7 juin 2016

Remerciments

Grace à " Dieu " le tout puissant, j'ai pu achever ce travail.

Mes vifs remerciements accompagnés de ma gratitude vont tout d'abord à mon encadreur Monsieur Pr.F.Madani. Je n'oublierai pas Monsieur le Pr.Idir Ouassou de l'université de Marrakech pour avoir dirigér mon travail, et qui a mis à ma disposition son précieux temps et ses connaissances scientifiques, et sa marque de confiance qu'il m'a manifestée durant mon séjour au laboratoire de mathématiques de l'université de Cadi ayyad, Marrakech, Maroc.

Toute ma gratitude aussi à Monsieur le Dr.F.Benziadi d'avoir accepté de présider le jury.

Mes sincères remerciements à Mme Dr.F.Mokhtari et Mlle Dr.N.Hachmi d'avoir bien voulu examiner ce travail.

Mes remerciements aussi à Monsieur Dr.A.Kandouci pour son aide précieuse, et tous les enseignants qui ont contribués de prés ou de loin à ma formation au sein de l'université de Saida.

Table des matières

Résumé

L'objectif de l'estimation statistique est d'évaluer certaines grandeurs associées à une population à partir d'observations faite sur un échantillon. Bien souvent, ces grandeurs sont des moyennes ou des variances. On prendra soin de distinguer ces grandeurs théoriques (inconnues θ et à estimer) de celles observées sur un échantillon. Dans ce travail on établit l'estimation de l'inconnu θ selon deux cas. Le première cas où θ est un paramètre réel inconnu, on utilise l'estimateur de maximum de vraisemblance qui donne une valeur approchée de θ, et pour le dexième cas, θ est une variable aléatoire définie par une loi $\pi(\theta)$ dite la loi a prioi. Selon la formule de Bayes on définie une loi dite la loi a posteriori, sous la condition de coût qaudratique, l'estimateur de Bayes est la moyenne de la distribution a posteriori. Finallement on définit l'estimateur de stein qui est la généralisation de L'estimateur de Bayes, cet estimateur résoudre le problème de l'admissibilité de l'estimateur de maximum de vraisemblane. On termine notre travail par une étude de comparaison par des simulation pour montrer que l'estimateur de James-Stein domine l'estimateur du maximumu de vraisemblance et son risque est meilleur que tout les risque présentés.

Introduction générale

Dans de nombreuses situations d'expériences aléatoires, il semble raisonnable d'imaginer que la praticien a une certaine idée du phénomène aléatoire qu'il est en train d'observer. Or, la démarche statistique classique repose essentiellement sur un principe de vraisemblance qui consiste à considérer que ce qui a été observé rend compte de manière exhaustive du phénomène. Mais l'observation ne fournit qu'une image et celle-ci peut être mauvaise. Certes cet inconvénient est en général gommé par les considérations asymptotiques et un certain nombre de théorèmes permettent d'évaluer la bonne qualité des estimateurs si le nombre d'observations est suffisant. L'analyse bayésienne des problèmes statistiques propose d'introduire dans la démarche d'inférence, l'information dont dispose a priori le praticien. Dans le cadre de la statistique paramétrique, ceci se traduira par le choix d'un loi sur le paramètre d'intérêt.

Dans l'approche classique, le modèle statistique est le triplet $(\mathfrak{X}, \mathcal{A}, \mathbb{P}_\theta, \theta \in \Theta)$. Ayant un a priori sur le paramètre, modélisé par une densité de probabilité que nous noterons $\pi(\theta)$, on "ré-actualise" cet a priori au vu de l'observation en calculant la densité a posteriori $\pi(\theta|x)$, et c'est partir de cette loi que l'on mène l'inférence.

On peut alors, par exemple, de manière intuitive pour le moment, retenir l'espérance mathématique ou encore le mode de cette densité a posteriori comme estimateur de θ.

Le paramètre θ devient donc en quelque sorte une variable aléatoire, à laquelle on associe une loi de probabilité dite **loi a priori.**

On sent bien d'emblée que les estimateurs bayésiens sont trés dépendants du choix de l'a priori. Différentes méthodes existent pour déterminer ces lois a priori. On peut se référer à des techniques bayésiennes empiriques, où l'on construit la loi a priori sur la base d'une expérience passée, usant de méthodes fréquentistes, pour obtenir forme et valeurs de paramètres au moyen des lois dites lois non informatives.

L'approche bayésienne se différencie donc de l'approche classique dans le sens où le paramètre θ n'est plus considéré comme étant totalement inconnu, il est devenu une v.a. dont le comportement est supposé connu. On fait intervenir dans l'analyse statistique une distribution associée à ce paramètre.
En 1956, Charles Stein [18] découvre un phénomène statistique auquel la communauté statisticienne donne son nom en parlant d'effet Stein ou de paradoxe de Stein. Ce phénomèn consiste en la non admissibilité de l'estimateur de maximum de vraisemblance de la moyenne d'une loi gaussienne n-dimensionnelle lorsque la dimension n est supérieure ou égale à 3. Le caractère "phénomènal" a trait à la coupure de dimension ainsi mise en évidence. En effet, si n= 1 ou si $n = 2$, le maximum de vraisemblance est admissible sous coût quadratique. En revanche, cette admissibilité est perdue pour tout $n > 2$ (cf. [13] et [19]). Ce phénomène signifie que, disposant d'un n-uplet de moyennes, estimer celles-ci individuellement (en les considérant l'une aprés l'autre) ou les estimer globalement (en les envisageant comme formant un vecteur) n'est pas équivalent. Une différence est mise en évidence selon que $n < 3$ ou que $n > 3$. Le "paradoxe" tient à ce que le regroupement des n estimateurs admissibles peut conduire à un estimateur inadmissible alors même que les observations sous-jacentes sont des quantités indépendantes (pouvant, en pratique, n'avoir aucun lien entre elles).
Le résultat de Stein a des conséquences fondamentales. Il met en évidence que le caractère sans biais d'un estimateur n'est pas (ou n'est plus) la propriété ultime qu'il faille rechercher. Le phénomène souligne le fait que la Théorie de la décision apporte des critères qui, via la fonction de coût, peuvent s'avérer incompatibles avec les critères statistiques classiques. L'objectif de recherche d'un éventuel "meilleur" estimateur (i.e. meilleur que tous les autres) paraît alors sans objet.
Le paradoxe de Stein a ouvert la voie à de nombreuses recherches en estimation. Ainsi la littérature sur les estimateurs dits "shrinkage estimators" (la traduction française, peu élégante, d'estimateurs à rétrecisseur été proposée) est elle abondante. La démarche conduisant à ces estimateurs repose tout d'abord sur le choix d'un estimateur standard (pour le problème d'estimation considéré). On entend par là que cet estimateur possède des propriétés telles que son usage s'impose naturellement (soit, par exemple, la minimaxité qui est présente dans le cas gaussien énoncé plus haut). Le but est alors de déterminer

des estimateurs qui améliorent l'estimateur standard, le phénomène étant encore illustré par une coupure de dimension.
Il s'avère que le phénomène de Stein survient pour la plupart des lois d'échantillonnage, pour la plupart des fonctions de coût et pour la plupart des problèmes d'estimation ; ainsi l'estimation de coût est aussi le lieu où il se rencontre (cf. [11] et [9])
Nous voulons mettre en évidence l'intervention du phénomène de Stein dans divers points de la Théorie de l'estimation : estimation ponctuelle, estimation par région de confiance, estimation d'un coût. l'apport de l'analyse Bayésienne (cf. [20] et [21]) sera souligné et des croisements seront effectués avec l'analyse fréquentiste.

Ce travail est composé de la manière suivant :

Dans le premier chapitre, on présente une littérature sur la théorie bayésienne et l'estimateur de James-Stein.

Le deuxiéme chapitre est consacré a la statistique bayésienne. Au début, on a commencé par quelque notion sur l'estimation ponctuelle à savoir, l'estimateur de maximum de vraisemblance et l'estimateur des moindre carrés. Par la suit on a présenté des résultat sur la statistique bayésienne.

Le dernier chapitre est consacré à l'estimateur de James-Stein. On a mis en évidence le fait que l'estimateur de stein domine l'estimateur de maximum de vraisemblance, cet démanstration à été approuvé par des simulation.

Chapitre 1

Inférence Bayésienne

1.1 Estimation ponctuelle

L'estimation consiste à donner des valeurs approximatives aux paramètres d'une population (cela peut être la moyenne μ, écart type σ, une proportion p), à l'aide d'un échantillon de n observations issues dons population. On peut se tromper sur la valeur exacte, mais on donne la meilleure valeur possible que l'on peut supposer.

Définition 1.1.1. *Soit X une variable aléatoire sur un référentiel Ω. Un échantillon de X de taille n est un $n-$uplet $(X_1, \ldots, X_n)$ de variables aléatoires indépendantes de même loi que X. La loi de X sera appelée loi mère. Une réalisation de cet échantillon est un $n-$uplet de réels $(x_1, \ldots, x_n)$ où $X_i(\omega) = x_i$ avec $\omega \in \Omega$.*

Définition 1.1.2. *Un estimateur T de θ est une statistique de l'échantillon aléatoire :*

$$T = h(X_1, \ldots, X_n), \qquad h : \mathbb{R}^n \longrightarrow \mathbb{R}^p, \quad p \geq 1$$

Une estimation est la valeur de L'estimateur correspondant à une réalisation de l'echantillon.

Exemples : Étant donné un échantillon $(X_1, \ldots, X_n)$ dépent du paramètre $\theta \in \Theta$ inconnu, on admet que :

– Un estimateur de la moyenne $\bar{X} = \mathbb{E}[X]$ du variable aléatoire réelle X est la moyenne empirique

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

– Un estimateur de la variance $\sigma^2 = Var(X)$ du variable aléatoire réelle X est la variance empirique

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Dans le cas particulier où le variable aléatoire réelle X suit une loi de Bernoulli $\mathcal{B}(p)$, comme la moyenne μ est égale à la proportion p, c'est une estimation de proportion (ou de fréquence) qu'on fait quand on estime sa moyenne $\mathbb{E}[X] = p$.

1.1.1 Quelques Propriéteés sur les estimateur

1.1.1.1 L'exhaustivité

Définition 1.1.3. *On appelle la fonction de vraisemblance du vecteur $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ la densité conditionnelle du variable aléatoire réelle $X|\theta$, défini par :*

$$L(x;\theta) = f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

Ou f est la densité de la variable aléatoire réelle $X_i,\ i = 1, \ldots, n$.

Définition 1.1.4. *La statistique T sera dite **exhaustive** si l'on a $L(x;\theta) = g(t,\theta)h(x)$ (principe de factorisatioll) en d'autres tenues si la densité conditionnelle de l'échatillon est indépendante du paramètre.*

Théorème 1.1.1. ***(Darmois)***
Soit une vecteur aléatoire X dont le domaine de définition ne dépend pas de paramétre θ. Une condition nécessaire et suffisante pour que l'échantillon $(X_1, \ldots, X_n)$ admette une statistique exhaustive est que la forme de la densité soit :

$$f(x,\theta) = \exp[a(x)\alpha(\theta) + b(x) + \beta(\theta)] \quad (famille\ exponentielle)$$

Si la densité est de cette forme et si de plus l'application $x_i \to \sum_{i=1}^{n} a(x_i)$ *est bijective et continûment diférentiable pour tout* i*, alors* $T = \sum_{i=1}^{n} a(X_i)$ *est une statistique exhaustive particulière.*

L'information de Fisher

Définition 1.1.5. *On appelle quantité d'information de fisher* $I_n(\theta)$ *apportée par n-echantillon sur le paramétre* θ*, la qantité positive ou nulle si elle existe*

$$I_n(\theta) = \mathbb{E}\left[\left(\frac{\partial \log L(X;\theta)}{\partial \theta}\right)^2\right].$$

Théorème 1.1.2. *Si le domaine de définition du vecteur de* X *ne dépend pas de paramètre* θ *alors :*

$$I_n(\theta) = -\mathbb{E}\left(\frac{\partial^2 \log L(X;\theta), \theta}{\partial \theta^2}\right).$$

Si cette quantité existe

Preuve La vraisemblance $L(x;\theta)$ étant une densité alors

$$\int_{\mathbb{R}^n} L(x;\theta)dx = 1 \tag{1.1}$$

En dérivant les deux membres de 1.1 par rapport à θ et en remarquant

$$\frac{\partial L(x,\theta)}{\partial \theta} = L(x;\theta)\frac{\partial \ln L(x;\theta)}{\partial \theta} \tag{1.2}$$

il vient que :

$$\int_{\mathbb{R}^n} \frac{\partial \ln L(x;\theta)}{\partial \theta} L(x;\theta)dx = 0, \tag{1.3}$$

ce qui prouve que la variable aléatoire $\frac{\partial \ln L(X;\theta)}{\partial \theta}$ est centrée et que $I_n(\theta) = Var(\frac{\partial \ln L}{\partial \theta})$. Dérivons une deuxième fois 1.3, on obtient

$$\int_{\mathbb{R}^n} \frac{\partial^2 \ln L(x;\theta)}{\partial \theta^2} L(x;\theta)dx + \int_{\mathbb{R}^n} \frac{\partial \ln L(x;\theta)}{\partial \theta}\frac{\partial L(x;\theta)}{\partial \theta}dx = 0$$

en utilisant à nouveau 1.2 sur $\frac{\partial L(x;\theta)}{\partial \theta}$, il vient :

$$\int_{\mathbb{R}^n} \frac{\partial^2 \ln L(x;\theta)}{\partial \theta^2} L(x;\theta)dx + \int_{\mathbb{R}^n} \left(\frac{\partial \ln L(x;\theta)}{\partial \theta}\right)^2 L(x;\theta)dx = 0$$

ce qui démontre le théoreme.

1.1.1.2 L'estimateur sans biais de variance minimale

Théorème 1.1.3. *S'il existe un estimateur de θ sans biais, de variance minimale, il est unique presque sûrement.*

Preuve Raisonnons par l'absurde on supposons qu'il existe deux estimateurs sans biais T_1 et T_2 de θ de variance minimale V.
On considère L'estimateur T_3 defini par :

$$T_3 = \frac{T_1 + T_2}{2}$$

alors L'estimateur T_3 est sans biais

$$\mathbb{E}(T_3) = \frac{\mathbb{E}(T_1) + \mathbb{E}(T_2)}{2} = \frac{\theta + \theta}{2} = \theta$$

et :

$$Var(T_3) = \frac{1}{4}[Var(T_1) + Var(T_2) + 2\rho\sigma_{T_1}\sigma_{T_2}]$$

où ρ est le coefficient de corrélation linéaire entre T_1 et T_2 par hypothèse $Var(T_1) = Var(T_2) = V$ alors $Var(T_3) = \frac{V}{2}(1+\rho)$. Si $\rho < 1$ on a $Var(T_3) < V$ ce qui est impossible, donc $\rho = 1$. C'est-à-dire $T_1 - \mathbb{E}(T_1) = \lambda(T_2 - \mathbb{E}(T_2))$ avec $\lambda > 0$. Comme $Var(T_1) = Var(T_2)$ il vient $\lambda = 1$ et puisque $\mathbb{E}(T_1) = \mathbb{E}(T_2) = \theta$ alors $T_1 = T_2$ (ps).

Théorème 1.1.4. ***Rao-Blackwell***
Soit T un estimateur quelconque sans biais de θ et U une statistique exhaustive pour θ. Alors $T^* = \mathbb{E}(T|U)$ est un estimateur sans biais de θ au moins aussi bon que T.

Remarques

1. Si $\mathbb{E}(Var(T|U)) = 0$ il y a une relation fonctionnelle entre T et U.
   Ce théorème fournit une méthode pour améliorer un estimateur sans biais donné.
2. Si U un statistique exhaustive , alors L'estimateur T sans biais de θ de variance minimale (unique d'après le théorème de darmois) ne dépend que de U.
3. Comme il peut exister plusieurs estimateurs sans biais de θ fonction de U, on n'est pas sûr que L'estimateur T^* obtenu par la méthode de Rao-blackwell soit le meilleur, il faut alors introduire la notion de statistique complète.

Définition 1.1.6. *On dit qu'une statistique U est complète pour ume famille de lois de probabilités $f(x, \theta)$ si $\mathbb{E}[h(U)] = 0 \ \forall\theta \Rightarrow h = 0$ p.s, ou h est une fonction mesurable.*

Théorème 1.1.5. ***Lehmann-scheffé***
Si T^* est une estimateur sans biais de θ dépendant d'une statistique exhaustive complète U alors T^* est l'unique estimateur sans biais de variance minimale de θ. En particulier si l'on dispose déjà de T estimateur sans biais de θ, $T^* = \mathbb{E}(T|U)$.

1.1.1.3 L'estimateur efficace

Proposition 1.1.1. ***Inégalité de Fréchet-Darmois-Cramer-Rao (FDCR)***
Le résultat suivant nous indique que la variance d'un estimateur ne peut être inférieure à une certaine borne, qui dépend de la quantité d'information de Fisher apportée par l'échantillon sur le paramètre θ.
Si le domaine de définition de X ne dépend pas de θ, on a pour tout estimateur T sans biais de θ :

$$V(T) \geq \frac{1}{I_n(\theta)},$$

et si T est un estimateur sans biais de $h(\theta)$:

$$V(T) \geq \frac{[h'(\theta)]^2}{I_n(\theta)}.$$

Preuve
Considérons :

$$Cov\left(T, \frac{\partial \ln L(X,\theta)}{\partial\theta}\right) = \mathbb{E}\left(T\frac{\partial \ln L(X,\theta)}{\partial\theta}\right)$$

puisque la variable $\frac{\partial \ln L}{\partial\theta}$ est centrée. Donc :

$$\begin{aligned} Cov(T, \frac{\partial \ln L(X,\theta)}{\partial\theta}) &= \int t\frac{\partial \ln L(X,\theta)}{\partial\theta}L(X,\theta)dx = \int t\frac{\partial L(X,\theta)}{\partial\theta}dx \\ &= \frac{\partial}{\partial\theta}\int tL(X,\theta)dx = \frac{\partial}{\partial\theta}\mathbb{E}(T) = h'(\theta) \end{aligned}$$

D'autre part l'inégalité de Schwarz donne :

$$\left[Cov(T, \frac{\partial \ln L(X,\theta)}{\partial\theta})\right]^2 \leq Var(T)Var\left(\frac{\partial \ln L(X,\theta)}{\partial\theta}\right)$$

c'est-à-dire :

$$[h'(\theta)]^2 \leq Var(T)I_n(\theta) \quad c.q.f.d.$$

Remarque Si l'on peut atteindre la borne minimale de la variance ; un tel estimateur sera efficace.

$*$ La définition de L'efficacité exige les conditions de régularité suivantes qui sont celles de FDCR

1. $\mathbb{E}_\theta$ est indépendant de θ.
2. $\frac{\partial L}{\partial\theta}$ existe et est continue par rapport à θ.
3. $I_n(\theta)$ est finie.
4. $\frac{\partial L}{\partial\theta}, T\frac{\partial L}{\partial\theta}$ sont intégrables par rapport à θ.

Dire que T est efficace c'est dire que sous ces conditions :

$$V(T) = \frac{[h'(\theta)]^2}{I_n(\theta)}$$

T est donc un estimateur sans biais de variance minimale de $h(\theta)$.

Théorème 1.1.6. ***(l'efficacité)***

- *La borne de Cramer-Rao ne peut être atteinte que si la loi de X est de la forme exponentielle :*

$$\ln f(x,\theta) = a(x)\alpha(\theta) + b(x) + \beta(\theta),$$

car T est nécessairement exhaustif pour θ.

- *Si la loi de X est bien de la forme précédente, il n'existe (à une transformation linéaire près) qu'une seule fonction $h(\theta)$ du paramètre qui puisse être estimée efficacement :*
  c'est $h(\theta) = -\frac{\beta'(\theta)}{\alpha'(\theta)}$
  L'estimateur de $h(\theta)$ est alors :

$$T = \frac{1}{n}\sum_{i=1}^{n} a(X_i).$$

La variance minimale est :

$$V(T) = -\frac{1}{n\alpha'(\theta)}\frac{d}{d\theta}\left(\frac{\beta'(\theta)}{\alpha'(\theta)}\right) = \frac{h'(\theta)}{n\alpha'(\theta)}$$

Exemples 1.1.1. *Dans la cas $X \sim \mathcal{P}(\theta)$ on a $S = \sum_{i=1}^{n} X_i$ est exaustive*
On peut le retrouver autrement : $\ln f(x,\theta) = -\theta + x\ln\theta - \ln(x!)$ donc la loi de poisson appartient à la famille exponentielle avec

$$a(x) = x,\ \ \alpha(x) = \ln(\theta),\ \ b(x) = -\ln(x!),\ \ \beta(\theta) = -\theta$$

Donc d'après le théorème de Darmois $S = \sum_{i=1}^{n} a(X_i) = \sum_{i=1}^{n} X_i$ est exhaustive.
D'après le théorème précédent la seule fonction qui puisse être estimée efficacement est $h(\theta) = -\frac{\beta'(\theta)}{\alpha'(\theta)} = \theta$ L'estimateur efficace est $\frac{S}{n}$

Remarque Si S est exhaustive, toute fonction déterministe de S est exhaustive. Par exemple, $\frac{S}{n}$ est exhaustive.

1.1.2 L'estimateur de maximum de vraisemblance

Quelques cas particuliers de la méthode du maximum de vraisemblance ont été connus depuis le XVIIIéme siécle, mais sa définition générale et l'argumentation de son rôle fondamental en Statistique sont dues à Fisher (1922).

Soit X une variable aléatoire réelle de loi paramétrique de variable aléatoire(discrète ou continue dépend du paramétre $\theta \in \Theta$ inconnu), dont on veut estimer le paramètre θ. Alors on définit une fonction f telle que

$$f(x,\theta) = \begin{cases} f_\theta(x) & \text{si X est une v.a continue de densité } f \\ \mathbb{P}_\theta(X = x) & \text{si X est une v.a discréte de probabilité ponctuelle } \mathbb{P}. \end{cases}$$

Définition 1.1.7. *La méthode de maximum de vraisemblance consiste à estimer θ par la valeur qui maximise $L(x,\theta)$, notée*

$$\hat{\theta}^{MV} = \sup_{\theta\in\Theta} L(x,\theta).$$

Remarque 1.1.1. *$\hat{\theta}^{MV}$ s'appelle estimateur de maximum de vraisemblance noté **EMV**, pour déterminer $\hat{\theta}^{MV}$ il suffit de chercher les valeurs critiques de vraisemblance de X. C'est à dire :*

$$\frac{\partial}{\partial\theta} L(X,\theta) = 0.$$

Exemples 1.1.2. *(Loi Normale)*
Soit X une variable de loi normale $\mathcal{N}(\mu,\sigma^2 I_n)$ on souhaite estimer la moyenne μ et la variance σ^2.
La fonction de vraisemblance pour une réalisation d'un échantillon de n variables indépendantes $(X_1,\ldots,X_n)$ est :

$$f(x_1,\ldots,x_n|\mu,\sigma^2) = \prod_{i=1}^{n} f(x_i|\mu,\sigma^2) = \left(\frac{1}{\sigma^2 2\pi}\right)^{\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right).$$

D'aprés le (théorème de König) on a $\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{x}+\overline{x}-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\overline{x}-\mu)^2$, où $\bar{x}$ représente la moyenne empirique de l'échantillon.

i) *EMV de la moyenne μ : pour maximiser la $f(x|\mu,\sigma^2)$ par rapport à μ équivalant à maximiser* $\log(f(x|\mu,\sigma^2))$ *d'ou,*

$$\begin{aligned}\frac{\partial}{\partial\mu}\log L(x_1,\ldots,x_n;\mu,\sigma^2) &= \frac{\partial}{\partial\mu}\log f(x_1,\ldots,x_n|\mu,\sigma^2)\\ &= \frac{\partial}{\partial\mu}\left(\ln\left(\frac{1}{\sigma^2 2\pi}\right)^{\frac{n}{2}} - \frac{\sum_{i=1}^n(x_i-\overline{x})^2+n(\overline{x}-\mu)^2}{2\sigma^2}\right)\\ &= 0-\frac{-2n(\overline{x}-\mu)}{2\sigma^2}.\end{aligned}$$

On obtient donc L'estimateur de maximum de vraisemblance de la moyenne μ :

$$\widehat{\mu}^{MV} = \overline{x} = \frac{1}{n}\sum_{i=1}^n x_i.$$

ii) *EMV de la variance σ^2 : de la même façon, on calcule les point critique de* $\log f(x_1,\ldots,x_n|\mu,\sigma^2)$ *par rapport à σ^2,*

$$\begin{aligned}\frac{\partial}{\partial\sigma}\log L(x;\mu,\sigma) &= \frac{\partial}{\partial\sigma}\left(\frac{n}{2}\log\left(\frac{1}{2\pi\sigma^2}\right)-\frac{\sum_{i=1}^n(x_i-\overline{x})^2+n(\overline{x}-\mu)^2}{2\sigma^2}\right)\\ &= \frac{-n}{\sigma}+\frac{\sum_{i=1}^n(x_i-\overline{x})^2+n(\overline{x}-\mu)^2}{\sigma^3}\end{aligned}$$

Pour $\mu=\hat{\mu}$ on obtient la constante σ^2 qui maximise la vraisemblance,

$$\widehat{\sigma^2}^{MV} = \frac{1}{n}\sum_{i=1}^n(x_i-\widehat{\mu})^2$$

que l'on peut traduire par

$$\widehat{\sigma^2}^{MV} = \frac{1}{n}\sum_{i=1}^n(x_i-\overline{x})^2.$$

Remarques

1. L'éstimateur θ^{MV} doit être unique s'il vérifie

$$\frac{\partial^2 \log L}{\partial \theta^2} < 0.$$

2. Si T est une statistique exhaustive pour θ alors θ^{MV} s'il existe, est fonction de T.
3. S'il existe un estimateur efficace de θ , alors il est égal à l'unique EMV de θ.
4. Si la famille de lois considerisé répond à certaines conditions de régularité et si elle admet un estimateur sans biais efficace pour θ, alors l'EMV existe et coinside avec cette estimateur sons biais.
5. Pour des échantillons suffisamment grands, l'EMV devient unique, et tend (en probabilité) vers la vraie valeur du paramètre θ. C'est donc un estimateur convergent :

$$\forall \epsilon > 0, \quad \lim_{n \to \infty} \mathbb{P}(\mid \theta_n^{MV} - \theta \mid > \epsilon) = 0.$$

6. L'EMV est asymptotiquement gaussien et asymptotiquement efficace (pour des échantillons suffisamment grands, sa variance est inférieure à celle de tout autre estimateur et est proche de la borne de Cramér-Rao) :

$$\sqrt{I_n(\theta)}(\theta_n^{MV} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

1.1.3 L'estimateur des moindre carée

La régression simple

Soit X et Y étant deux variables aléatoires, on cherche une fonction f telle que $f(X)$ soit aussi proche que possible de Y en moyenne quadratique. Au sens des moindres carrés :

- la meilleure approximation de Y par une constante est l'espérance mathématique $\mathbb{E}(Y)$,
- la meilleure approximation de Y par une fonction $f(X)$ est l'espérance conditionnelle $\mathbb{E}(Y|X)$:

$$\mathbb{E}[(Y - f(X))^2] \quad est\ minimale\ si\ \ f(X) = \mathbb{E}(Y|X).$$

Representation des variables

Si nous abordons le probléme d'un point de vue vectoriel, nous avons deux vecteurs à notre disposition :le vecteur $X = (x^1, \ldots, x_n)'$ des n observations pour la variable explicative et le vecteur $Y = (y_1, \ldots, y_n)'$ des n observation pour la variable à expliquer. Ces deux vecteurs appartiennent au même espace $\mathbb{R}^n$: l'espace des variables. Si on ajoute à cela le vecteur $1 = (1, \ldots, 1)'$, on voit tout d'abord que par l'hypothése selon laquelle tous les x_i ne sont pas égaux, les vecteurs 1 et X ne sont pas colinéaires : ils engendrent donc un sous-espace de $\mathbb{R}^n$ de dimension 2, note L^2_X. On peut projeter orthogonalement le vecteur Y sur le sous-espace L^2_X, notons provisoirement $\mathbb{E}(Y|X)$ ce projete :puisque $(1, X)$ forme une base de L^2_X, il existe une unique décomposition de la forme $\mathbb{E}(Y|X) = \alpha + \beta X$.

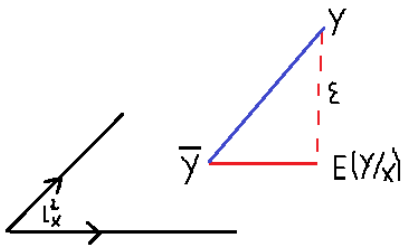

FIG. 1.1: Interprétation géométrique.

Par définition du projection orthogonal, $\mathbb{E}(Y|X)$ est défini comme l'unique vecteur de L^2_X minimisant la distance euclidienne $\|Y - \mathbb{E}(Y|X)\|$, ce qui revient au même que de minimiser son carré. Or on a :

$$\|Y - \mathbb{E}(Y|X)\|^2 = \sum_{i=1}^{n}(y_i - (\beta_1 + \beta_2 x_i))^2,$$

donc

$$f(X) = \mathbb{E}(Y|X) = \beta_1 + \beta_2 x.$$

La fonction $\mathbb{E}(Y|X)$ qui, à chaque valeur x de X, associe $E(Y|X = x)$ est la fonction de **régression** de Y en X et son graphe est la courbe de régression de Y en X.

I) Modèle linéaire simple

La variable Y est une variable aléatoire, mais la variable X n'est pas une variable aléatoire ; c'est une variable mesurée sans erreur ou à niveaux fixés. Dans ces conditions, on pose pour chaque valeur x_i de X :

$$y_i = \beta_1 + \beta_2 x_i + \epsilon_i$$

telle que

1. $Y = y_1, \ldots, y_n$ la variable endogène,
2. $X = x_1, \ldots, x_n$ la variable exogène,
3. β_1 est le paramétre à l'origine,
4. β_2 est le paramétre de pente,
5. $\epsilon_1, \ldots, \epsilon_n$ sont des variables aléatoires gaussienne de moyenne nulle et de variance σ^2(pas trop grand), ϵ_i est independantes, de même loi.

II) Estimation des paramètres

Soit $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$ équation de la droite des moindres carrés, $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$ la valeur calculée et $e_i = y_i - \hat{y}$ la valeur résiduèlle, ou écart. Les coéfficients $\hat{\beta}_1$ et $\hat{\beta}_2$ de la droite des moindres carres verifient la proprietes

$$\sum_{i=1}^{n}(y_i - \hat{y})^2 \quad minimum$$

Estimation de β_1 et β_2

$$\sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 x_i)^2 = S(\beta_1, \beta_2)$$

Le minimum de $S(\beta_1, \beta_2)$ est obtenu pour :

$$\frac{\partial S(a,b)}{\partial \beta_1} = 0 \quad \Rightarrow \quad \sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 x_i) = 0$$

$$\frac{\partial S(a,b)}{\partial \beta_2} = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} x_i(y_i - \beta_1 - \beta_2 x_i) = 0$$

La première équation à pour solution :

$$\bar{y} = \beta_1 + \beta_2 \bar{x}$$

et la deuxième, compte tenu de la première solution :

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(X,Y)}{Var(X)}.$$

Estimation de σ^2 : La variance de l'erreur s'estime par

$$\hat{\sigma}_\epsilon^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}.$$

III) Validation du modèle sur les données

Il faut que le modèle soit de bonne qualité (bon pouvoir explicatif et prédictif).

Le projection orthogonale du vecteur Y sur L_X^2 et

$$\hat{\epsilon} = Y - \hat{Y},$$

le vecteur des résidus déjà rencontrés. Le théoreme de Pythagore donne alors directement :

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{\epsilon}_i^2 \\ SCT &= SCE + SCR. \end{aligned}$$

Où
- $SCT = \sum_{i=1}^{n}(y_i - \bar{y})^2$ représente la somme des carrés totale,
- $SCE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ représente la somme des carrés expliquée,
- $SCR = \sum_{i=1}^{n}\hat{\epsilon}_i^2$ représente la somme des carrés résiduelle.

I)Indicateur principal de qualité du modèle : Le coéfficient de détermination $\mathcal{R}^2$ est défini par

$$\mathcal{R}^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

Ce coefficient donne la part de la variance de la variable Y expliquée par la régression. Il est compris entre 0 et 1.

II)Vérification des hypothèses sur les aléas ϵ : il faut que les aléas soient identiquement indépendante de meme loi gaussiens.

III)Validation du modèle sur la population :

Une fois la gaussianité vérifiée, on peut effectuer des tests afin d'asseoir la pertinence du modèle sur la population étudiée. Ces tests testent l'hypothèse :

$$H_0 : \beta = 0 \quad contre \quad H_1 : \beta \neq 0$$

($\beta = 0$ signifie absence de lien linéaire entre X et Y)

- *Teste de student.* Base sur la statistique

$$T = \frac{\hat{\beta}}{\sigma_{\hat{\beta}}} \sim \mathcal{T}(n-2) \quad sous \quad H_0$$

- *Teste de fisher.* Basé sur la statistique

$$F = \frac{SCE}{\hat{\sigma}_{\epsilon}^2}, \qquad F \sim \mathcal{F}(1, n-2) \ sous \ H_0.$$

On accepter l'hypothése H_0 si $T > t_{n-2}^{\alpha}$(resp $F > f_{1,n-2}^{\alpha}$).

Remarque 1.1.2. *Dans le cas ou ϵ suit une loi gaussien centré, L'estimateur de moindre carée est L'estimateur de maximum de vraisemblance.*

1.1.4 La qualités d'un estimateur

Définition 1.1.8. *Un estimateur T est dit convergent si $\mathbb{E}(T)$ tend vers θ lorsque n tend vers l'infini. Il sera dit consistant si T converge en probabilité vers θ lorsque n tend vers l'infini.*

Définition 1.1.9. *On appelle biais (erreur systématique)de T pour θ la valeur*

$$b_\theta = \mathbb{E}(T) - \theta.$$

Un estimateur T est dit sans biais si $\mathbb{E}(T) = \theta$.

Exemple : La variance emprique corrigée $S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ est un estimateur sans baias de la variance σ^2 car

$$\mathbb{E}(S_c^2) = \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \mathbb{E}\left(\frac{n}{n-1}S^2\right) \tag{1.4}$$

telle que $\mathbb{E}(S^2) = \frac{n-1}{n}\sigma^2$ on remplace dans 1.4

$$\mathbb{E}(S_c^2) = \sigma^2.$$

Définition 1.1.10. *soit T un estimateur de θ, le risque quadratique est défini par*

$$\mathcal{R}(T,\theta) = \mathbb{E}[(T-\theta)^2].$$

La précision d'un estimateur est mésurée par l'erreur quadratique moyenne.

Propriétés 1.1.1. *Si T est un estimateur de θ alors $\mathcal{R}(T,\theta) = Var(T) + (\mathbb{E}[T] - \theta)^2$.*

Preuve Notant $\mu = \mathbb{E}(T)$, on a

$$\begin{aligned}
\mathcal{R}(T,\theta) &= \mathbb{E}[(T-\theta)^2] \\
&= \mathbb{E}[(T-\mu+\mu-\theta)^2] \\
&= \mathbb{E}[(T-\mu)^2 + 2(T-\mu)(\mu-\theta) + (\mu-\theta)^2] \\
&= \mathbb{E}[(T-\mu)^2] + 2(\mu-\theta)\mathbb{E}[(T-\mu)] + (\mu-\theta)^2.
\end{aligned}$$

Le premier terme est la variance de T, le second est nul par définition de μ et le troisiéme est bien le carré du biais.

1.1.5 La comparison d'estimateurs

Définition 1.1.11. *Soient T_1, T_2 deux estimateur de θ, on dit que T_1 est meilleur estimateur que T_2 si*

$$\forall \theta \in \Theta, \quad \mathcal{R}(T_1, \theta) \leq \mathcal{R}(T_2, \theta).$$

Définition 1.1.12. *On dit q'un estimateur T_1 est **minimax** si pour tout autre estimateur T on a*

$$\sup_{\theta \in \Theta} \mathcal{R}(T_1, \theta) \leq \sup_{\theta \in \Theta} \mathcal{R}(T, \theta).$$

Définition 1.1.13. *On dit que L'estimateur T_1 **domine** L'estimateur T_2 si pour tout $\theta \in \Theta$, $\mathcal{R}(T_1, \theta) \leq \mathcal{R}(T_2, \theta)$, l'inégalité étant stricte pour au moins une valeur de θ.*

Définition 1.1.14. *On dit que estimateur T est **admissible** s'il n'existe aucun estimateur le dominont T.*

Définition 1.1.15. *On dit que L'estimateur T_1 est **UMVUE** pour θ (uniformly minimum variance unbiased estimators) s'il est sans biais pour θ et si pour toute autre estimateur T sans biais on a :*

$$Var_\theta(T^*) \leq Var_\theta(T), \quad \forall \theta \in \Theta.$$

1.2 Statistique Bayésienne

Définitions

Définition 1.2.1. *On appele **modèle statistique bayésien**, la donnée d'un madèl statistique paramétré $(\mathfrak{X}, \mathcal{A}, \mathbb{P}_\theta, \theta \in \Theta)$ avec $f(x|\theta)$ densité de $\mathbb{P}_\theta$ et d'une loi $\pi(\theta)$ sur le paramètre.*

1.2.1 La loi a priori et la loi a posteriori

Tandisque dans l'approche bayésienne, on considére θ un paramétre aléatoire et on associe à l'information tirée de l'echantillon, une information prevennant d'une autre source (avis d'analyse,d'experts,...). Cette information additionnelle sur θ est résume par une loi de probabilité $\pi(.)$ dite **loi a priori** du paramétre θ.

Définition 1.2.2. *Soient deux fonctions réelles f et g définies sur le même espace $\mathcal{Y}$. On dit que f et g sont proportionnelles, ce qu'on note $f \propto g$, si il existe une constante a tel que $f(y) = ag(y)$ pour tout $y \in \mathcal{Y}$.*

Définition 1.2.3. loi a posteriori
Selon la formule de ***Bayse*** *on définit la* ***loi a posteriori*** *de θ, c'est-à-dire aprés avoir pris connaissance des réalisations $(x_1, x_2, \ldots, x_n)$ de l'échantillon $(X_1, X_2, ..., X_n)$. Ci-aprés le vecteur des réalisations sera noté x et l'éhantillon sera noté X. Par transcription de la formule de Bayes la densité a posteriori est :*

$$\pi_{\theta|X=x}(\theta) = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)d\theta}.$$

La quantité $m_\pi(x) = \int_\Theta f(x|\theta)\pi(\theta)d\theta$ est la loi marginale de X ou prédictive et est une constante de normalisation de la loi a posteriori, indépendante de θ. Nous travaillerons donc très régulièrement à une constante multiplicative près : $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$. Nous ajoutons que par construction la loi a posteriori est absolument continue par rapport à la loi a priori.

Exemple

On considère un échantillon de Bernouli de paramètre θ d'un vecteur d'observations : $x = (x_1, \ldots, x_i, \ldots, x_n)$, et on considére le modéle bayésien suivant : $X_i|\theta \sim$ Bernoulli (θ) et $\theta \sim$Beta(a, b). On a :

$$L(x;\theta) = f(x|\theta) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i|\theta) = \theta^s(1-\theta)^{n-s}$$

où $s = \sum_{i=1}^n x_i$ comme $\theta \sim$Beta(a, b), on a :

$$\pi(\theta) = \frac{1}{\beta(a,b)}\theta^{a-1}(1-\theta)^{b-1}\mathbb{1}_{[0,1]}(\theta)$$

où $\beta(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

La vraisemblance marginale est :

$$\begin{aligned}
L(x) &= f(x) \\
&= \int_\Theta f(x|\theta)\pi(\theta)d\theta \\
&= \int_0^1 \theta^s(1-\theta)^{n-s}\frac{1}{\beta(a,b)}\theta^{a-1}(1-\theta)^{b-1}d\theta \\
&= \frac{1}{\beta(a,b)}\int_0^1 \theta^{s+a-1}(1-\theta)^{n-s+b-1}d\theta \\
&= \frac{\beta(s+a,n-s+b)}{\beta(a,b)}\int_0^1 \frac{\theta^{s+a-1}(1-\theta)^{n-s+b-1}}{\beta(s+a,n-s+b)}d\theta \\
&= \frac{\beta(s+a,n-s+b)}{\beta(a,b)}\int_0^1 f_{\beta_1((s+a,n-s+b))}(\theta)d\theta \\
&= \frac{\beta(s+a,n-s+b)}{\beta(a,b)}.
\end{aligned}$$

La loi a posteriori est déterminée par sa densité :

$$\begin{aligned}
\pi_{\theta|x}(\theta) &= \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)d\theta} \\
&= \frac{\beta(a,b)}{\beta(s+a,n-s+b)}\frac{1}{\beta(a,b)}\theta^{s+a-1}(1-\theta)^{n-s+b-1} \\
&= \frac{1}{\beta(s+a,n-s+b)}\theta^{s+a-1}(1-\theta)^{n-s+b-1} \qquad \textit{pour } \theta \in [0,1]
\end{aligned}$$

Par conséquent :

$$\theta|x \sim Beta(a+\sum_{i=1}^n x_i, b+n-\sum_{i=1}^n x_i).$$

Définition 1.2.4. Lois a priori conjuguées
une loi a priori $\pi(\theta)$ est conjuguée si $\pi(\theta)$ et $\pi(\theta|x)$ appartiennent à la même famille de lois.

Exemple.

Soit $X_1, \ldots, X_n$ un échantéllon, telle que $X_i|\theta$ suit une loi de Bernoulli de paramétre θ et que la loi a priori est une loi Beta. Comme $\theta|x$ suit aussi une loi Beta, on en déduit que la loi Beta est ici conjuguée, d'ou $\pi(\theta)$ est conjuguée.

La Relation entre la loi a priori et la loi a posteriori

La loi a priori sur θ : π

$+$

Observations suivant une loi $f(x|\theta)$

$\Downarrow$

On extrait des observations une information sur θ
On actualise la loi sur θ à partir des observations

$$\boxed{\pi(\theta|x) = f(x|\theta)\frac{\pi(\theta)}{m(x)}}$$

1.2.2 La fonction de coût et le risque

Pour le modèle $(\mathcal{X}, \mathcal{B}, \{\mathbb{P}_\theta, \ \theta \in \Theta\})$, on définit $\mathcal{D}$ l'ensemble des décisions possibles. C'est-à-dire l'ensemble des fonctions de θ dans $g(\theta)$ où g dépend du contexte :

-si le but est d'estimer θ alors $\mathcal{D} = \theta$

-pour un test, $D = \{0, 1\}$.

La fonction de coût(perte) est une fonction mesurable de $(\Theta \times \mathcal{D})$ à valeurs réelles positives : $L : \Theta \times \mathcal{D} \to \mathbb{R}_+$. Elle est définie selon le problème étudié et constitue l'armature du problème statistique.

$*$ Nous représentons quelques fonctions de perte rencontrées dans la littérature :

1. $\mathcal{L}_1(t, \theta) = c_1(t - \theta)\mathbb{1}_{\{\theta \leq t\}}(t) + c_2(\theta - t)\mathbb{1}_{\{\theta > t\}}(t)$.
2. $\mathcal{L}_2(t, \theta) = \varphi(\theta) \mid t - \theta \mid^r$ avec $\varphi(\theta) \geq 0$ et $\quad A > 0$

3. $\mathcal{L}_3(t,\theta) = \begin{cases} A, & \text{si } |t-\theta| > \epsilon \quad \epsilon > 0, \quad A > 0 \\ 0, & \text{si } |t-\theta| \leq \epsilon. \end{cases}$

Remarque

i) Si $\varphi(\theta) = 1$, et $r = 2$, $\mathcal{L}_2(t-\theta) = (t-\theta)^2$ est l'erreur quadratique.
ii) Si $c_1 = c_2, \mathcal{L}_1(t-\theta) = |t-\theta|$ est l'erreur absolue.

Exemples de construction de fontion de coût

1. (Berger)Pour décider de la commercialisation d'un nouveau médicament antalgique, une entreprise de l'industrie pharmaceutique s'intéresse en particulier à deux facteur susceptible d'affecter sa décision :
   θ_1 : la propartion d'individus sur laquelle l'antalgique sera efficace,
   θ_2 : la part de marché que le médicament est susceptible de prendre(demande).
   Ces deux paramètre sont inconnues. On pourra faire des expériences pour essayer d'obtenir des informations à leur sujet. On a ici un problème classique de théorie de la décision où le but ultime est de décider de mettre ou non le produit sur le marché, dans quelle proportion, à quel prix, etc.
   intéressons-nous a θ_1. θ_2 est une proportion.
   On a donc $\Theta = \{\theta_2 : 0 \leq \theta_2 \leq 1\}$ et une décision sera donc ici un nombre compris entre 0 et 1 ; une estimation que les experts veulent faire de la marché. L'espace des actions A est l'intervalle $[0,1]$.
   L'information dont on dispose est la suivante. Les pensent que le coût d'une surestimation de la demande est 2 fois plus élevé qu'une sous-estimation de celle-ci. Ce qui peut se traduire par une fonction de coût de la forme suivante :

$$L(\theta_2, a) = \begin{cases} \theta_2 - a, & \text{si } \theta_2 - a \geq 0 \quad \text{(sous estimation)} \\ 2(a-\theta_2), & \text{si } \theta_2 - a \leq 0 \quad \text{(sur estimation)} \end{cases}$$

2. Une fonction de coût classiquement utilisée est la fonction de coût quadratique : $(\theta - d)^2$. C'est le critère des moindres carée en régression.

1.2.3 L'estimateur de Bayes

Définition 1.2.5. *On appelle risque fréquentiste le coût moyen (l'espérance mathématique) du coût d'une règle de décision :*

$$\mathcal{R}(\theta,\delta) = \mathbb{E}_\theta[\mathcal{L}(\theta,\delta(X))] = \int_{\mathfrak{X}} \mathcal{L}(\theta,\delta(x))d\mathbb{P}_\theta(x).$$

Définition 1.2.6. *On dira que δ_1 est préférable à δ_2 et on note $\delta_1 \prec \delta_2$ si :*

$$\mathcal{R}(\theta,\delta_1) \leq \mathcal{R}(\theta,\delta_2), \qquad \forall\theta \in \Theta.$$

Cette définition permet d'établir un préordre sur l'ensemble $\mathcal{D}$ des décisions.
Cependant, ce préordre est partiel puisqu'il ne permet pas de comparer deux régles de décision telles que :

$$\mathcal{R}(\theta_1,\delta_1) < \mathcal{R}(\theta_1,\delta_2) \;\; et \;\; \mathcal{R}(\theta_2,\delta_1) > \mathcal{R}(\theta_2,\delta_2).$$

Définition 1.2.7. *Puisque l'approche Bayésienne met à la disposition du statisticien une loi a priori $\pi(\theta)$, on peut considérer la moyenne du risque fréquentiste i.e la moyenne du coût moyen suivant la loi a priori : $\mathbb{E}^\pi[\mathcal{R}(\theta,\delta(X))]$. Il s'agit du risque bayésien ou* ***risque de Bayes*** *que l'on note $r(\pi,\delta)$. On a :*

$$\begin{aligned} r(\pi,\delta) &= \mathbb{E}^\pi[\mathcal{R}(\theta,\delta)] \\ &= \int_\Theta \mathcal{R}(\theta,\delta)\pi(\theta)d\theta \\ &= \int_\Theta\int_{\mathfrak{X}} \mathcal{L}(\theta,\delta(x))f(x|\theta)dx\pi(\theta)d\theta \\ &= \int_\Theta\int_{\mathfrak{X}} \mathcal{L}(\theta,\delta(x))\pi(\theta|x)f(x)dxd\theta \end{aligned}$$

On définit alors le **coût a posteriori** $\rho(\pi,\delta(x))$ comme étant la moyenne du coût par rapport à la loi a posteriori :

$$\rho(\pi,\delta(x)) = \mathbb{E}^{\pi(.|x)}[\mathcal{L}(\theta,\delta(x))] = \int_\Theta \mathcal{L}(\theta,\delta(x))\pi(\theta|x)d\theta$$

Il s'agit d'une fonction de x.

Définition 1.2.8. *On appelle estimateur de Bayes associé à un coût $\mathcal{L}(\theta, \delta(x))$ et à une distribution a priori π, toute décision δ^π qui minimise le risque de Bayes $r(\pi, \delta)$.On a :*

$$\delta^\pi(x) = \arg\min_{\delta \in \mathcal{D}} r(\pi, \delta).$$

Proposition 1.2.1. *Sous l'hypothèse d'un coût quadratique, L'estimateur de Bayes $\delta^\pi(x)$ de θ associé à la loi a priori π est la moyenne a posteriori de θ :*

$$\delta^\pi(x) = \mathbb{E}^{\pi(.|x)}(\theta) = \int_{\theta \in \Theta} \theta \pi(\theta|x) d\theta.$$

Preuve Par définition, L'estimateur de Bayes minimise le coût a posteriori, où $\rho(\pi, \delta) = \mathbb{E}^{\pi(.|x)}[\mathcal{L}(\theta, \delta^\pi(x))]$

$$\begin{aligned} \frac{\partial}{\partial \delta} \mathbb{E}[\mathcal{L}(\theta, \delta^\pi(x))|x] &= \mathbb{E}\left[\frac{\partial}{\partial \delta} \|\theta - \delta^\pi(x)\|^2 |x\right] \\ &= 2\mathbb{E}[-(\theta - \delta^\pi(x))|x] \\ &= 2(\delta^\pi(x) - \mathbb{E}[\theta|x]) \end{aligned}$$

donc

$$\frac{\partial}{\partial \delta^\pi(x)} \mathbb{E}[\mathcal{L}(\theta, \delta^\pi(x))|x] = 0 \quad \Leftrightarrow \quad \delta^\pi(x) = \mathbb{E}[\theta|x]. \tag{1.5}$$

Exemples 1.2.1. *On considére le modéle bayésien suivant : $X_i|\theta \sim$ Bernoulli (θ) et $\theta \sim$Beta (a, b). Rappelons que : $\theta|x \sim$ Beta (α, β), où $\alpha = a + s$ et $\beta = b + n - s$ et $s = \sum_{i=1}^n x_i$. D'où L'estimateur bayésienn est :*

$$\delta^\pi(x) = \frac{a + \sum_{i=1}^n x_i}{a + b + n}.$$

Rappelons que L'estimateur de maximum de vraisemblance usuel est :

$$\hat{\theta}^{MV} = \frac{1}{n} \sum_{i=1}^n x_i.$$

On constate que les 2 estimateur sont équivalent quand on a beaucoup de données.

Remarque 1.2.1. *Dans un contexte multi-dimensionnel ou* $\theta = (\theta_j; \quad j = 1, \ldots, J)$ *la moyenne a posteriori* $\mathbb{E}[\theta|x]$ *est égale au vecteur* $(\mathbb{E}[\theta_j|x_j]; \quad j = 1, \ldots, J)$*, où*

$$\mathbb{E}[\theta_j|x] = \int_{\Theta_j} \theta_j \pi(\theta_j|x) d\theta_j$$

est obtenu en integrant $\pi(\theta_j|x)$ *sur toutes les composantes de* θ *autres que* θ_j*.*

1.2.4 L'estimateur de Bayes généralisé

Soit $\pi(\theta)$ une application de Θ dans $]0, +\infty[$ tel que :

$$\int_{\Theta} \pi(\theta) d\theta = +\infty$$

on parle alors de **loi a priori impropre**. Cette terminologie est bien sûr un abus de langage puisque $\pi(\theta)$ n'est pas une densité de probabilité. L'intérêt d'introduire une telle notion est de définir L'estimateur de Bayes généralisé comme suit.

– Soit $\pi(\theta)$ loi a priori impropre, telle que l'intégrale

$$\int_{\Theta} \mathrm{L}(x; \theta) \pi(\theta) d\theta$$

est convergente. On considére la densité de probabilité, en θ, définie par :

$$\frac{\mathrm{L}(x; \theta) \pi(\theta)}{\int_{\Theta} \mathrm{L}(x; \theta) \pi(\theta) d\theta}$$

Cette densité est notée $\pi(\theta|x)$ et est appelée densité de la loi a posteriori. Dans ces conditions, on appelle estimateur de Bayes généralisé de θ, la moyenne de cette loi a posteriori.

Remarque

– L'estimateur de Bayes est admissible, biaisé, converge en probabilité (quand la taille de l'échantillon $n \longrightarrow \infty$).

– La loi a posteriori peut être asymptotiquement approximée par une loi normale $\mathcal{N}(\mathbb{E}[\theta|x], Var[\theta|x])$.

1.2.5 L'estimateur du maximum a posteriori

Définition 1.2.9. *estimateur du maximum a posteriori (MAP)*
On appelle estimateur MAP tout estimateur $\delta^\pi(X) \in Arg\max_\theta \pi(\theta|X)$.
Cette notion est le pendant bayésien du maximum de vraisemblance fréquentiste. Il a le grand avantage de ne pas dépendre d'une fonction de perte et est utile pour les approches théoriques. Ses inconvénients sont les mêmes que L'estimateur du maximum de vraisemblance : non unicité, instabilité (dus aux calculs d'optimisation) et dépendance vis à vis de la mesure de référence (dominant Θ*). En autre, il ne vérifie pas la non invariance par reparamétrisation qui peut apparaître importante intuitivement.*
Le dernier point se formalise ainsi : pour g *un* $\mathcal{C}^1$*-difféomorphisme et* $\eta = g(\theta), \theta \in \Theta$, $Arg\max \pi'(\eta|X) \neq Arg\max \pi(\theta|X)$ *avec* $\pi'(\eta|X) \propto \pi(\theta(\eta), X).\left|\frac{d\theta}{d\eta}\right|$.

Exemples 1.2.2. *Modèles de mélange*
Ceci a pour but de modéliser des populations non distinguées.
Dans ce cas, la loi conditionnelle s'écrit comme suit :

$$f(X|\theta_k) = \sum_{j=1}^{k} p_j g(X|\gamma_j)$$

où

$$\sum_{j=1}^{k} p_j = 1 \text{ et pour tout } j \; p_j \geq 0.$$

En outre, le nombre de composantes k *n'est pas forcément connu.*
Les paramètres sont donc $(k, (p_j, \gamma_j)_{0\leq j\leq k})$ *et la loi a priori se met sous la forme* $d\pi(\theta) = p(k)\pi_k(p_1, \ldots, p_k, \gamma_1, \ldots, \gamma_k)$. *Un estimateur naturel maximise la loi a posteriori :*

$$\hat{k}^\pi = Arg\max_k \pi(k|X) = \frac{p(k)\int_{\Theta_k} f(X|\theta_k)d\pi_k(\theta_k)}{\sum_{k=1}^{k_{\max}} p(k)\int_{\Theta_k} f(X|\theta_k)d\pi_k(\theta_k)}.$$

1.2.6 Importance de la statistique exhaustive

Définition 1.2.10. *Statistique exhaustive*
On appelle statistique exhaustive une satistique (c'est à dire une fonction des données). $S(X)$ telle que la loi conditionnelle se décompose sous la forme :

$$f(X|\theta) = h(X|S(X)) \cdot \tilde{f}(S(X)|\theta).$$

Ceci se récrit $\pi(\theta|X) = \pi(\theta|S(X))$. En termes informationnels, ceci signifie que S résume l'information a priori.
Lorsqu'on utilise des lois impropres, le comportement d'une statistique peut être capricieux. Par exemple, si on considère un paramètre de la forme suivante $\theta = (\theta_1, \theta_2)$ avec $\pi(\theta)$ une loi impropre et pour S une statistique exhaustive pour θ_1, nous pouvons écrire :

$$f(X|\theta_1, \theta_2) = g(X|\theta_2) \cdot f(S|X).$$

Dans ce cas, il peut arriver que $\pi(\theta_1|X) = \pi(\theta_1|\delta)$ mais sans qu'il n'existe de $\pi_1(\theta_1)$ tel que $\pi(\theta_2, S) = \frac{f(S|X)\pi_1(\theta_1)}{\int_{\Theta_1} f(S|\theta_1)\pi_1(\theta_1)d\theta_1}$.

1.2.7 Prédiction

Le contexte du problème de la prédiction est le suivant : les observations X sont identiquement distribuées selon P_θ, absolument continue par rapport à une mesure dominante μ et donc qu'il existe une fonction de densité conditionnelle $f(\cdot|\theta)$. Par ailleurs on suppose que θ suit une loi a priori π.
Il s'agit alors à partir de n tirages $X_1, \ldots, X_n$ de déterminer le plus précisément possible ce que pourrait être le tirage suivant X_{n+1}.
Dans l'approche fréquentiste, on calcul $f(X_{n+1}|X_1, \ldots, X_n, \hat{\theta})$. Comme le révèle la notation $\hat{\theta}$, on ne connaît pas exactement θ. On doit donc l'estimer dans un premier temps et de ce fait, on utilise deux fois les données : une fois pour l'estimation du paramètrebet une nouvelle fois pour la prédiction dans la fonction f. En général, ceci amène à sous-estimer les intervalles de confiance.
La stratégie du paradigme bayésien, désormais bien comprise par la lectrice et peut-être

un peu assimilé par le lecteur, consiste à intégrer a prévision suivant une loi a priori sur θ et ce, afin d'avoir la meilleure prédiction compte tenu à la fois de notre savoir et de notre ignorance sur le paramètre.

La loi prédictive s'écrit ainsi :

$$f^{\pi}(X_{n+1}|X_1,\ldots,X_n) = \int_{\Theta} f(X_{n+1}|X_1,\ldots,X_n,\theta)\pi(\theta|X_1,\ldots,X_n)d\theta.$$

Dans le cas des tirages indépendants et identiquement distribués, ceci devient :

$$f^{\pi}(X_{n+1}|X_1,\ldots,X_n) = \int_{\Theta} f(X_{n+1}|\theta)\pi(\theta|X_1,\ldots,X_n)d\theta.$$

En considérant le coût quadratique $\mathcal{L}(\theta,\delta) = \|\theta-\delta\|^2$, on peut proposer le prédicteur :

$$\hat{X}^{\pi}_{n+1} = \mathbb{E}^{\pi}(X_{N+1}|X_1,\ldots,X_n) = \int X_{n+1} f(X_{n+1}|X_1,\ldots,X_n)dX_{n+1}.$$

Chapitre 2

La dérivation Bayésienne de l'estimateur de James-Stein

L'estimateur de James-Stein pour k moyennes de loi normale domine L'estimateur de maximum de vraisemblance si $k \geq 3$. Dans se chapitre, nous demandons si L'estimateur de stein est meilleur que L'estimateur de maximumu de vraisemblance. Notre réponse est oui car L'estimateur de stein est un membre d'une classe des meilleurs estimateurs qui ont des propriétés bayésiens et qui dominent aussi L'estimateur de maximum de vraisemblance. D'autre membres de cette classe sont également utiles dans des situation diverses. Notre approche est par L'estimateur de Bayes empirique .

2.1 L'estimateur de James-Stein

Soit $X = (X_1, \ldots, X_k)$ un vecteur Gaussien d'espérance $\theta \in \mathbb{R}^k$ et de matrice de covariance I_k. L'estimateur du maximum de vraisemblance est $\bar{X} \sim \mathcal{N}(\theta, (1/k)I_k)$.
L'objectif est d'exhiber un autre estimateur de la moyenne $\theta \in \mathbb{R}^k$ admettant un risque quadratique plus petit. La forme de cet estimateur sera

$$\tilde{\theta} = X + g(X),$$

pour une fonction mésurable $g = (g_1, \ldots, g_k) : R^k \to R^k$, à débattre. A cet effet, des lemmes préparatoires sont indispensables. [1]

[1] D'aprés Ibragimov et Hasminskii (1981) : Theoretical Statistical Estimation : asymp- totic theory. Editions Springer.

Lemme 2.1.1. ***(James-Stein)***
Cas univariée*; Soit* $X \sim \mathcal{N}(\theta, 1)$*, et* $g : \mathbb{R} \longrightarrow \mathbb{R}$ *une fonction dérivable presque partout, telle que* $\mathbb{E}[g^{'}(X)] < \infty$ *alors*

$$\mathbb{E}[g^{'}(X)] = \mathbb{E}(X - \theta)g(X).$$

Cas multivariée*; Soit* $X = (X_1, \ldots, X_k) \sim \mathcal{N}_k((\theta_1, \ldots, \theta_k), I_k)$*, et* $g : R^k \longrightarrow R^k$ *une fonction admettant presque partout une dérivée partielle par rapport à sa* j^{em} *variable* x_j *pour un entier* $1 \leq j \leq k$*, et telle que* $\mathbb{E}[\sum_{j=1}^{k} \frac{\partial g_j}{\partial x_j}(X)] < \infty$ *alors*

$$\mathbb{E}\left[\sum_{j=1}^{k} \frac{\partial g_j}{\partial x_j}(X) - (x_j - \theta_j)g(X)\right] = 0.$$

Preuve (**Lemme de James-Stein**)

– Soit $g : R \longrightarrow R$ une fonction dérivable presque partout, et telle que $\mathbb{E}[g^{'}(X)] < \infty$ pour $X \sim \mathcal{N}(\theta, 1)$ alors

$$\begin{aligned}
\mathbb{E}[(X - \theta)g(X)] &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}(x - \theta)g(x)e^{-\frac{1}{2}(x-\theta)^2}dx \\
&= \int_{-\infty}^{+\infty} -\frac{1}{\sqrt{2\pi}}g(x)\left(\frac{\partial}{\partial x}e^{-\frac{1}{2}(x-\theta)^2}\right)dx \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\left(\frac{\partial}{\partial x}g(x)\right)e^{-\frac{1}{2}(x-\theta)^2}dx \\
&= \mathbb{E}[g^{'}(X)].
\end{aligned}$$

L'hypothése rend ce calcul licite, par utilisation du théoréme de convergence dominée.

– Soit $g : R^k \longrightarrow R^k$ une fonction dérivable presque partout, et telle que $\mathbb{E}\left[\sum_{i=1}^{k} \frac{\partial g_j}{\partial x_j}(X)\right] < \infty$ pour $X \sim \mathcal{N}_k(\theta, I_k)$. De même que cas univarié x_j :

$$\begin{aligned}
\mathbb{E}[(X-\theta)g(X)] &= \mathbb{E}\left[\sum_{j=1}^{k}(x_j-\theta_j)g_j(x)\right] \\
&= \sum_{j=1}^{k}\mathbb{E}[(x_j-\theta_j)g_j(x)] \\
&= \sum_{j=1}^{k}\mathbb{E}\left[\frac{\partial}{\partial x_j}g_j(x)\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{k}\frac{\partial}{\partial x_j}g_j(x)\right]
\end{aligned}$$

Théorème 2.1.1. *Soit X un vecteur aléatoire de loi normale $\mathcal{N}(\theta, I_k)$, $X \in \mathbb{R}^k$ et $\theta \in \mathbb{R}^k$ telle que $k \geq 3$. Soit g une fonction de $\mathbb{R}^k \longrightarrow \mathbb{R}^k$ faiblement différentiable où la **divergence** de $g(X)$ est $\sum_{j=1}^{k}\frac{\partial}{\partial x_j}g_j(X)$, L'estimateur $\delta_g(X) = X + g(X)$ domine L'estimateur $\delta^{MV}(X) = X$ sous le coût quadratique si*

$$\|g(X)\|^2 + 2\boldsymbol{div}g(X) \leq 0.$$

Prueve Soit $g : \mathbb{R}^k \longrightarrow \mathbb{R}^k$ et $X \sim \mathcal{N}(\theta, I_k)$, soit $\Delta(\theta)$ la différence de risque

$$\begin{aligned}
\Delta(\theta) &= \mathcal{R}(\delta_g(X), \theta) - \mathcal{R}(\delta^{MV}(X), \theta) \\
&= \mathbb{E}[\|\delta_g(X) - \theta\|^2] - \mathbb{E}[\|\delta^{MV}(X) - \theta\|^2] \\
&= \mathbb{E}[\|X + g(X) - \theta\|^2 - \|X - \theta\|^2] \\
&= \mathbb{E}[\|g(X)\|^2 + 2(X-\theta)^t g(X)],
\end{aligned}$$

d'aprés le lemme de James-Stein (2.1.1), on a $\mathbb{E}[(x-\theta)g(X)] = \mathbb{E}[\mathbf{div}g(X)]$, d'où la différence du risque est

$$\Delta(\theta) = \mathbb{E}[\|g\|^2 + 2\mathbf{div}g(X)].$$

Si $\|g(X)\|^2 + 2\sum_{i=1}^{k} \frac{\partial}{\partial x_i} g_i(X) \leq 0$, alors

$$\Delta(\theta) \leq 0 \;\; \Rightarrow \;\; \mathcal{R}(\delta_g(X), \theta) \leq \mathcal{R}(\delta^{MV}(X), \theta)$$

d'où le résultat cherché.

Théorème 2.1.2. *(Estimateur de James-Stein)*
Soit X un vecteur aléatoire de loi gaussien de $\mathcal{N}(\theta, I_k)$, $X \in \mathbb{R}^k, \theta \in \mathbb{R}^k$ et $k \geq 3$, L'estimateur $\delta_{JS}(X) = \left(1 - \frac{a}{\|X\|}\right) X$ domine L'estimateur $\delta^{MV}(X)$ si $0 < a \leq 2(k-2)$.

Preuve D'aprés le théorème (2.1.1), on a la fonction $g(X) = -\frac{a}{\|X\|^2} X$, pour montre que $\delta_{JS}(X)$ domine que $\delta^{MV}(X)$ il suffit démontré que $\|g\|^2 + 2divg(X) \leq 0$ en effet :

$$\begin{aligned}
\mathbf{div}g(X) &= \sum_{i=1}^{k} \frac{\partial}{\partial x_i} \left(\frac{-a x_i}{\|X\|^2} \right) \\
&= \sum_{i=1}^{k} \left(\frac{-a}{\|X\|^2} + 2a \frac{x_i^2}{\|X\|^4} \right) \\
&= -k \frac{a}{\|X\|^2} + 2a \sum_{i=1}^{k} \frac{x_i^2}{\|X\|^4} \\
&= -\frac{ak}{\|X\|^2} + \frac{2a}{\|X\|^2} \\
&= \frac{a(2-k)}{\|X\|^2}
\end{aligned}$$

on a donc

$$\begin{aligned}
\|g(x)\|^2 + 2\mathbf{div}g(x) &= \frac{a^2}{\|x\|^4} \sum_{i=1}^{k} x_i^2 + 2\frac{a(2-k)}{\|x\|^2} \\
&= \frac{a^2}{\|x\|^2} + \frac{2a(2-k)}{\|x\|^2} \\
&= \frac{a}{\|x\|^2}(a + 2(2-k))
\end{aligned}$$

d'où

$$\begin{aligned} \|g(x)\|^2 + 2\mathbf{div}g(x) \leq 0 &\Leftrightarrow a + 2(2-k) \leq 0 \\ &\Leftrightarrow 0 < a \leq 2(k-2). \end{aligned}$$

Définition 2.1.1. *L'estimateur de James-Stein pour la moyenne θ d'une variable aléatoire $X \sim \mathcal{N}_k(\theta, I_k)$ est défini par :*

$$\delta^{JS} = \left(1 - \frac{k-2}{\|X\|^2}\right) X.$$

2.2 L'estimateur Bayesienne empirique

Dans cette section on suppose que l'echantillon $X_1, \ldots, X_n$ suit la loi normale de moyenne $\theta_1, \ldots, \theta_k$ et de variance 1, i.e $X_i/\theta_i \sim \mathcal{N}(\theta_i, 1)$. On suppose que la dimension $k \geq 3$, on désire estimer la moyenne $\theta = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$ sous le coût quadratique ;

$$\mathcal{L}(\theta, \delta) = \frac{1}{k}\|\delta - \theta\|^2 = \frac{1}{k}\sum_{i=1}^{k}(\delta_i - \theta_i)^2, \tag{2.1}$$

pour évaluer la performance d'une régle d'estimation

$$\delta(x) = (\delta_1(x), \delta_2(x), \ldots, \delta_k(x)),$$

où $\delta_i(x)$ est L'estimateur de θ_i. Nous permettons à notre estimation de θ_i, de dépondre de l'ensemble du vecteur d'observation $x = (x_1, \ldots, x_k)$.

Corollaire 2.2.1. *L'estimateur de maximum de vraisemblance $\delta_i^{MV}(X) = X_i$, a une fonction de risque $\mathcal{R}(\theta, \delta^{MV}) = \mathbb{E}_\theta[\mathcal{L}(\theta, \delta^{MV}(x))] = 1$ pour chaque valeur de $\theta \in \mathbb{R}^k$.*

Preuve

1. On calcule le maximum de la vraisemblance $x_i|\theta_i$:

$$\begin{aligned}
\frac{\partial}{\partial\theta_i} f_{\theta_i}(x_i) = 0 \quad &\Leftrightarrow \quad \frac{\partial}{\partial\theta_i}\left[\left(\frac{1}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\|x_i-\theta_i\|^2\right)\right] = 0 \\
&\Leftrightarrow \quad \frac{\partial}{\partial\theta_i}\left[-\frac{1}{2}\|x_i-\theta_i\|^2\right] = 0 \\
&\Leftrightarrow \quad (x_i-\theta_i) = 0 \\
&\Leftrightarrow \quad \delta_i^{MV}(x_i) = x_i.
\end{aligned}$$

2. Calcule de la fonction de risque $\mathcal{R}(\theta, \delta^{MV})$

$$\begin{aligned}
\mathcal{R}(\theta,\delta^0) &= \mathbb{E}_\theta[\mathcal{L}(\theta,\delta^{MV}(x))] \\
&= \mathbb{E}_\theta\left[\frac{1}{k}\|\delta^{MV}(x)-\theta\|^2\right] \\
&= \mathbb{E}_\theta\left[\frac{1}{k}\sum_{i=1}^{k}(\delta_i^{MV}(x_i)-\theta_i)^2\right] \\
&= \frac{1}{k}\mathbb{E}_\theta\left[\sum_{i=1}^{k}(\delta_i^{MV}(x_i)-\theta_i)^2\right] \\
&= \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}_\theta[x_i-\theta_i]^2 \\
&= Var(X/\theta) = 1.
\end{aligned}$$

Remarque 2.2.1. *L'estimateur $\delta_i^{MV}(X) = X_i$ de θ_i vérifie les propriétés suivant : L'estimateur $\delta_i^{MV}(X)$ est de variance minimal, sans biais et admissible pour θ_i.*

L'estimateur de James-Stein

James et Stein (1961) fournissent un estimateur explicite δ_{JS} qui domine strictement δ^{MV}, qui est connu comme L'estimateur de **James-Stein**, noté :

$$\delta_j^{JS}(X) = \left(1 - \frac{k-2}{\|X\|^2}\right) X_j \qquad pour\ (k \geq 3)$$

de $\mathcal{R}(\theta, \delta^{JS}) < 1$ pour toute les valeurs de θ. Une autre démonstration et instatée sous des hypothése bayésiennes, i.e on suppose que les θ_i sont des variable aléatoire indépendamments normalements distribués de moyenne 0 et de variance A,

$$\theta_i \sim \mathcal{N}(0, A) \qquad pour \quad i = 1, \ldots, k \tag{2.2}$$

Sous le coût qaudratique est la moyenne a postériori on a :

$$\delta_i^{\pi} = \mathbb{E}[\theta_i | x_i] = (1 - B)x_i. \tag{2.3}$$

En effet : d'aprés la formule de Bayes on a :

$$\begin{aligned}
f(\theta|x) &= \frac{f(x|\theta)\pi(\theta)}{f(x)} \\
&= \frac{(\frac{1}{2\pi})^{\frac{1}{2}} e^{-\frac{1}{2}(x-\theta)^2} (\frac{1}{A2\pi})^{\frac{1}{2}} e^{-\frac{1}{2}(\frac{\theta}{A})^2}}{\int_\Theta (\frac{1}{2\pi})^{\frac{1}{2}} e^{-\frac{1}{2}(x-\theta)^2} (\frac{1}{A2\pi})^{\frac{1}{2}} e^{-\frac{1}{2}(\frac{\theta}{A})^2} d\theta} \\
&= \frac{e^{-\frac{1}{2}[(x-\theta)^2 + (\frac{\theta}{A})^2]}}{\int_\Theta e^{-\frac{1}{2}[(x-\theta)^2 + (\frac{\theta}{A})^2]} d\theta} \\
&= \frac{e^{-\frac{1}{2}[(x-\theta)^2 + (\frac{\theta}{A})^2 - 2(x-\theta)\frac{\theta}{A}]}}{\int_\Theta e^{-\frac{1}{2}[(x-\theta)^2 + (\frac{\theta}{A})^2 - 2(x-\theta)\frac{\theta}{A}]} d\theta} \\
&= \frac{e^{-\frac{1}{2}[x - \theta(\frac{A+1}{A})]^2}}{\int_\Theta e^{-\frac{1}{2}[x - \theta(\frac{A+1}{A})]^2} d\theta} \\
&= \frac{e^{-\frac{1}{2}\left[\frac{[\theta - (\frac{A}{A+1})x]^2}{(\frac{A}{A+1})^2}\right]}}{\int_\Theta e^{-\frac{1}{2}\left[\frac{[\theta - (\frac{A}{A+1})x]^2}{(\frac{A}{A+1})^2}\right]} d\theta}
\end{aligned}$$

Alors

$$\begin{aligned} f(\theta|x) &= \frac{\left(\frac{1}{2\pi(\frac{A}{A+1})}\right)^{\frac{1}{2}} e^{-\frac{1}{2}\left[\frac{[\theta-(\frac{A}{A+1})x]^2}{(\frac{A}{A+1})^2}\right]}}{\int_\Theta \left(\frac{1}{2\pi\frac{A}{A+1}}\right)^{\frac{1}{2}} e^{-\frac{1}{2}\left[\frac{[\theta-(\frac{A}{A+1})x]^2}{(\frac{A}{A+1})^2}\right]} d\theta} \\ &= \left(\frac{1}{2\pi(\frac{A}{A+1})}\right)^{\frac{1}{2}} e^{-\frac{1}{2}\left[\frac{[\theta-(\frac{A}{A+1})x]^2}{(\frac{A}{A+1})^2}\right]} \end{aligned}$$

on pose

$$B = \frac{1}{(A+1)} \tag{2.4}$$

Donc $f(\theta|x) = \left(\frac{1}{2\pi(1-B)}\right)^{\frac{1}{2}} e^{-\frac{1}{2}\left[\frac{[\theta-(1-B)x]^2}{(1-B)^2}\right]}$, par la suit la loi conditionnelle de θ_i sachant x_i est

$$\theta_i|x_i \sim \mathcal{N}((1-B)x_i, (1-B)). \tag{2.5}$$

D'où

$$\delta^\pi(x) = \mathbb{E}(\theta_i|x_i) = (1-B)x.$$

Calculons le risque a posteriori :

Le risque a postériori de L'estimateur de Bayes δ^π est

$$\begin{aligned} \mathcal{R}(B, \delta^\pi) &= \mathbb{E}_x[\|\delta^\pi(x) - \theta\|^2] \\ &= \mathbb{E}_x[\|(1-B)x - \theta\|^2] \\ &= \mathbb{E}_x[\|\theta - \mathbb{E}(\theta|x)\|^2] \\ &= Var(\theta|x) \\ &= 1 - B. \end{aligned} \tag{2.6}$$

ou $\mathcal{R}(B, \delta^\pi)$ est la notation un peut abusé pour le risque de Bayes δ^π, qui est $\mathbb{E}_B[\mathcal{R}(\theta, \delta^\pi)]$. "$\mathbb{E}_B$" l'espérence par rapport à la loi a priori (2.2), avec $A = (1/B) - 1$. Si A est augment alors B est diminu et quand A s'approche de 0 alors B s'approche de 1.

$$\begin{aligned}\mathcal{R}(B,\delta^{\pi}) &= \mathbb{E}_B[\mathcal{L}(\theta,\delta^{\pi})] \\ &= \mathbb{E}_B[\mathbb{E}_{\theta}[\mathcal{L}(\theta,\delta^{\pi})]] \\ &= \mathbb{E}_B[\mathcal{R}(\theta,\delta^{\pi})].\end{aligned}$$

Si on ne connait pas B(ou de manière équivalent A), on ne peut pas utiliser L'estimateur de Bayes δ_i^{π}. Dans ce cas, il faut estimer B à partir de l'echantillon $\theta_1,\ldots,\theta_k$. Sous (2.2), $S=\sum_{i=1}^{k} x_i^2$ est un estimateur pour A, avec une distribution marginale $S\sim(1/B)\mathcal{X}_k^2$. On remplaçons B par son estimateur $\hat{B}(S)$, obteint :

$$\hat{\delta}_i^{\pi}=(1-\hat{B}(S))X_i. \tag{2.7}$$

On comparons $\mathcal{R}(B,\hat{\delta}^{\pi})$ avec le risque de Bayes de δ^{MV}, qui vaut 1, la différence du risque est $\mathcal{R}(B,\delta^{MV})-\mathcal{R}(B,\delta^{\pi})=B$. Le lemme suivant exprime ce risque en terme de la "**relative savings loss,**"

$$\begin{aligned}RSL(B,\hat{\delta}^{\pi}) &= \frac{\mathcal{R}(B,\hat{\delta}^{\pi})-\mathcal{R}(B,\delta^{\pi})}{\mathcal{R}(B,\delta^{MV})-\mathcal{R}(B,\delta^{\pi})} \\ &= \frac{\mathcal{R}(B,\hat{\delta}^{\pi})-(1-B)}{B}.\end{aligned} \tag{2.8}$$

Lemme 2.2.1. *Sous le modèle de Bayes* $\theta_i\sim\mathcal{N}(0,A), X_i|\theta\sim\mathcal{N}(\theta_i,1)$ *indépendante pour* $i=1,2,\ldots,k$, *et soit L'estimateur* $\hat{\delta}_i^{\pi}=(1-\hat{B})x_i$, *avec* $i=1,\ldots,k$ *alors*

$$RSL(B,\hat{\delta}^{\pi})=\tilde{\mathbb{E}}_B\left[\frac{\hat{B}-B}{B}\right]^2. \tag{2.9}$$

$\tilde{\mathbb{E}}$ *désigne l'espérance par rapport à la loi de* $S\sim(1/B)\mathcal{X}_{k+2}^2$.

Preuve
Le risque a postériori est :

$$\begin{aligned}
\mathbb{E}_B[\mathcal{L}(\theta,\hat{\delta}^\pi)|x] &= \frac{1}{k}\mathbb{E}_B[\|(1-\hat{B}(S))x-\theta\|^2|x] \\
&= \frac{1}{k}\mathbb{E}_B[\|((1-B)+(B-\hat{B}(S))x-\theta\|^2|x] \\
&= \frac{1}{k}\mathbb{E}_B[\|((1-B)x-\theta)+(B-\hat{B}(S))x\|^2|x] \\
&= \mathbb{E}_B[\frac{1}{k}\|((1-B)x-\theta)\|^2] + \frac{2(B-\hat{B}(S))x}{k}\underbrace{\mathbb{E}_B[((1-B)x-\theta)|x]}_{I=0} \\
&\quad +\frac{1}{k}\|(B-\hat{B}(S))x\|^2 \\
&= \mathbb{E}_B[\mathcal{L}(\theta,\delta^\pi)] + \frac{1}{k}(B-\hat{B}(S))^2 S \\
&= \mathcal{R}(B,\delta^\pi) + \frac{1}{k}(B-\hat{B}(S))^2 S \\
&= (1-B) + \frac{1}{k}(B-\hat{B}(S))^2 S
\end{aligned}$$

d'aprés (2.6). Alors la différence de risque entre $\hat{\delta}^\pi$ et δ^π est

$$\mathcal{R}(B,\hat{\delta}^\pi) - \mathcal{R}(B,\delta^\pi) = \frac{1}{k}\mathbb{E}_B(\hat{B}-B)^2 S,$$

d'où d'aprés (2.9),

$$RSL(B,\hat{\delta}^\pi) = \mathbb{E}_B\left[\frac{\hat{B}-B}{B}\right]^2 \frac{B.S}{k}.$$

Posson pour tout fonction $g(S)$ on a :

$$\mathbb{E}_B\left[g(S)\frac{B.S}{k}\right] = \tilde{\mathbb{E}}_B[g(S)].$$

$$\begin{aligned}
\mathbb{E}_B\left[g(S)\frac{B.S}{k}\right] &= \int_0^\infty g(S)\frac{B.s}{k}\left(\frac{B.s}{2}\right)^{\frac{k}{2}-1}\frac{e^{\frac{-Bs}{2}}}{\Gamma(\frac{k}{2})}ds \\
&= \int_0^\infty g(S)\left(\frac{B.s}{2}\right)^{\frac{(k+2)}{2}-1}\frac{e^{\frac{-Bs}{2}}}{\Gamma(\frac{k+2}{2})}ds \\
&= \tilde{\mathbb{E}}_B[g(S)].
\end{aligned}$$

Théorème 2.2.1. *(Sur L'estimateur de James-Stein)*
Sous les même hypothèses du Lemme 2.2.1, RSL de L'estimateur de James-Stein est :

$$RSL(B,\delta^{JS}) = \frac{2}{k} \tag{2.10}$$

pour chaque valeur de B. En terme de la fonction des risque de Bayes,

$$\mathcal{R}(B,\delta^{JS}) = 1 - \frac{k-2}{k}B \tag{2.11}$$

Preuve : Soit $\hat{B}(S) = (k-2)/S$, alors d'aprés le Lemme (2.2.1),

$$\begin{aligned}
RSL(B,\delta^{JS}) &= \tilde{\mathbb{E}}_{\mathbb{B}}\left[\frac{\hat{B}}{B}-1\right]^2 \\
&= \tilde{\mathbb{E}}_{\mathbb{B}}\left[\left(\frac{(k-2)}{BS}\right)^2 - 2\left(\frac{k-2}{BS}\right)+1\right] \\
&= \mathbb{E}\left[\frac{(k-2)^2}{(\mathcal{X}^2_{k+2})^2} - \frac{2(k-2)}{\mathcal{X}^2_{k+2}}+1\right] \\
&= \frac{(k-2)^2}{k(k-2)} - \frac{2(k-2)}{k}+1 = \frac{2}{k}
\end{aligned} \tag{2.12}$$

L'equation (2.11) résulte de l'équation (3.8).

Corollaire 2.2.2. *Le risque de L'estimateur de James-Stein en fonction de θ_i est donné par :*

$$\mathcal{R}(\theta, \delta^{JS}) = 1 - \frac{(k-2)}{k}\mathbb{E}_\theta\left[\frac{k-2}{S}\right]. \tag{2.13}$$

Preuve On remarque que $\mathcal{R}(\theta, \delta^{JS})$ est une fonction de $\|\theta\|^2$. Par définition

$$\mathcal{R}(B, \delta^{JS}) = 1 - \frac{k-2}{k}B = f(\|\theta\|^2).$$

aussi

$$\begin{aligned}
\mathbb{E}_B[\mathcal{R}(\theta, \delta^{JS})] &= 1 - \frac{(k-2)^2}{k}\mathbb{E}_B\left[\mathbb{E}_\theta\left[\frac{1}{S}\right]\right] \\
&= 1 - \frac{(k-2)^2}{k}\mathbb{E}_B\left[\frac{1}{S}\right] \\
&= 1 - \frac{(k-2)^2}{k}\mathbb{E}_B\left[\frac{B}{\mathcal{X}_k^2}\right] \\
&= g(\|\theta\|^2).
\end{aligned}$$

Donc

$$f(\|\theta\|^2) = g(\|\theta\|^2)$$

pour chaque valeur de B. Cela prouve que f et g sont tous en function de $\|\theta\|^2$, étant donné que la distribution de $\|\theta\|^2$, est $\|\theta\|^2 \sim A\mathcal{X}_k^2$, sont en fonction de A et donc aussi en fonction de B.

Remarque 2.2.2. *Le corollaire (2.2.2) montre que $\mathcal{R}(\theta, \delta^{JS}) < 1$ pour tout θ, avec*

$$1 - \mathcal{R}(\theta, \delta^{JS}) = \frac{(k-2)^2}{k}\mathbb{E}_\theta\left[\frac{1}{S}\right],$$

si on défini $\lambda = \|\theta\|^2/2$, alors $S \sim \mathcal{X}_k^2(2\lambda)$, suit la loi de khi-deux décentré de degrés de liberté k et de paramétre de noncentralité $\sum_{i=1}^k \theta_i^2 = 2\lambda$. Notons par

$$r(\lambda, \delta^{JS}) = \mathcal{R}(\theta, \delta^{JS}), \qquad et \ \frac{\|\theta\|^2}{2} = \lambda.$$

Corollaire 2.2.3. *La fonction de risque $r(\lambda, \delta^{JS})$ vérifieé :*

1. *$r(\lambda, \delta^{JS})$ est une fonction concave croissante*
2. *$r(0, \delta^{JS}) = \frac{2}{k}$*
3. *$r(\infty, \delta^{JS}) = 1$*
4. *$r(\lambda, \delta^{JS}) = 1 - \frac{k-2}{k} \sum_{j=0}^{\infty} \frac{\Gamma(k/2)}{\Gamma(k/2+j)}(-\lambda)^j$*
5. *$r(\lambda, \delta^{JS}) = 1 - \frac{k-2}{k} \left[\frac{(k/2-1)!}{(-\lambda)^{k/2-1}}\right] . \left[e^{-\lambda} - \sum_{j=0}^{k/2-2} \frac{(-\lambda)^j}{j!}\right]$*
6. *Elle admet aussi cette integrale de représentation*

$$r(\lambda, \delta^{JS}) = 1 - \frac{(k-2)^2}{2k} \int_0^1 e^{-\lambda t}(1-t)^{\frac{k}{2}-2} dt. \tag{2.14}$$

Preuve : Sous l'hypothèse bayésienne $\theta_i \sim \mathcal{N}(0, A)$, alors la variable aléatoire $\lambda = \frac{\|\theta\|^2}{2}$, ayant comme densité la loi $AG_{\frac{k}{2}}$ de paramétre $k/2$,

$$p_A(\lambda) = \left(\frac{\lambda}{A}\right)^{\frac{k}{2}-1} \frac{e^{-\frac{\lambda}{k}}}{A\Gamma(\frac{k}{2})}$$

par définition du risque

$$\begin{aligned} \mathcal{R}(B, \delta^{JS}) &= \mathbb{E}_B[r(\lambda, \delta^{JS})] \\ &= \int_0^{\infty} r(\lambda, \delta^{JS}) p_A(\lambda) d\lambda \\ &= \int_O^{\infty} \frac{r(A\lambda, \delta^{JS}) \lambda^{\frac{k}{2}-1} e^{-\lambda}}{\Gamma(\frac{k}{2})} d\lambda. \end{aligned}$$

La dérivée j^{me} par rapport de A de $\mathcal{R}(B, \delta^{JS})$ est

$$\frac{d^j \mathcal{R}(\theta, \delta^{JS})}{dA^j}|_{A=0} = \frac{\Gamma(\frac{k}{2}+j)}{\Gamma(\frac{k}{2})} \frac{d^j r(\lambda, \delta^{JS})}{d\lambda^j}|_{\lambda=0}$$

donc le développement en série exponensielle de $r(\lambda, \delta^{JS})$ a propos de $\lambda = 0$. À est une forme particuliére de la fonction hypergéométrique, et sa forme intégrale (2.14). Différenciant (2.14) par rapport a λ révéle la nature concave croissante de $r(\lambda, \delta^{JS})$. Par changement de variable, si on pose $u = 1 - t$, l'expression $e^{\lambda u}$ est une série exponensielle, et l'intégration du résultat de (2.14) terme a terme donne :

$$r(\lambda, \delta^{JS}) = 1 - \left(\frac{k-2}{k}\right)^2 \sum_{j=0}^{\infty} \frac{1}{k-2+2j} \frac{e^{-\lambda}\lambda^j}{j!}$$

2.3 L'approche Bayésienne empirique

On a utiliser l'hypothèse $\theta_i \sim \mathcal{N}(0, A)$, pour $i = 1, \ldots, k$, et le lemme 2.2.1, pour étudier la fonction de risque $\mathcal{R}(\theta, \delta^{JS})$ de L'estimateur de James-Stein. On va étudier des régles d'estimation plus compliquées $\hat{\delta}^{\pi}(x)$. Il se trouve qu'il est généralement beaucoup plus facile d'utiliser $\mathcal{R}(B, \hat{\delta}^{\pi})$ que $\mathcal{R}(\theta, \hat{\delta}^{\pi})$. En principe, tout les information sur $\mathcal{R}(\theta, \hat{\delta}^{\pi})$ sont contenues dans $\mathcal{R}(B, \hat{\delta}^{\pi})$, par la relation $\mathcal{R}(B, \hat{\delta}^{\pi}) = \mathbb{E}_B[\mathcal{R}(\theta, \hat{\delta}^{\pi})]$.

Il y a une autre façon pour montré le lemme 2.2.1. On peut prendre l'hypothése $\theta_i \sim \mathcal{N}(0, A)$, mais en supposant que A est inconnue et doit être estimé à partir des données. Qui peut être un estimateur **bayésienne empirique**.

La vertu de lemme 2.2.1 est qu'il réduit ce probléme de Bayes empirique plus familier on suppose que $S \sim (1/B)\mathcal{X}^2_{k+2}$ dans la suit, on désire estimer B sous le coût quadratique

$$\mathcal{L}(B, \hat{B}) = \left(\frac{\hat{B} - B}{B}\right)^2 .$$

La question est de montrer que L'estimateur de James-Stein domine l'EMV

Une réponse partielle est donnée par le théorème suivant,

Théorème 2.3.1. *Soit S une variable aléatoire de loi $(1/B)\mathcal{X}^2_{k+2}$, avec B inconnu,*

$0 < B \leq 1$,. *La loi a priori de* B *est* $\pi(B) = B^{-a}.[1-a]$. *Et la fonction du coût quadratique,* $\mathcal{L}(B, \hat{B}) = ((\hat{B} - B)/B)^2$, *on a :*

1. *L'estimateur de Bayes de* B *est*

$$\hat{B}(S) = \frac{\int_0^1 B^{k/2-a} e^{-BS/2} dB}{\int_0^1 B^{k/2-a-1} e^{-BS/2} dB}$$

2. *Le risque de Bayes de* B *est*

$$\mathbb{E}_a \left[\frac{\hat{B}_a(S) - B}{B} \right]^2 = \mathbb{E}_a \left[\tau^2(S) \left(\frac{\hat{B}^1(S)}{B} \right)^2 - 2\tau(S) \frac{\hat{B}^1(S)}{B} + 1 \right]$$

3. *La valeur qui maximise le risque est* $2/k$.

Preuve

Soit $g(B)$ la distribution a priori sur B ayant une densité $B^{-a}.[1-a]$ sur $(0, a]$, $a < 1$. Alors L'estimateur de Bayes est

$$\begin{aligned}
\hat{B}_a(S) &= \mathbb{E}(B|S) \\
&= \int_0^1 B f(B|S) dB \\
&= \int_0^1 B \frac{f(S|B)\pi(B)}{f(S)} \, dB \\
&= \frac{\int_0^1 B f(S|B)\pi(B) dB}{\int_0^1 f(S|B)\pi(B) dB} \\
&= \frac{\int_0^1 B^{\frac{k}{2}-a} e^{-\frac{BS}{2}} dB}{\int_0^1 B^{\frac{k}{2}-a-1} e^{-\frac{BS}{2}} dB}
\end{aligned} \tag{2.15}$$

On défini deux variables,

$$c_1 = \Gamma(\frac{k}{2} - a + 1) \int_0^1 \frac{B^{\frac{k}{2}-a} e^{-\frac{BS}{2}}}{\Gamma(\frac{k}{2} - a + 1)} dB$$

et

$$c_2 = \Gamma(\frac{k}{2} - a) \int_0^1 \frac{B^{\frac{k}{2}-a-1} e^{-\frac{BS}{2}}}{\Gamma(\frac{k}{2} - a)} dB$$

On remplace dans (2.15)

$$\begin{aligned} \hat{B}_a(S) &= \frac{c_1}{c_2} \\ &= \frac{\Gamma(\frac{k}{2} - a + 1)}{\Gamma(\frac{k}{2} - a)} \frac{(\frac{2}{S})^{\frac{k}{2}-a}}{(\frac{2}{S})^{\frac{k}{2}-a-1}} \frac{\int_0^1 \frac{(\frac{BS}{2})^{\frac{k}{2}-a} e^{-\frac{BS}{2}}}{\Gamma(\frac{k}{2}-a+1)} dB}{\int_0^1 \frac{(\frac{BS}{2})^{\frac{k}{2}-a-1} e^{-\frac{BS}{2}}}{\Gamma(\frac{k}{2}-a)} dB} \\ &= \frac{k-2a}{S} \frac{\int_0^1 \frac{(\frac{BS}{2})^{\frac{k}{2}-a} e^{-\frac{BS}{2}}}{\Gamma(\frac{k}{2}-a+1)} dB}{\int_0^1 \frac{(\frac{BS}{2})^{\frac{k}{2}-a-1} e^{-\frac{BS}{2}}}{\Gamma(\frac{k}{2}-a)} dB} \end{aligned} \tag{2.16}$$

Par le changement de variable $U = \frac{BS}{2}$, on a :

$$\begin{aligned} \hat{B}_a(S) &= \frac{k-2a}{S} \frac{\int_0^{\frac{S}{2}} \frac{U^{\frac{k}{2}-a} e^{-U}}{\Gamma(\frac{k}{2}-a+1)} dB}{\int_0^{\frac{S}{2}} \frac{U^{\frac{k}{2}-a-1} e^{-U}}{\Gamma(\frac{k}{2}-a)} dB} \\ &= \frac{k-2a}{S} \frac{I_{k/2-a+1}(S/2)}{I_{k/2-a}(S/2)} \end{aligned} \tag{2.17}$$

où

$$I_l(t) = \int_0^t \frac{s^{l-1} e^{-s}}{\Gamma(l)} ds,$$

on définit $\tau(S) = \hat{B}_a(S)/\hat{B}^1(S)$, ou $\hat{B}^1(S) = (k-2)/S$, de telle sorte que

$$\tau(S) = \frac{k-2a}{k-2} \frac{I_{k/2-a+1}(S/2)}{I_{k/2-a}(S/2)} \tag{2.18}$$

étant donné que le rapport des densités gamma

$$\frac{i_{k/2-a+1}(s)}{i_{k/2-a}(s)} = \frac{s}{(k/2-a)},$$

est une fonction croisante de s. Il est facile de montre que $\tau(S)$ est monotone, croissante de 0 à $(k-2a)/(k-2)$ pour S assez grand.
Nous avons

$$\mathbb{E}_a \left[\frac{\hat{B}_a(S) - B}{B} \right]^2 = \mathbb{E}_a \left[\tau^2(S) \left(\frac{\hat{B}^1(S)}{B} \right)^2 - 2\tau(S) \frac{\hat{B}^1(S)}{B} + 1 \right]. \tag{2.19}$$

Lorsque $a \longrightarrow 1$ alors $g_a(B)$ converge vers 0, et donc $S \sim (1/B)\mathcal{X}^2_{k+1}$ converge dans la distribution vers ∞. Par l'équation (2.18) nous avans $\tau(S)$ convergent vers 1, et par l'equation (2.19), on a une séquence de L'estimateur de Bayes appropriées avec un risque de Bayes qui s'approche de $2/k$, ce qui montre que la valeur est minimax dans cette situation.

Quelques Simulation

Une façon pour donner la trajetoire de la loi a posteriori d'une loi a priori(loi de beta, comme exemple), d'aprés la figure 2.1 on remarque que la lois a priori est conjuguées.

Script

> curve(dbeta(x,1,11), 0, 10)

> for (i in 2 :12)curve(dbeta(x, i, 12-i), 0, 10, add=TRUE)

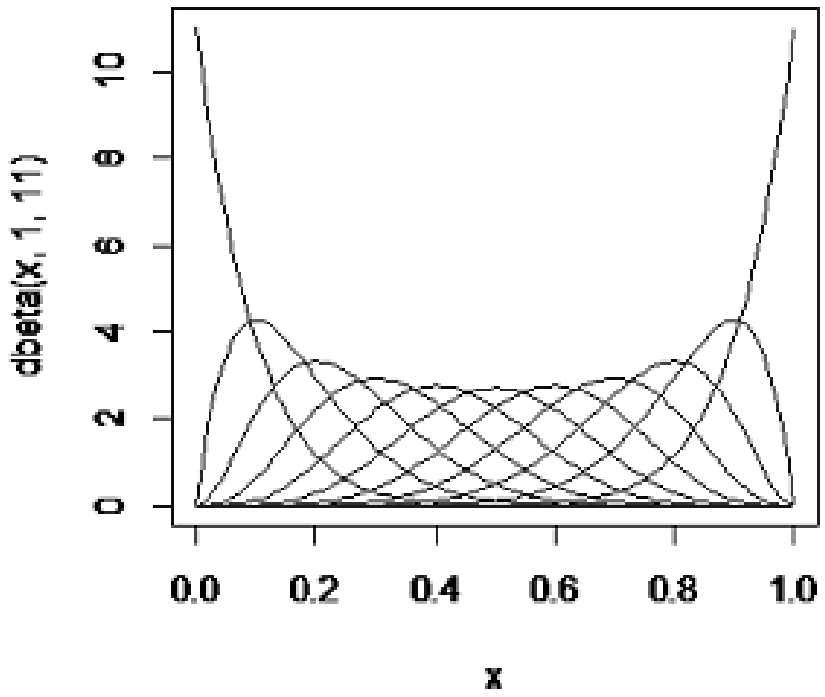

FIG. 2.1: loi $\pi_{\theta|x}$ d'une loi a priori $\pi(\theta)$ beta.

Une façon pour donner la trajectoire de la loi a postériori

1. Pour une taille d'echantillon égale 10 ;

   Script

   > curve(dnorm(x,0,1), 0, 10)

   > for (i in 1 :10){$curve(dnorm(x, (1/(i-1)), (i+1)), 0, 10, add = TRUE)$}

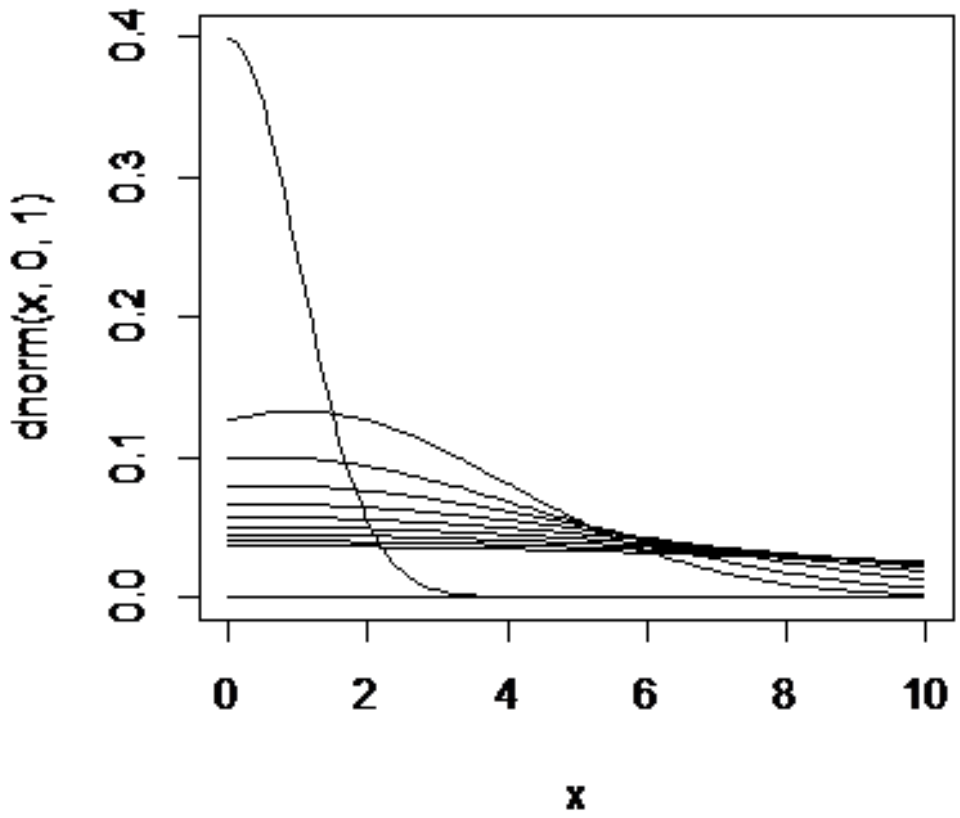

FIG. 2.2: loi $\pi_{\theta|x}$ d'une loi a priori $\pi(\theta)$ normale.

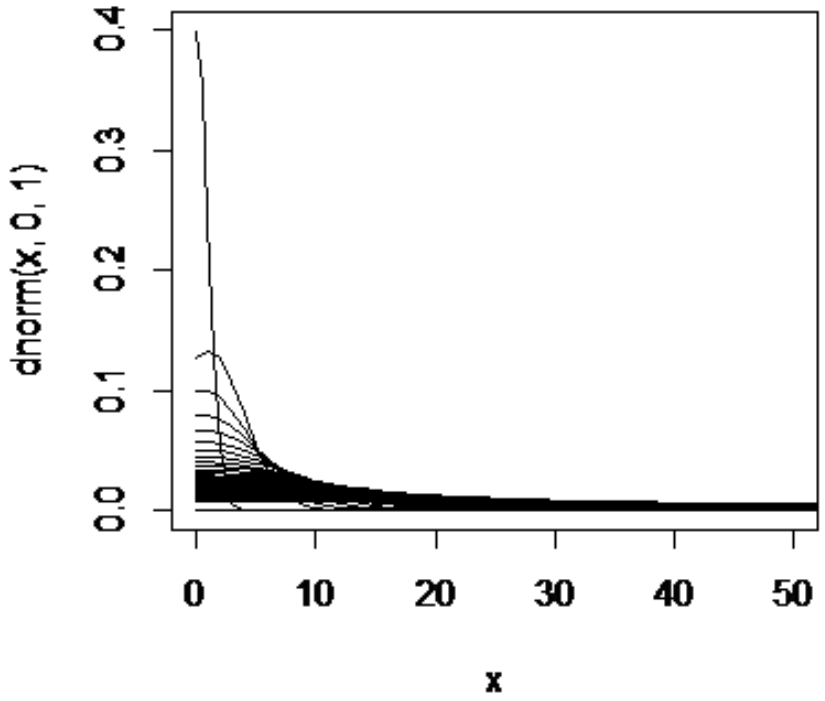

FIG. 2.3: loi $\pi_{\theta|x}$ d'une loi a priori $\pi(\theta)$ normale .

2. Pour une taille d'echantillon égale 50 ;

Script

> curve(dnorm(x,0,1), 0, 10)

> for (i in 1 :10){$curve(dnorm(x, (1/(i-1)), (i+1)), 0, 10, add = TRUE)$}

La Relation entre δ^{JS} et a ;

L'estimateur de stein varie l'orsque **a** varie, quant **a** augmente L'estimateur de stein ce déminue.

Script

> $x = matrix(1, 10, 1)$

> for(i in 1 :10)$\{x[i, 1] = rnorm(1, 0, 1)\}$

> fun=function(a)$\{stein = x - ((a/sum(x^2)) * x)\}$

> l=fun(0)

> k=fun(1)

> f=fun(2)

> p=fun(3)

> plot(l, type="l", xlab="k", ylab="stein", col = "black", lwd=2 , main="Estimateur de stein")

> lines(k, col="red")

> lines(f, col="blue")

> lines(p, col="green")

> ex.cs1 <- expression(plain(sin) * phi, paste("cos", phi))

> legend(0, 1.5, c("a=1", "a=2", "a=3", "a=4"), lty = 1, col = c("black", "red", "blue", "green"), adj = c(0, 0.7), lwd=2).

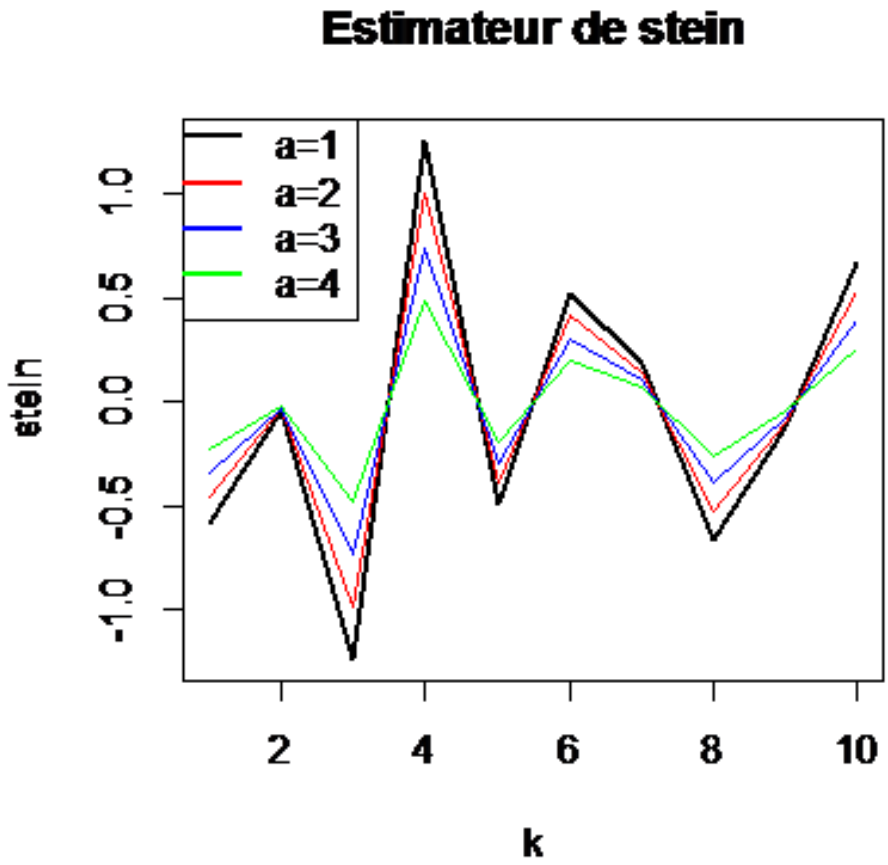

FIG. 2.4: L'estimateur de stein (**a** varie)

Comparaison des risque δ^{MV} et δ^{JS}

Soit δ^{JS} le risque de James-Stein, soit δ^{MV} le risque du maximum de vraisemblance. Dans ce graphe, on constate que δ^{JS} est moins important que δ^{MV} à partir de $k \geq 3$.

Script

> x=$matrix(1, 20, 1)$

> for(i in 1 :20)$\{x[i, 1] = rnorm(1)\}$

> m=$mean(x^2)$

> fun=function(a)$\{g = (-a/sum(x^2)) * x$

> risque=$m + mean(g^2) + 2 * mean(x * g)\}$

> l=fun(0)

> k=fun(1)

> f=fun(3)

> p=fun(5)

> b=fun(6)

> z=fun(c(7,8))

> u=fun(9)

> r=fun(11)

> plot(c(l,k,f,p,b,z,u,r),ylim=c(0,1),type = "l", col = "black",lwd=2,xlab="k",ylab="Risque", main="Comparaison des Risques EMV et JS")

> abline(h=1, lwd=2, col = "red", lty = 2)

> ex.cs1 <- expression(plain(sin) * phi, paste("cos", phi))

> legend(1,0.6, c("RJS","RMV"), lty = 1 :2, col = c("black","red"), adj = c(0, 0.6),lwd=2)

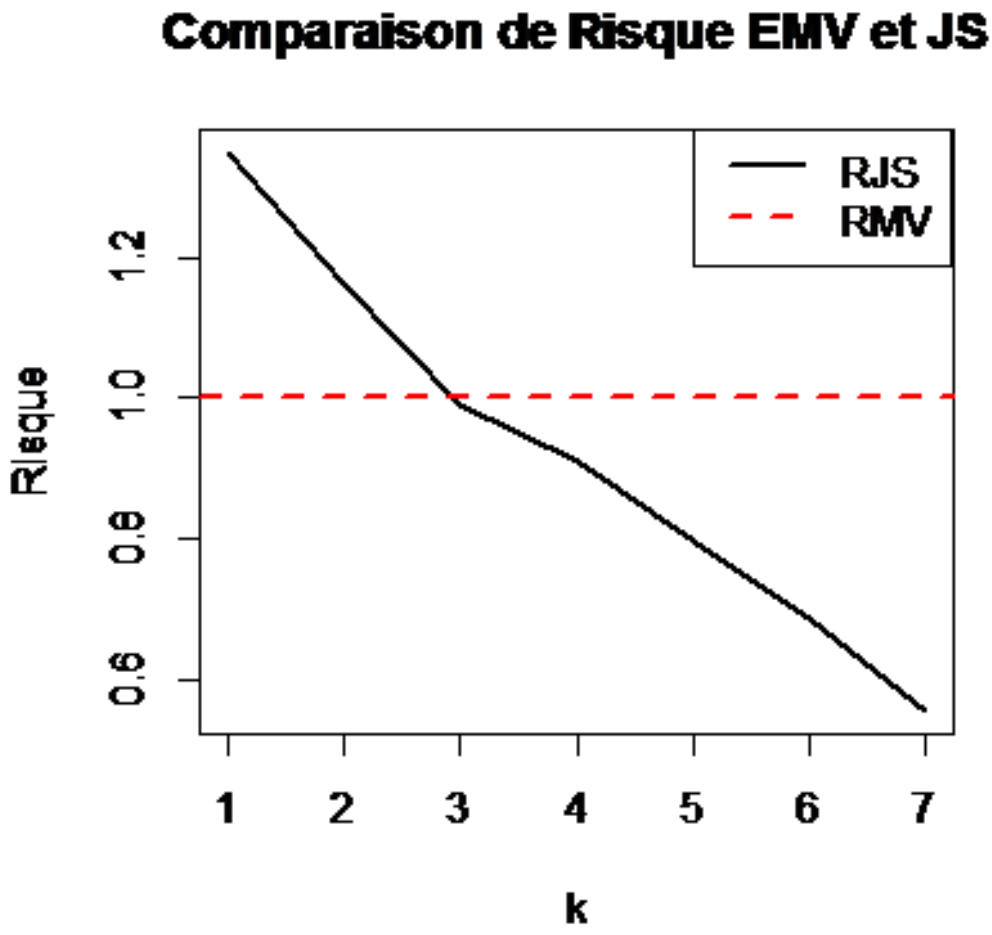

FIG. 2.5: le risque de δ^{JS} et δ^{MV}

Ici, on constate qu'il y a un lien direct entre le risque bayésiene étudié dans la deuxieme partie et la variance, c.à.d l'orsque la variance augmente, le risque augmente vers la valeur 1.

Script

> g=function(A){
> o=matrix(1, 10, 1)
> X=matrix(1, 10, 1)
> for(i in1 :10) $\{o[i,1] = rnorm(1,0,A)$
> $X[i] = rnorm(1, o[i], 1)\}$
> B=1/(1+A)
> pos=$rnorm(1,(1-B)*X,(1-B))$
> $risque = (1-B)\}$
> l=g(1)
> k=g(2)
> f=g(3)
> m=g(4)
> t=g(5)
> p=g(6)
> z=g(7)
> e=g(8)
> r=g(9)
> plot(c(l, k, f, m, t, p, z, e, r), xlim=c(1,8), type="l", xlab="A",ylab="Bayes", col = "green" ,lwd=2, main="Risque de Bayes")

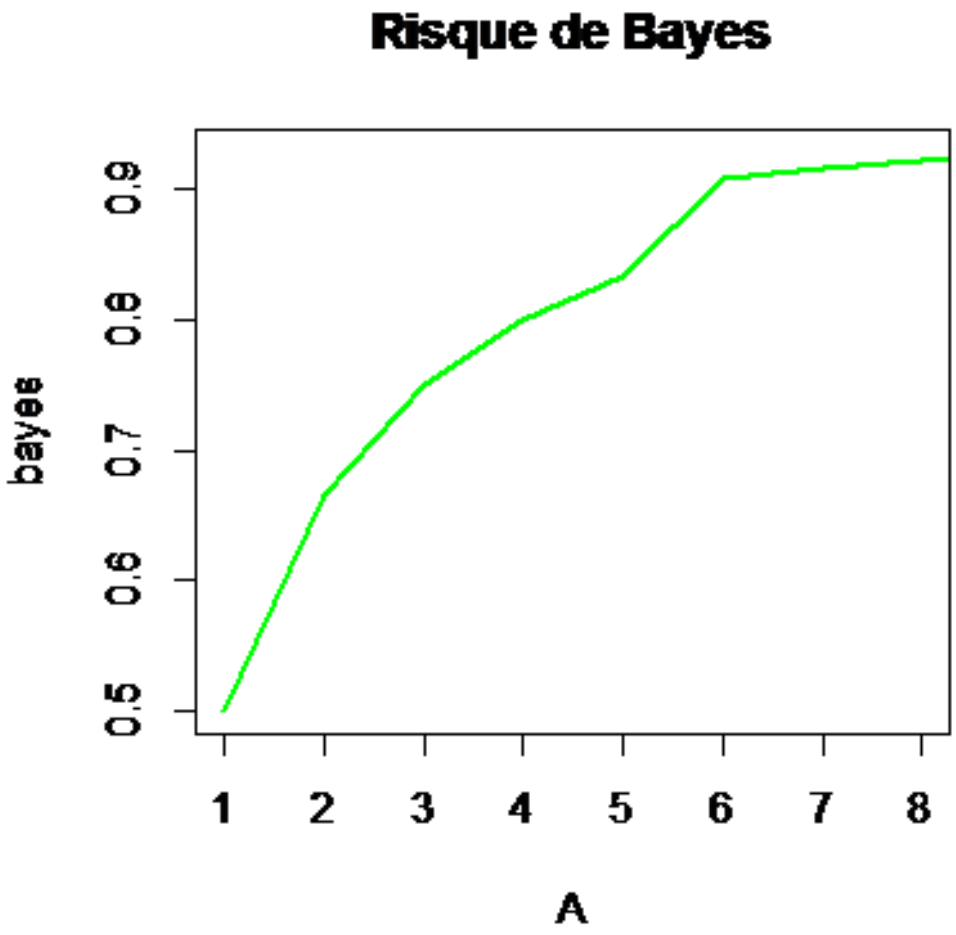

FIG. 2.6: Le risque Bayéssienne

on compare les trois risque, le risque de James-Stein, le risque de Bayes et le risque de Maximum de Vraisumblance, on conclut L'estimateur de stein et minimale que les deux autres risque. Donc L'estimateur de James-Stein est meilleur que les autres.

Script

> fun=function(a){$g = (-a/sum(x^2)) * x$

> risque=$m + mean(g^2) + 2 * mean(x * g)$}

> $l = fun(0)$ > $k = fun(1)$ > $f = fun(3)$ > $p = fun(5)$

> $b = fun(6)$ > $z = fun(7)$ > $l = fun(8)$ > $r = fun(9)$

> plot(c(l,k,f,p,b,z,l,r),type = "l", col = "black",lwd=2,xlab="k",ylab="Risque", main="Comparaison de Risque")

> abline(h=1, lwd=2, col = "red", lty = 2)

> g=function(A){$o = matrix(1, 10, 1)$ > $X = matrix(1, 10, 1)$

> for(i in 1 :10){$o[i, 1] = rnorm(1, 0, A)$

> $X[i] = rnorm(1, o[i], 1)$}

> B=1/(1+A)

> pos=$rnorm(1, (1 - B) * X, (1 - B))$

> $risque = (1 - B)$}

> $l = g(1)$ > $k = g(2)$ > $f = g(3)$

> $m = g(4)$ > $t = g(5)$ > $p = g(6)$

> $z = g(7)$ $e = g(8)$ $r = g(9)$

> lines(c(l,k,f,m,t,p,z), xlim=c(1,8), type="l", xlab="A", ylab="Bayes", col = "green", lwd=2, main="Risque de Bayes")

> ex.cs1 <- expression(plain(sin) * phi, paste("cos", phi))

> legend(5,1.3, c("RJS","RMV","RBS"), lty = c(1,2,1), col = c("black","red","green"), adj = c(0, 0.6),lwd=2).

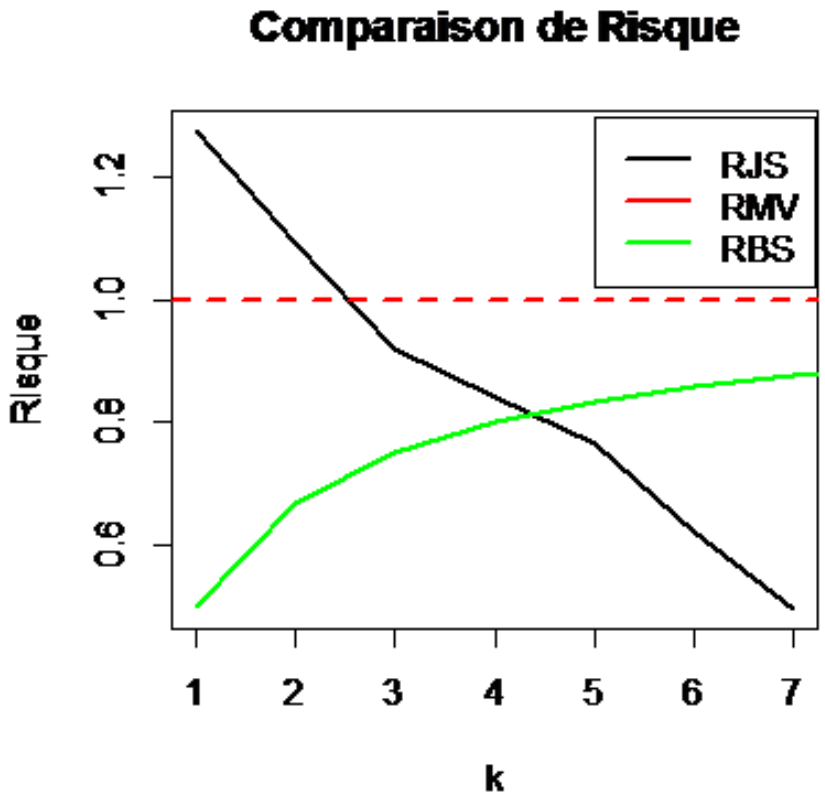

FIG. 2.7: le risque de δ^{JS} et δ^{MV} et le risque δ^{π}

Conclusion

Dans ce travail, on a étudié un estimateur biaisé(estimateur de James-Stein), ce dernier estime la moyenne d'un échantillon de taille élevé et précisement l'orsque $n \geq 3$.
Par la suit, on a montre que L'estimateur de James-Stein domine L'estimateur de maximum de vraisemblance et L'estimateur de moindre carée.
Dans un premier temp, on a montrer que les deux estimateurs coïncident c'est à dire L'estimateur des moindre carés ordinaire est égale à L'estimateur de maximum de vraisemblance pour le cas gaussiene.
En second term, on a généraliser L'estimateur de bayes, puis on se compare avec L'estimateur de maximum de vraisemblance en utilisant le critère du risque.

Mais le probléme c'est que L'estimateur de stein n'est pas admissible (probléme ouvert).

Les chercheurs essaient d'étendre cette problématique (L'estimateur de James-Stein) a l'estimation non paramétrique.

Bibliographie

[1] Bradly et Morris,*Stein's estimation rule and its competitors an empirical bayes approch,* Journal of the american statistical association, March 1973, Vol 68, Number 341.

[2] Cellier, D., Fourdrinier, D. and Robert,*shrinkage estimators of the location parameter for elliptically symmetric distribution,* Journale of Multivariate Analysis, 29 :39–52,1989.

[3] Cellier, D. and Fourdrinier,*D. Shrinkage estimators under spherical symmetry for the general linear model. Journal of Multivariate Analysis,* 52(2) :338–351, 1995.

[4] Dellacherie, C. and Mayer,*P.A. Probabilities and Potential,* Narth Holland, Amsterdam.

[5] Donoho, D.L., and Johntone,*I.Adapting to unknown smoothness via wavelet shrinkage,* Journal of the American Statistical Association, 90(432) :1200–1224, 1995.

[6] Didier Concordet,*Introduction à la statistique inférentielle,* Ecole Vétérinaire de Toulouse.

[7] Efron, B.Morris, *C. (1977). Le " paradoxe de Stein dans les statistiques ",* Américain scientifique 238 (5) : 119-127.

[8] Fourdrinier,*D. and Wells, M.T. Estimation of a loss function for spherically symmetric distributions in the general linear model. The Annals of Statictics,* 23(2) :571–592. 1995.

[9] Fourdrinier,*D. and Wells, M.T. Loss estimation for spherically symmetric distributions. Journal of Multivariate Analysis, 53(2) :311–331, 1995. [6] Dellacherie,* North Holland, Amsterdam, 1978.

[10] Gilbert Saporta,*Probabilités analyse données est statistique,* 27 Rue Ginoux, 75737 Paris Cedex 15, France.

[11] Guillaume Obozinski,*Estimateur de Stein, régularisation et pénalisation,* 18 avril 2013.

[12] Judith Rousseau,*Statistique Bayésienne,* Université de Nantes.

[13] James, W. and Stein,*C. Estimation with quadratic loss. Proc. Fourth Berkeley Symp. Math. Statist. Prob., volume 1, pages 361–380. Berkeley,* University of California Press, 1961.

[14] Peter D. Hoff,*A First Course in Bayesian Statistical Methods,* University of Washington Seattle WA 98195-4322 USA.

[15] Paul Doukhan,*Cours de Statistiques, M1,* Université de Cergy Pontoise.

[16] Renée Veysseyre,*Statistique et probabilités pour l'ingénieur,* Dunod, Paris, 2001, 2006 ISBN 2 10 0499947.

[17] Symour Geisser,*Fisher and making of maximum likelihood 1912-1922,* 27 Université de Minnesota.

[18] Stein,*C. Inadmissibility of the usual estimator for the mean of a multivariate distribution,* University of California Press, 1956.

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

[20] Strawderman,*W.E. On the existence of proper Bayes minimax estimators of the mean of a multivariate normal distribution. Proc. Sixth Berkeley Symp. Math. Statist. Prob., volume 1, pages 51–55. Berkeley,* University of California Press, 1970.

[21] Strawderman,*W.E. Proper bayes minimax estimators of the multivariate normal mean,* Ann. Math. Statis., 42 :385–388, 1971.