theta\|^2$:

$$\frac{\nabla\pi(\theta)}{\pi(\theta)} = -\frac{2}{\|\theta\|^2}\,\theta\ ,\ \frac{\Delta\pi(\theta)}{\pi(\theta)} = -\frac{2\,(p-4)}{\|\theta\|^2}\ ,\ \frac{\nabla(\Delta\pi(\theta))}{\pi(\theta)} = \frac{8(p-4)}{\|\theta\|^4}\,\theta \tag{2.10}$$

and

$$\frac{\Delta^{(2)}\pi(\theta)}{\pi(\theta)} = \frac{8\,(p-4)\,(p-6)}{\|\theta\|^4}.$$

First, we will check that the finitness risk condition (2.2) is satisfied for the marginal density associated to that prior. Using Corollary 5.1 the term $\Delta m(x)/m(x)$ in (2.2) can be rewritten through E^x, the expectation with respect to the posterior expectation given $X = x$, so that

$$\left(\frac{\Delta m(x)}{m(x)}\right)^2 = \left(E^x\left[\frac{\Delta\pi(\theta)}{\pi(\theta)}\right]\right)^2 \leq E^x\left[\left(\frac{\Delta\pi(\theta)}{\pi(\theta)}\right)^2\right] \tag{2.11}$$

by the Jensen inequality. Then it follows from (2.10) and (2.11) that $(\Delta m(x)/m(x))^2$ is bounded from above by a quantity proportional to

$$E^x\left[\frac{1}{\|\theta\|^4}\right] = \frac{E_x\left[\frac{1}{\|\theta\|^6}\right]}{E_x\left[\frac{1}{\|\theta\|^2}\right]} \tag{2.12}$$

by definition of E^x (here E_x denotes the expectation with respect to the normal distribution $\mathcal{N}(x, I_p)$). Now as, for any $b > 0$, we have

$$\frac{1}{\|\theta\|^{2b}} = \frac{1}{\Gamma(b)\,2^b}\int_0^\infty u^{b-1}\,\exp\left(-\frac{\|\theta\|^2}{2}\,u\right)du \tag{2.13}$$

it follows, through routine calculations, that

$$E_x\left[\frac{1}{\|\theta\|^{2b}}\right] = \frac{1}{\Gamma(b)\,2^b}\int_0^1 t^{b-1}\,(1-t)^{p/2-b-1}\,\exp\left(-\frac{\|x\|^2}{2}\,t\right)dt\,. \tag{2.14}$$

Thus the expectation in (2.14) appears as being proportional to the Laplace transform at $s = \|x\|^2/2$ of the beta distribution function $\alpha = \mathcal{B}e(b, p/2-b)$, that is,

$$E_x\left[\frac{1}{\|\theta\|^{2b}}\right] = \frac{\Gamma(p/2-b)}{\Gamma(p/2)\,2^b} f_b(\|x\|^2/2) \tag{2.15}$$

where

$$f_b(s) = \int_0^\infty e^{-ts}\,d\alpha(t)\,.$$

Its behavior for t in a neighbourhood of ∞ can be derived from an Abelian theorem. Indeed, according to Corollary 1.a page 182 of Widder [Wid46], as

$$\alpha(t) \sim \frac{\Gamma(p/2)}{\Gamma(b+1)\,\Gamma(p/2-b)}\,t^b \qquad (t \to 0_+)$$

it follows that

$$f_b(s) \sim \frac{\Gamma(p/2)}{\Gamma(p/2-b)}\,s^b \qquad (s \to \infty)\,.$$

Therefore we have

$$\frac{f_3(s)}{f_1(s)} \sim \frac{\Gamma(p/2-1)}{\Gamma(p/2-3)}\,\frac{1}{s^2} \qquad (s \to \infty)\,. \tag{2.16}$$

Finally, gathering (2.12), (2.14) and (2.16) shows that $(\Delta m(x)/m(x))^2$ is bounded from above by a quantity which goes to 0 when $\|x\|$ goes to ∞. It follows that Condition (2.2) is satisfied.

To get the desired domination result, we will use the unbiased estimator of the risk difference in Lemma 2.1 and prove that, for any $x \in \mathbb{R}^p$, $\zeta(x) \leq 0$. First, we will rely on the fact that

$$\frac{\nabla m(x)}{m(x)} \cdot \nabla \left[\frac{\Delta m(x)}{m(x)} \right] = -\frac{\Delta m(x)}{m(x)} \left\| \frac{\nabla m(x)}{m(x)} \right\|^2 + \frac{\nabla(\Delta m(x))}{m(x)} \cdot \frac{\nabla(m(x))}{m(x)}$$

to reexpress $\zeta(x)$ in (2.1) as

$$\zeta(x) = A(x) + B(x) \tag{2.17}$$

where

$$A(x) = 3 \left(\frac{\Delta m(x)}{m(x)} \right)^2 - 2 \, \frac{\Delta^{(2)} m(x)}{m(x)} - 4 \, \frac{\Delta m(x)}{m(x)} \left\| \frac{\nabla m(x)}{m(x)} \right\|^2 \tag{2.18}$$

and

$$B(x) = 4 \, \frac{\nabla(\Delta m(x))}{m(x)} \cdot \frac{\nabla(m(x))}{m(x)} \, . \tag{2.19}$$

Interpreting, as above through Corollary 5.1, the three first terms of (2.18) through E^x, we have

$$\left(E^x \left[\frac{\Delta \pi(\theta)}{\pi(\theta)} \right] \right)^2 \leq E^x \left[\left(\frac{\Delta \pi(\theta)}{\pi(\theta)} \right)^2 \right] \tag{2.20}$$

by Jensen's inequality. Likewise

$$\left\| E^x \left[\frac{\nabla \pi(\theta)}{\pi(\theta)} \right] \right\|^2 \leq E^x \left[\left\| \frac{\nabla \pi(\theta)}{\pi(\theta)} \right\|^2 \right]$$

which implies

$$E^x \left[\frac{\Delta \pi(\theta)}{\pi(\theta)} \right] \left\| E^x \left[\frac{\nabla \pi(\theta)}{\pi(\theta)} \right] \right\|^2 \geq E^x \left[\frac{\Delta \pi(\theta)}{\pi(\theta)} \right] \, E^x \left[\left\| \frac{\nabla \pi(\theta)}{\pi(\theta)} \right\|^2 \right] \tag{2.21}$$

since, according to the expression of $\Delta \pi(\theta)$ below, the prior density π is superharmonic for $p \geq 5$. Then $A(x)$ in (2.18) is bounded above by

$$A(x) \leq 3 \, E^x \left[\left(\frac{\Delta \pi(\theta)}{\pi(\theta)} \right)^2 \right] - 2 \, E^x \left[\frac{\Delta^{(2)} \pi(\theta)}{\pi(\theta)} \right] - 4 \, E^x \left[\frac{\Delta \pi(\theta)}{\pi(\theta)} \right] \, E^x \left[\left\| \frac{\nabla \pi(\theta)}{\pi(\theta)} \right\|^2 \right] . \tag{2.22}$$

Now, using the expressions in (2.10) we see that (2.22) becomes

$$\begin{aligned} A(x) &\leq [12\,(p-4)^2 - 16\,(p-4)\,(p-6)]\,E^x\left[\frac{1}{\|\theta\|^4}\right] + 32\,(p-4)\left(E^x\left[\frac{1}{\|\theta\|^2}\right]\right)^2 \\ &\leq (p-4)\,[-4\,(p-4)+32] + 32\,(p-4)\,E^x\left[\frac{1}{\|\theta\|^4}\right] \\ &= -(p-4)\,(p-20)\,E^x\left[\frac{1}{\|\theta\|^4}\right], \end{aligned} \tag{2.23}$$

the second inequality following from Jensen's inequality. Consequently $A(x) \leq 0$ for $p \geq 20$.

We turn now our attention to the term $B(x)$ in (2.19); it equals

$$B(x) = 4\,E^x\left[\frac{\nabla(\Delta\pi(\theta))}{\pi(\theta)}\right] \cdot E^x\left[\frac{\nabla\pi(\theta)}{\pi(\theta)}\right] = -64\,(p-4)\,E^x\left[\frac{\theta}{\|\theta\|^4}\right] \cdot E^x\left[\frac{\theta}{\|\theta\|^2}\right] \tag{2.24}$$

according to (2.10). Note that, Lemma 5.1 applies to the two last expectations, with $f(t) \propto \exp(-t/2)$, which is non increasing, and $g(u) = 1/u^2$ and $g(u) = 1/u$ respectively. Thus there exist two non negative functions $\Gamma_1(x)$ and $\Gamma_2(x)$ such that

$$E_x\left[\frac{\theta}{\|\theta\|^2}\right] = \Gamma_1(x)\,x \quad \text{and} \quad E_x\left[\frac{\theta}{\|\theta\|^4}\right] = \Gamma_2(x)\,x\,.$$

It is then clear that $B(x) \leq 0$ and hence that $\zeta(x)$ in (2.17) satisfies $\zeta(x) \leq 0$, which is the desired result. □

The method of proof used for Proposition 2.1 is not accurate enough and leads to a high lower bound beyond which the improvement of the generalized Bayes estimator over δ_0 is obtained: p has to be greater than or equal to 20. We give below sharper conditions for which such an improvement is guaranteed and which give rise to a class of generalized Bayes estimators with smaller lower bound of improvement for p. In the following, we assume that the priors $\pi(\theta)$ are spherically symmetric, that is, with an abuse of notation are of the form $\pi(\|\theta\|^2)$.

Theorem 2.1 *Let $\pi(\|\theta\|^2)$ be a spherically symmetric prior density which is four times weakly differentiable. Assume that, as a function of $t = \|\theta\|^2$, the following monotonicity conditions are satisfied:*

1. *$\pi(t)$ is superharmonic,*
2. *$\pi'(t)/\pi(t)$ is non decreasing,*
3. *$\Delta\pi(t)/\pi(t)$ is non decreasing*
4. *$3\,(\Delta\pi(t)/\pi(t))^2 - 2\,\Delta^{(2)}\pi(t)/\pi(t) \leq 0$ for all $t \in \mathbb{R}_+$.*

Then the generalized Bayes estimator δ_π associated to $\pi(\|\theta\|^2)$ dominates δ_0.

PROOF of THEOREM 2.1 Let $x \in \mathbb{R}^p$ fixed. The proof consits in showing that the unbiased estimator of the risk difference $\zeta(x)$ in Lemma 2.1 is non positive. First, note that Condition 4 of the theorem guarantees that the sum of the two first terms of the right hand side of (2.1) is non positive. Indeed, using Corollary 5.1, we have

$$\begin{aligned} 3\left(\frac{\Delta m(x)}{m(x)}\right)^2 - 2\,\frac{\Delta^{(2)} m(x)}{m(x)} &= 3\left[E^x\left(\frac{\Delta\pi(\|\theta\|^2)}{\pi(\|\theta\|^2)}\right)\right]^2 - 2\,E^x\left[\frac{\Delta^{(2)}\pi(\|\theta\|^2)}{\pi(\|\theta\|^2)}\right] \\ &\leq E^x\left[3\left(\frac{\Delta\pi(\|\theta\|^2)}{\pi(\|\theta\|^2)}\right)^2 - 2\,\frac{\Delta^{(2)}\pi(\|\theta\|^2)}{\pi(\|\theta\|^2)}\right] \\ &\leq 0 \end{aligned}$$

applying Jensen's inequality to obtain the second inequality. Hence it remains to show that the third term of the right hand side of (2.1) is non positive, which reduces to,

$$\nabla m(x) \cdot \nabla\left[\frac{\Delta m(x)}{m(x)}\right] \leq 0\,. \tag{2.25}$$

Since the prior $\pi(||\theta)||^2$ is spherically symmetric, it follows that the coresponding marginal $m(x)$ is also spherically symmetric, that is, depends on x only through $\|x\|^2$. With an abuse of notation, we denote $m(x) = m(\|x\|^2)$. Also $\Delta m(x)$ depends only on $\|x\|^2$ so that $\Delta m(x)/m(x) = F(\|x\|^2)$ for some function F.

Now, as $\nabla m(\|x\|^2) = 2\,m'(\|x\|^2)\,x$ and $\nabla F(\|x\|^2) = 2\,F'(\|x\|^2)\,x$, Inequality (2.25) is equivalent to $m'(\|x\|^2)\,F'(\|x\|^2) \leq 0$. It has been shown in [FS08] that, as soon as $\pi(t)$ is non increasing (which follows from Condition 1 by the mean value property of superharmonic functions), $m'(\|x\|^2) \leq 0$, so that it remains to prove that $F'(\|x\|^2) \geq 0$, that is, $F(t)$ is non decreasing in t.

Note that, according to Lemma 5.2,

$$F(t) = \frac{\int_0^\infty e^{-r^2/2}\,r^{p-2}\,\psi(t,r)\,dr}{\int_0^\infty e^{-r^2/2}\,r^{p-2}\,\varphi(t,r)\,dr} \tag{2.26}$$

where

$$\varphi(t,r) = \int_0^\infty G(t,\eta)\,\pi(\eta^2 + r^2)\,d\eta \tag{2.27}$$

and

$$\psi(t,r) = \int_0^\infty G(t,\eta)\,\Delta\pi(\eta^2 + r^2)\,d\eta \tag{2.28}$$

with

$$G(t,\eta) = 2\,e^{-t^2}e^{-\eta^2}\,\cosh(t\,\eta)\,. \tag{2.29}$$

Thus $F(t)$ in (2.26) appears as the expectation

$$F(t) = E_t\left[\frac{\psi(t,R)}{\varphi(t,R)}\right]$$

where R is a random variable with density

$$f_t : r \mapsto \frac{e^{-r^2/2}\, r^{p-2}\, \varphi(t,r)}{\int_0^\infty e^{-r^2/2}\, r^{p-2}\, \varphi(t,r)\, dr}\,.$$

Now let $0 \le t_1 < t_2$. By Lemma 5.3 (i), for any fixed r, the function $\psi(t,r)/\varphi(t,r)$ is non decreasing in t. Hence

$$F(t_1) = E_{t_1}\left[\frac{\psi(t_1,R)}{\varphi(t_1,R)}\right] \le E_{t_1}\left[\frac{\psi(t_2,R)}{\varphi(t_2,R)}\right]\,. \tag{2.30}$$

Also, according to Lemma 5.3 (ii), the function $\psi(t_2,r)/\varphi(t_2,r)$ is non decreasing in r and Lemma 5.3 (iii) implies that the family of densities $(f_t)_{t\in\mathbb{R}_+}$ has non decreasing monotone likelihood ratio in t. Therefore

$$E_{t_1}\left[\frac{\psi(t_2,R)}{\varphi(t_2,R)}\right] \le E_{t_2}\left[\frac{\psi(t_2,R)}{\varphi(t_2,R)}\right] = F(t_2)\,. \tag{2.31}$$

Therefore (2.30) and (2.31) give $F(t_1) \le F(t_2)$, which is the desired monotonicity of F.

Finally the theorem will be proved showing that Condition (2.2) is satisfied. To this end, note that, as the prior $\pi(\|\theta\|^2)$ is superharmonic by Condition 1, the marginal density m is superharmonic as well. Hence the function $F(\|x\|^2) = \Delta m(x)/m(x)$ is non positive and, as it is non decreasing in $\|x\|^2$, its square $(\Delta m(x)/m(x))^2$ is non increasing in $\|x\|^2$ and is bounded. Therefore Condition (2.2) follows and the theorem is proved. □

3 Examples

In this section, we consider the family of spherically symmetric priors

$$\pi_{a,b}(\|\theta\|^2) = \left(\frac{1}{\|\theta\|^2 + a}\right)^b \tag{3.1}$$

where a is a non negative number and b is a positive number. We will prove that the Bayes estimator of loss δ_π corresponding to the prior in (3.1) dominates δ_0 if

$$0 < b < 2 \quad \text{and} \quad p > 2\,\frac{(b-4)\,(b+1)}{b-2}\,. \tag{3.2}$$

Note that the last term in (3.2), that is, $f(b) = 2\,(b-4)\,(b+1)/(b-2)$ is a continuous function in b which strictly increases from 4 to ∞ when b varies in $]0, 2[$. Hence, for any $p \geq 5$ fixed, there exists $b_0 \in]0, 2[$ such that $p = f(b_0)$ and such that, for any $b \in]0, b_0]$, we have $p \geq f(b)$. More precisely, it can be shown through routine calculations that $b_0 = (p + 6 - \sqrt{p^2 - 4\,p + 100})/4$.

It is clear that, when $a > 0$, the prior $\pi_{a,b}(\|\theta\|^2)$ in (3.1) is infinitely differentiable and hence is four time weakly differentiable. However, when $a = 0$, it is four time weakly differentiable if and only if $b < p/2 - 2$ (see Lemma 5.4 and its comment afterwards), which is compatible with the above remark since it can be easily shown that $b_0 \leq p/2 - 2$ for $p \geq 5$. Note also that, setting $t = \|\theta\|^2$, the function $\pi(t) = 1/(t+a)^b$ is non increasing in t and $\pi'(t)/\pi(t) = -b/(t + a)$ is non decreasing in t so that Condition 2 of Theorem 2.1 is satisfied.

Now, using straightforward algebra we have

$$\frac{\Delta\pi(t)}{\pi(t)} = -\frac{2\,b}{t+a}\left(p - 2\,(b+1)\frac{t}{t+a}\right)$$

with derivative equal to

$$\frac{2\,b}{(t+a)^3}\,\Big(\big(p - 2\,(b-1)\big)\,t + a\,\big(p + 2\,(b+1)\big)\Big)$$

and non negative for $b \leq p/2 - 1$, so that Conditions 1 and 3 of Theorem 2.1 are satisfied. Note that, when $a = 0$, this last condition on b is automatically satisfied under the above weak differentiability condition.

We now turn our attention to Condition 4 of Theorem 2.1. Similar calculations as those in Subsection 2.1 allow us to write

$$\eta(t) = 3\left(\frac{\Delta\pi(t)}{\pi(t)}\right)^2 - 2\,\frac{\Delta^{(2)}\pi(t)}{\pi(t)}$$

as

$$\eta(t) = -\frac{4b}{(t+a)^2}\left\{A\left(\frac{t}{t+a}\right)^2 + B\left(\frac{t}{t+a}\right) + C\right\} \tag{3.3}$$

where

$$A = -4\,(b+1)(b^2 - 7\,b - 12), \quad B = 4\,(b+1)\left[(b-4)\,p - 4\,(b+2)\right]$$

and

$$C = -p\left[(b-2)p - 4\,(b+1)\right].$$

When $a = 0$, $\eta(t)$ in (3.3) reduces to $\eta(t) = Q(p)/t^2$ where

$$Q(p) = 4\,b\left[p - 2\,(b+1)\right]\left[(b-2)\,p - 2\,(b+1)(b-4)\right]. \tag{3.4}$$

Condition 4 of Theorem 2.1 is equivalent to $Q(p) \leq 0$. As the above weak differentiability condition can be written as $p > 2\,(b+2)$, we have $p-2\,(b+1) > 2$, so that the polynomial $Q(p)$ in (3.4) is of the sign of $R(p) = (b-2)\,p - 2\,(b+1)(b-4)$. Then it is clear that we cannot have $b = 2$ since this value imposes $p > 8$ and gives $R(p) = 6 > 0$. We cannot have $b > 2$ either since $R(p) \leq 0$ is equivalent to $p \leq 2\,(b+1)\,(b-4)/(b-2)$, which contradicts the above condition $p > 2\,(b+2)$ in so far as $b + 2 < (b+1)\,(b-4)/(b-2)$ leads to $b < 0$.

Finally we should consider the case where $0 < b < 2$ for which $Q(p) \leq 0$ is equivalent to $p \geq 2\,(b+1)\,(b-4)/(b-2)$. Clearly, this lower bound for p is greater than $2\,(b+2)$ and hence is compatible with the weak differentiability condition. Therefore Condition 4 of Theorem 2.1 is satisfied in this situation and, finally, the generalized Bayes estimator associated to the prior for $0 < b < 2$ dominates δ_0.

In the case where $a > 0$, setting $z = t/(t+a)$ in (3.3), we see that Condition 4 in Theorem 2.1 is equivalent to

$$Q(z) = Az^2 + Bz + C \geq 0 \tag{3.5}$$

for any $0 \leq z \leq 1$. Note that $A + B + C \geq 0$ is a necessary condition in order that Inequality (3.5) is satisfied. Actually we will prove that this is a sufficient condition showing that, for any $z \in [0,1]$, $Q'(z) \leq 0$, which implies that

$$Q(z) \geq Q(1) = A + B + C \geq 0\,.$$

Note that, clearly, $A > 0$ for $0 < b < 2$, so that, for $z \in [0,1]$, we have $Q'(z) = 2\,A\,z + B \leq 2\,A + B$ and it suffices to show that $2\,A + B \leq 0$. To this end, express

$$2\,A + B = -4\,(b+1)\left[(4-b)\,p + 2b^2 - 10\,b - 16\right]$$

so that $2A + B \leq 0$ if and only if $p \geq (-2b^2 + 10b + 16)/(4-b)$. Now that condition on p is entailed by Condition (3.2) since

$$2\,\frac{(b-4)\,(b+1)}{b-2} - \frac{-2b^2 + 10b + 16}{4-b} = \frac{12b}{(b-2)(b-4)} > 0\,.$$

Hence, under Condition (3.2), we have $2A + B \leq 0$, which is the desired result.

The above result expresses that, for the choice of a value of b between 0 and 2, the dimension p has to be large enough to obtain the desired improvement of the Bayes estimator: p should be such that $p > 2\,(b+2)$. Thus, when $b = 1$, this domination is guaranteed for $p \geq 12$, which shows an improvement of Proposition 2.1, where the domination result was only obtained $p \geq 20$, by Theorem 2.1. However, as Theorem 2.1 only needs that $p \geq 5$, we may wonder if such a result remains valid for lower dimensions. To clarify this point, with the help of MAPLE®, we have expended the expression of the unbiased estimator $\zeta(X)$ of the risk difference between δ_m and δ_0 given in 2.1 using the form of the prior in (2.13). We have checked that $\zeta(X) \leq 0$ when $6 \leq p \leq 11$.

4 Concluding remarks and perspectives

In this paper, for $X \sim \mathcal{N}_p(\theta, I_p)$, where θ is unknown, and for any observation x from X, we were interested in estimating the quadratic loss $\|x-\theta\|^2$. We found improvements, with respect to the loss $\big(\delta(X) - \|x-\theta\|^2\big)^2$, over the standard unbiased estimator $\delta_0(X) = p$ by generalized Bayes estimators δ_π associated to spherical priors $\pi(\|\theta\|^2)$. After providing an unbiased estimator of the risk difference between δ_π and δ_0 (also valid for pseudo-Bayes estimators), we gave sufficient conditions on the function $\pi(\cdot)$ which guarantee the domination δ_π over δ_0. The examples of priors $\pi(\|\theta\|^2) = (\|\theta\|^2 + a)^{-b}$ with $a \geq 0$ and $b > 0$ which illustrate our theory came from pseudo-marginals m with a similar form for which domination of δ_m over δ_0 occurs (under diferent conditions on the constants a and b). Our technique of proof relies on expressing the various differential quantities involving the marginal densities $m(x)$ in terms of expectations with respect to the a posteriori distribution given $X = x$.

Among the domination conditions on π given in Theorem 2.1, the most difficult to fulfil is Condition 4. Thus, as our examples of priors are all scale mixtures of normal densities, we may think to express that condition through the corresponding mixing distribution. It turns out that it is a difficult task so that a prospect would be to simplify this condition.

Here, the scale parameter is assumed to be known. In practice this is often non realistic. Hence it is of interest to consider the case where $X \sim \mathcal{N}_p(\theta, \sigma^2 I_p)$ with σ^2 unknown. More generally, it would be interesting to extend the distributional context. A natural extension is the class of the spherically symmetric distributions (see [FW95b] and [FW95a] where loss estimation is considered under spherical symmetry). We plan to investigate this issues in the near future.

5 Appendix

5.1 First, we prove that $\eta(y)$ in (2.9) is non positive under the conditions stated at the end of Subsection 2.1.

When $a = 0$, the expression of $\eta(y)$ in (2.9) reduces to

$$\eta(y) = \frac{-b}{(y+a)^2}(A + B + C) = \frac{-b}{(y+a)^2}(p - 2 - 2\,b)\,Q(b)$$

where

$$Q(b) = 2\,b^2 - (p-2)\,b + 2\,(p-4)\,,$$

according to the values of A, B and C. That polynomial of degree 2 in b has discriminant $\Delta(p) = p^2 - 20p + 68$ which is non positive if $10 - 4\sqrt{2} \leq p \leq 10 + 4\sqrt{2}$ and non negative

if $p \leq 10 - 4\sqrt{2}\,(< 5)$ or $p \geq 10 + 4\sqrt{2}\,(> 15)$. Now remind that, in our context, $p \geq 5$. Therefore, if $5 \leq p \leq 15$, then $\Delta(p) < 0$ and hence $Q(b) > 0$, which implies that $\eta(y) \leq 0$ if and only if $0 < b \leq (p-2)/2$. Also, if $p \geq 16$, then $\Delta(p) > 0$, and we cannot have $\eta(y) \leq 0$ and $Q(b) < 0$ in so far as this last condition implies $b < (p-2)/2$. Indeed $Q(b) < 0$ entails that $(p-2-\sqrt{\Delta(p)})/4 < b < (p-2+\sqrt{\Delta(p)})/4$ and it is easy to check that $(p-2+\sqrt{\Delta(p)})/4 < (p-2)/2$. So $\eta(y)$ will be non positive if and only if $Q(b) \geq 0$, that is, if and only if $b \leq (p-2-\sqrt{\Delta(p)})/4$ or $(p-2+\sqrt{\Delta(p)})/4 \leq b \leq (p-2)/2$.

When $a > 0$, we will limit ourselves to prove that $\eta(y) \leq 0$ in the case where $0 < b \leq 1$. To this end, note that , setting $z = y/(y+a)$ in in (2.9), we have to prove that $Az^2 + Bz + C \geq 0$ for any $z \in]0,1]$ where

$$A = -4\,(b-3)\,(b+1)\,(b+4), \quad B = 4\left[p\,b^2 - (p+8)\,b - 4\,(p+2)\right]$$

and

$$C = p\left[-(p-4)\,b + 2\,(p+2)\right]\,.$$

Through straightforward but tedious calculations, it can be shown that the discriminant $D = B^2 - 4\,A\,C$ of that polynomial of degree 2 in z is, up to a positive multiplicative constant, equal to

$$D(p) = \alpha(b)\,p^2 + \beta(b)\,p + \gamma(b)$$

where

$$\alpha(b) = -b^3 + 4\,b^2 - b - 4\,,\ \beta(b) = 2\,b^4 - 2\,b^3 - 18\,b^2 - 6\,b + 8 \text{ and } \gamma(b) = 32\,(b+1)^2\,.$$

We will obtain the desired result in proving that $D(p) \leq 0$.

Since $0 < b \leq 1$, it is clear that $A = 4\,(3-b)\,(b+1)\,(b+4)$ is positive and hence that it suffices to prove that $D(p) \leq 0$. Now, for $0 < b \leq 1$, the derivative of $D(p)$ with respect to p

$$D'(p) = 2\,\alpha(b)\,p + \beta(b)$$

is decreasing in p since $\alpha(b) < 0$ ($\alpha'(b) \geq 0$ for $(-4+\sqrt{13})/3 \leq b \leq (4+\sqrt{13})/3$ and $\alpha(1) = -2$). Also

$$D'(5) = 2\,(b-4)\,\epsilon(b) \leq 0$$

where $\epsilon(b) = b^3 - 2\,b^2 + 3\,b + 4$ ($\epsilon'(b) = 3\,b^2 - 4\,b + 3 > 0$ for any b and $\epsilon(0) = 4$). Therefore, for $p \geq 5$ and for $0 < b \leq 1$, we have $D'(p) \leq 0$ and, consequently, $D(p)$ is also decreasing in p. Now, from the expression of $D(p)$, it follows that

$$\begin{aligned} D(5) &= 10\,b^4 - 35\,b^3 + 42\,b^2 + 9\,b - 28 \\ &= (b-1)(10\,b^3 - 25\,b^2 + 17\,b + 26) - 2 \\ &= (b-1)(10\,b^3 + 17\,b + 1 + 25\,(1-b^2)) - 2 \\ &\leq -2. \end{aligned}$$

since $0 < b \leq 1$. Finally, the monotonicity of $D(p)$ guarantees that, for any $p \geq 5$, $D(p) \leq 0$, which gives the desired result. □

5.2 The proof of the next lemma borrows from the proof of Theorem 2.1 of [CFS95] to which we refer to. However, as the lines of the two results differ from each other, for completeness, we give an independent statement.

Lemma 5.1 *Let $x \in \mathbb{R}^p$ fixed and let Θ a random vector in $\mathbb{R}^p$ with unimodal spherically symmetric density $f(\|\theta - x\|^2)$. Denote by E_x the expectation with respect to that density and let g a function from $\mathbb{R}_+$ into $\mathbb{R}$.*

Then there exists a function Γ from $\mathbb{R}^p$ into $\mathbb{R}$ such that

$$E_x\left[\Theta\, g(\|\Theta\|^2)\right] = \Gamma(x)\cdot x\,, \tag{5.1}$$

provided this expectation exists. Moreover, if the function f is non increasing and if the function g is non negative then the function Γ is non negative.

PROOF We will use the orthogonal decomposition $\theta = \alpha + \beta$ with $\alpha \in \triangle_x$ and $\beta \in \triangle_x^\perp$ where $\triangle_x$ denotes the linear space spanned by x and $\triangle_x^\perp$ is its orthogonal subspace in $\mathbb{R}^p$. We have

$$E_x\left[\Theta\, g(\|\Theta\|^2)\right] = \int_{\mathbb{R}^p} \theta\, g(\|\theta\|^2)\, f(\|x-\theta\|^2)\, d\theta = A(x) + B(x) \tag{5.2}$$

where

$$A(x) = \int_{\triangle_x} \alpha \left[\int_{\triangle_x^\perp} g(\|\alpha\|^2 + \|\beta\|^2)\, f(\|\alpha - x\|^2 + \|\beta\|^2)\, d\beta\right] d\alpha$$

and

$$B(x) = \int_{\triangle_x^\perp} \beta \left[\int_{\triangle_x} g(\|\alpha\|^2 + \|\beta\|^2)\, f(\|\alpha - x\|^2 + \|\beta\|^2)\, d\alpha\right] d\beta$$

Now note that $B(x) = 0$ since the most inner integral a function of $\|\beta\|^2$ so that the integrand of the most outer integral is the product of a real valued function of $\|\beta\|^2$ and β. In [FS08], it is proved that such a quantity as $A(x)$ can be written as

$$A(x) = \Gamma(x)\, x$$

with

$$\Gamma(x) = \int_0^\infty z \int_0^\infty C(x,r,z)\, D(x,r,z)\, dr\, dz$$

where

$$C(x,r,z) = f(\{z-1\}^2\, \|x\|^2 + r^2) - f(\{z+1\}^2\, \|x\|^2 + r^2)$$

and

$$D(x,r,z) = \int_{S_r} g(z\, x + \beta)\, \sigma_r(d\beta)$$

(here, σ_r denotes the area measure on the sphere S_r in $\triangle_x^\perp$ of radius r and centered at 0), so that the first part of the result is obtained.

Finally, since the density $f(\|\theta - x\|^2)$ is unimodal, the function f is non increasing and, as $g \geq 0$, we obtain that $\Gamma(x) \geq 0$, which is the desired result. □

The two following lemmas deal with the expressions (2.27) and (2.28) which are involved in the marginal densities corresponding to spherically symmetric priors.

Lemma 5.2 *Let $X \sim N(\theta, I_p)$. For any spherically symmetric prior $\pi(\|\theta\|^2)$ on θ, the marginal density m is spherically symmetric and equals*

$$m(x) = \frac{1}{2^{p/2-1}\,\pi^{1/2}\,\Gamma((p-1)/2)} \int_0^\infty e^{-r^2/2}\, r^{p-2}\, \varphi(\|x\|, r)\, dr$$

where $\varphi(t, r)$ is defined in (2.27).

PROOF The spherical symmetry of the marginal density m around 0 is well known and follows from the invariance of the norm and of the Lebesgue measure under orthogonal transformations. Hence it suffices to consider the value of m at a point $x = (t, 0, \ldots, 0) \in \mathbb{R}^p$ with $t \geq 0$ so that $t = \|x\|$.

Now, setting $\theta = (\eta, \theta_2, \ldots, \theta_p)$ and $\theta_{-1} = (\theta_2, \ldots, \theta_p)$, the marginal is expressed as

$$m(t, 0, \ldots, 0) = \frac{1}{(2\pi)^{p/2}} \int_{\mathbb{R}} \exp\left(-\frac{1}{2}|t - \eta|^2\right) M(\eta)\, d\eta \tag{5.3}$$

where

$$M(\eta) = \int_{\mathbb{R}^{p-1}} \exp\left(-\frac{1}{2}\|\theta_{-1}\|^2\right)\, \pi(\eta^2 + \|\theta_{-1}\|^2)\, d\theta_{-1}\,. \tag{5.4}$$

Expressing (5.4) through the uniform measures σ_r on the spheres of radius r centered at 0 in $\mathbb{R}^{p-1}$ gives

$$\begin{aligned} M(\eta) &= \int_0^\infty \int_{S_r} \exp\left(\frac{\|\theta_{-1}\|^2}{2}\right) \pi(\eta^2 + \|\theta_{-1}\|^2)\, d\sigma_r(\theta_{-1})\, dr \\ &= C_{p-1} \int_0^\infty e^{-r^2/2}\, r^{p-2}\, \pi(\eta^2 + r^2)\, dr \end{aligned} \tag{5.5}$$

where $C_{p-1} = 2\,\pi^{(p-1)/2}/\Gamma((p-1)/2)$ is the area measure of the unit sphere in $\mathbb{R}^{p-1}$. Then, gathering (5.3), (5.4) and (5.5) and using Fubini's theorem give

$$m(t, 0, \ldots, 0) = \frac{1}{2^{p/2-1}\,\pi^{1/2}\,\Gamma((p-1)/2)} \int_0^\infty e^{-r^2/2}\, r^{p-2}\, \varphi(t, r)\, dr \tag{5.6}$$

where

$$\varphi(t,r) = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}|t-\eta|^2\right) \pi(\eta^2+r^2)\, d\eta\,. \tag{5.7}$$

Clearly we have

$$\begin{aligned}
\varphi(t,r) &= \int_0^{\infty} \left[\exp\left(-\frac{1}{2}\,|t+\eta|^2\right) + \exp\left(-\frac{1}{2}\,|t-\eta|^2\right)\right] \pi(\eta^2+r^2)\, d\eta \\
&= \int_0^{\infty} e^{-t^2/2}\, e^{-\eta^2/2}\, [e^{t\,\eta} + e^{-t\,\eta}]\, \pi(\eta^2+r^2)\, d\eta \\
&= \int_0^{\infty} G(t,\eta)\, \pi(\eta^2+r^2)\, d\eta\,,
\end{aligned} \tag{5.8}$$

according to the definition of the hyperbolic cosine and the expression of the function G in (2.28). □

The expression of the marginal in Lemma 5.2 allows to find again the fact proved in [FS08] (and mentioned in the proof of Theorem 2.1) that, if the prior $\pi(\|\theta\|^2)$ is unimodal, then the marginal density $m(\|x\|^2)$ is unimodal as well. Indeed this follows from the fact that $\varphi(t,r)$ is the expectation of the non increasing function $\eta \mapsto \pi(\eta^2+r^2)$ with respect to the density proportional to $G(t,r)$, which is shown in the proof of Lemma 5.3 below, to have monotone likelihood ratio in t.

Lemma 5.3 *Let $(t,r) \in \mathbb{R}_+ \times \mathbb{R}_+$. If the conditions of Theorem 2.1 are satisfied then*

(i) for any fixed $r \in \mathbb{R}_+$, the function $\psi(t,r)/\varphi(t,r)$ is non decreasing in t;

(ii) for any fixed $t \in \mathbb{R}_+$, the function $\psi(t,r)/\varphi(t,r)$ is non decreasing in r;

(iii) for any fixed $(t_1,t_2) \in \mathbb{R}_+ \times \mathbb{R}_+$ such that $t_1 < t_2$, the function $\varphi(t_2,r)/\varphi(t_1,r)$ is non decreasing in r.

PROOF (i) Clearly we have

$$\frac{\psi(t,r)}{\varphi(t,r)} = E_{t,r}\left[\frac{\Delta\pi(\eta^2+r^2)}{\pi(\eta^2+r^2)}\right] \tag{5.9}$$

where $E_{t,r}$ denotes the expectation with respect to the density

$$\eta \mapsto g(\eta|t,r) = \frac{G(t,\eta)\ \pi(\eta^2+r^2)}{\int_0^{\infty} G(t,\eta)\ \pi(\eta^2+r^2)\, d\eta}\,. \tag{5.10}$$

Condidering r as fixed, for $0 \le t_1 < t_2$ and for any $\eta \ge 0$, this density satisfies

$$\frac{g(\eta|t_2,r)}{g(\eta|t_1,r)} \propto \frac{G(t_2,\eta)}{G(t_1,\eta)} \propto \frac{h(t_2\,\eta)}{h(t_1\,\eta)} \tag{5.11}$$

where $h(y) = \cosh(y)$. It can be easily shown that the non decreasing monotonicity in η of the last ratio in (5.11) is equivalent to the non decreasing monotonicity in y of $y\,h'(y)/h(y)$ (actually, it was noticed in Maruyama [Mar03] and Fourdrinier, Kortbi and Strawderman [FKS08] that, if a density h has monotone non decreasing likelihood ratio when considered as a scale parameter family, $y\,h'(y)/h(y)$ is non increasing in y; here, η is the inverse of the scale parameter). As, for $y \geq 0$,

$$(y\,\cosh'(y)/\cosh(y))' = (y\,\tanh(y))' = \tanh(y) + y/\cosh^2(y) \geq 0\,,$$

it follows that the family of densities $(g(\eta|t,r))_{t\geq 0}$ defined in (5.10) has monotone non decreasing likelihood ratio in t. Hence Condition 3 of Theorem 2.1 and Formula (5.9) guarantee that the function $\psi(t,r)/\varphi(t,r)$ is non decreasing in t.

(ii) Now consider in (5.10) the parameter t as fixed. Then, for $0 \leq r_1 < r_2$ and for any $\eta \geq 0$, the ratio density in (5.11) can be viewed as

$$\frac{g(\eta|t,r_2)}{g(\eta|t,r_1)} \propto \frac{\pi(\eta^2+r_2^2)}{\pi(\eta^2+r_1^2)} \tag{5.12}$$

Here, the monotonicity in η of this ratio can be studied directly through differentiating with respect to $\tau = \eta^2$. We have

$$\frac{\partial}{\partial\tau}\frac{\pi(\tau+r_2^2)}{\pi(\tau+r_1^2)} = \frac{\pi(\tau+r_2^2)}{\pi(\tau+r_1^2)}\ \left[\frac{\pi'(\tau+r_2^2)}{\pi(\tau+r_2^2)} - \frac{\pi'(\tau+r_1^2)}{\pi(\tau+r_1^2)}\right] \geq\ 0$$

by Condition 2 of Theorem 2.1. Thus, by (5.12), $(g(\eta|t,r))_{r\geq 0}$ is a family of densities with non decreasing likelihood ratio in r. Hence Condition 3 of Theorem 2.1 and Formula (5.9) insure that the function $\psi(t,r)/\varphi(t,r)$ is non decreasing in r.

(iii) Similarly to the above, for $r \geq 0$ and for $0 \leq t_1 < t_2$, it is easily seen that

$$\frac{\varphi(t_2,r)}{\varphi(t_1,r)} = E_{t_1,r}\left[\frac{G(t_2,\eta)}{G(t_1,\eta)}\right] \tag{5.13}$$

where $E_{t_1,r}$ is the expectation with respect to the density $g(\eta|t_1,r)$ given in (5.10). In proving in (i) that the family of densities $(g(\eta|t,r))_{t\geq 0}$ defined in (5.10) has monotone non decreasing likelihood ratio in t, it has actually been shown, through (5.11), that $G(t_2,\eta)/G(t_1,\eta)$ is non decreasing in η. Also we have seen in (ii) that $(g(\eta|t,r))_{r\geq 0}$ is a family of densities with non decreasing likelihood ratio in r. Therefore the expectation in (5.13) is a non decreasing function of r so that it is likewise for $\varphi(t_2,r)/\varphi(t_1,r)$.

The next lemma follows from Stein's identity (see [Ste81]) and its use was central in [FS03]. Note that, for his identity, Stein highlighted the fact that the notion of "almost differentiablity" (which allows to include certain functions which are not differentiable in the usual sense) is at the heart of his integration by part technique. However, we refer here to the equivalent notion of weak differentiability (this equivalence was noticed by Johnstone [Joh88]) since it is of a more common use in analysis (for more details, see after the statement of lemma 5.4).

Lemma 5.4 *For $x \in \mathbb{R}^p$ let $\theta \sim \mathcal{N}(x, I_p)$ and denote by E_x the expectation with respect to that distribution. Let $k \in \mathbb{N}$.*

For any k-times weakly differentiable function f from $\mathbb{R}^p$ into $\mathbb{R}$ and for any $i = 1, \ldots, p$, we have

$$\frac{\partial^k}{\partial x_i^k} E_x\big[f(\theta)\big] = E_x\left[\frac{\partial^k}{\partial \theta_i^k} f(\theta)\right]$$

provided that

$$E_x\left[\left|\frac{\partial^k}{\partial \theta_i^k} f(\theta)\right|\right] < \infty\,.$$

Recall that a locally integrable function $f : \mathbb{R}^p \to \mathbb{R}$ is said weakly differentiable if there exist p locally integrable functions $g_1, \ldots g_p$ such that, for any $i = 1, \ldots, p$,

$$\int_\Omega f(x)\, \frac{\partial \varphi}{\partial x_i}(x)\, dx = -\int_\Omega g_i(x)\, \varphi(x)\, dx \tag{5.14}$$

for any function φ indefinitely differentiable with compact support. The functions g_i are the i-th partial weak derivatives of f and are denoted, as the usual derivatives, by $g_i = \partial f/\partial x_i$. They are unique in the sense that any function $\tilde{g}_i$ which satisfies (5.14) is equal almost everywhere to g_i. Naturally, the vector $\nabla f = \big(\partial f/\partial x_1, \ldots, \partial f/\partial x_p\big)$ is referred to the weak gradient of f.

Also, if the weak derivatives are themselves weakly differentiable, then twice weak differentiability is involved. This process can, of course, be reiterated and, for any integer k, this notion can be extended to the notion of k-times weak differentiability in a natural way. Thus, for instance, for $k = 2$, we refer to $\Delta f(x) = \sum_{i=1}^p \partial^2 f/\partial x_i^2$ as the weak Laplacian of f and, for for $k = 4$, we refer to $\Delta^{(2)} f(x) = \sum_{i=1}^p \partial^2 \Delta f/\partial x_i^2$ as the weak bi-Laplacian of f .

As an example, consider the function $\theta \mapsto \|\theta\|^{-2b}$. It is easy to check that it is a locally integrable function as soon as $p > 2\,b$ and that, for $i = 1, \ldots, p$, the function $-2\,b\,\theta_i\,\|\theta\|^{-2(b+1)}$ is locally integrable for $p > 2\,b+1$ and is a i-th partial weak derivative. More generally, for any integer k, the k-times weak differentiability of the above function is guaranteed as soon as $p > 2\,b + k$.

Let $X \sim \mathcal{N}_p(\theta, I_p)$ where $\theta \in \mathbb{R}^p$ has prior density π (non necessarily spherically symmetric). For $x \in \mathbb{R}^p$, denote by E^x the expectation with respect to the posterior expectation given $X = x$ and by $m(x)$ the corresponding marginal at x. Then the following corollary of Lemma 5.4 is immediate.

Corollary 5.1 *Assume that the prior density π is k-weakly differentiable and such that, for any $i = 1, \ldots, p$, $E_x[|\partial^k / \partial \theta_i^k f(\theta)|] < \infty$ successively for $k = 1, 2, 3, 4$. Then we have*

$$\begin{aligned} \frac{\nabla m(x)}{m(x)} &= E^x\left[\frac{\nabla \pi(\theta)}{\pi(\theta)}\right], \ \frac{\Delta m(x)}{m(x)} = E^x\left[\frac{\Delta \pi(\theta)}{\pi(\theta)}\right], \\ \frac{\nabla(\Delta m(x))}{m(x)} &= E^x\left[\frac{\nabla(\Delta \pi(\theta))}{\pi(\theta)}\right] \quad \text{and} \quad \frac{\Delta^{(2)} m(x)}{m(x)} = E^x\left[\frac{\Delta^{(2)} \pi(\theta)}{\pi(\theta)}\right]. \end{aligned}$$

Acknowledgments We are grateful to Bill Strawderman for helpful discussions.

References

[Boc88] M. E. Bock. Shrinkage estimator: Pseudo-bayes estimators for normal mean vectors. In S. S. Gupta and J. 0. Berger, editors, *Statistical Decision Theory and Related Topics 4*, volume 1, pages 281–298. Springer-Verlag, New York, 1988.

[CFS95] D. Cellier, D. Fourdrinier, and W. E. Strawderman. Shrinkage positive rule estimators for spherically symmetric distributions. *Journal of Multivariate Analysis*, 53:194–209, 1995.

[FKS08] D. Fourdrinier, O. Kortbi, and W.E. Strawderman. Bayes minimax estimators of the mean of a scale mixture of multivariate normal distributions. *Journal of Multivariate Analysis*, 99(1):74–93, 2008.

[FS03] D. Fourdrinier and W. E. Strawderman. On Bayes and unbiased estimators of loss. *Annals of the Institute of Statistical Mathematics*, 55:803–816, 2003.

[FS08] D. Fourdrinier and W. E. Strawderman. Generalized bayes minimax estimators of location vector for spherically symmetric distributions. *Journal of Multivariate Analysis*, 99:735–750, 2008.

[FW95a] D. Fourdrinier and M. T. Wells. Estimation of a loss function for spherically symmetric distributions in the general linear model. *Annals of Statistics*, 23:571–592, 1995.

[FW95b] D. Fourdrinier and M. T. Wells. Loss estimation for spherically symmetric distributions. *Journal of Multivariate Analysis*, 53:311–331, 1995.

[Joh88] I. Johnstone. On inadmissibility of some unbiased estimates of loss. In S. S. Gupta and J. 0. Berger, editors, *Statistical Decision Theory and Related Topics 4*, volume 1, pages 361–379. Springer-Verlag, New York, 1988.

[Mar03] Y. Maruyama. Admissible minimax estimators of a mean vector of scale mixtures of multivariate normal distribution. *Journal of Multivariate Analysis*, 21:69–78, 2003.

[Ste56] C. Stein. Inadmissibility of the usual estimator of the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability 1*, pages 197–206. University of California Press, Berkeley, 1956.

[Ste81] C. Stein. Estimation of the mean of multivariate normal distribution. *Annals of Statistics*, 9:1135–1151, 1981.

[Wid46] D. V. Widder. *The Laplace Transform*. Princeton University Press, Princeton, 1946.