Thèse de doctorat

en vue de l'obtention du grade de

Docteur de l'Université de Toulouse

Spécialité Mathématiques Appliquées

délivrée par l'Université Toulouse III - Paul Sabatier.

Présentée et soutenue publiquement par

Paul Rochet

le 6 décembre 2011.

Régularisation de problèmes inverses linéaires avec opérateur inconnu

Thèse dirigée par

Jean-Michel Loubes et Jean-Pierre Florens,

et présentée devant le jury composé par

Rapporteurs :	Laurent Cavalier
	Xiahong Chen
Examinateurs :	Jean-Marc Azaïs
	Gérard Biau
	Eric Gautier
Directeurs:	Jean-Michel Loubes
	Jean-Pierre Florens

ED Mathématiques Informatique Télécommunication de Toulouse,
Université Toulouse III - Paul Sabatier,
118 route de Narbonne,
31062 Toulouse, France.

Institut de Mathématiques de Toulouse,
UMR CNRS 5219,
Université Toulouse III - Paul Sabatier,
118 route de Narbonne,
31062 Toulouse, France.

Remerciements

Mes remerciements s'adressent avant tout à mon directeur Jean-Michel Loubes qui m'a encadré durant ces trois années de thèse et les six mois de Master 2 qui les ont précédées. Je lui suis très reconnaissant pour ses encouragements, sa confiance, son enthousiasme et sa générosité ainsi que pour ses nombreux enseignements, tant sur le plan mathématique que sur le monde de la recherche.

Je souhaite également remercier vivement Jean-Pierre Florens pour sa disponibilité et pour avoir accepté de faire parti de mon jury de thèse. Ses conseils avisés et la pertinence de ses réponses m'ont donné l'envie de m'ouvrir à de nouvelles perspectives de travail.

Je remercie Laurent Cavalier et Xiaohong Chen qui m'ont fait l'honneur d'accepter d'évaluer ce travail.

Je tiens à remercier Jean-Marc Azaïs, Gérard Biau et Eric Gautier pour avoir accepté de faire partie de mon jury.

Je tiens à saluer mes collègues et amis de l'université avec qui j'éprouve un réel plaisir à travailler. Je remercie en particulier mes collègues de bureau Thibault, Yohann, Victor et Hung pour leur sympathie et pour nos nombreuses discussions enrichissantes. Un grand merci également à Rim, Matthieu, Marie-Laure, Flavia, Manu, Julie, Guillaume, Adrien pour les bons moments passés à l'université et en dehors. Je remercie également les membres de l'ESP, notamment Jean-Marc, Philippe, Pattrick, Fabrice, Thierry, Monique et Nicolas qui contribuent tous les jours à instaurer un climat détendu et sympathique au sein du laboratoire de statistiques et probabilité.

Mes remerciements s'adressent enfin à ma famille, mes amis et à Aurélie pour son soutien.

Résumé

Dans cette thèse, nous étudions des méthodes de résolution pour différents types de problèmes inverses linéaires. L'objectif est d'estimer un paramètre de dimension infinie (typiquement une fonction ou une mesure) à partir de l'observation bruitée de son image par un opérateur linéaire. Nous nous intéressons plus précisément à des problèmes inverses dits *discrets*, pour lesquels l'opérateur est à valeurs dans un espace de dimension finie. Pour ce genre de problème, la non-injectivité de l'opérateur rend impossible l'identification du paramètre à partir de l'observation. Un aspect de la régularisation consiste alors à déterminer un critère de sélection d'une solution parmi un ensemble de valeurs possibles. Nous étudions en particulier des applications de la méthode du *maximum d'entropie sur la moyenne*, qui est une méthode Bayésienne de régularisation permettant de définir un critère de sélection à partir d'information a priori. Nous traitons également des questions de stabilité en problèmes inverses sous des hypothèses de compacité de l'opérateur, dans un problème de régression non-paramétrique avec observations indirectes.

Mots-clefs

Problèmes inverses, régularisation, entropie

Regularization methods for linear inverse problems

We study regularization methods for different kinds of linear inverse problems. The objective is to estimate an infinite dimensional parameter (typically a function or a measure) from the noisy observation of its image through a linear operator. We are interested more specifically to discret inverse problems, for which the operator takes values in a finite dimensional space. For this kind of problems, the non-injectivity of the operator makes impossible the identification of the parameter from the observation. An aspect of the regularization is then to determine a criterion to select a solution among a set of possible values. We study in particular some applications of the *maximum entropy on the mean* method, which is a Bayesian regularization method that allows to choose a solution from prior informations. We also treat stability issues in inverse problems under compacity assumptions on the operator, in a general nonparametric regression framework with indirect observations.

Keywords

Inverse problems, regularization, entropy

Contents

Introduction générale

L'omniprésence des problèmes inverses dans les disciplines scientifiques en fait l'un des domaines les plus importants des statistiques. Les problèmes inverses linéaires jouent notamment un rôle fondamental de par leurs domaines d'application mais également car ils constituent le point de départ pour l'étude de problèmes inverses plus généraux. Dans cette thèse, nous étudions des méthodes de résolution pour différents types de problèmes inverses linéaires. Nous nous intéressons à l'estimation d'un paramètre de dimension infinie f (typiquement f est une fonction ou une mesure) à partir de l'observation bruitée g de son image par un opérateur linéaire A. Le modèle est le suivant,

$$g = Af + \varepsilon,$$

où ε est une variable aléatoire qui représente le bruit. Ce problème est étudié dans la littérature théorique [Cav08], [CGPT00], [EHN96], [GG97] mais également dans de nombreux domaines appliqués comme l'imagerie médicale [Gzy02], la restoration d'image [HN00], les sondages [Thé99] ou encore l'économétrie [CFR06], [Kit06]. La résolution d'un problème inverse ne pose généralement de difficultés que dans des cas où l'opérateur est peu "régulier". D'un point de vue général, un problème inverse sera dit *mal posé* si A n'est pas bijectif ou si son inverse, quand il existe, n'est pas continu. Dans ce cas, l'estimation de f nécessite que le problème soit *régularisé*.

Dans cette thèse, nous nous intéressons plus précisément à des problèmes inverses dits *discrets*, pour lesquels l'opérateur A est à valeurs dans un espace de dimension finie. Pour ce genre de problème, la non-injectivité de l'opérateur rend impossible l'identification du paramètre f à partir de l'observation g. Un aspect de la régularisation consiste alors à déterminer un critère de sélection d'une solution parmi un ensemble de valeurs possibles. Il existe pour cela des méthodes Bayésiennes de régularisation qui permettent de définir un critère de sélection à partir d'information a priori. Nous nous intéresserons en particulier à la méthode du *maximum d'entropie sur la moyenne* [GG97], qui s'applique à l'estimation d'une mesure sous des conditions de moments.

Dans les problèmes inverses pour lesquels l'opérateur A n'est pas injectif, la composante de f appartenant au noyau $K = \ker(A)$ est indépendante de l'observation g. Ainsi, lorsqu'aucune information a priori sur f n'est disponible, il est alors nécessaire de restreindre le problème à l'estimation de la projection orthogonale de f sur l'orthogonal du noyau (sous réserve que f soit définie sur un espace de Hilbert). Cette projection, que l'on note $f^\dagger$, est appelée *meilleure solution approchée* (voir [EHN96]). L'espace des solutions est ainsi restreint à l'orthogonal de K, ce qui permet de résoudre le problème de non-injectivité de A.

Un estimateur sans biais de la meilleure solution approchée, est donnée par l'image de g par le pseudo-inverse de A. Cependant, dans les cas où le signal perceptible est fortement atténué par l'opérateur, une faible perturbation sur l'observation peut engendrer un fort changement sur l'estimation, ce qui la rend instable. Pour ce genre de problèmes inverses, les méthodes classiques de régularisation consistent alors à utiliser une version "lissée" du pseudo-inverse, de manière à contrôler la variance de l'estimateur, quitte à augmenter le biais.

Dans ce mémoire, nous traitons dans un premier temps de la question d'identifiabilité d'une solution dans un problème inverse particulier qui fait intervenir des conditions de moments. Différents cadres et applications de problèmes de conditions de moment sont étudiés dans les chapitres 2,3 et 4 de cette thèse, ainsi que des extensions à des situations où l'opérateur est inconnu. Dans un second temps, nous nous intéressons à la question de la stabilité de la solution dans un problème de régression non-paramétrique avec observations indirectes, où l'instabilité est due à des hypothèses de compacité sur l'opérateur.

Le premier chapitre présente quelques outils statistiques utiles à l'étude des problèmes inverses. Dans le chapitre 2, nous introduisons une généralisation de la méthode du *maximum d'entropie sur la moyenne* (MEM) à des situations où l'opérateur est inconnu, mais est estimé indépendamment. Dans le chapitre 3, nous nous intéressons à un modèle paramétrique de conditions de moments, étudié notamment en économétrie. Nous donnons une justification Bayésienne des méthodes classiques d'estimation par le biais du maximum d'entropie sur la moyenne. Le chapitre 4 est consacrée à une application de la méthode MEM dans le cadre de l'estimation d'un paramètre linéaire en sondages. Enfin, nous étudions dans le chapitre 5, des modèles de régression non-paramétrique dans le cadre des problèmes inverses.

Chapitre 2

Nous cherchons à estimer une mesure finie μ_0, définie sur un ouvert non-vide $\mathcal{X} \subset \mathbb{R}^d$, à partir de l'observation bruitée y de moments généralisés de μ_0

$$y = \int_{\mathcal{X}} \Phi(x) d\mu_0(x) + \varepsilon,$$

où Φ est une fonction continue à valeurs dans $\mathbb{R}^k$. Ce problème rentre dans le cadre des problèmes inverses linéaires discrets. L'opérateur correspondant est l'application linéaire qui à une mesure μ sur $\mathcal{X}$ associe le vecteur de ses moments généralisés $\int \Phi d\mu \in \mathbb{R}^k$. Cet opérateur est clairement non-injectif, ce qui rend le problème mal posé.

Soit $X_1, ..., X_n$ une discrétisation de l'espace $\mathcal{X}$ dont la mesure empirique associée $\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ converge étroitement vers une mesure $\mathbb{P}_X$. L'estimateur du maximum de d'entropie sur la moyenne, introduit par Gamboa et Gassiat [GG97], est construit comme une version pondérée $\mathbb{P}_n(w) := \frac{1}{n}\sum_{i=1}^n w_i \delta_{X_i}$ de la mesure empirique, où $w = (w_1, ..., w_n)^t \in \mathbb{R}^n$ est un vecteur de poids. Après avoir fixé un a priori ν_0 sur le vecteur de poids w, l'estimateur MEM est obtenu comme l'espérance d'une mesure discrète à poids aléatoires $\mathbb{P}_n(W)$, qui minimise une fonctionnelle sous des contraintes convexes. La fonctionnelle est déterminée à l'aide d'information a priori intégrée au problème par le choix de la distribution ν_0.

Lorsque la fonction de contrainte Φ est connue, Gamboa et Gassiat ont montré un résultat de convergence de l'estimateur MEM. Ici, la fonction de contrainte Φ est supposée inconnue mais une suite d'approximations $\{\Phi_m\}_{m\in\mathbb{N}}$ est observée indépendamment. Ainsi, le critère de régularisation est construit à partir de l'approximation Φ_m, ce qui cause une erreur supplémentaire dans l'estimation. La méthode du maximum d'entropie sur la moyenne sous contraintes approchées (AMEM) a été introduite par Loubes et Pelletier [LP08], sous des hypothèses de convergence uniforme de la suite $\{\Phi_m\}$. Ici, nous montrons que le résultat obtenu par Loubes et Pelletier reste vrai sous l'hypothèse plus faible de convergence quadratique de la fonction de contrainte. Sous certaines conditions de régularité, nous obtenons une borne uniforme sur l'erreur causée par l'approximation Φ_m.

Théorème 2.2.1 (Convergence de l'estimateur AMEM) *Pour toute suite $\{\varphi_m\}_{m\in\mathbb{N}}$ telle que $\varphi_m\|\Phi_m(X)-\Phi(X)\|^2_{\mathbb{L}^2(\mathbb{P}_X)} = O(1)$ et pour toute fonction g continue bornée, l'estimateur AMEM $\hat{\mu}_{m,n}$ vérifie*

$$\left|\int_{\mathcal{X}} g(x)d\hat{\mu}_{m,n}(x) - \int_{\mathcal{X}} g(x)d\mu^*(x)\right| = O(\varphi_m^{-1}) + \kappa_{m,n},$$

avec $\sup_{m\in\mathbb{N}} \kappa_{m,n} = O_P(n^{-1/2})$.

Le cadre des modèles de conditions de moments avec contrainte approchée permet de formaliser des problèmes pour lesquels la connaissance exacte de Φ n'est pas une hypothèse réaliste. On rencontre ce genre de situations par exemple lorsque la fonction de contrainte Φ n'a pas de forme analytique simple et nécessite d'être évaluée numériquement ou d'être estimée. Il est connu que des méthodes d'estimation non-paramétrique telles que les méthodes à noyaux permettent de construire un estimateur de Φ qui converge en norme quadratique. Ainsi, nous montrons qu'il est possible d'utiliser une contrainte de moments obtenue à partir d'une estimation à noyaux afin d'estimer μ_0 par la méthode AMEM. Nous traitons en particulier un problème lié à l'utilisation de variables instrumentales qui peut être interprété commes un modèle de conditions de moments avec contrainte approchée.

Chapitre 3

Nous nous intéressons à un modèle similaire pour lequel la fonction de contrainte Φ dépend d'un paramètre $\theta_0 \in \Theta$ inconnu. Soit $X = (X_1, ..., X_n)$ un échantillon i.i.d. dont la loi μ_0 vérifie la condition de moments suivante

$$\int_{\mathcal{X}} \Phi(\theta_0, x)d\mu_0(x) = 0.$$

Ce modèle très étudié en économétrie se rencontre notamment dans le cadre de l'utilisation de variables instrumentales (voir par exemple [Cha87], [Han82] et [QL94]). Pour estimer le paramètre θ_0, une des principales méthodes développées dans la littérature est la vraisemblance empirique généralisée [QL94]. Le paramètre θ_0 est estimé par le biais d'une mesure discrète

$\hat{\mu}$ qui minimise un critère empirique appelé f-divergence, sous des contraintes de moments. Précisemment, le critère s'exprime sous la forme

$$\mu \mapsto \mathcal{D}_f(\mu, \mathbb{P}_n) = \int f\left(\frac{d\mu}{d\mathbb{P}_n}\right) d\mathbb{P}_n,$$

où f est une fonction convexe qui atteint son minimum en 1. Ainsi, $\hat{\mu}$ est obtenue comme la mesure la plus proche de $\mathbb{P}_n$ (au sens de la f-divergence) qui vérifie la condition de moment $\int \Phi(\hat{\theta}, .)d\hat{\mu} = 0$ pour un $\hat{\theta} \in \Theta$. Cette valeur $\hat{\theta}$ est utilisée pour estimer le paramètre θ_0. Initialement, l'estimateur du maximum de vraisemblance de Owen [Owe91] est construit à partir de la divergence de Kullback, remarquant que minimiser cette divergence est équivalent à maximiser la vraisemblance parmi les lois multinomiales. La méthode a été généralisée en remplaçant la divergence de Kullback par d'autres types de f-divergences, donnant naissance à des estimateurs comme le *continuous updating* de Hansen, Yeaton et Yaron [HHY96] où l'*exponential tilting* de Kitamura et Stutzer [KS97].

Dans ce chapitre, nous étudions une approche Bayésienne basée sur la méthode du maximum d'entropie sur la moyenne, et qui s'avère être étroitement liée à la méthode de vraisemblance empirique généralisée. L'estimation du paramètre se fait à partir de mesures discrètes de la forme $\mathbb{P}_n(w) = \frac{1}{n}\sum_{i=1}^n w_i \delta_{X_i}$. Un a priori ν_0 est fixé sur le vecteur de poids, ce qui permet de considérer chaque mesure discrète $\mathbb{P}_n(w)$ comme la réalisation d'une mesure à poids aléatoires $\mathbb{P}_n(W)$ où W est un vecteur aléatoire de loi ν_0. On définit ensuite la distribution *a posteriori* ν^* comme la projection entropique de ν_0 sur l'ensemble des mesures ν sur $\mathbb{R}^n$ pour lesquelles $\mathbb{P}_n(W)$ vérifie en moyenne sous ν la condition de moment $\mathbb{E}_\nu[\int \Phi(\theta, .)d\mathbb{P}_n(W)] = 0$, pour un $\theta \in \Theta$. La valeur $\hat{\theta}$ pour laquelle la contrainte de moment est vérifiée est l'estimateur MEM de θ_0. Nous montrons qu'il peut s'exprimer de la façon suivante.

Théorème 3.2.2 (Caractérisation de l'estimateur MEM) L'estimateur MEM de θ est donné par

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \sup_{(\gamma,\lambda)\in\mathbb{R}\times\mathbb{R}^k} \left\{\gamma - \Lambda_{\nu_0}(\gamma\mathbb{1} + \lambda^t\Phi(\theta, X))\right\},$$

où $\mathbb{1} = (1, ..., 1)^t \in \mathbb{R}^n$ et $\Lambda_\nu(s) = \log \int \exp\langle s, t\rangle \ d\nu(t)$.

Pour certains choix d'a priori ν_0, il s'avère que l'estimateur MEM correspond à l'estimateur de vraisemblance empirique généralisée, où la f-divergence est entièrement déterminée par le choix de l'a priori. Ainsi, la méthode MEM offre une justification Bayésienne aux f-divergences utilisées pour la méthode de vraisemblance empirique généralisée.

La principale alternative à la vraisemblance empirique est la méthode des moments généralisée, introduite dans ce contexte par Hansen [Han82]. L'estimateur de θ_0 est construit en minimisant le critère empirique

$$\theta \mapsto \left\|\int \Phi(\theta, .)d\mathbb{P}_n\right\|_M^2 = \left[\int \Phi(\theta, .)d\mathbb{P}_n\right]^t M \left[\int \Phi(\theta, .)d\mathbb{P}_n\right]$$

où M est une matrice symétrique définie positive, choisie par le statisticien. Hansen [Han82] et Chamberlain [Cha87] ont montré qu'un choix optimal pour M est la matrice associée au produit scalaire sur $\mathbb{R}^M$ qui rend les composantes de $\Phi(\theta_0, .)$ orthonormées dans $\mathbb{L}^2(\mu_0)$. Nous donnons une nouvelle preuve de ce résultat, qui s'inspire de travaux faits notamment dans [AC03], [AC09] et [CHT08] sur le calcul de bornes d'efficacité dans des modèles de conditions de moments.

Chapitre 4

Nous nous intéressons à un problème de *calage* (calibration en anglais) dans le cadre des sondages. Nous voulons estimer le total t_y d'une variable y dans une population U, donné par $t_y = \sum_{i \in U} y_i$. Les valeurs de y sont observées dans un sous-échantillon $s \subset U$, tiré aléatoirement selon une loi d'échantillonnage $p(.)$. Pour estimer le total t_y, l'estimateur d'Horvitz-Thompson est défini comme le total de y sur l'échantillon s, renormalisé par les inverses des probabilités d'inclusion $d_i = 1/p(i \in s)$,

$$\hat{t}_y^{HT} = \sum_{i \in s} d_i y_i,$$

ce qui en fait un estimateur sans biais sous la loi $p(.)$. Le calage a pour objectif d'améliorer l'estimateur d'Horvitz-Thompson en se servant de variables dites *auxiliaires* $\mathbf{x}_i, i \in U$, observées sur toute la population. On construit un estimateur sous la forme $\hat{t}_y = \sum_{i \in s} w_i y_i$ où les poids w_i sont choisis proches des poids d'Horvitz-Thompson, tout en étant calibrés sur la variable auxiliaire $\mathbf{x}$ en imposant la condition

$$\sum_{i \in s} w_i \mathbf{x}_i = \sum_{i \in U} \mathbf{x}_i.$$

Ici, la variable $\mathbf{x}$ étant observée sur l'ensemble de la population, la quantité $t_{\mathbf{x}} = \sum_{i \in U} \mathbf{x}_i$ est connue par le statisticien. Ainsi, l'estimateur par calage est défini à partir de poids $\hat{w}_i$ minimisant un certain critère arbitraire avec les poids d'Horvitz-Thompson, sous la contrainte de calage. Ce problème peut être interprété comme un problème inverse pour lequel on cherche à retrouver une mesure discrète sur s, identifiée au vecteur des poids $\{w_i\}_{i \in s}$, satisfaisant une contrainte linéaire. Il est alors possible d'appliquer la méthode du maximum d'entropie sur la moyenne.

Nous montrons que les méthodes usuelles de calage conduisent à des estimateurs obtenus par maximum d'entropie sur la moyenne, pour des choix particuliers d'a priori. Ainsi, cette approche permet d'apporter une justification Bayésienne à la plupart des critères arbitrairement définis en calage. En étudiant le comportement asymptotique de ces estimateurs, nous établissons des conditions d'efficacité en fonction de l'a priori choisi. Nous nous intéressons également à la question de la variable auxiliaire optimale, où l'on s'autorise à modifier la variable auxiliaire afin qu'elle explique au mieux la variable d'intérêt. Cette approche, connue sous le nom de *model calibration* [WS01], consiste à estimer dans un cadre paramétrique ou non-paramétrique, la variable auxiliaire optimale en parallèle du procédé de calage. Nous montrons que ce modèle est lié à des problèmes inverses avec opérateur approché, qui peuvent justifier l'utilisation d'un point de vue Bayésien.

Chapitre 5

Nous nous intéressons à un problème de régression non-paramétrique avec données indirectes. Nous cherchons à estimer une fonction x_0 appartenant à un espace de Hilbert $\mathcal{X}$ à partir de l'observation bruitée

$$y = A_n x_0 + \varepsilon,$$

où $A_n : \mathcal{X} \to \mathbb{R}^n$ est un opérateur linéaire discret. De nombreuses méthodes de régularisation font intervenir la *décomposition en valeurs singulières* de l'opérateur A_n, où l'estimation de la fonction d'intérêt se fait à partir des coefficients de x_0 dans la base de vecteurs propres de l'opérateur auto-adjoint $A_n^* A_n$. De cette façon, le modèle peut s'écrire comme un modèle de régression hétéroscédastique où chaque coefficient x_i peut être estimé indépendamment des autres,

$$y_i = x_i + \eta_i, i = 1, ..., n,$$

où les η_i sont des bruits centrés, orthogonaux deux à deux et de variances respectives $\sigma_i^2 \sim 1/nb_i^2$, inversement proportionnelles aux valeurs propres b_i^2 de $A_n^* A_n$. Ainsi, le signal étant très affecté par le bruit dans les directions correspondantes à des petites valeurs propres, estimer x_i par y_i s'avère inefficace. La plupart des méthodes de régularisation consistent alors à considérer des estimateurs de la forme $\widehat{x}_i = \lambda_i y_i$ où $\lambda_i \in [0; 1]$ est un coefficient appelé *filtre*. C'est le cas par exemple de la méthode de Tikhonov [TA77] associée aux filtres dits de Wiener $\lambda_i = (1 + \theta\sigma_i^2)^{-1}$ ou du *spectral cut-off* [Han87] correspondant à des filtres de la forme $\lambda_i = \mathbb{1}\{\sigma_i^2 \leq \theta\}$ où θ est un paramètre d'ajustement. Par ailleurs, le choix du paramètre d'ajustement peut se faire par des méthodes permettant de manipuler des classes générales de filtres, telles que l'*enveloppe du risque* ou l'*estimation sans biais du risque* étudiées dans [Cav08], [CG06] et [CGPT00].

Dans ce chapitre, nous nous intéressons à des estimateurs dits de *projection* liés à des méthodes de seuillage, obtenus avec des filtres de la forme $\lambda_i = \mathbb{1}\{i \in m\}$, pour un certain modèle $m \subset \{1, ..., n\}$. Ainsi, la méthode revient à sélectionner les variables pertinentes y_i. Pour un modèle m donné, notons $\widehat{x}_m$ l'estimateur associé. Nous montrons facilement que le modèle m^* qui minimise le risque quadratique $m \mapsto \mathbb{E}\|x_0 - \widehat{x}_m\|^2$ est donné par $m^* = \{i : x_i^2 \geq \sigma_i^2\}$. Cependant, l'*oracle* m^* étant inconnu, il s'agit d'une certaine façon de l'estimer. Pour cela, nous considérons un modèle de la forme $\widehat{m} = \{i : y_i^2 \geq c_i\}$ où c_i est un seuil à déterminer. Alors que des seuils proportionnels à la variance σ_i^2 ont été étudiés dans la littérature (voir [AS98], [CGPT00] ou [Lou08]), nous montrons qu'un terme logarithmique de la variance permet de contrôler le risque de l'estimateur dans les cas de faible régularité.

Théorème 5.3.1 (Inégalité oracle) Supposons qu'il existe des constantes positives K, β telles que $\mathbb{E}[\exp(\eta_i^2/\beta\sigma_i^2)] \leq K$. Soit $\theta > 0$, et $c_i = 4\sigma_i^2\beta\log(e + \theta\sigma_i^2)$. L'estimateur $\hat{x}_{\widehat{m}}$ vérifie

$$\mathbb{E}\|\hat{x}_{\widehat{m}} - x_0\|^2 \leq \mathbb{E}\|\hat{x}_{m^*} - x_0\|^2 + \big(6\beta\log(e + \theta\|x_0\|^2) + 2\big)\sum_{i\in m^*}\sigma_i^2 + \frac{2K\beta n}{\theta}.$$

De la même manière que dans [CH05], nous étudions également l'application de cette méthode dans une situation où nous observons une version bruitée $\hat{b}_i = b_i + \xi_i$ de chaque valeur propre. En s'inspirant des résultats de Cavalier et Hengartner [CH05], nous proposons un estimateur de projection construit à partir d'un seuillage simultané sur les coefficients y_i et les valeurs propres

$\hat{b}_i$. Nous obtenons une majoration du risque de l'estimateur conditionnellement aux observations $\hat{b}_i$ et une vitesse de convergence proche de celle de l'oracle.

Conclusion

Deux types de problèmes inverses sont étudiés dans cette thèse. Dans un premier temps, nous nous intéressons dans les chapitres 2, 3 et 4, à des problèmes liés aux modèles de condition de moments. Les contributions faites dans ces chapitres sont de deux ordres. Premièrement, nous fournissons une justification Bayésienne aux méthodes naturelles utilisées pour la résolution des problèmes inverses de conditions de moment, par le biais du maximum d'entropie sur la moyenne. Deuxièmement, nous tentons de généraliser ces modèles à des situations où l'opérateur n'est pas entièrement connue. Nous traitons notamment le cadre des conditions de moments paramétriques rencontré en économétrie, faisant le lien entre les méthodes de vraisemblance empirique et le maximum d'entropie. Nous étudions également des problèmes d'estimation de paramètres linéaires (e.g. une moyenne ou un total sur une population) en sondages, pour lesquels les méthodes usuelles de calages s'interprètent également comme des méthodes de maximum d'entropie.

Le chapitre 5 de cette thèse traite d'un problème inverse plus général qui peut se rencontrer notamment en régression non-paramétrique avec observations indirectes. Nous proposons un estimateur adaptatif basé sur une méthode de seuillage. Alors que la plupart des méthodes de seuillages étudiées dans la littérature utilisent un seuil proportionnel à la variance, nous proposons un seuil différent qui fait faisant intervenir un terme logarithmique de la variance. Nous montrons notamment que ce seuillage permet de s'adapter à tous les degrés de régularité en problèmes inverses, ainsi qu'à des situations où l'opérateur est inconnu.

Chapter 1

Preliminary results

1.1 Generalities on inverse problems

For the statistician, an inverse problem deals with the estimation of a parameter f from a noisy observation of the image of f through an operator A:

$$g = Af + \varepsilon. \tag{1.1}$$

To estimate efficiently f from g relies on the regularity of the operator, as well as on the distribution of the noise. In the case where A is a linear operator, an inverse problem is defined in [EHN96] as *well-posed* if it satisfies the three Hadamard's conditions,

H1. For all observation g, a solution h such that $g = Ah$ exists.

H2. For all observation g, the solution is unique.

H3. The solution depends continuously of the observations.

These conditions can be expressed in term of assumptions on the operator A. Let $\mathcal{F}$ denote the space of solutions, $\mathcal{G}$ the space of the observations g and $A : \mathcal{F} \to \mathcal{G}$. The conditions of existence and uniqueness of a solution for a given observation g correspond respectively to assumptions of surjectivity and injectivity of A. In the same way, the third Hadamard condition is equivalent to the continuity of A^{-1}, provided that the inverse exists. Thus, it appears that the problem (1.1) is well-posed if the operator A is sufficiently regular. Of course, the notion of regularity depends on the spaces $\mathcal{F}$ and $\mathcal{G}$ on which the operator A is defined, as well as the topologies on these spaces, that define the continuity. When one or several Hadamard's conditions are not met, the inverse problem is said to be *ill-posed*. In this case, estimating f requires the problem to be *regularized*, i.e., to be modified in order to satisfy all three Hadamard's conditions.

In this thesis, we are interested in the regularization of *discrete* inverse problems, for which the space $\mathcal{G}$ has finite dimension. In such problems, Hadamard's first condition of well-posedness is generally verified or can be regularized without difficulty. Indeed, if the condition is not met, one may simply reduce the dimension of the observation g in order to make the operator surjective. The second condition of regularity causes more problems. In particular, if f belongs to an infinite dimensional space, a discrete inverse problem is *always* ill-posed, due to the non-injectivity of the operator A. An observation g may correspond to an infinite number of solutions which are impossible to distinguish without further information on the true value f. In such cases, the regularization of Hadamard's condition of injectivity generally reduces to finding a criterion to select, more or less arbitrarily, of a solution in a set of possible candidates. The choice can be made for computational reasons (e.g. the solution with minimal norm) or based on prior information using Bayesian methods.

1.2 Moment condition models

A particular example of discrete inverse problems is given by moment condition models, which are discussed in the chapters 2, 3 and 4 of this thesis. In the literature, moment condition models are studied in [GG91], [GG97] and in many applied fields such that image denoising [HN00], spectroscopy [Ski88], crystallography [DHLN92], tomography [FLLn06], survey sampling [DS92] and Econometry [Cha87], [Han82], [Owe91].

Let $\mathcal{X}$ be an open subset of $\mathbb{R}^p$ endowed with its Borel field $\mathcal{B}(\mathcal{X})$ and let $\mathcal{P}(\mathcal{X})$ (resp. $\mathcal{F}(\mathcal{X})$) denote the set of all probability measures (resp. finite measures) on $\mathcal{X}$. We are interested in recovering a finite measure μ_0 on $\mathcal{X}$ from generalized moments of μ_0,

$$y = \int_{\mathcal{X}} \Phi(x) d\mu_0(x),$$

where $y \in \mathbb{R}^k$ and $\Phi : \mathcal{X} \to \mathbb{R}^k$ is a known continuous vector valued map. In most cases, the measure μ_0 to recover is a probability measure although, for sake of generality, we shall only assume here that μ_0 is a finite measure. Remark that imposing to μ_0 to be a probability measure can be easily integrated as a moment condition. This framework can be viewed as an inverse problem with parameter of interest the finite measure μ_0 and with linear operator $\mu \mapsto \int \Phi d\mu$. The existence of several finite measures μ such that $\int \Phi d\mu = y$ makes the problem ill-posed. On the other hand, the surjectivity condition is fulfilled as soon as the components of Φ are linearly independent, which we assume here. Hence, in this framework where the injectivity of the operator is the only unverified Hadamard condition, to regularize the ill-posed inverse problem means to associate to one (possibly noisy) observation y, a unique solution $\hat{\mu}$. To this purpose, many regularization techniques have been suggested in the literature, where a solution is obtained as the minimizer μ of a convex functional subject to the linear constraint $\int \Phi d\mu = y$ when y is observed, or more generally, subject to a convex constraint of the form $\int \Phi d\mu \in K_Y$ if we observe a noisy version of y, for some convex set K_Y reflecting the uncertainty due to the noise. Here and in the sequel, we study regularization methods that rely on a discretization of the space $\mathcal{X}$. More precisely, assume we observe n points $X_1, ..., X_n$ in $\mathcal{X}$ and denote by $\mathbb{P}_n$ the empirical distribution

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

where δ stands for the Dirac mass. We assume that the discrete measure $\mathbb{P}_n$ converges weakly toward a possibly unknown measure $\mathbb{P}_X$, having full support on $\mathcal{X}$. The discrete measure $\mathbb{P}_n$

is used as a reference measure to estimate μ_0. In this framework, there are basically two types of situations. The X_i's may be i.i.d. realizations of a random variable X with distribution $\mathbb{P}_X$, or a deterministic design, in which case $\mathbb{P}_X$ is generally known by the statistician. In the case of a random design, the limit distribution is implicitly assumed to be "close" to the true measure μ_0, to justify the use of $\mathbb{P}_n$ as a reference (a common situation is the convenient case $\mathbb{P}_X = \mu_0$). If on the contrary, no random design is suggested by the problem, one can use a uniform discretization of the space $\mathcal{X}$ or a design that reflects some prior knowledge on μ_0.

A particular example of moment condition model involving a discrete space is encountered in linear parameter estimation problems in survey sampling. Consider a large population $U = \{1, ..., N\}$ and an unknown variable $y = (y_1, ..., y_N) \in \mathbb{R}^N$. The objective is to estimate its total $t_y := \sum_{i \in U} y_i$ when only a random subsample $s \subset U$ of size n is available. An unbiased estimate of t_y is given by the Horvitz-Thompson estimator, which can be expressed as the renormalized sum $\hat{t}_y^{HT} = \frac{N}{n} \sum_{i \in s} y_i$ if the sample s is drawn uniformly over U, or more generally by

$$\hat{t}_y^{HT} = \sum_{i \in s} d_i y_i,$$

where $d_i = 1/\mathbb{P}(i \in s)$ is the inverse of the inclusion probability (the Horvitz-Thompson weight). Suppose that it exists an auxiliary vector variable $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_N)$ observed over the entire population and set $t_{\mathbf{x}} = \sum_{i \in U} \mathbf{x}_i \in \mathbb{R}^k$. If the sample is not representative, it is likely that the Horvitz-Thompson estimator of $t_{\mathbf{x}}$ will be far from its true value. Thus, to deal with possible bad sample effects, a natural idea is to modify the Horvitz-Thompson weights in order that the weighted total of $\mathbf{x}$ over s is equal to the total over the whole population U. This is the underlying idea of the calibration method developped in [DS92]. An estimator of t_y is constructed as a weighted total $\sum_{i \in s} w_i y_i$ over the sample s, with the weights $\{w_i\}_{i \in s}$ satisfying the condition

$$\sum_{i \in s} w_i \mathbf{x}_i = t_{\mathbf{x}}.$$

Viewing the collection $\{w_i\}_{i \in s}$ as a discrete measure on the sample s, the previous equality can be interpreted as a moment condition.

1.2.1 Entropic projection

In this section, we discuss entropic projections as solutions to the moment condition problem. Let us first introduce some definitions and notations. We endow the set of finite measures $\mathcal{F}(\mathcal{X})$ with the total variation topology, induced by the following distance

$$d(\mu, \nu) = \|\mu - \nu\|_{TV} = \sup_{A \in \mathcal{B}(\mathcal{X})} |\mu(A) - \nu(A)|, \ \mu, \nu \in \mathcal{F}(\mathcal{X}).$$

Definition Let ν and μ denote two finite measures on $\mathcal{X}$, the *relative entropy*, or *Kullback divergence*, of ν with respect to μ is defined as

$$\mathcal{K}(\nu|\mu) = \int_{\mathcal{X}} \log\left(\frac{d\nu}{d\mu}\right) d\nu + \mu(\mathcal{X}) - \nu(\mathcal{X}) \ \text{ if } \nu \ll \mu, \quad \mathcal{K}(\nu|\mu) = +\infty \text{ otherwise.}$$

Although the relative entropy is not a metric in the mathematical sense (in particular it is not symmetric), it somehow measures the "closeness" between two finite measures. The functional $\nu \mapsto \mathcal{K}(\nu|\mu)$ is non-negative, strictly convex on the set $\{\nu \in \mathcal{F}(\mathcal{X}) : \nu \ll \mu\}$ and is null only for $\nu = \mu$ (see [Csi75]). Csiszár generalizes the relative entropy by introducing the notion of f-divergences.

Definition Let ν and μ denote two finite measures on $\mathcal{X}$. Let $f : \mathbb{R} \to \overline{\mathbb{R}}_+$ be a strictly convex function, twice differentiable on its domain, such that $f(1) = f'(1) = 0$, the f-divergence of ν with respect to μ is defined as

$$\mathcal{D}_f(\nu|\mu) = \int_{\mathcal{X}} f\left(\frac{d\nu}{d\mu}\right) d\mu \ \text{ if } \nu \ll \mu, \quad \mathcal{D}_f(\nu|\mu) = +\infty \ \text{ otherwise.}$$

In particular, the relative entropy $\mathcal{K}(\nu|\mu)$ is the f-divergence associated to the function $f(x) = x \log x - x + 1$ if $x > 0$ and $f(x) = +\infty$ if $x \leq 0$, while the relative entropy with reversed arguments $\mathcal{K}(\mu|\nu)$ is the f-divergence corresponding to $f(x) = -\log x + x - 1$ if $x > 0$ and $f(x) = +\infty$ if $x \leq 0$. In both cases, the function satisfies $f(1) = f'(1) = 0$. The conditions made on f warrant to general f-divergences the same properties as relative entropy, namely $\nu \mapsto \mathcal{D}_f(\nu|\mu)$ is non-negative, strictly convex on the set $\{\nu \in \mathcal{F}(\mathcal{X}) : \nu \ll \mu\}$ and null only for $\nu = \mu$. For all subset C of $\mathcal{F}(\mathcal{X})$, we shall note

$$\mathcal{D}_f(C|\mu) = \inf_{\nu \in C} \mathcal{D}_f(\nu|\mu).$$

Definition Let μ be a probability measure on $\mathcal{X}$ and C a subset of $\mathcal{P}(\mathcal{X})$ such that $\mathcal{D}_f(C|\mu)$ is finite. We call *entropic projection* of μ onto C relative to $\mathcal{D}_f$, any measure ν in the closure of C for the total variation topology, such that $\mathcal{D}_f(\nu|\mu) = \mathcal{D}_f(C|\mu)$.

If C is convex, the entropic projection is unique as a consequence of Csiszár's Theorem 2.1 in [Csi75]. The entropic projection relative to the Kullback divergence is called *I-projection.*

In moment conditions models, f-divergences appear as natural tools for regularization methods. We recall that our problem is to recover a probability measure μ_0 satisfying a moment condition $\int \Phi d\mu = y$ to which we observe a noisy version $y^{obs} = y + \varepsilon$. Under a certain control over the noise, we may consider a moment condition of the form $\int \Phi d\mu \in K_Y$ for some arbitrary set K_Y containing y^{obs}. If K_Y is appropriately chosen, the true measure μ_0 satisfies the moment condition with high probability, or with probability one if ε is bounded almost surely. Thus, the solution to the inverse problem is sought in the set

$$\mathcal{M} = \{\mu \in \mathcal{F}(\mathcal{X}) : \textstyle\int \Phi d\mu \in K_Y\},$$

which represents in some way the set of possibles values for μ_0. The regularization now consists in choosing a unique solution in the set $\mathcal{M}$ of candidates. Based on the idea that μ_0 must be close to the empirical distribution $\mathbb{P}_n$, a natural solution to the inverse problem can be obtained as an entropic projection of $\mathbb{P}_n$ onto $\mathcal{M}$. By construction an entropic projection is absolutely continuous with respect to the reference measure. Therefore, an entropic projection of $\mathbb{P}_n$, if one exists, is expressed as a weighted version $\mathbb{P}_n(w)$ of $\mathbb{P}_n$, where

$$\mathbb{P}_n(w) = \frac{1}{n} \sum_{i=1}^{n} w_i \delta_{X_i},$$

for $w = (w_1, ..., w_n)^t \in \mathbb{R}^n$ a collection of weights with $w_i \geq 0$ for all $i = 1, ..., n$. In some cases, relaxing the positivity condition on the weights might ease the computation of the solution, at the cost of allowing it to be a signed measure.

Definition The problem is said to be *feasible* relative to $\mathcal{D}_f$ if there exists a vector $\delta = (\delta_1, ..., \delta_n)^t \in \mathbb{R}^n$ such that $f(\delta_i)$ is finite for all $i = 1, ..., n$ and $\mathbb{P}_n(\delta) \in K_Y$.

The feasibility condition warrants the existence of an entropic projection of $\mathbb{P}_n$ onto $\mathcal{M}$. It is more likely to be fulfilled if the domain of f, $\text{dom}(f) = \{x \in \mathbb{R} : f(x) < \infty\}$ is widespread. On the other hand, the uniqueness of the solution relies on the convexity of the set K_Y, as we see in the following lemma.

Lemma 1.2.1 *Assume that the problem is feasible relative to $\mathcal{D}_f$. If K_Y is a closed convex subset of $\mathbb{R}^k$, then the solution $\hat{\mu} = \arg\min_{\mu \in \mathcal{M}} \mathcal{D}_f(\mu|\mathbb{P}_n)$ exists and is unique.*

Proof. Consider the set $\mathcal{M}_n$ of discrete probability measures satisfying the moment condition $\mathcal{M}_n = \{\mathbb{P}_n(w) : \frac{1}{n}\sum_{i=1}^n w_i \Phi(X_i) \in K_Y\}$. The solution $\hat{\mu}$ is the entropic projection of $\mathbb{P}_n$ onto $\mathcal{M}_n$. Using Csiszár's Theorem 2.1 in [Csi75], it suffices to show that $\mathcal{M}_n$ is convex, closed for the total variation topology and satisfies $\mathcal{D}_f(\mathcal{M}_n|\mathbb{P}_n) < \infty$. For this, first remark that the total variation distance defines a norm on the finite dimensional the linear space $\{\mathbb{P}_n(w) : w \in \mathbb{R}^n\}$. So, we deduce that $\mathcal{M}_n$ is closed and convex, as the inverse image of K_Y through the linear operator $\mathbb{P}_n(w) \mapsto \frac{1}{n}\sum_{i=1}^n w_i \Phi(X_i)$. Moreover, the feasibility conditions yields

$$\mathcal{D}_f(\mathcal{M}_n|\mathbb{P}_n) \leq \mathcal{D}_f(\mathbb{P}_n(\delta)|\mathbb{P}_n) = \frac{1}{n}\sum_{i=1}^{n} f(\delta_i) < \infty,$$

which ends the proof.

One important aspect of the regularization by entropic projection is the choice of the f-divergence. The relative entropy plays an important role among f-divergences as it can be given a particular interpretation via empirical likelihood. If we assume that the discretization $X_1, ..., X_n$ is an i.i.d. sample drawn from μ_0, the probability of observing the sample $X_1, ..., X_n$ is $\prod_{i=1}^n \mu_0(X_i)$. Thus, to estimate μ_0, one may reasonably be concerned with finding the measure maximizing this probability, or equivalently, maximizing the nonparametric log-likelihood function defined on $\mathcal{P}(\mathcal{X})$ by

$$\mu \mapsto \ell(\mu, X_1, ..., X_n) = \sum_{i=1}^{n} \log \mu(X_i) = -n\left(\mathcal{K}(\mathbb{P}_n|\mu) + \log n\right).$$

This equality points out that maximizing the empirical likelihood reduces to minimizing the relative entropy $\mu \mapsto \mathcal{K}(\mathbb{P}_n|\mu)$. When no restriction are made on μ, the maximum empirical likelihood in the nonparametric model $\mathcal{P}(\mathcal{X})$ is achieved for the empirical distribution $\mathbb{P}_n$. When applied to our model, this equality tells us that the regularization solution obtained with the Kullback divergence is actually the probability measure maximizing the empirical likelihood in the set $\mathcal{M}$ of possible values of μ_0.

For $f : \mathbb{R}^q \to \overline{\mathbb{R}}_+$ a convex function, we denote by f^* its *convex conjugate* (or *Fenchel-Legendre transform*) given by

$$f^*(u) = \sup_{v \in \mathbb{R}^q} \{u^t v - f(v)\},$$

where u^t denotes the transpose of u and $u^t v$ is the usual inner product on $\mathbb{R}^q$.

Theorem 1.2.2 (Characterization of the entropic projection) *Suppose that the problem is feasible relative to $\mathcal{D}_f$ and let K_Y be a closed convex subset of $\mathbb{R}^k$. Let $\hat{\mu} = \mathbb{P}_n(\hat{w})$ be the entropic projection of $\mathbb{P}_n$ onto $\mathcal{M}$ relative to $\mathcal{D}_f$. If the solution lies in the interior of the domain of f, then we have*

$$d\hat{\mu}(x) = f^{*\prime}(\hat{\lambda}^t \Phi(x))\, d\mathbb{P}_n(x),$$

where $\hat{\lambda}$ is the maximizer over $\mathbb{R}^k$ of

$$\lambda \mapsto \frac{1}{n} \sum_{i=1}^{n} f^*(\lambda^t \Phi(X_i)) - \inf_{z \in K_Y} \lambda^t z.$$

Proof. This is a classical convex optimization problem. The proof relies on some properties of the convex conjugate f^* which are given in [Roc97]. For a fixed $z \in K_Y$, note $\hat{w}(z)$ the solution of the minimization problem subject to the constraint $\frac{1}{n}\sum_{i=1}^n w_i \Phi(X_i) = z$. To avoid considering trivial cases, we take $z \in K_Y$ such that $\hat{w}(z)$ exists and has all its components $\hat{w}_i(z)$ in the interior of the domain of f. The Lagrangian is given by

$$\mathcal{L}_z(\lambda, w) = \sum_{i=1}^{n} f(w_i) - \lambda^t \left(\sum_{i=1}^{n} w_i \Phi(X_i) - nz \right),$$

where λ is the Lagrange multiplier. The first order conditions are $f'(w_i) = \lambda^t \Phi(X_i)$. The function f' is increasing by assumptions, yielding $\hat{w}_i(z) = f'^{-1}(\lambda_z^t \Phi(X_i)) = f^{*\prime}(\lambda_z^t \Phi(X_i))$ where λ_z satisfies $\sum_{i=1}^n [f^{*\prime}(\lambda_z^t \Phi(X_i))\Phi(X_i) - z] = 0$. By convexity of f^*, λ_z is the unique minimizer over $\mathbb{R}^k$ of

$$\lambda \mapsto \sum_{i=1}^{n} f^*(\lambda^t \Phi(X_i)) - n\lambda^t z.$$

Using that $f(f^{*\prime}(x)) = x f^{*\prime}(x) - f^*(x)$, we find

$$\begin{aligned} \mathcal{D}_f(\mathbb{P}_n(\hat{w}(z)) | \mathbb{P}_n) &= \lambda_z^t \sum_{i=1}^{n} f^{*\prime}(\lambda_z^t \Phi(X_i))\Phi(X_i) - \sum_{i=1}^{n} f^*(\lambda_z^t \Phi(X_i)) \\ &= \sup_{\lambda \in \mathbb{R}^k} n\lambda^t z - \sum_{i=1}^{n} f^*(\lambda^t \Phi(X_i)). \end{aligned}$$

Writing $\mathcal{D}_f(\mathcal{M}|\mathbb{P}_n) = \inf_{z \in K_Y} \mathcal{D}_f(\mathbb{P}_n(\hat{w}(z))|\mathbb{P}_n)$, we conclude using Sion's minimax theorem.

Entropic projections are natural regularization solutions to moment condition problems. However, while the use of the Kullback entropy can be justified by a maximum likelihood argument, other f-divergences usually lack a probabilistic justification. In the next section, we see that f-divergences may be given an interpretation in a Bayesian setting.

1.2.2 Maximum entropy on the mean

Maximum entropy on the mean (MEM) is a Bayesian approach to inverse problems. As previously, a solution is sought as a weighted version $\mathbb{P}_n(w)$ of the empirical measure. The problem is treated as a parametric estimation problem where the parameter of interest is the vector of weights w. The procedure is the following. Fix a prior distribution ν_0 on w, viewing each discrete measure $\mathbb{P}_n(w)$ as a realization of the random measure $\mathbb{P}_n(W)$, where W is drawn from ν_0 (see [Goz05]). The distribution ν_0 must be chosen to reflect some prior knowledge on the shape or support of μ_0 (see for instance the discussion in [GG97]). Here, the support $\{X_1, ..., X_n\}$ of the random measure $\mathbb{P}_n(W)$ is considered fixed, only the variable W is random. The next step of the MEM procedure is to build a posterior distribution ν^*. The idea is to slightly modify the prior ν_0 so that the expectation of $\mathbb{P}_n(W)$ satisfies the moment condition in mean. In this way, the posterior distribution ν^* remains close to the prior while providing a suitable solution to the inverse problem. Denote by $\mathcal{P}(\mathbb{R}^n)$ the set of probability measures on $\mathbb{R}^n$ and let

$$\Pi(\mathcal{M}) = \{\nu \in \mathcal{P}(\mathbb{R}^n) : \mathbb{E}_\nu \left[\mathbb{P}_n(W)\right] \in \mathcal{M}\},$$

where $\mathbb{E}_\nu \left[\mathbb{P}_n(W)\right]$ is the expectation of $\mathbb{P}_n(W)$ when W is drawn from ν. We define the posterior distribution ν^* as the I-projection of ν_0 onto $\Pi(\mathcal{M})$, that is, ν^* is the minimizer of the Kullback entropy $\mathcal{K}(.|\nu_0)$ subject to $\nu^* \in \Pi(\mathcal{M})$. The MEM estimator $\hat{\mu}$ is then obtained as

$$\hat{\mu} = \mathbb{E}_{\nu^*} \left[\mathbb{P}_n(W)\right].$$

The choice of the Kullback divergence to construct the posterior distribution is motivated by large deviation principles, stating that under regularity conditions, the MEM estimator converges weakly toward the same limit as the somehow more classical Bayesian construction

$$\hat{\mu}^{bay} = \mathbb{E}_{\nu_0} \left[\mathbb{P}_n(W)|\mathbb{P}_n(W) \in \mathcal{M}\right],$$

whenever it is well defined, where for an event $\mathcal{A}$, $\mathbb{E}\left[X|\mathcal{A}\right]$ denotes the expectation of X conditionally to $\mathcal{A}$. The proof of this result is given in Theorem 2.3 in [GG97].

Maximum entropy on the mean provides a natural and easy way of incorporating some prior information to the problem. The next theorem, due to Gamboa and Gassiat [GG91], shows that specific choices of prior in the MEM methodology lead to entropic projection solutions. For some probability measure ν on $\mathbb{R}^q$, $q \geq 1$, we shall note $\Lambda_\nu : \mathbb{R}^q \to \overline{\mathbb{R}}$ the log-Laplace transform of ν:

$$\Lambda_\nu(s) = \log \int_{\mathbb{R}^q} \exp(s^t x)\; d\nu(x), \;\; s \in \mathbb{R}^q.$$

The convex conjugate of Λ_ν, noted Λ_ν^*, is called the *Cramer transform* of ν.

Theorem 1.2.3 (Expression of the MEM estimate) *Take ν_0 with orthogonal marginals, i.e. $\nu_0 = \nu^{\otimes n}$ for ν a probability measure on $\mathbb{R}$ with mean 1. Assume there exists $\delta = (\delta_1, ..., \delta_n)^t \in \mathbb{R}^n$ with the δ_i's in the interior of the convex hull of the support of ν, such that $\mathbb{P}_n(\delta) \in K_Y$. Then, the MEM estimator $\hat{\mu}$ is given by*

$$\hat{\mu} = \arg\min_{\mu \in \mathcal{M}} \int_{\mathcal{X}} \Lambda_\nu^* \left(\frac{d\mu}{d\mathbb{P}_n} \right) d\mathbb{P}_n.$$

As pointed out in this theorem, the MEM solution for a prior of the form $\nu_0 = \nu^{\otimes n}$ is an entropic projection, with f-divergence associated to the Cramer transform of ν. Remark that the assumptions made on ν correspond to the required regularity conditions for Λ_ν^* to define a f-divergence. For instance, assuming that ν is a probability measure with mean 1 warrants that $\Lambda_\nu^*(1) = {\Lambda_\nu^*}'(1) = 0$ and therefore $\Lambda_\nu^* \geq 0$ by convexity. Moreover, the existence of $\delta \in \mathbb{R}^n$ satisfying the assumption of the theorem is equivalent to the feasibility condition, as the interior of the convex hull of the support of ν is actually the interior of the domain of Λ_ν^*.

Proof. Write $\hat{\mu} = \mathbb{P}_n(\hat{w})$, we shall prove the equivalent statement

$$\hat{w} = \arg\min_{w \in S} \sum_{i=1}^n \Lambda_\nu^*(w_i),$$

where $S = \{w \in \mathbb{R}^n : \frac{1}{n}\sum_{i=1}^n w_i \Phi(X_i) \in K_Y\}$. For $w \in \mathbb{R}^n$, let ν_w be the I-projection of ν_0 onto the convex set $\mathcal{F}_w = \{\mu \in \mathcal{P}(\mathbb{R}^n) : \mathbb{E}_\mu(W) = w\}$. First, we want to prove the preliminary result

$$\mathcal{K}(\nu_w|\nu_0) = \inf_{\mu \in \mathcal{F}_w} \mathcal{K}(\mu|\nu_0) = \Lambda_{\nu_0}^*(w).$$

Note $\rho_w = d\nu_w/d\nu_0$, we write the Lagrangian

$$\mathcal{L}(\lambda, \gamma, \rho) = \int_{\mathcal{X}} \rho(\log\rho - 1) d\nu_0 - \lambda \left[\int_{\mathcal{X}} \rho \; d\nu_0 - 1 \right] - \gamma^t \left[\int_{\mathcal{X}} \tau\rho(\tau) d\nu_0(\tau) - w \right],$$

where $(\gamma, \lambda) \in \mathbb{R} \times \mathbb{R}^n$ is the Lagrange multiplier. The first order condition gives $\log\rho(\tau) = \gamma + \lambda^t\tau$. We deduce $\rho_w(\tau) = \exp(\gamma_w + \lambda_w^t\tau)$ where $\gamma_w = -\Lambda_{\nu_0}(\lambda_w)$ is the normalizing constant and λ_w is such that $\int \tau \exp(-\Lambda_{\nu_0}(\lambda_w) + \lambda_w^t\tau) d\nu_0(\tau) = \Lambda'_{\nu_0}(\lambda_w) = w$. We get after integrating with respect to λ_w,

$$\lambda_w = \arg\max_{\lambda \in \mathbb{R}^n} \; \lambda^t w - \Lambda_{\nu_0}(\lambda).$$

Writing $\mathcal{K}(\nu_w|\nu_0) = \int \rho_w(\log\rho_w - 1)d\nu_0 + 1$, we find

$$\mathcal{K}(\nu_w|\nu_0) = \int_{\mathcal{X}} (\gamma_w + \lambda_w^t\tau - 1)\exp(\gamma_w + \lambda_w^t\tau) d\tau + 1 = -\Lambda_{\nu_0}(\lambda_w) + \lambda_w^t w = \Lambda_{\nu_0}^*(w).$$

The prior being of the form $\nu_0 = \nu^{\otimes n}$, we have that $\Lambda_{\nu_0}^*(w) = \sum_{i=1}^n \Lambda_\nu^*(w_i)$. We conclude noticing that $\hat{w} = \mathbb{E}_{\nu^*}(W) = \arg\min_{w \in S} \; \mathcal{K}(\nu_w|\nu_0)$.

1.3 Nonparametric indirect regression

In this section, we are interested in nonparametric regression in inverse problems. Precisely, assume we want to estimate a function x_0 in a Hilbert space $\mathcal{X}$, in the following model

$$y(t) = (Ax_0)(t) + \varepsilon(t),$$

where the observation $y(.)$ belongs to a functional space $\mathcal{Y}$, $\varepsilon(.)$ is the noise and $A : \mathcal{X} \to \mathcal{Y}$ is a known linear operator. This model has been extensively studied in the literature in inverse problem, see for instance [BHMR07], [CGPT00], [FLn08], [HO93], [Lou08], [TA77] and [Tsy03].

Here, we focus on the regularization of Hadamard's third condition, i.e., we consider situations where the ill-posedness is due to the instability of the solution. Assume that $\mathcal{Y}$ is a locally convex topological vector space and A is an injective compact operator, with discontinuous pseudo inverse $A^\dagger = (A^*A)^{-1}A^*$, A^* standing for the adjoint of A. The direct use of the pseudo inverse leads to a solution $\hat{x} = A^\dagger y = x_0 + A^\dagger \varepsilon$ having a potentially infinite variance. To overcome this issue, regularization methods aim to replace the discontinuous pseudo inverse $A^\dagger$ by a continuous version. The definitions and results presented in this section can be found in more details in [EHN96].

1.3.1 The discrete case

We are interested in a discretized version of the problem, where we observe only the evaluation of $y(.)$ at a finite number of points $t_1, ..., t_n$. This problem is studied in the chapter 5 of this thesis. It can be formalized as a discrete inverse problem, involving the operator $A_n : x \mapsto (Ax(t_1), ..., Ax(t_n))^t$. So, note $y = (y(t_1), ..., y(t_n))^t$ and $\varepsilon = (\varepsilon(t_1), ..., \varepsilon(t_n))^t$, we want to recover the function x_0 from the observation

$$y = A_n x_0 + \varepsilon.$$

We assume that the noise ε is centered with known variance $\sigma^2 I$ where I denotes the identity matrix in $\mathbb{R}^n$. Moreover, we assume for simplicity that the design $(t_1, ..., t_n)$ is such that A_n is surjective (if not, one may simply remove some observations to make it surjective). So, let K_n denote the kernel of A_n, its orthogonal $K_n^\perp$ is of dimension n. Clearly, the discretized problem no longer satisfies the injectivity condition. However, this issue is quite easily solved by restricting the set of solutions to $K_n^\perp$, making the operator one-to-one. Actually, if A_n is a sufficiently close approximation of A, the main issues for estimating x_0 are caused by the compacity of the operator A.

In the sequel, the norm and inner product on $\mathcal{X}$ are noted $\|.\|$ and $\langle ., . \rangle$ and we endow $\mathbb{R}^n$ with the inner product $\langle u, v \rangle_n = n^{-1} u^t v$ and the associated norm $\|.\|_n$. We shall now give some useful definitions.

Definition The generalized Moore-Penrose inverse of A_n, noted $A_n^\dagger$, is defined by

$$A_n^\dagger z = \arg\min_{x \in \mathcal{X}} \|x\| \quad \text{subject to} \quad A_n x = z, \; z \in \mathbb{R}^n.$$

The generalized Moore-Penrose inverse $A_n^\dagger$ can be defined more intuitively as the inverse of the restriction of A_n to $K_n^\perp$. It satisfies

$$A_n^\dagger A_n = I \quad \text{and} \quad A_n A_n^\dagger = \Pi_{K_n^\perp},$$

where $\Pi_{K_n^\perp}$ denotes the orthogonal projector onto $K_n^\perp$ in $\mathcal{X}$. This definition extends the notion of inverse to non-invertible operators. We refer to [EHN96] for more details.

Definition We call *best approximate solution*, the orthogonal projection $x^\dagger$ of x_0 onto $K_n^\perp$. It is the image of $A_n x_0$ through the Moore-Penrose inverse $A_n^\dagger$.

The best approximate solution is generally unknown to the statistician, due to the presence of noise ε in the observation. In a certain way, it is the best approximation of x_0 one can get in this model, as no information on the components of x_0 in the kernel of A_n can be deduced from the observation. Thus, when no external information on x_0 is available, the implicit objective in such problems is to estimate the best approximate solution $x^\dagger$. In this purpose, a natural idea is to apply the generalized inverse $A_n^\dagger$ to the observation y. In this way, we obtain an estimator of $x^\dagger$ by

$$y^\dagger = A_n^\dagger y = x^\dagger + \eta,$$

where we set $\eta = A_n^\dagger \varepsilon \in K_n^\perp$. Although unbiased, this estimator is generally inefficient, especially if A_n highly attenuates the signal x_0, which results in the generalized inverse $A_n^\dagger$ emphasizing the effect of the noise. To deal with this issue, regularization methods aim to replace the generalized inverse $A_n^\dagger$ by a smoothed version, thus reducing the variance at the cost of adding a bias to the estimator.

Let $(\mathcal{E}, \langle .,. \rangle)$ be an Hilbert space and $M : \mathcal{E} \to \mathcal{E}$ a diagonalizable linear map with spectral decomposition $\{\rho_i, v_i\}_i$, i.e. $M(x) = \sum_i \rho_i \langle x, v_i \rangle v_i,\ x \in \mathcal{E}$. For f a function defined on the spectrum of M, we note $f(M)$ the operator with spectral decomposition $\{f(\rho_i), v_i\}_i$. Define the function $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ by $\Phi(x) = 1/x$ if $x > 0$ and $\Phi(0) = 0$, the operator $A_n^\dagger$ can be expressed as

$$A_n^\dagger = \Phi(A_n^* A_n) A_n^*,$$

where A_n^* denotes the adjoint of A_n. Because the compacity of A may result in a high valued spectrum of the pseudo inverse $\Phi(A_n^* A_n)$, the signal x_0 can be heavily affected with noise in the directions corresponding to large eigenvalues. A solution is to replace $\Phi(A_n^* A_n)$ by a smoothed version whose spectral radius does not exceed a certain limit α. Hence, consider a bounded function Φ_α approximating the inverse function and define

$$\widehat{x}_\alpha = \Phi_\alpha(A_n^* A_n) A_n^* y.$$

We assume that the smoothing function Φ_α satisfies the following conditions

$$\sup_{t \geq 0} |t \Phi_\alpha(t)| \leq 1 \quad \text{and} \quad \sup_{\alpha > 0} \sup_{t \geq 0} |\alpha \Phi_\alpha(t)| \leq 1. \tag{1.2}$$

In a more general setting, the constant 1 in the two conditions can be replaced by arbitrary positive constants C_1, C_2 (see for instance [BHMR07]) although most usual smoothing functions

Φ_α satisfy the condition for $C_1 = C_2 = 1$. The distance of the estimator to the best approximate solution can be decomposed into a bias term and a variance term

$$\mathbb{E}\|x^\dagger - \widehat{x}_\alpha\|^2 = \|x^\dagger - \Phi_\alpha(A_n^* A_n)A_n^* A_n x^\dagger\|^2 + \mathbb{E}\|\Phi_\alpha(A_n^* A_n)A_n^* \varepsilon\|^2.$$

The variance term can be controlled in function of the parameter α writing

$$\mathbb{E}\|\Phi_\alpha(A_n^* A_n)A_n^* \varepsilon\|^2 = \mathbb{E}\left\langle \Phi_\alpha(A_n^* A_n)A_n^* A_n \Phi_\alpha(A_n^* A_n)\varepsilon, \varepsilon\right\rangle_n \leq \frac{\sigma^2}{n\alpha},$$

where the inequality holds as a consequence of (1.2). On the other hand, the bias term can be controlled by *source conditions*, relating the regularity of the input function $x^\dagger$ to the behavior of the operator A_n. Precisely, assume there exists a continuous, strictly increasing function $\Lambda : \mathbb{R}_+ \to \mathbb{R}_+$ with $\Lambda(0) = 0$ and a *source* $w_n \in \mathcal{X}$ with bounded norm $\|w_n\| \leq \overline{w}$ for some $\overline{w} > 0$, such that $x^\dagger = \Lambda(A_n^* A_n)w_n$. Moreover, assume there exists a constant $C > 0$ such that the function Λ satisfies for all t in the spectrum of $A_n^* A_n$,

$$\forall \alpha \geq 0, \; |\Lambda(t)(1 - t\Phi_\alpha(t))| \leq C\Lambda(\alpha). \tag{1.3}$$

The function Λ characterizes the regularity of $x^\dagger$ with respect to the smoothing properties of A_n (see [EHN96]). The faster the function Λ goes to zero as $t \to 0$, the more regular is the problem. Usual situations treated in the literature consider polynomial source conditions for $\Lambda : t \mapsto t^\nu$ or exponential source conditions for $\Lambda : t \mapsto (-\log t)^{-\nu}$ for some $\nu > 0$. Remark however that there is no reason to assume that the function Λ is known since it relies on the regularity of $x^\dagger$. Nevertheless, source conditions enable to obtain a bound on the bias term by writing

$$\|x^\dagger - \Phi_\alpha(A_n^* A_n)A_n^* A_n x^\dagger\| = \|(I - \Phi_\alpha(A_n^* A_n)A_n^* A_n)\Lambda(A_n^* A_n)w_n\| \leq C\; \overline{w}\; \Lambda(\alpha).$$

We obtain the following bound for the risk

$$\mathbb{E}\|x_0 - \widehat{x}_\alpha\|^2 \leq \|x_0 - x^\dagger\|^2 + C^2\; \overline{w}\;{}^2\Lambda(\alpha)^2 + \frac{\sigma^2}{n\alpha}.$$

An optimal value of α must be chosen in order to find a good balance between the bias and variance in this expression. Of course, the optimal value of α is unknown since it depends on the regularity function Λ.

1.3.2 Spectral value decomposition

For the regularization of this kind of inverse problems, a convenient tool is to use the *singular value decomposition* of the operator A_n. Let $b_1^2 \geq ... \geq b_n^2 > 0$ be the ordered eigenvalues of the self-adjoint operator $A_n A_n^*$. For all $i = 1, ..., n$, let ϕ_i (resp. ψ_i) denote the eigenvector of $A_n^* A_n$ (resp. $A_n A_n^*$) associated to b_i^2. The collections $\{\phi_i\}_{i=1,...,n}$ and $\{\psi_i\}_{i=1,...,n}$ form an orthogonal system of $K_n^\perp$ and $\mathbb{R}^n$ respectively and we have $A_n\phi_i = b_i\psi_i$ and $A_n^*\psi_i = b_i\phi_i$, for all $i = 1, ..., n$. Using the singular value decomposition of A_n enables to rewrite the problem in a more convenient way. Let $y_i = \langle y, \psi_i\rangle_n$, $x_i = \langle x_0, \phi_i\rangle$ and $\varepsilon_i = \langle \varepsilon, \psi_i\rangle_n$ for $i = 1, ..., n$, we have the following relation

$$y_i = b_i x_i + \varepsilon_i, \; i = 1, ..., n. \tag{1.4}$$

In this setting, it now suffices to divide each term y_i by the known singular value b_i to observe the coefficient x_i, up to a noise term $\eta_i := b_i^{-1}\varepsilon_i$ which is equivalent to look at the decomposition of $y^\dagger = A_n^\dagger y \in K_n^\perp$ in the singular system $\{\phi_i\}_{i=1,...,n}$, yielding

$$y_i^\dagger = x_i + \eta_i, \ i = 1, ..., n,$$

where $y_i^\dagger = \langle y^\dagger, \phi_i \rangle = b_i^{-1} y_i$. The problem is then turned into an heteroscedastic model where the variance of η_i is inversely proportional to b_i^2, as we have $\mathbb{E}(\eta_i^2) = n^{-1} b_i^{-2} \sigma^2$. This representation points out the effect of the decay of the singular values b_i on the noise level. The estimation of x_0 is inefficient in the directions corresponding to small eigenvalues b_i, i.e. when the variance of η_i is large. To control the high frequency noises, a solution is to consider weighted versions of $y^\dagger$. So, for some collection of weights $\lambda = (\lambda_1, ..., \lambda_n)^t \in \mathbb{R}^n$ called *filter*, we define the estimator $\hat{x}(\lambda)$ as the function in $K_n^\perp$ such that $\langle \hat{x}(\lambda), \phi_i \rangle = \lambda_i y_i^\dagger$ for all $i = 1, ..., n$. The purpose of filter estimators is to cancel-out the high frequency noises by allocating low weights to the components $y_i^\dagger$ corresponding to small singular values. This approach is actually more general than the approach discussed in the previous section, as we can show that the estimator $\widehat{x}_\alpha = \Phi_\alpha(A_n^* A_n) A_n^* y$ is the estimator associated to the filters $\lambda_i = \Phi_\alpha(b_i^2) b_i^2$.

Examples of filter methods

1. Spectral cut-off. An estimator of x_0 is obtained as a truncated version of $y^\dagger$, where we change to zero all coefficients $y_i^\dagger$ corresponding to arbitrarily small singular values. This approach can be viewed as a principal component analysis, where only the highly explanatory directions are selected. The spectral cut-off estimator is associated to the filters $\lambda_i = \mathbb{1}\{b_i^2 \le \alpha^{-1}\}$, where $\mathbb{1}\{.\}$ denotes the indicator function, which corresponds to the smoothing function $\Phi_\alpha : t \mapsto t^{-1}\mathbb{1}\{t \le \alpha^{-1}\}$. We refer to [BHMR07], [EHN96] and [Han87].

2. Projection filters. While the spectral cut-off only deals with monotonic sequences of filters, a natural generalization is to consider unrestricted binary filters $\lambda_i = \mathbb{1}\{i \in m\}$, for $m \subseteq \{1, ..., n\}$. For instance, hard-thresholding procedures are related to projection filters, choosing a subset m of the form $m = \{i : y_i^2 \ge c_i\}$ for some collection of thresholds $\{c_i\}_{i=1,...,n}$. Such estimators are studied in [CGPT00], [LL08] and [LL10].

3. Tikhonov. The Tikhonov regularization [TA77] is associated to the so-called Wiener filters $\lambda_i = b_i^2(\alpha^{-1} + b_i^2)^{-1}$ and smoothing function $\Phi_\alpha : t \mapsto (t + \alpha^{-1})^{-1}$. The solution is obtained as the minimizer of the functional

$$x \mapsto \|y - A_n x\|^2 + \alpha^{-1}\|x\|^2, \ x \in \mathcal{X},$$

   which makes the method particularly convenient in cases where the SVD of $A_n^* A_n$ or the coefficients $y_i^\dagger$ are not easily computable. This approach is the analog of ridge regression for the inverse problem framework.

4. Iterative Landweber. Like the Tikhonov regularization, the iterative Landweber method does not require the calculation of the SVD of the operator A_n, which makes it generally

easily feasible computationally. Define the sequence $\{x_n\}_{n\geq 1}$ recursively by $x_{k+1} = x_k - \tau A_n^*(A_n x_k - y)$ where $\tau > 0$ is a scaling parameter such that $\sup_{i=1,\dots,n} \tau b_i^2 < 1$ and the initial value x_1 is an arbitrary guess on the true function. Without prior knowledge, the Landweber method with initial choice $x_1 = 0$ leads, after k iterations, to the estimator associated to the filters $\lambda_i = 1 - (1 - \tau b_i^2)^k$. The stopping time k plays the role of a tuning parameter to be chosen by the practitioner, keeping in mind that the solution tends to the fixed point $y^\dagger$ as $k \to \infty$.

In each one of the regularization methods, the choice of the *tuning parameters* plays a crucial role. In the literature, different parameter selection techniques have been implemented, relying for instance on cross-validation [DRM96], the discrepancy principle [EHN96] or the L-curve [HO93]. There exist also penalized procedures for selecting an appropriate choice of the parameters such as Stein's *unbiased risk estimation* [Ste81] or the *risk hull method* studied in [CGPT00].

Chapter 2

Quadratic approximate maximum entropy on the mean

We are interested in recovering an unknown finite measure μ_0 from a noisy observation of generalized moments of μ_0, defined as the integral of a continuous function Φ with respect to μ_0. When only a quadratic approximation Φ_m of the operator is known, we introduce the $\mathbb{L}^2$ approximate maximum entropy solution as a minimizer of a convex functional subject to a sequence of convex constraints. Under several regularity conditions, we establish the convergence of the approximate solution and provide its rate of convergence.

2.1 Introduction

We tackle the inverse problems of reconstructing an unknown finite measure μ_0 on a set $\mathcal{X} \subset \mathbb{R}^d$, from observations of generalized moments of μ_0,

$$y = \int_{\mathcal{X}} \Phi(x) d\mu_0(x),$$

where $\Phi : \mathcal{X} \to \mathbb{R}^k$ is a given map. Such problems are encountered in various fields of sciences, like medical imaging, time-series analysis, speech processing, image denoising, spectroscopy, geophysical sciences, crystallography, and tomography, see for example [DHLN92], [HN00] and [Ski88]. This problem has also been extensively studied in the literature in Econometry, some of the main references are [Cha87], [Han82] and [Owe91]. The problem of recovering the unknown measure μ_0 is *ill-posed*, in particular because a solution to the equation $y = \int \Phi \, d\mu_0$ is not unique. For inverse problems with known operator Φ, regularization techniques have been implemented in order to turn the problem into a convex optimization program for which a solution is uniquely defined. Precisely, a solution is obtained as the minimizer of a convex functional $\mu \mapsto J(\mu)$ subject to the linear constraint $\int \Phi d\mu = y$ when y is observed, or more generally, subject to a convex constraint of the form $\int \Phi d\mu \in K_Y$ in presence of noise, for some convex set K_Y. Several types of regularizing functionals have been introduced in the literature. In this general setting, the inversion procedure is deterministic, i.e. the noise distribution is not used in the definition of the regularized solution. Bayesian approaches to inverse problems allow one to handle the noise distribution, provided it is known, yet in general, a distribution like the normal distribution is postulated (see [ES02] for a survey). However in many real-world inverse problems, the noise distribution is unknown, and only the output y is easily observable, contrary to the input to the operator. Consequently very few paired data are available to reliably estimate the noise distribution, thereby causing robustness deficiencies on the retrieved parameters. Nonetheless, even if the noise distribution is unavailable to the practitioner, she often knows the noise level, i.e., the maximal magnitude of the disturbance term, say $\eta > 0$, and this information may be reflected by taking a constraint set K_Y of diameter 2η.

As an alternative to standard regularization methods such as Tikhonov and Galerkin (see for instance [EHN96]), we focus on a regularization functional with grounding in information theory, leading to maximum entropy solutions to the inverse problem. The method, known as *maximum entropy on the mean* (MEM), provides a very simple and natural manner to incorporate constraints on the support and the range of the solution, as discussed in [GG97]. In a deterministic framework, maximum entropy solutions have been studied in [BLN96], [BL91], while some others study exist in a Bayesian setting [Gam99], [GG97], in seismic tomography [FLLn06], in image analysis [GZ02] and in survey sampling [GLR11].

While the literature in this domain has focused on inverse problems with complete knowledge of the operator, it appears that many actual situations do not allow the operator to be exactly known by the statistician, whether because of noise in the observations or for computational feasibility reasons. Thus, in many actual situations, the map Φ is unknown and only an approximation Φ_m is available. In this paper, we introduce an approximate maximum entropy on the mean (AMEM) estimate $\hat{\mu}_{m,n}$ of the measure μ_0 to be reconstructed. This estimate is expressed in the form of a discrete measure concentrated on n points of $\mathcal{X}$. In our main result, we prove that the convergence in $\mathbb{L}^2$-norm of the sequence $\{\Phi_m\}_{m\in\mathbb{N}}$ toward Φ is sufficient to ensure the weak convergence of the estimator $\hat{\mu}_{m,n}$ to the solution of the initial inverse problem as $m \to \infty$ and $n \to \infty$. Moreover, we provide a rate of convergence for this estimator.

A natural field of applications arises from the use of instrumental variables in Econometry (see for instance [Flo03]). We will provide a new estimation procedure in this setting.

The chapter is organized as follows. Section 2 introduces some notations and the definition of the AMEM estimate. We state our main result (Theorem 2.2.1) in Section 3. Applications to remote sensing and instrumental variable estimation are studied in Section 4, while Section 5 is devoted to the proofs of our results.

2.2 The AMEM estimate

2.2.1 Problem setting

Let Φ be a continuous map defined on a subset $\mathcal{X}$ of $\mathbb{R}^d$ with values in $\mathbb{R}^k$. We note $\mathcal{B}(\mathcal{X})$ the Borel σ-field of $\mathcal{X}$ and $\mathcal{F}(\mathcal{X})$ the set of finite measures on $\mathcal{X}$. Let $\mu_0 \in \mathcal{F}(\mathcal{X})$ be an unknown measure satisfying the constraint $y = \int \Phi d\mu_0$. Assume we observe a perturbed version y^{obs} of y:

$$y^{obs} = \int_{\mathcal{X}} \Phi(x) d\mu_0(x) + \varepsilon, \tag{2.1}$$

where ε is an error term supposed bounded in norm from above by some positive constant η, representing the maximal noise level. Based on the data y^{obs}, we aim at reconstructing the measure μ_0 with a maximum entropy procedure. In image analysis this measure may be viewed as the intensity at each pixel of the image, blurred by an unknown filter. Other applications in seismic tomography can be found in [FLLn06], while we discuss an application to Econometry in Section 2.3.2.

For two probability measures ν, μ, we recall that the relative entropy of ν with respect to μ is given by

$$\mathcal{K}(\nu|\mu) = \int_{\mathcal{X}} \log\left(\frac{d\nu}{d\mu}\right) d\nu + \mu(\mathcal{X}) - \nu(\mathcal{X}) \text{ if } \nu \ll \mu, \quad \mathcal{K}(\nu|\mu) = +\infty \text{ otherwise.}$$

We denote by K_Y the closed ball of $\mathbb{R}^k$ centered at the observation y^{obs} and of radius η. The true measure μ_0 is known to satisfy the moment condition $\int \Phi d\mu_0 \in K_Y$, however, the map Φ being unknown, we consider the approximate moment condition

$$\int_{\mathcal{X}} \Phi_m(x) d\mu_0(x) \in K_Y. \tag{2.2}$$

Moreover, we note $\mathcal{M}_m = \{\mu \in \mathcal{F}(\mathcal{X}) : \int \Phi_m d\mu\}$ the set of finite measures satisfying the approximate moment condition. Let us now explain the construction of the AMEM estimator. Let $X_1, \ldots, X_n$ be a discretization of the space $\mathcal{X}$, for which the associated empirical measure $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is assumed to converge weakly to some distribution $\mathbb{P}_X$ having full support on $\mathcal{X}$. The X_i's may be i.i.d. realizations of a random variable X with distribution $\mathbb{P}_X$, or a deterministic design, in which case $\mathbb{P}_X$ is known by the statistician. We search for an estimator of μ_0 that can be written as a weighted version of the empirical measure $\mathbb{P}_n$

$$\mathbb{P}_n(w) := \frac{1}{n}\sum_{i=1}^n w_i \delta_{X_i},$$

for some vector $w = (w_1, ..., w_n)^t \in \mathbb{R}^n$. Moreover, we want the estimator to satisfy to approximate moment condition. Let $W = (W_1, ..., W_n)^t$ be a vector of n i.i.d. realizations drawn from a measure ν and consider the random measure $\mathbb{P}_n(W)$. Seeing each weighted measure $\mathbb{P}_n(w)$ as a realization of $\mathbb{P}_n(W)$, the measure $\nu^{\otimes n}$ can be interpreted as a *prior* distribution on the parameter w. The posterior distribution ν^* is defined as the probability measure minimizing the relative entropy $\mathcal{K}(.|\nu^{\otimes n})$ under the constraint that the approximate moment condition (2.2) holds in mean,

$$\mathbb{E}_{\nu^*}[\mathbb{P}_n(W)] \in \mathcal{M}_m.$$

The estimator $\hat{\mu}_{m,n}$ is obtained as the expectation of $\mathbb{P}_n(W)$ under ν^*,

$$\hat{\mu}_{m,n} = \mathbb{E}_{\nu^*}[\mathbb{P}_n(W)] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\nu^*}(W_i)\delta_{X_i}.$$

The existence of ν^* requires the feasibility of the problem, i.e. the existence of a vector δ in the convex hull of the support of $\nu^{\otimes n}$ such that $\int \Phi_m d\mathbb{P}_n(\delta) \in K_Y$. It is shown in [LP08] that under the Assumptions of Theorem 2.2.1, this condition tends to be verified with probability 1 as $m \to \infty$ and $n \to \infty$. Hence for m and n large enough, the AMEM estimate $\hat{\mu}_{m,n}$ is well defined with high probability, and asymptotically with probability 1.

2.2.2 Convergence of the AMEM estimate

We recall that for ν a probability measure on $\mathbb{R}$, Λ_ν and Λ_ν^* denote respectively the log-Laplace and Cramer transforms of ν, given by

$$\Lambda_\nu(s) = \log \int_{\mathbb{R}} e^{sx} d\nu(x) \quad \text{and} \quad \Lambda_\nu^*(s) = \sup_{u \in \mathbb{R}} \{su - \Lambda_\nu(u)\}, \; s \in \mathbb{R}.$$

We define the functional

$$\mu \mapsto I_\nu(\mu|\mathbb{P}_X) = \int_{\mathcal{X}} \Lambda_\nu^* \left(\frac{d\mu}{d\mathbb{P}_X} \right) d\mathbb{P}_X \text{ if } \mu \ll \mathbb{P}_X, \quad I_\nu(\mu|\mathbb{P}_X) = +\infty \text{ otherwise,}$$

which is the *f-divergence* of μ with respect to $\mathbb{P}_X$ associated to the convex function Λ_ν^*. We note $\mathcal{C}_b$ the set of continuous bounded functions on $\mathcal{X}$. For all $g \in \mathcal{C}_b$, we denote by $| \, . \, |_g$ the semi-norm defined for $\mu \in \mathcal{F}(\mathcal{X})$ by $|\mu|_g = \left| \int g d\mu \right|$. We recall that the family of semi-norms $\{| \, . \, |_g, \; g \in \mathcal{C}_b\}$ defines the weak topology: a sequence $\{\mu_n\}_{n \in \mathbb{N}}$ converges weakly toward μ if, and only if, $\lim_{n \to \infty} |\mu_n - \mu|_g = 0$, for all $g \in \mathcal{C}_b$.
We make the following assumptions.

A2.1. The minimization problem is feasible, i.e., there exists a continuous function g_0 defined on the convex hull of the support of ν such that $\int \Phi g_0 \, d\mathbb{P}_X \in K_Y$.

A2.2. The function Λ_ν'' is bounded by a constant $K > 0$.

A2.3. The approximating sequence Φ_m converges to Φ in $\mathbb{L}^2(\mathbb{P}_X)$. Its rate of convergence is given by

$$\|\Phi_m - \Phi\|_{\mathbb{L}^2} := \sqrt{\mathbb{E}\|\Phi_m(X) - \Phi(X)\|^2} = O(\varphi_m^{-1}),$$

for some growing sequence $\{\varphi_m\}_{m \in \mathbb{N}}$.

A2.4. The function $G : x \mapsto \sup_{m \in \mathbb{N}} \|\Phi_m(x)\|$ is square integrable: $\int G^2 d\mathbb{P}_X < \infty$.

A2.5. For all $m \in \mathbb{N}$, the components of Φ_m are linearly independent.

We are now in a position to state our main result.

Theorem 2.2.1 (Convergence of the AMEM estimate) *Suppose that* A2.1 *and* A2.2 *hold and let μ^* be the minimizer of the functional $\mu \mapsto I_\nu(\mu|\mathbb{P}_X)$ subject to the constraint $\int \Phi d\mu \in K_Y$.*

- *The AMEM estimate $\hat{\mu}_{m,n}$ is given by*

$$d\hat{\mu}_{m,n}(x) = \Lambda_\nu'(\hat{v}_{m,n}^t \Phi_m(x)) d\mathbb{P}_n(x),$$

where $\hat{v}_{m,n}$ minimizes over $\mathbb{R}^k$, $H_{m,n}(v) = \mathbb{P}_n \Lambda_\nu(v^t \Phi_m) - \inf_{y \in K_Y} v^t y$.

- *If* A2.3 *to* A2.5 *also hold,* $\hat{\mu}_{m,n}$ *converges weakly in probability to* μ^* *as* $m, n \to \infty$ *and its rate of convergence is expressed as follows,*

$$\forall g \in \mathcal{C}_b, \; |\hat{\mu}_{m,n} - \mu^*|_g = O(\varphi_m^{-1}) + \kappa_{m,n},$$

with $\sup_{m \in \mathbb{N}} \kappa_{m,n} = O_P(n^{-1/2})$.

The condition A2.2 is a rather strong requirement on the choice of the prior ν. It is equivalent to assuming that Λ_ν is dominated on $\mathbb{R}$ by a quadratic function. This condition is satisfied for instance for Gaussian priors or if ν has compact support. As a result, the function $H : v \mapsto \mathbb{P}_X \Lambda_\nu(v^t \Phi) - \inf_{y \in K_Y} v^t y$ attains its minimum at a unique point v^* belonging to the interior of its domain $\mathbb{R}$. If this assumption is not met, it is shown in [BL93] and [GG97] that the minimizers of $I_\nu(.|\mathbb{P}_X)$ over the set of finite measures satisfying the moment constraint may have a singular part with respect to $\mathbb{P}_X$.

The construction of the AMEM estimate relies on a discretization of the space $\mathcal{X}$ according to the probability $\mathbb{P}_X$. Therefore by varying the support of $\mathbb{P}_X$, the practitioner may easily incorporate some a-priori knowledge concerning the support of the solution. Similarly, the AMEM estimate also depends on the measure ν, which determines the domain of Λ_ν^*, and so the range of the solution.

2.2.3 Perspectives

The convergence of the estimator is obtained under some restrictions on the prior, which lead to strong conditions on the regularization criterion. In particular, the proof of the result imposes to choose a sub-Gaussian prior. Inspection of the proofs shows that relaxing this assumption would require stronger assumptions on the convergence of $\{\Phi_m\}$, such as the convergence in $\mathbb{L}^p$-norm, for some $p > 2$. To study the problem under this alternate assumption could allow to extend the result to a wider choice of priors. It is also assumed that the sequence $\{\Phi_m\}$ is independent from the model. In practice however, for instance in the application to survey sampling treated in Chapter 4, the problem may lead to a situation for which the approximation Φ_m is estimated from the observations. Maybe it is possible to generalize the result to a dependent data framework, in order to obtain a larger field of applications.

2.3 Applications

2.3.1 Remote sensing

In remote sensing of aerosol vertical profiles, one wishes to recover the concentration of aerosol particles from noisy observations of the radiance field (i.e., a radiometric quantity), in several spectral bands (see e.g. [GKP99], [MRVP97]). More specifically, at a given level of modeling, the noisy observation y^{obs} may be expressed as

$$y^{obs} = \int_{\mathcal{X}} \Phi(x; t^{obs}) d\mu_0(x) + \varepsilon, \tag{2.3}$$

where $\Phi : \mathcal{X} \times \mathcal{T} \to \mathbb{R}^k$ is a given operator, and where t^{obs} is a vector of angular parameters observed simultaneously with y^{obs}. The aerosol vertical profile is a function of the altitude x and is associated with the measure μ_0 to be recovered, i.e., the aerosol vertical profile is the Radon-Nykodim derivative of μ_0 with respect to a given reference measure (e.g., the Lebesgue measure on $\mathbb{R}$). The analytical expression of Φ is fairly complex as it sums up several models at the microphysical scale, so that basically Φ is available in the form of a computer code. So this problem motivates the introduction of an efficient numerical procedure for recovering the unknown μ_0 from y^{obs} and arbitrary t^{obs}.

More generally, the remote sensing of the aerosol vertical profile is in the form of an inverse problem where some of the inputs (namely t^{obs}) are observed simultaneously with the noisy output y^{obs}. Suppose that random points $X_1, \dots, X_n$ of $\mathcal{X}$ have been generated. Then, applying the maximum entropy approach would require the evaluations of $\Phi(X_i, t^{obs})$ each time t^{obs} is observed. If one wishes to process a large number of observations, say (y_i^{obs}, t_i^{obs}), for different values t_i^{obs}, the computational cost may become prohibitive. So we propose to replace Φ by an approximation Φ_m, the evaluation of which is faster in execution. To this aim, suppose first that $\mathcal{T}$ is a subset of $\mathbb{R}^p$. Let $T_1, ..., T_m$ be random points of $\mathcal{T}$, independent of $X_1, \dots, X_n$, and drawn from some probability measure μ_T on $\mathcal{T}$ admitting a density f_T with respect to the Lebesgue measure on $\mathbb{R}^p$ such that $f_T(t) > 0$ for all $t \in \mathcal{T}$. Next, consider the operator

$$\Phi_m(x, t) = \frac{1}{f_T(t)} \frac{1}{m} \sum_{i=1}^{m} K_{h_m}(t - T_i) \Phi(x, T_i),$$

where $K_{h_m}(.)$ is a symmetric kernel on $\mathcal{T}$ of smoothing sequence h_n. It is a classical result to prove that Φ_m converges to Φ in quadratic norm provided h_m tends to 0 at a suitable rate, which ensures that A2.4 is satisfied for Theorem 2.2.1. Since the T_i's are independent from the X_i, one may see that Theorem 2.2.1 applies, and so the solution to the approximate inverse problem

$$y^{obs} = \int_{\mathcal{X}} \Phi_m(x; t^{obs}) d\mu_0(x) + \varepsilon,$$

will converge to the solution to the original inverse problem in (2.3). In terms of computational complexity, the advantage of this approach is that the construction of the AMEM estimate requires, for each new observation (y^{obs}, t^{obs}), the evaluation of the m kernels at t^{obs}, i.e., $K_{h_m}(t^{obs} - T_i)$, the $m \times n$ outputs $\Phi(X_i, T_j)$ for $i = 1, \dots, n$ and $j = 1, \dots, m$ being evaluated once and for all.

2.3.2 Instrumental variable estimation

A natural field of application is given by nonparametric regression models involving instrumental variables. This kind of problem has been extensively studied in the literature in Econometry, we refer for instance to [Flo03], [HS82] and [New90] . In some cases, the instrumental variable estimation framework can be viewed as an inverse problem with unknown operator that can be solved using the AMEM procedure.

Let $X_1, ..., X_n$ be here a discretization of the space $\mathcal{X}$ such that the associated empirical distribution $\mathbb{P}_n$ converges weakly toward a known distribution $\mathbb{P}_X$ having full support on $\mathcal{X}$. Let $g : \mathcal{X} \to \mathbb{R}_+$ be an unknown function for which we observe a noisy evaluation at each point X_i,

$$Y_i = g(X_i) + U_i, \ i = 1, ..., n,$$

where the U_i's are centered real valued random variables. Contrary to the classical regression framework, we suppose here that the noises U_i are correlated with the X_i's (i.e. $\mathbb{E}(U_i|X_i) \neq 0$), which causes identification issues. This kind of model is used for instance to deal with simultaneous causality between supply and demand in economic markets. Assume we want to make a nonparametric regression of the price Y of a good with respect to its production X, the noise U in the corresponding model turns out to be correlated with X due to the mutual influence between the price and the production. To overcome this difficulty, econometricians assume there exist *instrumental variables*, that affect the price only through the produced quantity (for example, the amount of rain in the case of an agricultural product). Hence, we assume we observe simultaneously with (X_i, Y_i), an additional variable $W_i \in \mathbb{R}^k$ such that $\mathbb{E}(W_i|X_i) \neq 0$ and $\mathbb{E}(U_i|W_i) = 0$. In particular, we have the relation

$$y := \mathbb{E}(WY) = \mathbb{E}(Wg(X)). \tag{2.4}$$

In most cases, using the instrumental variable W is not sufficient to solve the identification issue, but it still provides some information that may be rendered in the form of linear constraints on g. Indeed, setting $\Phi : x \mapsto \mathbb{E}(W|X = x)$ and $d\mu_0(x) = gd\mathbb{P}_X(x), x \in \mathcal{X}$, the equation (2.4) can be written as

$$y = \int \Phi(x) d\mu_0(x).$$

Here, y is unknown but we observe a noisy version $y^{obs} = n^{-1} \sum_{i=1}^n W_i X_i$ that is close to y with high probability and asymptotically with probability one. The conditional expectation Φ is also unknown but can be estimated from the data by nonparametric procedures, yielding a converging sequence $\{\Phi_n\}$. As a result, estimating the measure μ_0 can be made using the AMEM procedure by considering an approximate moment condition of the form $\int \Phi_n d\mu \in K_Y$. We obtain a sequence of estimators $\hat{\mu}_n$, which is shown in Theorem 2.2.1 to converge weakly toward the minimizer μ^* of the convex functional $I_\nu(.|\mathbb{P}_X)$ subject to the moment constraint. Equivalently, the method ensures the convergence in a weak sense of the density $\hat{g} = d\hat{\mu}_n/d\mathbb{P}_n$ of the AMEM estimator toward the function $g^* := d\mu^*/d\mathbb{P}_X$. In particular, the identification issue on g is solved by incorporating some prior knowledge on μ_0 through the choice of the design $X_1, ..., X_n$ and the limit distribution $\mathbb{P}_X$.

2.4 Proofs

2.4.1 Technical Lemmas

For P a measure and g a function, we shall use the notation $Pg = \int g dP$. Moreover let

$$\begin{aligned} v_m^* &= \arg\min_{v\in\mathbb{R}^k} H_m(v) = \arg\min_{v\in\mathbb{R}^k} \left\{\mathbb{P}_X\Lambda_\nu(v^t\Phi_m) - \inf_{y\in K_Y} v^t y\right\}, \\ \hat{v}_{m,n} &= \arg\min_{v\in\mathbb{R}^k} H_{m,n}(v) = \arg\min_{v\in\mathbb{R}^k} \left\{\mathbb{P}_n\Lambda_\nu(v^t\Phi_m) - \inf_{y\in K_Y} v^t y\right\}, \\ v^* &= \arg\min_{v\in\mathbb{R}^k} H(v) = \arg\min_{v\in\mathbb{R}^k} \left\{\mathbb{P}_X\Lambda_\nu(v^t\Phi) - \inf_{y\in K_Y} v^t y\right\}. \end{aligned}$$

Lemma 2.4.1 *If Assumptions 1 to 5 hold,*

$$\sup_{m\in\mathbb{N}} \|\hat{v}_{m,n} - v_m^*\| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

Proof. For all $x \in \mathcal{X},\ v \in \mathbb{R}^k$, set

$$h_m(v,x) = \Lambda_\nu(v^t\Phi_m(x)) - \inf_{y\in K_Y} v^t y.$$

The parameter $\hat{v}_{m,n}$ is defined as the minimizer of the empirical contrast function $v \mapsto H_{m,n}(v) = \mathbb{P}_n h_m(v,.)$. To prove the result, we need to show that $h_m(v,x)$ satisfies the conditions of Corollary 5.53 in [vdV98]. First remark that $H_{m,n}$ is convex, which ensures the convergence in probability of its minimizer $\hat{v}_{m,n}$ toward v_m^*. Since K_Y is the ball centered in y^{obs} and of radius η, we may write

$$h_m(v,x) = \Lambda_\nu(v^t\Phi_m(x)) - v^t y^{obs} + \eta\|v\|.$$

By A2.2, we know that $\Lambda'_\nu(s) \le Ks + 1$ for all $s \in \mathbb{R}$. For all v_1, v_2 in a neighborhood $\mathcal{N}$ of v_m^*, we have by the triangular inequality and the mean value theorem

$$\begin{aligned} |h_m(v_1,.) - h_m(v_2,.)| &\le \left|\Lambda_\nu(v_1^t\Phi_m) - \Lambda_\nu(v_2^t\Phi_m)\right| + \left|\langle v_1 - v_2, y^{obs}\rangle + \eta\left|\|v_1\| - \|v_2\|\right|\right| \\ &\le \left[K\|v_2\|\ \|\Phi_m\| + 1 + \|y^{obs}\| + \eta\right] \|v_1 - v_2\| \\ &\le \left[K\delta\ G + 1 + \|y^{obs}\| + \eta\right] \|v_1 - v_2\|, \end{aligned}$$

where G is the function defined in A2.4 and where we set $\delta = \sup_{v\in\mathcal{N}}\|v\|$. Since v_m^* converges toward v^*, we may assume without loss of generality that $\mathcal{N}$ and δ are fixed for m sufficiently large. Hence the function h_m satisfies the first condition of Corollary 5.53 in [vdV98],

$$|h_m(v_1,.) - h_m(v_2,.)| \le \dot{h}\|v_1 - v_2\|,$$

where $\dot{h} : x \mapsto K\delta\ G(x) + 1 + \|y^{obs}\| + \eta$ does not depend and m and is such that $P_X\dot{h}^2 < \infty$. For all $v \in \mathbb{R}^k$, let $V_m(v)$ be the Hessian matrix of H_m at point v, which is well defined for all $v \neq 0$. Assume that $v_m^* \neq 0$, we need to prove that $V_m(v_m^*)$ is non-negative definite. The case $v_m^* = 0$ can be treated separately without difficulty using Theorem 5.52 in [vdV98], by considering the derivative at 0^+ of the functions $t \mapsto V_m(tv)$, $v \in \mathbb{R}^k$. Let ∂_i denote the derivative with respect to the i-th component. For $v \neq 0$, we have

$$\begin{aligned}[V_m(v)]_{ij} = \partial_i\partial_j H_m(v) &= \mathbb{P}_X[\partial_i\partial_j h_m(v,.)] \\ &= \mathbb{P}_X[\Phi_m^i \Phi_m^j \Lambda_\nu''(v^t\Phi_m)] + \eta\ \partial_i\partial_j N(v)\end{aligned}$$

where we set $N : v \mapsto \|v\|$. Thus, $V_m(v_m^*)$ can be split into the sum $A_m + \eta B_m$, with

$$(A_m)_{ij} = P_X[\Phi_m^i \Phi_m^j \Lambda_\nu''(v_m^{*t}\Phi_m)], \quad (B_m)_{ij} = \partial_i\partial_j N(v_m^*).$$

A_m is a Gram matrix, therefore it is positive definite by A2.5. Moreover, since A_m converge toward the positive-definite matrix $A = (P_X[\Phi^i\Phi^j\Lambda_\nu''(v^{*t}\Phi)])_{1\leq i,j\leq k}$, we conclude there exist an integer M and a constant $c > 0$ such that, for all $a \in \mathbb{R}^k$,

$$\inf_{m\geq M} a^t A_m a \geq c\|a\|^2.$$

By convexity of the map $N(.)$ on $\mathbb{R}^k$, the matrix B_m is non-negative definite and so is $V_m(v_m^*) = A_m + \eta B_m$. Hence, H_m undergoes the assumptions of Corollary 5.53 in [vdV98], uniformly for $m \in \mathbb{N}$, which proves the result.

Lemma 2.4.2 *If Assumptions 1 to 5 hold,*

$$\|v_m^* - v^*\| = O(\varphi_m^{-1}).$$

Proof. Using successively the mean value theorem and Cauchy-Schwarz's inequality, we find

$$\begin{aligned}|H_m(v) - H(v)| &= |\mathbb{P}_X[\Lambda_\nu(v^t\Phi_m) - \Lambda_\nu(v^t\Phi(x))]| \\ &\leq (K\|v\|^2\ \|G\|_{\mathbb{L}^2} + \|v\|)\|\Phi_m - \Phi\|_{\mathbb{L}^2}.\end{aligned}$$

We deduce that H_m converges uniformly on every compact set toward H as $m \to \infty$. By convexity of H_m, this warrants the convergence of v_m^* toward v^*. Moreover,

$$\begin{aligned}\nabla H_m(v) - \nabla H(v) &= \mathbb{P}_X\left[\Phi_m\Lambda_\nu'(v^t\Phi_m) - \Phi\Lambda_\nu'(v^t\Phi)\right] \\ &= \mathbb{P}_X\left[(\Phi_m - \Phi)\Lambda_\nu'(v^t\Phi_m) + \Phi[\Lambda_\nu'(v^t\Phi_m) - \Lambda_\nu'(v^t\Phi)]\right].\end{aligned}$$

In the same way as previously, we find

$$\|\nabla H_m(v) - \nabla H(v)\| \leq \|\Phi_m - \Phi\|_{\mathbb{L}^2}\|v\|\left(K\|\Phi_m\|_{\mathbb{L}^2} + 1 + K\|\Phi\|_{\mathbb{L}^2}\right),$$

which proves that ∇H_m converges toward ∇H, uniformly on every compact set. Noticing that $\nabla H(v_m^*) = \nabla H(v_m^*) - \nabla H_m(v_m^*)$, it follows that $|\nabla H(v_m^*)| = O(\varphi_m^{-1})$. Note $V(v^*)$ the Hessian matrix of H at v^*. We know it is positive definite by a similar reasoning as in the proof of Lemma 2.4.1. Writing the Taylor expansion

$$\nabla H(v_m^*) = V(v^*)(v^* - v_m^*) + o(\|v^* - v_m^*\|),$$

we conclude $\|v^* - v_m^*\| = O(\varphi_m^{-1})$.

2.4.2 Proof of Theorem 2.2.1

The first part of the theorem is proved in Theorem 3.1 in [LP08]. We here focus on the proof of the second part. We use the following notations

$$\hat{\mu}_{m,n} = \Lambda'_\nu(\hat{v}^t_{m,n}\Phi_m)\mathbb{P}_n \quad \text{and} \quad \mu^*_m = \Lambda'_\nu(v^{*t}_m\Phi_m)\mathbb{P}_X.$$

For $g \in \mathcal{C}_b$, write $|\hat{\mu}_{m,n} - \mu^*|_g \leq |\hat{\mu}_{m,n} - \mu^*_m|_g + |\mu^*_m - \mu^*|_g$. We shall bound each term separately. We have

$$\begin{aligned} |\hat{\mu}_{m,n} - \mu^*_m|_g &= \left|\Lambda'_\nu(\hat{v}^t_{m,n}\Phi_m)\mathbb{P}_n - \Lambda'_\nu(v^{*t}_m\Phi_m)\mathbb{P}_X\right|_g \\ &\leq \left|\Lambda'_\nu(\hat{v}^t_{m,n}\Phi_m)\mathbb{P}_n - \Lambda'_\nu(v^{*t}_m\Phi_m)\mathbb{P}_n\right|_g + \left|\Lambda'_\nu(v^{*t}_m\Phi_m)\mathbb{P}_n - \Lambda'_\nu(v^{*t}_m\Phi_m)\mathbb{P}_X\right|_g. \end{aligned}$$

We obtain for all $x \in \mathcal{X}$,

$$|\Lambda'_\nu(\hat{v}^t_{m,n}\Phi_m(x)) - \Lambda'_\nu(v^{*t}_m\Phi_m(x))| \leq K\|\Phi_m(x)\| \, \|\hat{v}_{m,n} - v^*_m\|,$$

by Cauchy-Schwarz's inequality. We get

$$\left|\Lambda'_\nu(\hat{v}^t_{m,n}\Phi_m)\mathbb{P}_n - \Lambda'_\nu(v^{*t}_m\Phi_m)\mathbb{P}_n\right|_g \leq K\|g\|_\infty\| \, \|\hat{v}_{m,n} - v^*_m\| \, \mathbb{P}_n G.$$

Using Slutsky's lemma and Lemma 2.4.1, we conclude

$$\sup_{m\in\mathbb{N}} \left|\Lambda'_\nu(\hat{v}^t_{m,n}\Phi_m)\mathbb{P}_n - \Lambda'_\nu(v^{*t}_m\Phi_m)\mathbb{P}_n\right|_g = O_P\left(\frac{1}{\sqrt{n}}\right).$$

The rate of convergence of the term $\left|\Lambda'_\nu(v^{*t}_m\Phi_m)\mathbb{P}_n - \Lambda'_\nu(v^{*t}_m\Phi_m)\mathbb{P}_X\right|_g$ follows directly from the uniform law of large numbers. We obtain

$$\sup_{m\in\mathbb{N}} |\hat{\mu}_{m,n} - \mu^*_m|_g = O_P\left(\frac{1}{\sqrt{n}}\right).$$

The second step is to bound the term $|\mu^*_m - \mu^*|_g$. We follow the same guidelines,

$$\begin{aligned} |\mu^*_m - \mu^*|_g &= \left|\Lambda'_\nu(v^{*t}_m\Phi_m)\mathbb{P}_X - \Lambda'_\nu(v^{*t}\Phi)\mathbb{P}_X\right|_g \\ &\leq \left|\Lambda'_\nu(v^{*t}_m\Phi_m)\mathbb{P}_X - \Lambda'_\nu(v^{*t}\Phi_m)\mathbb{P}_X\right|_g + \left|\Lambda'_\nu(v^{*t}\Phi_m)\mathbb{P}_X - \Lambda'_\nu(v^{*t}\Phi)\mathbb{P}_X\right|_g. \end{aligned}$$

In the same way as previously, the first term is bounded as follows

$$\left|\Lambda'_\nu(v^{*t}_m\Phi_m)\mathbb{P}_X - \Lambda'_\nu(v^{*t}\Phi_m)\mathbb{P}_X\right|_g \leq K\|g\|_\infty \, \mathbb{E}\|\Phi_m(X)\| \, \|v^*_m - v^*\|,$$

which is shown to be of order $O(\varphi_m^{-1})$ in Lemma 2.4.2. For the last term, we have in the same way

$$\left|\Lambda'_\nu(v^{*t}\Phi_m)\mathbb{P}_X - \Lambda'_\nu(v^{*t}\Phi)\mathbb{P}_X\right|_g \leq K\|v^*\| \, \|g\|_\infty \, \mathbb{E}\|\Phi_m(X) - \Phi(X)\|.$$

Regrouping all the terms, we get

$$|\hat{\mu}_{m,n} - \mu^*|_g = \kappa_{m,n} + O(\varphi_m^{-1}),$$

where $\kappa_{m,n} \leq |\hat{\mu}_{m,n} - \mu^*_m|_g$ satisfies $\sup_{m\in\mathbb{N}} \kappa_{m,n} = O_P\left(\frac{1}{\sqrt{n}}\right)$.

Chapter 3

Bayesian interpretation of GEL by maximum entropy

We study a parametric estimation problem related to moment condition models. As an alternative to the generalized empirical likelihood (GEL) and the generalized method of moments (GMM), a Bayesian approach to the problem can be adopted, extending the MEM procedure to parametric moment conditions. We show in particular that a large number of GEL estimators can be interpreted as a maximum entropy solution. Moreover, we provide a more general field of applications by proving the method to be robust to approximate moment conditions.

3.1 Introduction

We consider a parametric estimation problem in a moment condition model. Assume we observe an i.i.d. sample $X_1, ..., X_n$ drawn from an unknown probability measure μ_0, we are interested in recovering a parameter $\theta_0 \in \Theta \subset \mathbb{R}^d$, defined by a set of moment conditions

$$\int \Phi(\theta_0, x) d\mu_0(x) = 0, \tag{3.1}$$

where $\Phi : \theta \times \mathcal{X} \to \mathbb{R}^k$ is a known map. This model is involved in many problems in Econometry, notably when dealing with instrumental variables. We refer to [Cha87], [Han82], [QL94], [Owe91] and [DIN09]. Two main approaches to the problem have been studied in the literature, namely the generalized method of moments (GMM) and the generalized empirical likelihood (GEL). While the main advantage of GMM relies in its computational feasibility, likelihood-related methods have appeared to be the most efficient in term of small-sample properties. In its original form, the empirical likelihood (EL) of Owen [Owe91] defines an estimator by a maximum likelihood procedure on a discretized version of the model. As an alternative, GEL replaces the Kullback criterion relative to EL by a f-divergence, thus providing a large choice of solutions. A number of estimators corresponding to particular choices of f-divergences have emerged in the literature over the last decades, such as the exponential tilting (ET) of Kitamura and Stutzer [KS97] and the continuous updating estimator (CUE) of Hansen, Yeaton and Yaron [HHY96].

While an attractive feature of GEL is its wide range of solutions, a number of f-divergence used in the computation of the GEL estimators are mainly justified by empirical studies and lack a probabilistic interpretation. This issue can be solved by incorporating some prior information to the problem using a Bayesian point of view, as made in [PR94]. In this paper, we investigate a different Bayesian approach to the inverse problem, known as *maximum entropy on the mean* (MEM). Although the method was originally introduced in the frame of exact moment condition models (as opposed here to parametric moment conditions), it appears to provide a natural solution to the problem, expressed as the minimizer of a convex functional on a set of discrete measures and subject to linear constraints. When applied in a particular setting, we show that the MEM approach leads to a GEL solution for which the f-divergence is determined by the choice of the prior. As a result, the method gives an alternate point of view on some widely spread estimators such as EL, ET or CUE, as well as a general Bayesian background to GEL.

In many actual situations, the true moment condition is not exactly known to the statistician and only an approximation is available. It occurs for instance when Φ has a complicated form that must be evaluated numerically. Simulation-based methods have been implemented to deal with approximate constraints in [CF00] and [McF89], in the frame of the generalized method of moments. To our knowledge, the efficiency of GEL in a similar situation has not been studied. In [LP08], the MEM procedure is shown to be robust to approximate moment conditions, introducing the approximate maximum entropy on the mean estimator. Seeing GEL as a particular case of MEM, we extend the model in a situation where only an approximation Φ_m of the true constraint function Φ is available. We provide sufficient conditions under which the GEL method remains efficient asymptotically when replacing Φ by its approximation.

This chapter falls into the following parts. Section 3.2 is devoted to the position of the problem. We introduce the maximum entropy method for parametric moment condition models and discuss its close relationship with generalized empirical likelihood in Section 3.2.2. In Section 3.3, we discuss the asymptotic efficiency of the method when dealing with an approximate constraint. Proofs are postponed to the Appendix.

3.2 Estimation of the parameter

Let $\mathcal{X}$ be an open subset of $\mathbb{R}^q$, endowed with its Borel field $\mathcal{B}(\mathcal{X})$. We observe an i.i.d. sample $X_1, ..., X_n$ drawn from the unknown distribution μ_0. We want to estimate the parameter $\theta_0 \in \Theta \subset \mathbb{R}^d$ defined by the moment condition

$$\int_{\mathcal{X}} \Phi(\theta_0, x) d\mu_0(x) = 0, \tag{3.2}$$

where $\Phi : \Theta \times \mathcal{X} \to \mathbb{R}^k$ $(k \geq d)$ is a known map. To avoid a problem of identifiability, we assume that θ_0 is the unique solution to (3.2). This problem has many applications in Econometry, see for instance [Cha87], [Han82] and [QL94]. The information given by the moment condition (3.2) can be interpreted to determine the set $\mathcal{M}$ of possible values for μ_0 (the model). The true value of the parameter being unknown, the distribution of the observations can be any probability measure μ for which the map $\theta \mapsto \mu[\Phi(\theta, .)]$ is null for a unique $\theta = \theta(\mu) \in \Theta$. The model is therefore defined as

$$\mathcal{M} = \left\{\mu \in \mathcal{P}(\mathcal{X}) : \exists!\ \theta = \theta(\mu) \in \Theta, \int \Phi(\theta, .)d\mu = 0\right\},$$

where the map $\mu \mapsto \theta(\mu)$, defined on $\mathcal{M}$, is the parameter of interest. Let us introduce some notations and assumptions. For μ a measure and g a function, we shall note $\mu[g] = \int g d\mu$. Let E be an Euclidean space and let $\|.\|$ denote an Euclidean norm in E. For a function $f : \Theta \to E$ and a set $\mathcal{S} \subseteq \Theta$, we note

$$\|f\|_{\mathcal{S}} = \sup_{\theta \in \mathcal{S}} \|f(\theta)\|.$$

We assume that the following conditions are fulfilled.

A3.1. Θ is a compact subset of $\mathbb{R}^d$.

A3.2. The true value θ_0 of the parameter lies in the interior of Θ.

A3.3. For all $x \in \mathcal{X}$, $\theta \mapsto \Phi(\theta, x)$ is continuous on Θ and the map $x \mapsto \|\Phi(., x)\|_\Theta$ is dominated by a μ_0-integrable function.

A3.4. For all $x \in \mathcal{X}$, $\theta \mapsto \Phi(\theta, x)$ is twice continuously differentiable in a neighborhood $\mathcal{N}$ of θ_0 and we note $\nabla\Phi(\theta, x) = \partial\Phi(\theta, x)/\partial\theta \in \mathbb{R}^{d\times k}$ and $\Psi(\theta, x) = \partial^2\Phi(\theta, x)/\partial\theta\partial\theta^t \in \mathbb{R}^{d\times d\times k}$. Moreover, we assume that $x \mapsto \|\nabla\Phi(., x)\|_{\mathcal{N}}$ and $x \mapsto \|\Psi(., x)\|_{\mathcal{N}}$ are dominated by a μ_0-integrable function.

A3.5. The matrices

$$D := \int_{\mathcal{X}} \nabla\Phi(\theta_0, x)d\mu_0(x) \in \mathbb{R}^{d\times k} \quad \text{and} \quad V := \int_{\mathcal{X}} \Phi(\theta_0, x)\Phi^t(\theta_0, x)d\mu_0(x) \in \mathbb{R}^{k\times k}$$

are of full rank.

Some issues for estimating θ_0 may be due to the indirect definition of the parameter and these assumptions ensure that the map $\theta(.)$ is sufficiently smooth in a neighborhood of μ_0 for the total variation topology, which will make the asymptotic properties of the GEL estimator easily tractable and allow the calculation of efficiency bounds. The next theorem, due to Qin and Lawless [QL94], provides the efficiency bound for estimating θ in the model $\mathcal{M}$.

Theorem 3.2.1 (Theorem 3, [QL94]) *Suppose that Assumptions 1 to 5 hold. The efficiency bound in this model for estimating θ_0 is*

$$B = \left[DV^{-1}D^t\right]^{-1}.$$

An efficiency bound is to be understood as a variance lower bound for asymptotically Gaussian regular estimators of θ_0. We refer to the Appendix for more details. This theorem tells us that if $\hat{\theta}$ is a regular estimator of θ_0, then

$$\liminf_{n\to\infty}\ n \operatorname{var}(\hat{\theta}) \geq B.$$

Efficiency bounds have been derived in more general moment restriction frameworks, such as conditional moment restriction in [Cha87], sequential conditional moment restriction models in [Cha92b], as well as conditional moment conditions with unknown functions in [AC03], [Cha92a], [CP08], [CHT08] and [AC09]. Once the efficiency bound is calculated, the objective is to build an estimator for which the bound is achieved asymptotically.

3.2.1 Generalized empirical likelihood

Generalized empirical likelihood (GEL) was first applied to this problem in [QL94], generalizing an idea of [Owe91]. An estimate $\hat{\mu}$ of μ is obtained as an entropic projection of the empirical measure $\mathbb{P}_n$ onto the model $\mathcal{M}$. Hence, the measure $\hat{\mu}$ is the element of the model that minimizes a given f-divergence $\mathcal{D}_f(.,\mathbb{P}_n)$ with respect to the empirical distribution. We refer to Section 1.2.1 for more precisions. Setting $\mathcal{M}_\theta := \{\mu \in \mathcal{P}(\mathcal{X}) : \mu[\Phi(\theta,.)] = 0\}$, the model can be written as $\mathcal{M} = \cup_{\theta\in\Theta}\mathcal{M}_\theta$. Thus, the GEL estimator $\hat{\theta} = \theta(\hat{\mu})$ follows by

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \ \mathcal{D}_f(\mathcal{M}_\theta, \mathbb{P}_n).$$

Since the set of discrete measures in $\mathcal{M}_\theta$ is closed and convex, the entropy $\mathcal{D}_f(\mathcal{M}_\theta, \mathbb{P}_n)$ is reached for a unique measure $\hat{\mu}(\theta)$ in $\mathcal{M}_\theta$, provided that $\mathcal{D}_f(\mathcal{M}_\theta, \mathbb{P}_n)$ is finite. Then, it appears that computing the GEL estimator involves a two-step procedure. First, build for each $\theta \in \Theta$, the entropic projection $\hat{\mu}(\theta)$ of $\mathbb{P}_n$ onto $\mathcal{M}_\theta$. Then, minimize $\mathcal{D}_f(\hat{\mu}(\theta), \mathbb{P}_n)$ with respect to θ. Since $\hat{\mu}(\theta)$ is absolutely continuous w.r.t. $\mathbb{P}_n$ by construction, minimizing $\mathcal{D}_f(.,\mathbb{P}_n)$ reduces to finding the proper weights $p_1, ..., p_n$ to allocate to the observations $X_1, ..., X_n$. This turns into a finite dimensional problem, which can be solved by classical convex optimization tools (see for instance [Kit06]). In fact, the GEL estimator $\hat{\theta}$ can be expressed as the solution to the saddle point problem

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \ \sup_{(\gamma,\lambda)\in\mathbb{R}\times\mathbb{R}^k} \ \gamma - \mathbb{P}_n\left[f^*(\gamma + \lambda^t\Phi(\theta,.))\right],$$

where $f^*(x) = \sup_y \{xy - f(y)\}$ denotes the convex conjugate of f.

Note that if the choice of the f-divergence plays a key role in the construction of the estimator, it has no influence on its asymptotic efficiency. Indeed, it is shown in [QL94] that all GEL estimators are asymptotically efficient, regardless of the f-divergence used for their computation. Nevertheless, some situations justify the use of specific f-divergences. The empirical likelihood estimator introduced by Owen in [Owe91] uses the Kullback entropy $\mathcal{K}(.,.)$ as f-divergence, pointing out that minimizing $\mathcal{K}(.,\mathbb{P}_n)$ reduces to maximizing likelihood among multinomial distributions. Newey and Smith [NS04] remark that a quadratic f-divergence leads to the CUE estimator of Hansen Heaton and Yaron [HHY96]. In the next section, we discuss a general interpretation of f-divergence in a Bayesian framework.

3.2.2 Maximum entropy on the mean

In this section, we study a Bayesian approach to the inverse problem, known as maximum entropy on the mean (MEM) [GG97]. The method was developed to estimate a measure μ_0 based the observation of some of its moments. In this framework, it turns out that the MEM estimator of μ_0 can be used to estimate efficiently the parameter θ_0. We shall briefly recall the MEM procedure. Consider an estimator of μ_0 in the form of a weighted version of the empirical measure $\mathbb{P}_n$,

$$\mathbb{P}_n(w) = \frac{1}{n}\sum_{i=1}^{n} w_i \ \delta_{X_i},$$

for $w = (w_1, ..., w_n)' \in \mathbb{R}^n$ a collection of weights. Then, fix a prior distribution ν_0 on the vector of weight w so that each solution $\mathbb{P}_n(w)$ can be viewed as a realization of the random measure $\mathbb{P}_n(W)$, where W is drawn from ν_0. This setting enables to incorporate some prior knowledge on the shape or support of μ_0 through the choice of the prior ν_0, as discussed in [GG97]. Here, the observations $X_1, ..., X_n$ are considered fixed. Actually, it is the moment condition that is used to built the estimator *a posteriori*. In this framework where the true value θ_0 of the parameter is unknown, the information provided by the moment condition reduces to the statement $\mu_0 \in \mathcal{M}$. So, in order to take this information into consideration, the underlying idea of MEM is to build the estimator $\hat{\mu}$ as the expectation of $\mathbb{P}_n(W)$ conditionally to the event $\{\mathbb{P}_n(W) \in \mathcal{M}\}$. However, we may encounter some difficulties if this conditional expectation is not properly defined. To deal with this issue, the MEM method replaces the possibly ill-defined conditional expectation by a well-defined estimator, whose construction is motivated by large deviation principles. Precisely, construct the *posterior* distribution ν^* as the entropic projection of ν_0 onto the set

$$\Pi(\mathcal{M}) = \{\mu \in \mathcal{P}(\mathbb{R}^n),\ \mathbb{E}_\mu\left[\mathbb{P}_n(W)\right] \in \mathcal{M}\},$$

where $\mathbb{E}_\mu\left[\mathbb{P}_n(W)\right]$ denotes the expectation of $\mathbb{P}_n(W)$ when W has distribution μ. The MEM solution to the inverse problem is defined as the expectation of $\mathbb{P}_n(W)$ under the posterior distribution ν^*,

$$\hat{\mu} = \mathbb{E}_{\nu^*}\left[\mathbb{P}_n(W)\right] = \mathbb{P}_n(\mathbb{E}_{\nu^*}(W)).$$

This construction is justified by Theorem 2.3 in [GG97], which establishes the asymptotic equivalence between $\hat{\mu}$ and the conditional expectation $\mathbb{E}_{\nu_0}(\mathbb{P}_n(W)|\ \mathbb{P}_n(W) \in \mathcal{M})$, whenever it is well defined. The existence of the MEM estimator requires the problem to be *feasible* in the sense that there exists at least one solution δ in the interior of the convex hull of the support of ν_0, such that $\mathbb{P}_n(\delta) \in \mathcal{M}$. This assumption warrants that the set $\Pi(\mathcal{M})$ is non-empty and therefore allows the construction of the posterior distribution ν^*.

The MEM estimator $\hat{\mu}$ lies in the model $\mathcal{M}$ by construction. As a result, there exists a solution $\hat{\theta}$ to the moment condition $\hat{\mu}[\Phi(\theta, .)] = 0$. this solution is precisely the MEM estimator of θ_0. In Theorem 3.2.2 below, we give an explicit expression for the MEM estimator $\hat{\theta}$. We note $\mathbb{1} = (1, ..., 1)^t \in \mathbb{R}^n$, $\Phi(\theta, X) = (\Phi(\theta, X_1), ..., \Phi(\theta, X_n))^t \in \mathbb{R}^{n\times k}$ and as previously, Λ_ν denotes the log-Laplace transform of ν.

Theorem 3.2.2 *If the problem is feasible, the MEM estimator $\hat{\theta}$ is given by*

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \sup_{(\gamma,\lambda)\in\mathbb{R}\times\mathbb{R}^k} \{n\gamma - \Lambda_{\nu_0}(\gamma\mathbb{1} + \Phi(\theta, X)\lambda)\}.$$

In particular, if ν_0 has orthogonal marginals, i.e. $\nu_0 = \nu^{\otimes n}$ for some probability measure ν on $\mathbb{R}$, then

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \sup_{(\gamma,\lambda)\in\mathbb{R}\times\mathbb{R}^k} \left\{\gamma - \mathbb{P}_n\left[\Lambda_\nu(\gamma + \lambda^t\Phi(\theta, .))\right]\right\}.$$

The MEM estimator $\hat{\theta}$ can be expressed as the solution to a saddle point problem, specific to generalized empirical likelihood. Actually, this result points out that maximum entropy on the mean with a particular form of prior $\nu_0 = \nu^{\otimes n}$ leads to a GEL procedure, for which the criterion is the log-Laplace transform of ν. This approach provides a general Bayesian interpretation of GEL. Regularity conditions on the criterion Λ_ν in the GEL framework are reflected through conditions on the prior ν. Indeed, the usual normalization conditions $\Lambda_\nu'(0) = \Lambda_\nu''(0) = 1$ corresponds to taking a prior ν with mean and variance equal to one, while the normalization $\Lambda_\nu(0) = 0$ is imposed by the condition $\nu \in \mathcal{P}(\mathbb{R})$.

An interesting value of the prior is the exponential distribution $d\nu(x) = e^{-x}\mathbb{1}\{x > 0\}dx$. Observe that if the W_i are i.i.d. with exponential distribution, the likelihood of $\mathbb{P}_n(W)$ is constant over the set of probability discrete measures $\{\mathbb{P}_n(w) : \sum_{i=1}^n w_i = n\}$. Hence, an exponential prior can be roughly interpreted as a non-informative prior in this framework. The discrepancy associated to this prior is $\Lambda_\nu(s) = -\log(1-s),\ s < 1$, which corresponds to the empirical likelihood estimator of Owen [Owe91].

The MEM approach also provides a new probabilistic interpretation of some commonly used specific GEL estimators. The exponential tilting of Kitamura and Stutzer [KS97] is obtained for a Poisson prior of parameter 1, for which we have $\Lambda_\nu(s) = e^s - 1$. Another example is the Gaussian prior $\nu \sim \mathcal{N}(1,1)$, leading to the continuous updating estimator of Hansen, Yeaton and Yaron [HHY96], as we have in this case $\Lambda_\nu(s) = \frac{1}{2}(s-1)^2$. The Gaussian prior allows the discrete measure $\mathbb{P}_n(W)$ to have negative weights w_i and must be handled with care. Remark however that this is generally not an issue in practice since the solution $\hat{\mu}$ is implicitly chosen close to the empirical distribution $\mathbb{P}_n$ and will have all its weights w_i positive with high probability. More examples of classical priors leading to usual discrepancies can be found in [GG97].

3.3 Dealing with an approximate operator

In many actual applications, only an approximation of the constraint function Φ is available to the practitioner. This occurs for instance if the moment condition takes a complicated form that can only be evaluated numerically. In [McF89], McFadden suggested a method dealing with approximate constraint in a similar situation, introducing the method of simulated moments (see also [CF00]). In [LP08] and [LR09b], the authors study a MEM procedure for linear inverse problems with approximate constraints. Here, we propose to extend the results of [LP08] and [LR09b] to the GEL framework, using the connections between GEL and MEM.

We assume that we observe a sequence $\{\Phi_m\}_{m\in\mathbb{N}}$ of approximate constraints, independent with the original sample $X_1, ..., X_n$ and converging toward the true function Φ at a rate φ_m. We are interested in exhibiting sufficient conditions on the sequence $\{\Phi_m\}_{m\in\mathbb{N}}$ under which estimating θ_0 by the GEL procedure remains efficient when the constraint is replaced by its approximation. We discuss the asymptotic properties of the resulting estimates in a framework where both indices n and m simultaneously grow to infinity.

The approximate estimator is obtained by the GEL methodology, replacing the constraint function Φ by its approximation Φ_m,

$$\hat{\theta}_m = \arg\min_{\theta\in\Theta} \sup_{(\gamma,\lambda)\in\mathbb{R}\times\mathbb{R}^k} \left\{\gamma - \mathbb{P}_n\left[\Lambda(\gamma + \lambda^t\Phi_m(\theta, .))\right]\right\}, \tag{3.3}$$

where $\Lambda : \mathbb{R} \to \overline{\mathbb{R}}$ is a strictly convex, twice differentiable function such that $\Lambda'(0) = \Lambda''(0) = 1$ and $\Lambda(0) = 0$. As previously, the existence of $\hat{\theta}_m$ requires the feasibility condition that the supremum of $\gamma - \mathbb{P}_n \left[\Lambda(\gamma + \lambda^t \Phi_m(\theta, .))\right]$ is reached for a finite value of $(\gamma, \lambda) \in \mathbb{R} \times \mathbb{R}^k$, for at least one value of $\theta \in \Theta$. This condition relies essentially on the domain of Λ being sufficiently widespread. We make the following additional assumptions.

A3.6. The functions $x \mapsto \|\Phi(., x)\|_\Theta$, $x \mapsto \|\nabla\Phi(., x)\|_{\mathcal{N}}$ and $x \mapsto \|\Psi(., x)\|_{\mathcal{N}}$ are dominated by a function κ such that $\int \kappa^4(x) d\mu_0(x) < \infty$.

A3.7. For all $x \in \mathcal{X}$ and for sufficiently large m, the map $\theta \mapsto \Phi_m(\theta, .)$ is twice continuously differentiable in $\mathcal{N}$ and we note $\nabla\Phi_m(\theta, .) = \partial\Phi_m(\theta, .)/\partial\theta$ and $\Psi_m(\theta, .) = \partial^2\Phi_m(\theta, .)/\partial\theta\partial\theta^t$.

A3.8. The functions $x \mapsto \|\Phi_m(., x) - \Phi(., x)\|_\Theta$, $x \mapsto \|\nabla\Phi_m(., x) - \nabla\Phi(., x)\|_{\mathcal{N}}$ and $x \mapsto \|\Psi_m(., x) - \Psi(., x)\|_{\mathcal{N}}$ are dominated by a function κ_m such that $\int \kappa_m^4(x) d\mu_0(x) = O(\varphi_m^{-4})$.

A3.9. The function Λ'' is bounded by a constant $K < \infty$.

Assumptions A3.6 to A3.8 are made to obtain a uniform control over $\|\hat{\theta}_m - \hat{\theta}\|$ for all $n \in \mathbb{N}$. The condition A3.9 implies that Λ is dominated by a quadratic function. In the MEM point of view, this condition is fulfilled for the log-Laplace transform Λ_ν of sub-Gaussian priors ν.

Theorem 3.3.1 (Robustness of GEL) *If Assumptions 1 to 9 hold,*

$$n\|\hat{\theta}_m - \hat{\theta}\|^2 = O_P(n\varphi_m^{-2}) + o_P(1).$$

In particular, $\hat{\theta}_m$ is $\sqrt{n}$-consistent and asymptotically efficient if $n\varphi_m^{-2}$ tends to zero.

By considering a situation with approximate operator, we extend the GEL model to a more general framework that gives a more realistic formulation of actual problems. The previous theorem gives an upper bound of the error caused by the use of the approximation Φ_m in place of the true function Φ. By this result, we aim to provide an insight on convergence conditions that are necessary for asymptotic efficiency when dealing with an approximate operator.

3.4 Proofs

3.4.1 Proof of Theorem 3.2.2

Let $\mathcal{S}_\theta = \{w \in \mathbb{R}^n : \mathbb{P}_n(w) \in \mathcal{M}_\theta\}$ and $\mathcal{F}_w = \{\mu \in \mathcal{P}(\mathbb{R}^n) : \mathbb{E}_\mu(W) = w\}$. We use that $\inf_{\mu\in\mathcal{F}_w} \mathcal{K}(\mu, \nu_0) = \Lambda^*_{\nu_0}(w)$, as shown in the proof of Theorem 1.2.3 in Section 1.2.2. Let $\Pi(\mathcal{M}_\theta) = \{\mu \in \mathcal{P}(\mathbb{R}^n),\ \mathbb{E}_\mu\left[\mathbb{P}_n(W)\right] \in \mathcal{M}_\theta\}$, we have the equality

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \inf_{\mu\in\Pi(\mathcal{M}_\theta)} \mathcal{K}(\mu, \nu_0) = \arg\min_{\theta\in\Theta} \inf_{w\in\mathcal{S}_\theta} \inf_{\mu\in\mathcal{F}_w} \mathcal{K}(\mu, \nu_0),$$

which can be written

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \inf_{w\in\mathcal{S}_\theta} \Lambda^*_{\nu_0}(w) = \arg\min_{\theta\in\Theta} \inf_{w\in\mathcal{S}_\theta} \sup_{\tau\in\mathbb{R}^n} \{\tau^t w - \Lambda_{\nu_0}(\tau)\}.$$

The feasibility assumption warrants that the extrema are reached. Hence, using Sion's minimax Theorem, we find

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \sup_{\tau\in\mathbb{R}^n} \inf_{w\in\mathcal{S}_\theta} \{\tau^t w - \Lambda_{\nu_0}(\tau)\},$$

We know that $w = (w_1, ..., w_n)^t \in \mathcal{S}_\theta$ if and only if $\sum_{i=1}^n w_i = n$ and $\sum_{i=1}^n w_i \Phi(\theta, X_i) = 0$. Thus, for a fixed value of τ, the map $w \mapsto \tau^t w - \Lambda_{\nu_0}(\tau)$ can be arbitrarily close to $-\infty$ on $\mathcal{S}_\theta$ whenever τ is not orthogonal to $\mathbb{1}$ and $\Phi(\theta, X)$. As a result, we may assume that $\tau = \gamma\mathbb{1} + \Phi(\theta, X)\lambda$ for some $(\gamma, \lambda) \in \mathbb{R} \times \mathbb{R}^k$ without loss of generality. In this case, the map $w \mapsto \tau^t w - \Lambda_{\nu_0}(\tau)$ is constant over $\mathcal{S}_\theta$, equal to $n\gamma - \Lambda_{\nu_0}(\gamma\mathbb{1} + \Phi(\theta, X)\lambda)$, which ends the proof. If $\nu_0 = \nu^{\otimes n}$, then $\Lambda_{\nu_0}(w) = \sum_{i=1}^n \Lambda_\nu(w_i)$ and we conclude easily.

3.4.2 Proof of Theorem 3.3.1

The proof of the results relies mainly on the uniform law of large numbers, using that the set $\{\|\Phi_m(\theta, .)\|, \|\nabla\Phi_m(\theta, .)\|, \|\Psi_m(\theta, .)\|,\ \theta \in \Theta, m \in \mathbb{N}\}$ is a Glivenko-Cantelli class of functions, consequently to A3.6 and A3.8. For all $\theta \in \Theta$, $v \in \mathbb{R}^k$, $x \in \mathcal{X}$, let

$$\begin{aligned} h_n(\theta, v) &= \begin{pmatrix} \mathbb{P}_n\left[\Phi(\theta, .)\Lambda'(v^t\Phi(\theta, .))\right] \\ \mathbb{P}_n\left[v^t\nabla\Phi^t(\theta, .)\Lambda'(v^t\Phi(\theta, .))\right] \end{pmatrix} \\ h_{m,n}(\theta, v) &= \begin{pmatrix} \mathbb{P}_n\left[\Phi_m(\theta, .)\Lambda'(v^t\Phi_m(\theta, .))\right] \\ \mathbb{P}_n\left[v^t\nabla\Phi_m^t(\theta, .)\Lambda'(v^t\Phi_m(\theta, .))\right] \end{pmatrix}. \end{aligned}$$

The pair $(\hat{\theta}_m, \hat{v}_m)$ (resp. $(\hat{\theta}, \hat{v})$) is defined as the unique zero over $\Theta \times \mathbb{R}^k$ of $h_{m,n}$ (resp. h_n). The condition A3.9 implies that there exists a constant $K > 0$ such that $\Lambda'(s) \le Ks + 1$ for all $s \in \mathbb{R}$. Hence, using successively the mean value theorem and Cauchy-Schwarz's inequality, we show that the contrast function $h_{m,n}$ converges uniformly on every compact set toward h_n as $m \to \infty$, which warrants the convergence of $(\hat{\theta}_m, \hat{v}_m)$ toward $(\hat{\theta}, \hat{v})$. For all $v \in \mathbb{R}^k$, the application $\theta \mapsto \nabla h_{m,n}(\theta, v)$ is continuous in a neighborhood on θ_m^* for sufficiently large values of m by the condition A3.7, as explicit calculation gives

$$\nabla h_{m,n}(\theta, v) = \begin{pmatrix} A_{m,n}(\theta, v) & D_{m,n}(\theta, v) \\ D_{m,n}^t(\theta, v) & V_{m,n}(\theta, v) \end{pmatrix},$$

where

$$\begin{aligned} A_{m,n}(\theta, v) &= \mathbb{P}_n\left[\Psi_m(\theta, .)v\Lambda'(v^t\Phi_m(\theta, .)) + \nabla\Phi_m(\theta, .)v\ v^t\nabla\Phi_m^t(\theta, .)\Lambda''(v^t\Phi_m(\theta, .))\right] \\ D_{m,n}(\theta, v) &= \mathbb{P}_n\left[\nabla\Phi_m(\theta, .)\Lambda'(v^t\Phi_m(\theta, .)) + \nabla\Phi_m(\theta, .)v\Phi_m^t(\theta, .)\Lambda''(v^t\Phi_m(\theta, .))\right] \\ V_{m,n}(\theta, v) &= \mathbb{P}_n\left[\Phi_m(\theta, .)\Phi_m^t(\theta, .)\Lambda''(v^t\Phi_m(\theta, .))\right]. \end{aligned}$$

We define in the same way $A_n(\theta, v)$, $D_n(\theta, v)$ and $V_n(\theta, v)$ by replacing Φ_m by Φ in the expressions above. Using Cauchy-Schwarz's inequality, A3.8 ensures the uniform convergence of $\nabla h_{m,n}$ toward ∇h_n on every compact set at the rate φ_m. Note ρ_n the smallest eigenvalue of $\nabla h_n(\hat{\theta}, \hat{v})$, we know from Theorem 3.2 in [NS04] that $\mathbb{P}(\rho_n > \eta) = o(n^{-1})$ for sufficiently small $\eta > 0$, since A3.5 ensures that the limit of $\nabla h_n(\hat{\theta}, \hat{v})$ as $n \to \infty$ is positive definite. Thus, for $c > 0$ sufficiently small, consider the event $\Omega = \{\rho_n > c\}$. Writing the Taylor expansion

$$h_{m,n}(\hat{\theta}, \hat{v}) = \nabla h_{m,n}(\hat{\theta}_m, \hat{v}_m) \begin{pmatrix} \hat{\theta} - \hat{\theta}_m \\ \hat{v} - \hat{v}_m \end{pmatrix} + o(\|\hat{\theta}_m - \hat{\theta}\|),$$

we deduce that on Ω,

$$\begin{pmatrix} \hat{\theta}_m - \hat{\theta} \\ \hat{v}_m - \hat{v} \end{pmatrix} = -\left[\nabla h_n(\hat{\theta}, \hat{v})\right]^{-1} h_{m,n}(\hat{\theta}, \hat{v}) + O_P(\varphi_m^{-1}).$$

The Schur complement formula gives in particular

$$\hat{\theta}_m - \hat{\theta} = -\left[\hat{D}_n \hat{V}_n^{-1} \hat{D}_n^t\right]^{-1} \hat{D}_n \hat{V}_n^{-1}\ \mathbb{P}_n[\Phi_m(\hat{\theta}, .)\Lambda'(\hat{v}^t \Phi_m(\hat{\theta}, .))] + O_P(\varphi_m^{-1}) + o_P(n^{-1}),$$

where $\hat{D}_n = D_n(\hat{\theta}, \hat{v})$ and $\hat{V}_n = V_n(\hat{\theta}, \hat{v})$ and where we used that $\hat{v} = O_P(n^{-1})$ (see for instance Theorem 3.2 in [NS04]). Thus, on the event Ω,

$$\|\hat{\theta}_m - \hat{\theta}\| \leq c \left\|\mathbb{P}_n[\Phi_m(\hat{\theta}, .)\Lambda'(\hat{v}^t \Phi_m(\hat{\theta}, .))]\right\| + O_P(\varphi_m^{-1}) + o_P(n^{-1}).$$

By construction, $\mathbb{P}_n[\Phi(\hat{\theta}, .)\Lambda'(\hat{v}^t \Phi(\hat{\theta}, .))] = 0$, which yields

$$\begin{aligned}
& \left\|\mathbb{P}_n[\Phi_m(\hat{\theta}, .)\Lambda'(\hat{v}^t \Phi_m(\hat{\theta}, .))]\right\| \\
\leq\ & \mathbb{P}_n\left[\|(\Phi_m(\hat{\theta}, .) - \Phi(\hat{\theta}, .))\Lambda'(\hat{v}^t \Phi_m(\hat{\theta}, .))\| + \|\Phi(\hat{\theta}, .)[\Lambda'(\hat{v}^t \Phi_m(\hat{\theta}, .) - \Lambda'(\hat{v}^t \Phi(\hat{\theta}, .))]\|\right] \\
\leq\ & K\|\hat{v}\|\ \mathbb{P}_n\left[\|\Phi_m(\hat{\theta}, .)\|\ \|\Phi_m(\hat{\theta}, .) - \Phi(\hat{\theta}, .)\|\right] + \mathbb{P}_n\|\Phi_m(\hat{\theta}, .) - \Phi(\hat{\theta}, .)\| \\
& +K\|\hat{v}\|\ \mathbb{P}_n\left[\|\Phi(\hat{\theta}, .)\|\ \|\Phi_m(\hat{\theta}, .) - \Phi(\hat{\theta}, .)\|\right],
\end{aligned}$$

as a consequence of A3.9. We conclude that $\|\hat{\theta}_m - \hat{\theta}\|^2 \mathbb{1}_\Omega = O_P(\varphi_m^{-2}) + o_P(n^{-1})$ by the condition A3.8. On the complement of Ω, $\|\hat{\theta}_m - \hat{\theta}\|$ can be bounded by the diameter δ of Θ, yielding $\|\hat{\theta}_m - \hat{\theta}\|\mathbb{1}_{\Omega^c} = o_P(n^{-1})$, which ends the proof.

3.5 Efficiency of the generalized method of moments

The main alternative to GEL is the generalized method of moments (GMM). Although GEL and GMM aim to solve the same estimation problem, the methods rely on different semiparametric models. By calculating efficiency bounds relative to the GMM model, we provide a new proof Hansen's result in [Han82] on optimal GMM.

3.5.1 Information theory in semiparametric models

This section is devoted to some results on Information theory and efficiency in semiparametric models, for which a further study can be found in [vdV98] and [RB90]. This theory aims to quantify the amount of information available in a given statistical problem. It applies mostly to classical sampling models, in which the objective is to estimate a parameter $\psi(\mu_0)$ from i.i.d. observations $X_1, ..., X_n$ drawn from an unknown measure μ_0.

Definition A *model* $\mathcal{M}$ is a set of probability measures, i.e. it is a subset of $\mathcal{P}(\mathcal{X})$.

For a given problem, the model is the set of possible for the distribution μ_0 of the observations. We generally assume that the model $\mathcal{M}$ is suitably defined in the sense that it contains the true measure μ_0. There exist in the literature three main kinds of models:

- Parametric models, $\mathcal{M} = \{\mu_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$. The distribution of the observations is assumed to belong to a parametric class of measure, which reflects a significant knowledge on μ_0. To restrict the set of possible values of μ_0 to a parametric model is usually a strong assumption, although it may be reasonable in many situations (Gaussian models, Poisson models, etc...).

- Nonparametric model, $\mathcal{M} = \mathcal{P}(\mathcal{X})$. We have no available information on μ_0, which forces us to consider any possible probability measure as a possible solution.

- Semiparametric models, $\mathcal{M}$ is neither parametric, nor nonparametric. We have some information on μ_0 although the set of possible values of μ_0 can not be identified with a finite dimensional space (e.g. density measures, Cox model, centered measures, etc...).

The description of the model helps characterizing the available information on μ_0, which is all the more important that the model is restricted.

Definition A *parameter* with values in a space $\mathcal{H}$ is a map $\psi : \mathcal{M} \to \mathcal{H}$.

A value of the parameter is defined for any element of the model. Classical examples of parameters are the mean, the median or the variance but we may also consider infinite dimensional parameters such the density with respect to Lebesgue measure $\psi : \nu \mapsto d\nu/d\lambda$ or any reference measure in a suitable model. The measure itself can also be seen as a parameter, the map ψ being the identity in this case.

Definition A model $\{\mu_t\}_{t\geq 0}$ is *differentiable in quadratic mean* at μ_0 if there exists $g : \mathcal{X} \to \mathbb{R}$ such that $\int_{\mathcal{X}} g^2 \, d\mu_0 < \infty$ and

$$\lim_{t\to 0} \int_{\mathcal{X}} \left[\frac{1}{t} \left(\sqrt{\frac{d\mu_t}{d\tau_t}} - \sqrt{\frac{d\mu_0}{d\tau_t}} \right) - \frac{1}{2}\, g \right]^2 d\tau_t = 0,$$

setting for all $t \geq 0$, $\tau_t = \mu_t + \mu_0$.

The function g is called the *score* of $\{\mu_t\}_{t\geq 0}$ and it has zero mean under μ_0. For any function $T_n : \mathcal{X}^n \to \Theta$ of the observations and μ a measure, we denote by $\mathcal{L}(T_n|\ \mu)$ the law of $T_n(X_1, ..., X_n)$ when $X_1, ..., X_n$ are independent with distribution μ. Following the notations in [RB90], we denote by $\dot{\mathcal{M}}$ the tangent space of $\mathcal{M}$ at μ_0, defined as the set of all score functions of differentiable submodels in $\mathcal{M}$.

Definition A parameter $\psi : \mathcal{M} \to \mathcal{H} \subset \mathbb{R}^p$ defined on a model $\mathcal{M}$ is said to be *differentiable* at μ_0 if there exists a function $\dot{\psi} : \mathcal{X} \to \mathbb{R}^p$ such that, for all differentiable submodels $\{\mu_t\}_t \subset \mathcal{M}$,

$$\lim_{t\to 0} \frac{\psi(\mu_t) - \psi(\mu_0)}{t} = \int_{\mathcal{X}} \dot{\psi}\ g\ d\mu_0, \tag{3.4}$$

where g is the score $\{\mu_t\}_t$.

A function $\dot{\psi}$ satisfying (3.4) (called *influence function*) is not uniquely defined, as the differentiability property only involves components of ψ that are orthogonal to the tangent space $\dot{\mathcal{M}}$. However, there is only one influence function with minimal norm, called *efficient influence function* and which we note $\tilde{\psi}$.

Definition An estimator $\hat{\psi} = \hat{\psi}(X_1, ..., X_n)$ of a parameter $\psi : \mathcal{M} \to \mathcal{H}$ is *locally Gaussian regular* at $\psi(\mu_0)$ if for all differentiable submodel $\{\mu_t\}_{t\geq 0} \subset \mathcal{M}$ and for all positive sequence $(t_n)_{n\in\mathbb{N}}$ such that $\sqrt{n}t_n$ is bounded, $\mathcal{L}(\sqrt{n}(\hat{\psi} - \psi(\mu_{t_n}))|\ \mu_{t_n})$ converges weakly toward a Gaussian distribution as $n \to \infty$.

The following theorem, which is a consequence of the convolution Theorem in [RB90], establishes a lower bound on the variance of locally Gaussian regular estimators.

Theorem 3.5.1 *Let T_n be a locally Gaussian regular estimator of ψ in $\mathcal{M}$. Then*

$$\liminf_{n\to\infty} n \ \mathrm{var}(T_n) \geq \int_{\mathcal{X}} \tilde{\psi}\tilde{\psi}^t d\mu_0.$$

The quantity $\int \tilde{\psi}\tilde{\psi}^t d\mu_0 \in \mathbb{R}^{p\times p}$ is called *efficiency bound* to estimate ψ in $\mathcal{M}$.

3.5.2 Semiparametric efficiency of GMM

As discussed in Chapter 3 in [BKRW93], there exist two natural, although seemingly different approaches to estimate a parameter $\theta_0 = \theta(\mu_0)$ based on i.i.d. observations drawn from μ_0 in a statistical model $\mathcal{M}$. Roughly speaking, the two approaches are based on the idea of building a preliminary estimate $\hat{\mu}$ of μ_0 in order to define the estimator $\hat{\theta} = \theta(\hat{\mu})$. In this process, we can face two different situations. On one hand, if one uses a classical estimate $\hat{\mu}$ such as the empirical distribution, the estimator $\hat{\theta} = \theta(\hat{\mu})$ is defined only if $\hat{\mu} \in \mathcal{M}$. Since this is not true in most cases, the map $\theta(.)$ needs to be extended to a larger model that contains the estimator $\hat{\mu}$. Precisely, one considers a model $\mathcal{P} \supseteq \mathcal{M}$ and an extension $\overline{\theta}$ of θ on $\mathcal{P}$ in order to construct the estimator $\hat{\theta} = \overline{\theta}(\hat{\mu})$. The main advantage with this approach is that few conditions are needed on the preliminary estimate $\hat{\mu}$. The efficiency of the method will essentially rely on the construction of the map $\overline{\theta}$, which must be sufficiently smooth to avoid a loss of information. A second idea is to impose that the preliminary estimate $\hat{\mu}$ lies in the model $\mathcal{M}$ so that θ_0 can be obtained directly as the image of $\hat{\mu}$ through $\theta(.)$. This approach is arguably the most natural one, although it may involve constrained optimization problems, making the method generally more expensive computationally.

In the parametric moment condition model, the two main methods that have been implemented for this problem, namely the generalized empirical likelihood and the generalized method of moments, provide a good illustration of each situation. The generalized empirical likelihood is a good example of the second idea, where one considers a preliminary estimate $\hat{\mu}$ as a discrete measure lying in the model $\mathcal{M}$. Here, we discuss the generalized method of moments which illustrates the first procedure of estimation. We shall see that the method is efficient as soon as the extended map $\overline{\theta}$ is sufficiently smooth on the initial model $\mathcal{M}$.

The generalized method of moments consists in replacing in the moment constraint, the true measure μ_0 by its empirical approximation $\mathbb{P}_n$. Then, find the value of θ for which the empirical moment condition $\mathbb{P}_n[\Phi(\theta, .)] = \frac{1}{n}\sum_{i=1}^{n} \Phi(\theta, X_i)$ is the closest to 0, according to a given norm of $\mathbb{R}^k$. Precisely, define for M a symmetric positive definite $k \times k$ matrix and $a \in \mathbb{R}^k$, $\|a\|_M^2 = a^t M a$. The GMM estimator $\hat{\theta}$ of θ_0 associated to the norm $\|.\|_M$ is given by

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \; \|\mathbb{P}_n[(\Phi(\theta, .)]\|_M.$$

If the minimizer is not unique, one may select one arbitrarily among all minimizers. In practice, the matrix M may have a dependency in n, in which case it is generally chosen to converge toward a symmetric positive definite matrix. Nevertheless, replacing the matrix by its limit leads to the same first order asymptotic properties of the estimate, under regularity conditions. Here, we will assume for simplicity that M is fixed, this being sufficient for our purposes.

The generalized method of moments is a good illustration of the first procedure. Indeed, the GMM estimator $\hat{\theta}$ can be seen as the image of the empirical distribution $\mathbb{P}_n$ by the function

$$\overline{\theta}_M(\mu) = \arg\min_{\theta \in \Theta} \; \|\mu[\Phi(\theta, .)]\|_M, \;\; \mu \in \mathcal{P},$$

where $\mathcal{P}$ is an extension of the original model $\mathcal{M}$, containing $\mathbb{P}_n$. For example, we may take $\mathcal{P}$ as the set of all probability measures μ for which the map $\theta \mapsto \mu[\Phi(\theta, .)]$ can take finite values on Θ.

3.5.3 A proof of optimal GMM

The generalized method of moments may seem inefficient since it involves a larger model, thus decreasing the amount of available information. Actually, to be able to provide an efficient estimator, the extension $\overline{\theta}_M$ must be "smooth" enough so that differentiable submodels in $\mathcal{P}$ carry at least as much information as the original model. Basically, we want the efficiency bound $\overline{B}_M$ for estimating $\overline{\theta}_M$ over $\mathcal{P}$ not to be higher than the original bound B. Since it obviously can not be lower, the objective is to find an efficient extension, for which $\overline{B}_M = B$. In the next theorem, we show that the smoothness of the extension $\overline{\theta}_M$ can be measured in function of the scaling matrix M, via the direct calculation of the efficiency bound $\overline{B}_M$.

Theorem 3.5.2 *Suppose that Assumption 1 to 5 hold. The efficiency bound for estimating* $\overline{\theta}_M$ *in* $\mathcal{P}$ *is*

$$\overline{B}_M = \left[DMD^t\right]^{-1}\left[DMVMD^t\right]\left[DMD^t\right]^{-1}.$$

Proof. Note $\mathcal{T}$ the set of bounded functions with zero mean under μ_0. For any $g \in \mathcal{T}$ and $t > 0$, the measure $\mu_t := (1 + tg)\mu_0$ lies in $\mathcal{P}$ provided that t is small enough and the path $\{\mu_t\}_{t\geq 0}$ is differentiable with score g. The uniform convergence of $\theta \mapsto \mu_t[\Phi(\theta, .)]$ toward $\theta \mapsto \mu_0[\Phi(\theta, .)]$ as $t \to 0$ (which follows from A1 and A2) ensures the existence of a minimizer $\theta(t)$ of $\theta \mapsto \mu_t[\Phi(\theta, .)]$ continuously close to θ_0 as $t \to 0$ and satisfying the first order condition $\gamma_M(\theta(t), \mu_t) = 0$ where

$$\gamma_M(\theta, \mu) = \left[\int \nabla\Phi(\theta, .)d\mu\right] M \left[\int \Phi(\theta, .)d\mu\right], (\theta, \mu) \in \Theta \times \mathcal{P}.$$

Under A2, A3 and A4, the implicit functions theorem applied to the map $(\theta, t) \mapsto \gamma_M(\theta, \mu_t)$ in a neighborhood of $(\theta_0, 0)$ warrants the uniqueness of the minimum $\theta(t) = \overline{\theta}_M(\mu_t)$. Note $\dot{\ell} = (\dot{\ell}_1, ..., \dot{\ell}_d)^t$ the efficient influence function of $\overline{\theta}_M$. By a Taylor expansion of $\Phi(\theta, .)$ at θ_0 and using that $\gamma_M(\overline{\theta}_M(\mu_t), \mu_t) = 0$, we get

$$\left[\int \nabla\Phi(\theta_0, .)d\mu_t\right] M \left[\int \Phi(\theta_0, .)(1 + tg)d\mu_0 + \left[\int \nabla\Phi^t(\theta_0, .)d\mu_t\right](\overline{\theta}_M(\mu_t) - \theta_0)\right] = o(t).$$

Since $\overline{\theta}_M(\mu_t) - \theta_0 = t\int \dot{\ell}gd\mu_0 + o(t)$ by definition of $\dot{\ell}$, we obtain after dividing each term by t and making t tend to zero

$$DM\left[\int \Phi(\theta_0, .)g\ d\mu_0\right] = -DMD^t\left[\int \dot{\ell}\ g\ d\mu_0\right].$$

Since this holds for all $g \in \mathcal{T}$, we deduce using A5

$$\dot{\ell}(.) = -\left[DMD^t\right]^{-1} DM\Phi(\theta_0, .),$$

checking beforehand that $\dot{\ell}$ lies in the closure of $\mathcal{T}$. The efficiency bound is the variance of $\dot{\ell}(X)$ which proves the result.

The following lemma proves that, as expected, the efficiency bound $\overline{B}_M$ in the extended model $\mathcal{P}$ is larger than in the original model $\mathcal{M}$. We note $\mathrm{Im}(D^t) = \{D^t u,\ u \in \mathbb{R}^d\} \subset \mathbb{R}^k$.

Lemma 3.5.3 *For all symmetric positive semidefinite matrix M such that MD^t is of full rank,*

$$DMD^t \left[DMVMD^t\right]^{-1} DMD^t \leq DV^{-1}D^t,$$

with equality if and only if $\mathrm{Im}(D^t)$ *is stable through VM, i.e. $\forall v \in \mathrm{Im}(D^t),\ VMv \in \mathrm{Im}(D^t)$.*

Proof. Set $A = V^{1/2}MD^t$, $A[A^tA]^{-1}A^t$ is an orthogonal projection matrix with in particular $A[A^tA]^{-1}A^t \leq I$. Thus, we deduce

$$DV^{-1/2}\ A[A^tA]^{-1}A^t\ V^{-1/2}D^t \leq DV^{-1}D^t,$$

proving the inequality. To have the equality, we need that $V^{-1/2}D^t = AX$ for some $X \in \mathbb{R}^{d\times d}$, or equivalently $D^t = VMD^tX$. Since D, V and MD^t are assumed of full rank, it follows that X is invertible. Finally, the condition $MD^t = V^{-1}D^tX^{-1}$ is sufficient to have the equality, proving the lemma.

This result provides necessary and sufficient conditions for the optimality of GMM. The asymptotic variance of the GMM estimator is precisely the lower bound $\overline{B}_M$, as shown in [Han82], which proves the efficiency of the method. The theorem and lemma point out in particular that the optimality of GMM is achieved for $M = V^{-1}$, as we have in this case $\overline{B}_M = B$, recovering the efficiency bound of Theorem 3.2.1. Note that the matrix V is generally unknown, since it depends on the true measure μ_0. However, it can be replaced by a consistent estimate, leading to the same asymptotic properties of the GMM estimator under mild conditions. Here again, several approaches are possible. In the two-step GMM procedure, the estimate $\tilde{V}$ is built using a preliminary estimator $\tilde{\theta}$ of θ_0 obtained by a GMM procedure with known scaling matrix (in general, the identity matrix). As a result, $\tilde{\theta}$ is not in general asymptotically efficient, however, it is $\sqrt{n}$-consistent and enables to construct a consistent estimate of V. The resulting estimator can be viewed as a two-step semiparametric estimator, for which the optimal scaling matrix V^{-1} is estimated in a first step and is used for the estimation of θ_0 in a second step (see for instance [ACH11]). Another solution is to minimize simultaneously over Θ

$$\theta \mapsto \mathbb{P}_n[\Phi^t(\theta,.)]\ \hat{V}^{-1}(\theta)\ \mathbb{P}_n[\Phi(\theta,.)], \tag{3.5}$$

where $\hat{V}^{-1}(\theta)$ denotes here an arbitrary consistent estimate of $V^{-1}(\theta)$, for all $\theta \in \Theta$. In particular, taking $\hat{V}^{-1}(\theta)$ as the inverse of the empirical variance of $\Phi(\theta, X)$ recovers the continuous updating estimator of [HHY96].

Remark that the choice $M = V^{-1}$ is not the only one achieving optimality. For instance, the efficiency bound can be shown to be minimal if $d = k$, for any full rank matrix M. Moreover, the lemma does not rule out that M is positive semidefinite. Actually, if $d < k$, we can find non-invertible symmetric positive semidefinite matrices M for which the efficiency bound $\overline{\theta}_M$ is minimal. Examples are $M = V^{-1}D^tDV^{-1}$ or $M = V^{-1}\Pi_DV^{-1}$ where Π_D denotes the orthogonal projector onto $\mathrm{Im}(D^t)$, $\Pi_D = D^t[DD^t]^{-1}D$. These matrices have rank d and satisfy all the required conditions of the lemma, as well as the equality $\overline{B}_M = B$. The use of a scaling matrix M different from V^{-1} achieving the optimal efficiency bound might be interesting if M is easier to estimate than V^{-1}. Since the computation of the GMM estimator relies on the preliminary estimation of M, this may improve the small sample properties of the estimator.

3.6 Extension: An efficiency bound for non-smooth models

In this section, we introduce a variance lower bound for unbiased estimators in general statistical models. The construction of the efficiency bound is based on the same idea as the Cramer-Rao inequality, however it does not require differentiability conditions, which enables the calculation of the efficiency bound in a larger class of models. Moreover, we show that in many situations, this efficiency bound is actually larger than the Cramer-Rao bound, and therefore provides a sharper result.

3.6.1 Preliminary results

Let $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ be an open subset of $\mathbb{R}^p$ endowed with its Borel field, we denote by $\mathcal{M}(\mathcal{X})$ the set of all probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Let $\{\mu_\theta\}_{\theta\in\Theta}$ be a family of probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ with $\mu_{\theta_0} = \mu_0$ and with $\theta \in \Theta \subset \mathbb{R}^d$.

Definition We say that $\{\mu_\theta\}_{\theta\in\Theta}$ is *differentiable in* $\mathbb{L}^2(\mu_0)$ at θ_0, if there exists $g : \mathcal{X} \to \mathbb{R}^d$ such that $\int_{\mathcal{X}} g^t g \, d\mu_0 < \infty$ and such that for all $a \in \mathbb{R}^d$,

$$\lim_{t\to 0} \int_{\mathcal{X}} \left[\frac{1}{t} \left(\frac{d\mu_{\theta_0+ta}}{d\mu_0}(x) - 1 \right) - a^t g(x) \right]^2 d\mu_0(x) = 0.$$

This notion of differentiability is actually less general than the differentiability in quadratic mean, although in most situations, it leads to the same results in "large" models (semiparametric or nonparametric) but also in most exponential parametric models. We call g the score function of the model $\{\mu_\theta\}_\theta$ at $\theta = \theta_0$. The variance of $g(X)$ is the Fisher information matrix of the model.

Theorem 3.6.1 (Cramer-Rao inequality) *Let $\mathcal{M} = \{\mu_\theta\}_{\theta\in\Theta}$ be a differentiable model with invertible Fisher information matrix. Let $\psi : \mathcal{M} \to \mathcal{H} \subset \mathbb{R}^q$ be a map such that $\theta \mapsto \psi(\mu_\theta)$ is differentiable at θ_0 with $d \times q$ derivative matrix $\dot{\psi}_0$. Then, if $T = T(X_1, ..., X_n)$ is an unbiased estimator of $\psi(\mu_0)$,*

$$n \text{ var}(T) \geq \dot{\psi}_0^t \left[\int_{\mathcal{X}} g g^t d\mu_0 \right]^{-1} \dot{\psi}_0.$$

The Cramer-Rao bound in the model $\mathcal{M}$ for estimating $\psi(\mu_0)$ will be noted $B_\psi(\mathcal{M})$ (the dependency in $\mathcal{M}$ in the notation may be dropped if there is no possible confusion). This well-known result that applies to parametric models can be extended to larger models. Assume that $\mathcal{M}$ is a semiparametric model, for every smooth submodel $\{\mu_\theta\}_{\theta\in\Theta} \subset \mathcal{M}$ such that $\mu_{\theta_0} = \mu_0$, we can calculate the Fisher information for estimating $\psi(\mu_0)$. Moreover, the information for estimating $\psi(\mu_0)$ over the whole model is certainly not bigger than the information of any submodel. We shall simply define the information for the whole model as the infimum (if one exists) of the informations over all smooth submodels. Equivalently, the lower bound for the variance of an unbiased estimator of $\psi(\mu_0)$ is defined as the supremum of Cramer-Rao bounds over smooth submodels.

Remark The maximum is unique whenever the parameter to estimate $\psi(\mu_0)$ is real. However, the uniqueness may become an issue for a vectorial parameter. If $\psi : \mathcal{M} \to \mathbb{R}^d$ with $d > 1$, a variance lower bound for estimating $\psi(\mu_0)$ would have to be a $d \times d$ matrix. Then, it may occur that several maxima exist (a maximum is meant in the sense of the quadratic forms: we say that two $d \times d$ matrices M and N are such that $M \geq N$ if $M - N$ is nonnegative-definite). In such situations, we consider a set $\mathcal{B}_\psi$ of maxima where each element may need be studied separately.

In the sequel, a maximum will be meant as an element of the set $\mathcal{B}_\psi$. Let $\mathcal{T}$ be an open subset of $\mathbb{R}$ containing 0, we denote by $\mathcal{M}_{\psi,\mathcal{T}}$ the set of all differentiable one-dimensional submodels $\{\mu_t\}_{t\in\mathcal{T}}$ equal to μ_0 at 0 and with score $g \neq 0$. Assume that the map ψ is differentiable at μ_0, that is, $t \mapsto \psi(\mu_t)$ is differentiable at 0 for all $\{\mu_t\}_t \in \mathcal{M}_{\psi,\mathcal{T}}$. We define the efficiency bound as

$$B_\psi(\mathcal{M}) = \sup_{(\mu_t)_t \in \mathcal{M}_{\psi,\mathcal{T}}} B_\psi(\{\mu_t\}_t).$$

By convention, we extend the definition of the Cramer-Rao bound to any model $\mathcal{M}$ by stating that $B_\psi(\mathcal{M}) = 0$ as soon as $\mathcal{M}_{\psi,\mathcal{T}} = \emptyset$. A CR bound can thus be defined for any statistical model as a supremum over all CR bounds of one-dimensional smooth submodels. A submodel for which the supremum is reached is called a *least favorable path*.

An interesting example to illustrate this definition is to consider a multi-dimensional parametric model. Let Θ be an open set of $\mathbb{R}^q$ and let $\{\mu_\theta\}_{\theta\in\Theta}$ be a parametric model with $\mu_{\theta_0} = \mu_0$ and score $g : \mathcal{X} \to \mathbb{R}^q$. The parameter is defined by the map $\psi : \{\mu_\theta\}_{\theta\in\Theta} \to \mathbb{R}$. One-dimensional smooth submodels are of the form $\mu_t^a = \mu_{\theta_0+ta+o(t)}$ for $a \in \mathbb{R}^q$, with associated score function $a^t g$. So, the generalized CR bound in the model is given by

$$B_\psi(\mathcal{M}) = \sup_{a\in\mathbb{R}^{q*}} \frac{a^t \dot\psi(\theta_0)\dot\psi^t(\theta_0)a}{a^t \left[\int_\mathcal{X} gg^t d\mu_0\right] a} = \dot\psi_0^t \left[\int_\mathcal{X} gg^t d\mu_0\right]^{-1} \dot\psi_0.$$

We here recover the Cramer-Rao bound of Theorem 3.6.1.

3.6.2 Construction of the efficiency bound

Definition Let μ and ν be two probability measures on $\mathcal{X}$. The quadratic divergence (or Q-divergence) of ν with respect to μ is given by:

$$\mathcal{D}(\mu,\nu) = \int_\mathcal{X} \left(1 - \frac{d\mu}{d\nu}\right)^2 d\mu \ \text{ if } \nu \ll \mu, \ \ \mathcal{D}(\mu,\nu) = +\infty \text{ otherwise.}$$

Moreover, for A a subset of $\mathcal{M}(\mathcal{X})$, we define $\mathcal{D}(\mu, A) = \inf_{\nu\in A} \mathcal{D}(\mu,\nu)$. A measure $\mu^* \in A$ for which the infimum is reached is call the Q-projection of μ on A.

In the next theorem we show that to each measure of a model but the true measure μ_0, can be associated a variance lower bound for an unbiased estimator of a parameter $\psi(\mu_0) \in \mathbb{R}^q$. We shall use the convention $1/\infty = 0$.

Theorem 3.6.2 *Let $X_1, ..., X_n$ be an i.i.d sample with distribution μ_0. Let $T = T(X_1, ..., X_n)$ be an unbiased estimate of $\psi(\mu_0) \in \mathbb{R}^q$ in the model $\mathcal{M}$. Then, $\forall \mu \in \mathcal{M} \setminus \{\mu_0\}$:*

$$\mathrm{var}(T) \geq \frac{(\psi(\mu_0) - \psi(\mu))(\psi(\mu_0) - \psi(\mu))^t}{(\mathcal{D}(\mu_0, \mu) + 1)^n - 1}.$$

Proof. First assume that $\psi(\mu_0) \in \mathbb{R}$. If $\mathcal{D}(\mu_0, \mu) = +\infty$, the inequality is trivially verified. If not, write

$$\psi(\mu_0) - \psi(\mu) = \mathbb{E}\left((T - \psi(\mu_0))\left(1 - \frac{d\mu^{\otimes n}}{d\mu_0^{\otimes n}}\right)\right)$$

where the expectation is meant under the distribution of the observations $\mu_0^{\otimes n}$. Applying Cauchy-Schwarz inequality, we get

$$\psi(\mu_0) - \psi(\mu) \leq \sqrt{\mathrm{var}(T)}\sqrt{\mathcal{D}(\mu_0^{\otimes n}, d\mu^{\otimes n})}.$$

Moreover, we see after calculation that $\mathcal{D}(\mu_0^{\otimes n}, \mu^{\otimes n}) = (\mathcal{D}(\mu_0, \mu) + 1)^n - 1$. Hence,

$$\forall \mu \in \mathcal{M}^*, \ \mathrm{var}(T) \geq \frac{(\psi(\mu_0) - \psi(\mu))^2}{(\mathcal{D}(\mu_0, \mu) + 1)^n - 1}.$$

If $\psi(\mu_0) \in \mathbb{R}^q$ with $q > 1$, we apply the previous result to the estimator $\alpha^t T \in \mathbb{R}$ for some $\alpha \in \mathbb{R}^q$. We get for all $\mu \neq \mu_0$:

$$\mathrm{var}(\alpha^t T) = \alpha^t \mathrm{var}(T)\alpha \geq \frac{(\alpha^t\psi(\mu_0) - \alpha^t\psi(\mu))^2}{(\mathcal{D}(\mu_0, \mu) + 1)^n - 1} = \alpha^t \frac{(\psi(\mu_0) - \psi(\mu))(\psi(\mu_0) - \psi(\mu))^t}{(\mathcal{D}(\mu_0, \mu) + 1)^n - 1}\alpha.$$

The inequality is true for all $\alpha \in \mathbb{R}^q$, which proves the result.

To each element of the model $\mathcal{M}$ corresponds a variance lower bound. Given a map $\psi : \mathcal{M} \to \mathbb{R}^q$, we shall denote by $H_\psi^n(\mu_0, .)$ the entropy function defined on $\mathcal{M}^* = \mathcal{M} \setminus \{\mu_0\}$ as

$$H_\psi^n(\mu_0, \mu) = n\frac{(\psi(\mu_0) - \psi(\mu))^t(\psi(\mu_0) - \psi(\mu))}{(\mathcal{D}(\mu_0, \mu) + 1)^n - 1}.$$

As the Cramer-Rao bound in the case of a parametric model, $H_\psi^n(\mu_0, \mu)$ provides a lower bound for n times the variance of an unbiased estimate. Since $H_\psi^n(\mu_0, \mu)$ is zero for all measure μ such that $d\mu/d\mu_0 \notin \mathcal{F} = \{f \in \mathbb{L}^2(\mu_0) : f\mu_0 \in \mathcal{M}\}$ (we admit for convenience that $d\mu/d\mu_0 = +\infty$ if μ is not absolutely continuous w.r.t. μ_0), we can settle for the set $\mathcal{F}$ of all densities w.r.t. μ_0 to determine the variance lower bound. The main advantage is that $\mathcal{F}$ being a subspace of $\mathbb{L}^2(\mu_0)$, it can be endowed with its natural Hilbert space topology.

In a statistical model $\mathcal{M}$, we denote by $B_\psi^n(\mathcal{M})$ the bound:

$$B_\psi^n(\mathcal{M}) := \sup_{\mu \in \mathcal{M}^*} H_\psi^n(\mu_0, \mu) = \sup_{f \in \mathcal{F}\setminus\{1\}} H_\psi^n(\mu_0, f\mu_0).$$

As for the Cramer-Rao bound, a maximum is not necessarily unique if the parameter to estimate is vectorial. For sake of simplicity, most of our results will be given assuming that the maximum B_ψ^n is real, and therefore, properly defined. Cases with several maxima are harder to handle but may be reduced to the unique maximum case by studying each maximum separately.

To fully understand the previous theorem, consider the binary model $\mathcal{M} = \{\mu_0, \mu\}$. Assume that we want to estimate the parameter $\psi(\mu_0)$ based on an i.i.d sample $X = (X_1, ..., X_n)$ and knowing that the true measure is either μ_0 or μ. In that case, the variance of an estimator $T = T(X)$ of $\psi(\mu_0)$ depends on two considerations. First, the information given by the observations: if μ and μ_0 are "far" from each other in some sense, it is likely that the observations will easily permit to decide which distribution they are drawn from. This is all the more true that the number of observations is large. To illustrate this, take the extreme case where the support of μ is not contained in the support of μ_0. Then, we could immediately tell that the observations are not drawn from μ as soon as some observation X_i does not lie in the support of μ. The distribution μ_0 of the observations and therefore the value of $\psi(\mu_0)$, would then be known. Second, if $\psi(\mu)$ is close to the true value $\psi(\mu_0)$, the error made by choosing μ instead of μ_0 will not have so much influence on the estimation. A trivial example is to consider the case where $\psi(\mu) = \psi(\mu_0)$. Here, the known value $\psi(\mu_0) = \psi(\mu)$ is actually an unbiased estimate of $\psi(\mu_0)$ with null variance.

Hence, $H_\psi^n(\mu_0, \mu)$ can be seen as a measure of entropy between μ_0 and μ for estimating $\psi(\mu_0)$ and for a number n of observations. It takes account of the two aspects mentioned above, namely the distance between $\psi(\mu_0)$ and $\psi(\mu)$ through the term $(\psi(\mu_0) - \psi(\mu))^2$ and the entropy of μ with respect to μ_0 for a number n of observations, through the value $(\mathcal{D}(\mu_0, \mu) + 1)^n$. The more the entropy $H_\psi^n(\mu_0, \mu)$ is large, the less information is given by μ for estimating $\psi(\mu_0)$. For instance we naturally observe a null entropy when μ is not absolutely continuous w.r.t. μ_0 or if $\psi(\mu) = \psi(\mu_0)$. An element μ for which the supremum of $H_\psi^n(\mu_0, .)$ is reached can be seen as the *least favorable measure*, that is, the element of the model that has the worst influence on the variance of the estimator T.

Proposition 3.6.3 *Let $\{\mu_t\}_{t\in\mathcal{T}}$ be a differentiable path. Let $\psi : \{\mu_t\}_t \to \mathbb{R}$ be a map such that $t \mapsto \psi(\mu_t)$ is differentiable at 0. Then, $B_\psi(\{\mu_t\}_t) = \lim_{t\to 0} H_\psi^n(\mu_0, \mu_t)$ for all $n \in \mathbb{N}$. It follows that $B_\psi^n(\{\mu_t\}_t) \geq B_\psi(\{\mu_t\}_t)$ for all $n \in \mathbb{N}$.*

Proof. First remark that if $\{\mu_t\}_{t\in\mathcal{T}}$ is differentiable in $\mathbb{L}^2(\mu_0)$ at 0 with corresponding score function g, the limit as $t \to 0$ of $\mathcal{D}(\mu_0, \mu_t)/t^2$ exists and is equal to the Fisher information $\int g^2 d\mu_0$. Furthermore, since $\lim_{t\to 0} \mathcal{D}(\mu_0, \mu_t) = 0$, it follows that $(\mathcal{D}(\mu_0, \mu_t) + 1)^n - 1 = n\mathcal{D}(\mu_0, \mu_t) + o(t)$. Hence,

$$\lim_{t\to 0} H_\psi^n(\mu_0, \mu_t) = \lim_{t\to 0} \frac{(\psi(\mu_0) - \psi(\mu_t))^2}{t^2} \frac{t^2}{\mathcal{D}(\mu_0, \mu_t)} = B_\psi(\{\mu_t\}_t).$$

The proposition establishes that in a one-dimensional smooth parametric model $\{\mu_t\}_{t\in\mathcal{T}}$, the efficiency bound is larger than the Cramer-Rao bound. Indeed, it is defined as the supremum of the entropy function over the whole model, while the CR bound is the limit as $\mu_t \to \mu_0$. Hence, the entropy function can be extended by continuity at μ_0 taking the value $H_\psi^n(\mu_0, \mu_0) = B_\psi$. It is not rare that the two bounds are equal (i.e. the maximum of $H_\psi^n(\mu_0, \mu)$ is reached for $\mu = \mu_0$). It occurs for example as soon as the Cramer-Rao bound can be reached.

Example 1 (Gaussian model). Consider the model $\{\mu_\theta\}_{\theta\in\mathbb{R}}$, where $\mu_\theta \sim \mathcal{N}(\theta, 1)$ and with the map $\psi : \mu_\theta \mapsto e^\theta$. The true measure is $\mu_0 \sim \mathcal{N}(\theta_0, 1)$, the Cramer-Rao bound $B_\psi(\{\mu_\theta\}_\theta) = e^{2\theta_0}$. On the other hand, we have $\mathcal{D}(\mu_0, \mu_\theta) = \exp(\theta_0 - \theta)^2 - 1$ yielding $H_\psi^n(\mu_0, \mu_\theta) = (e^{\theta_0} - e^\theta)^2/(\exp[n(\theta_0 - \theta)^2] - 1)$ for all $\theta \in (-1; +\infty)$. The supremum is taken for $\theta = \theta_0 + 1/n$, which gives $B_\psi^n = ne^{2\theta_0}(\exp(1/n) - 1)$. We then have a strict inequality $B_\psi^n > B_\psi$ for all $n \in \mathbb{N}$.

Example 2 (Exponential model). We consider the model $\{\mu_t\}_{t\in(-1;+\infty)}$, where μ_t is an exponential distribution with parameter $t + 1$, i.e. $d\mu_t(x) = (t + 1) \exp[-(t + 1)x]\mathbb{1}\{x \geq 0\}dx$. We want to estimate the parameter defined by the map $\psi : \mu_t \mapsto t$, the true measure being μ_0. Calculating the Cramer-Rao bound in this model, we get $B_\psi = 1$. On the other hand, the Q-divergence of μ_t w.r.t. μ_0 is $\mathcal{D}(\mu_0, \mu_t) = (t + 1)^2/(2t + 1) - 1$ for all $t > -1/2$ and $\mathcal{D}(\mu_0, \mu_t) = +\infty$ otherwise. It follows that

$$H_\psi^n(\mu_0, \mu_t) = \frac{t^2(2t + 1)^n}{(t + 1)^{2n} - (2t + 1)^n}\mathbb{1}\{t > -1/2\}.$$

A first (quite useless) remark is that we can not build an unbiased estimator of $\psi(\mu_0)$ with a finite variance, based on a single observation X_1 with distribution μ_0. Indeed, the entropy $H_\psi^n(\mu_0, .)$ is not bounded on the model if $n = 1$. We now see the plot of the entropy function for $n = 4$ to 15.

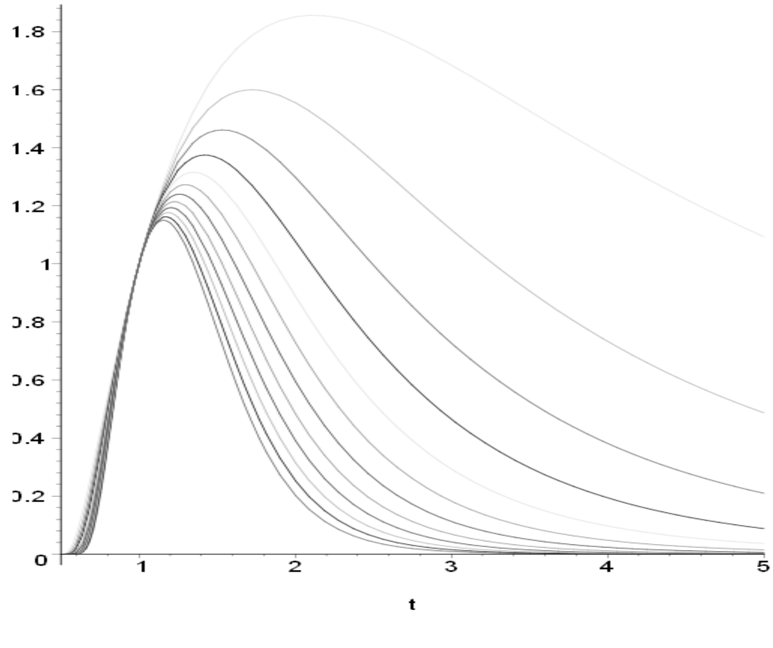

Figure 3.1: Plot of $t \mapsto H_\psi^n(\mu_0, \mu_t)$ for $n = 4$ to 15

The curves are decreasing as n grows (the curve on the top represents H_ψ^n for $n = 4$ while the lowest curve is for $n = 15$). This property is always true as shown in Section 3.6.3. The functions are not defined at $t = 0$ but they can be extended by continuity taking the value $H^n_\psi(\mu_0, \mu_0) = B_\psi = 1$ for all n. This corresponds on the graph to the intersection point of all the curves at $t = 0$. We observe that for all n, the supremum is larger than the Cramer-Rao bound $B_\psi = 1$.

In the two previous examples, we observe the strict inequality $B^n_\psi > B_\psi$, proving that the Cramer-Rao bound is not optimal in these parametric models. In larger models, the construction of the efficiency bound relies on the behavior of one-dimensional smooth submodels. Therefore, we easily generalize Proposition 3.6.3 to any model $\mathcal{M}$.

Corollary 3.6.4 *In a model $\mathcal{M}$ with $\psi : \mathcal{M} \to \mathcal{H} \subset \mathbb{R}$ a map differentiable at μ_0 in $\mathcal{M}$, $B^n_\psi(\mathcal{M}) \geq B_\psi(\mathcal{M})$ for all $n \in \mathbb{N}$.*

Proof. By Proposition 3.6.3, we know that $B^n_\psi(\{\mu_t\}_t) \geq B_\psi(\{\mu_t\}_t)$ for all submodel $\{\mu_t\}_{t\in\mathcal{T}} \in \mathcal{M}_{\psi,\mathcal{T}}$. Furthermore,

$$B^n_\psi(\mathcal{M}) \geq \sup_{\{\mu_t\}_t \in \mathcal{M}_{\psi,\mathcal{T}}} B^n_\psi(\{\mu_t\}_t) \geq \sup_{\{\mu_t\}_t \in \mathcal{M}_{\psi,\mathcal{T}}} B_\psi(\{\mu_t\}_t) = B_\psi(\mathcal{M}).$$

Remark Proposition 3.6.3 and Corollary 3.6.4 are proved for a real parameter. In the vectorial case, similar results are obtained for any element B^n_ψ of the set of suprema $\mathcal{B}^n_\psi$ following the same pattern of proof. If $\mathcal{B}^n_\psi$ contains more than one element, we show that any maximum $B^n_\psi \in \mathcal{B}^n_\psi$ is not smaller than B_ψ in the sense that $B_\psi - B^n_\psi$ is positive semidefinite. Both statements are obviously equivalent in the one-dimensional case.

In semiparametric and non-parametric models, Fisher information is calculated by studying the behavior of least favorable paths. We recall that a least favorable path in a model $\mathcal{M}$ is a differentiable submodel $\{\mu_\theta\}_{\theta\in\Theta}$ such that $B_\psi(\mathcal{M}) = B_\psi(\{\mu_\theta\}_\theta)$. Such a submodel may not exist, but we can always find a collection of submodels $\{\mu_\theta\}_{\theta,m}, m \in \mathbb{N}$ for which $B_\psi(\{\mu_\theta\}_{\theta,m})$ can get as close as possible of $B_\psi(\mathcal{M})$ as m range over $\mathbb{N}$. The entropy function $H^n_\psi(\mu_0, .)$ turns out to be an efficient tool to construct a least favorable path. To see it, consider the sets $\mathcal{F}_\eta = \left\{f \in \mathbb{L}^2(\mu_0) : \ \psi(f\mu_0) = \eta\right\}, \eta \in \mathcal{H}$ which we assume to be non-empty for simplicity, and rewrite the expression of the efficiency bound as follows

$$B^n_\psi = \sup_{\eta\neq\eta_0} \sup_{f\in\mathcal{F}_\eta} H^n_\psi(\mu_0, f\mu_0) = \sup_{\eta\neq\eta_0} \frac{n(\eta-\eta_0)(\eta-\eta_0)^t}{(\mathcal{D}(\mu_0,\mathcal{F}_\eta)+1)^n - 1} \tag{3.6}$$

where $\eta_0 = \psi(\mu_0)$. In these settings, we see that calculating the efficiency bound reduces to minimizing a function with respect to η. Our insight is that if we choose in each set $\mathcal{F}_\eta$ the least favorable density, that is, the function f maximizing the entropy $H^n_\psi(\mu_0, f\mu_0)$, the resulting submodel would have to be a least favorable path (if a least favorable measure can not be reached, we consider a proper collection of densities arbitrarily close to the least favorable measure in each set $\mathcal{F}_\eta$, leading to a collection of submodels). Since the term $\psi(f\mu_0) - \psi(\mu_0)$ is constant when f ranges over $\mathcal{F}_\eta$, a density maximizing the entropy on $\mathcal{F}_\eta$ is in fact a minimizer of $f \mapsto \mathcal{D}(\mu_0, f\mu_0)$, which explains the term $\mathcal{D}(\mu_0, \mathcal{F}_\eta)$ in (3.6).

Definition We call *quadratic projection path* (or *Q-projection path*) a submodel $(\mu_\eta)_{\eta\in\mathcal{H}}$ such that $\mathcal{D}(\mu_0,\mu_\eta)=\mathcal{D}(\mu_0,\mathcal{F}_\eta)$ and $\psi(\mu_\eta)=\eta$ for all $\eta\in\mathcal{H}$.

A Q-projection path does not necessarily exist, for instance if the infimum of $\mathcal{D}(\mu_0,.)$ on $\mathcal{F}_\eta$ is not reached for some values of η. However, a Q-projection path does exist as soon as the map $f\mapsto\psi(f\mu_0)$ is continuous on $\mathcal{F}$. By making this assumption, we avoid considering trivial cases, the efficiency bound being infinite if $f\mapsto\psi(f\mu_0)$ is not continuous as f tends to 1.

If the sets $\mathcal{F}_\eta$ are convex, $\{\mu_\eta\}_{\eta\in\mathcal{H}}$ is unique, μ_η being defined as the quadratic projection of μ_0 on $\mathcal{M}_\eta=\{\mu\in\mathcal{M}:\psi(\mu)=\eta\}$. This is illustrated in Examples 3 and 4. A Q-projection path does not depend on the number of observations, although it contains a maximizer of $H_\psi^n(\mu_0,.)$ for all $n\in\mathbb{N}$. In a certain way, it carries the whole information of the model, as we see in the next proposition.

Proposition 3.6.5 *Let $\eta_0=\psi(\mu_0)\in\mathcal{H}\subset\mathbb{R}^q$ be the parameter to estimate. A Q-projection path $\{\mu_\eta\}_{\eta\in\mathcal{H}}$ satisfies $B_\psi^n(\mathcal{M})=B_\psi^n(\{\mu_\eta\}_\eta)$ for all $n\in\mathbb{N}$. Moreover, $\{\mu_\eta\}_{\eta\in\mathcal{H}}$ is a least favorable path if and only if it is differentiable at η_0.*

Example 3 (Linear parameters). Consider the model $\mathcal{M}=\left\{\mu\in\mathcal{M}(\mathcal{X}):\ \int_{\mathcal{X}}\Phi d\mu=0\right\}$ for $\Phi:\mathcal{X}\to\mathbb{R}^k$ a known map. We aim to estimate $\eta_0=\int_{\mathcal{X}}hd\mu_0\in\mathbb{R}$ for $h\in\mathbb{L}^2(\mu_0)$ a given function. For all $\eta\in\mathbb{R}$, $\mathcal{F}_\eta$ is an affine subspace of $\mathbb{L}^2(\mu_0)$ of finite dimension, it is therefore closed and convex. Hence, there exists a unique path $\{\mu_\eta\}_{\eta\in\mathbb{R}}$ satisfying the conditions of Proposition 3.6.5, we denote by f_η its density w.r.t. μ_0. Finally, we note $h^\perp$ the part of h that is orthogonal with Φ in $\mathbb{L}^2(\mu_0)$: $h^\perp=h-(\int_{\mathcal{X}}h\Phi d\mu_0)^t[\int_{\mathcal{X}}\Phi\Phi^t d\mu_0]^{-1}\Phi$. We have:

$$f_\eta=\arg\min_{f\in\mathcal{F}_\eta}\ \mathbb{E}(1-f(X))^2=1-(\eta_0-\eta)V^{-1}(h^\perp-\eta),$$

with $V=\mathrm{var}(h^\perp(X))$. Moreover, $\mathcal{D}(\mu_0,\mu_\eta)=\mathbb{E}(1-f_\eta(X))^2=(\eta_0-\eta)^2V^{-1}$, yielding

$$B_\psi^n=\sup_{\eta\neq\eta_0}\frac{n(\eta_0-\eta)^2}{((\eta_0-\eta)^2V^{-1}+1)^n-1}=V.$$

Note that the model $\{\mu_\eta\}_{\eta\in\mathbb{R}}$ is smooth, with Cramer-Rao bound $B_\psi=B_\psi^n=V$ for all integer n.

Example 4 (Moment condition model). Assume that the true measure μ_0 satisfies the constraint $\int_{\mathcal{X}}\Phi_{\eta_0}d\mu_0=0$ for some known collection of maps $\{\Phi_\eta:\eta\in\mathcal{H}\}$ and where $\eta_0\in\mathcal{H}$ is the parameter we want to estimate. The sets $\mathcal{F}_\eta=\left\{f\in\mathbb{L}^2(\mu_0):\ \int\Phi_\eta fd\mu_0=0\right\}$ are closed and convex for all $\eta\in\mathcal{H}$. We note $\{\mu_\eta\}_\eta$ the Q-projection path satisfying

$$\frac{d\mu_\eta}{d\mu_0}=\arg\min\nolimits_{f\in\mathcal{F}_\eta}\ \mathbb{E}\left(1-f(X)\right)^2=1-\left(\int\Phi_\eta d\mu_0\right)^t[\mathrm{var}(\Phi_\eta(X))]^{-1}\left(\Phi_\eta-\int\Phi_\eta d\mu_0\right)$$

for all $\eta\in\mathcal{H}$. If we assume that $\eta\mapsto\Phi_\eta$ is differentiable in a neighborhood of η_0, with derivative $\nabla\Phi(.)$, it follows that the path $\{\mu_\eta\}_\eta$ is also differentiable. We calculate $\mathcal{D}(\mu_0,\mu_\eta)=\left(\int\Phi_\eta d\mu_0\right)^t[\mathrm{var}(\Phi_\eta(X))]^{-1}\left(\int\Phi_\eta d\mu_0\right)$. We get

$$B_\psi^n=\sup_{\eta\neq\eta_0}\frac{n(\eta_0-\eta)^2}{(\alpha_\eta+1)^n-1}\overset{n\to\infty}{\longrightarrow}\left[\left(\int\nabla\Phi(\eta_0)d\mu_0\right)^t\left[\int\Phi_{\eta_0}\Phi_{\eta_0}^t d\mu_0\right]^{-1}\left(\int\nabla\Phi(\eta_0)d\mu_0\right)\right]^{-1}.$$

3.6.3 Asymptotic properties

We are now interested in the asymptotic behavior of the entropy function. Expanding the term $(\mathcal{D}(\mu_0,\mu)+1)^n = 1+n\mathcal{D}(\mu_0,\mu)+\frac{n(n-1)}{2}\mathcal{D}(\mu_0,\mu)^2...$ in the denominator of $H^n_\psi(\mu_0,\mu)$, it appears that the sequence $\{H^n_\psi(\mu_0,.)\}_{n\in\mathbb{N}}$ is decreasing and converges pointwise toward 0 as $n\to\infty$. So, the sequence $\{B^n_\psi\}_{n\in\mathbb{N}}$ is also decreasing and since it is positive for all $n\in\mathbb{N}$, it converges (or is equal to $+\infty$). We now aim to determine its limit, noted B^∞_ψ. In this section, our results are given assuming for writing convenience that $\psi(\mu_0)$ is real. They can be generalized to vectorial parameters by studying each maximum at a time if there are several.

Lemma 3.6.6 *Assume that $B^{n_0}_\psi<\infty$ for some $n_0\in\mathbb{N}$. Then, for all $\varepsilon>0$, $H^n_\psi(\mu_0,.)$ converges uniformly toward 0 on the set $\{\mu\in\mathcal{M}:\mathcal{D}(\mu_0,\mu)>\varepsilon\}$ as $n\to\infty$.*

Proof. For all $\mu\neq\mu_0$, we know that $H^{n_0}_\psi(\mu_0,\mu)\leq B^{n_0}_\psi$. Since the bound is decreasing as $n\to\infty$, we have

$$\forall\mu\neq\mu_0,\ \forall n>n_0,\ H^n_\psi(\mu_0,\mu)\leq\frac{nB^{n_0}_\psi}{n_0}\frac{(\mathcal{D}(\mu_0,\mu)+1)^{n_0}-1}{(\mathcal{D}(\mu_0,\mu)+1)^n-1}.$$

Since the function $x\mapsto((x+1)^{n_0}-1)/((x+1)^n-1)$ is decreasing on the interval $(\varepsilon;+\infty)$ as soon as $n\geq n_o(\varepsilon+1)^{n_0}/((\varepsilon+1)^{n_0}-1)$, we conclude that for large enough values of n

$$\forall\varepsilon>0,\sup_{\mathcal{D}(\mu_0,\mu)>\varepsilon}H^n_\psi(\mu_0,\mu)\leq\frac{nB^{n_0}_\psi}{n_0}\frac{(\varepsilon+1)^{n_0}-1}{(\varepsilon+1)^n-1}.$$

The right-hand term tends to 0 as $n\to\infty$ for all $\varepsilon>0$, which proves the lemma.

Remark The condition that $B^{n_0}_\psi$ is finite for some integer n_0 is necessary to ensure the existence of an unbiased estimator with finite variance, even asymptotically. It may occur that this condition is not fulfilled while the Cramer-Rao bound exists and is finite.

An interpretation of Lemma 3.6.6 is the following. A measure $\mu\in\mathcal{M}$ far from μ_0 will no longer have any influence on the variance of an estimator as soon as the number of observations is large enough. A set of measures μ such that the entropy $H^n_\psi(\mu_0,\mu)$ is arbitrarily large will be contained in a ball of radius $\varepsilon_n\in\overline{\mathbb{R}}_+$ centered on μ_0 with ε_n tending to zero as n tends to infinity. Roughly speaking, for any element μ of the model with a non zero distance with μ_0 (basically, any $\mu\neq\mu_0$), it ends up as $n\to\infty$ that the observations give too much information so that its distribution can not be mistaken with μ. Thus, only the behavior of the measures of the model in the neighborhood of μ_0 matters asymptotically.

Theorem 3.6.7 *Assume that $B^{n_0}_\psi(\mathcal{M})<\infty$ for some $n_0\in\mathbb{N}$. If there exists a Q-projection path $\{\mu_\eta\}_{\eta\in\mathcal{H}}$ differentiable at μ_0, then $\lim_{n\to\infty}B^n_\psi(\mathcal{M})=B_\psi(\mathcal{M})$.*

Proof. The theorem is true if $B^\infty_\psi=0$. Now, assume that $B^\infty_\psi>0$. By Proposition 3.6.5, we know that $B^n_\psi(\mathcal{M})=B^n_\psi(\{\mu_\eta\}_\eta)$ for all $n\in\mathbb{N}$. Let $\{\mu_n\}_{n\in\mathbb{N}}$ be a sequence of measures in $\{\mu_\eta\}_\eta$, suitably chosen so that $\lim_{n\to\infty}H^n_\psi(\mu_0,\mu_n)=B^\infty_\psi$. We want to prove that $\lim_{n\to\infty}\mathcal{D}(\mu_0,\mu_n)=$

0. By contradiction, if there exists $\varepsilon > 0$ and an increasing sequence of integers $\{n_k\}_{k\in\mathbb{N}}$ such that $\forall k \in \mathbb{N}, \mathcal{D}(\mu_0, \mu_{n_k}) > \varepsilon$, then:

$$H_\psi^{n_k}(\mu_0, \mu_{n_k}) \leq \sup_{\mathcal{D}(\mu_0,\mu)>\varepsilon} H_\psi^{n_k}(\mu_0, \mu) \overset{k\to\infty}{\longrightarrow} 0$$

by Lemma 3.6.6. Since, $\lim_{k\to\infty} H_\psi^{n_k}(\mu_0, \mu_{n_k}) = B_\psi^\infty > 0$ there is a contradiction. So, we conclude that $\lim_{n\to\infty} \mathcal{D}(\mu_0, \mu_n) = 0$. Since $H_\psi^n(\mu_0, .)$ is decreasing as $n \to \infty$, we get that for all $n \in \mathbb{N}$,

$$B_\psi^\infty = \lim_{n\to\infty} H_\psi^n(\mu_0, \mu_n) \leq \lim_{\eta\to\eta_0} H_\psi^n(\mu_0, \mu_\eta) = B_\psi(\{\mu_\eta\}_\eta).$$

By Proposition 3.6.5, $\{\mu_\eta\}_{\eta\in\mathcal{H}}$ is a least favorable path of the model and therefore satisfies $B_\psi(\mathcal{M}) = B_\psi(\{\mu_\eta\}_\eta)$. So, $B_\psi^\infty(\mathcal{M}) \leq B_\psi(\mathcal{M})$. The inverse inequality being an obvious consequence of Corollary 3.6.3, we conclude that $B_\psi^\infty(\mathcal{M}) = B_\psi(\mathcal{M})$.

This result is not surprising when we know that the efficiency bound only depends asymptotically on the behavior of the model in the neighborhood of μ_0. See for instance Examples 1 and 2 where the convergence of the efficiency bound toward the CR bound is shown in the first example and is observed graphically in the second. However, the theorem is not always true if the model does not contain a differentiable Q-projection path. This is mainly due to the strong assumptions needed to build the CR bound. Basically, in some parametric model $\{\mu_t\}_t$, the efficiency bound has positive limit $B_\psi^\infty(\{\mu_t\}_t)$ in non-trivial cases as soon as the map $t \mapsto \sqrt{\mathcal{D}(\mu_0, \mu_t)}$ is differentiable at 0. This is true in particular if $\{\mu_t\}_t$ is differentiable in $\mathbb{L}^2(\mu_0)$. The differentiability in $\mathbb{L}^2(\mu_0)$ which is required to construct the CR bound is therefore a stronger condition. So, if $t \mapsto \sqrt{\mathcal{D}(\mu_0, \mu_t)}$ is differentiable at 0, but the model is not differentiable in $\mathbb{L}^2(\mu_0)$, $B_\psi^\infty(\{\mu_t\}_t)$ might be positive while the CR bound is zero. More generally in larger models, B_ψ^∞ may be larger than B_ψ since B_ψ^∞ is calculated from a larger set of submodels. We here see one example of model where we observe the strict inequality $B_\psi^\infty > B_\psi$.

Example 5. Let μ_0 be the uniform distribution on $(0;1)$. Define $f_t(.) = \mathbb{1}(0;\ t^2/2)(.) - \mathbb{1}(1 - t^2/2; 1)(.)$ and $\mu_t = (1 + f_t)\mu_0$ for all $t \in (-1; 1)$. Now assume that we want to estimate the parameter defined by the map $\psi : \mu_t \mapsto t$ in the model $(\mu_t)_{t\in(-1;1)}$ with true measure for $t = 0$. The model is clearly not differentiable in $\mathbb{L}^2(\mu_0)$, we then have $B_\psi = 0$. Furthermore, for all $t \in (-1; 1)$, $\mathcal{D}(\mu_0, \mu_t) = t^2$, yielding $H_\psi^n(\mu_0, \mu_t) = nt^2/[(t^2 + 1)^n - 1]$. So, in particular, $B_\psi^n \geq \lim_{t\to 0} H_\psi^n(\mu_0, \mu_t) = 1$ for all $n \in \mathbb{N}$, which implies that $B_\psi^\infty \geq 1$ (here, we actually have $B_\psi^n = B_\psi^\infty = 1$). Hence $B_\psi^\infty > B_\psi$.

Example 6. Consider the Gaussian model $\{\mu_t\}_{t\in\mathbb{R}}$ where $\mu_t \sim \mathcal{N}(t, 1)$. We aim to estimate a differentiable function $f : \mathbb{R} \to \mathbb{R}$ of the mean, $\psi : \mu_t \mapsto f(t)$. The true value of the parameter is $\psi(\mu_0) = f(0)$. The Cramer-Rao bound is $B_\psi(\{\mu_t\}_t) = f'(0)$ which we assume finite. On the other hand, the entropy function $H_\psi^n(\mu_0, .)$ may take larger values for some elements in the model. It may even occur that for all $n \in \mathbb{N}$, the entropy $H_\psi^n(\mu_0, .)$ is not bounded in the model. If so, an unbiased estimator with finite variance of $\psi(\mu_0)$ is impossible to compute, even asymptotically. In our example, the entropy is given by

$$\forall t \in \mathbb{R},\ H_\psi^n(\mu_0, \mu_t) = \frac{n\,[f(t) - f(0)]^2}{\exp nt^2 - 1}.$$

Unbiased estimators of $\psi(\mu_0)$ with finite variance do not exist as soon as $f(t) \geq \exp(\alpha(t)t^2)$ for some function α unbounded on $\mathbb{R}$, i.e. if $t \mapsto \log f(t)/t^2$ is not bounded on $\mathbb{R}$. In these cases, we have $B_\psi^\infty = \infty$ while the Cramer-Rao bound is finite.

3.6.4 Efficiency bounds for centered moments of order q

In this section, we provide efficiency bounds for centered moments of order $q > 1$ of unbiased estimators. Following the proof of Theorem 3.6.2 which gives a variance efficiency bound, we can easily generalize this result to moments of order $q \neq 2$.

Definition The p-divergence between of ν with respect to μ is given by:

$$\mathcal{D}_p(\mu,\nu) = \int_{\mathcal{X}} \left(1 - \frac{d\nu}{d\mu}\right)^p d\mu \text{ if } \nu \ll \mu, \quad \mathcal{D}_p(\mu,\nu) = +\infty \text{ otherwise.}$$

We define for a subset A of $\mathcal{M}(\mathcal{X})$, $\mathcal{D}_p(\mu, A) = \inf_{\nu \in A} \mathcal{D}_p(\mu,\nu)$.

Theorem 3.6.8 *Let $X_1, ..., X_n$ be an i.i.d sample with distribution μ_0. Let T be an unbiased estimate of $\psi(\mu_0) \in \mathbb{R}$ in the model $\mathcal{M}$. Then, $\forall \mu \in \mathcal{M}^*$:*

$$\mathbb{E}(T - \psi(\mu_0))^q \geq \frac{(\psi(\mu_0) - \psi(\mu))^q}{\mathcal{D}_p(\mu_0^{\otimes n}, \mu^{\otimes n})^{q-1}},$$

with $p + q = pq$.

The proof of this theorem follows from Hölder's inequality while Cauchy-Scharwz's was used in the proof of Theorem 3.6.2.

Remark One may reasonably think that a similar result could be extended to Bregman informations of an estimator T, defined for $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ a convex function with $\phi(0) = \phi'(0) = 0$, by

$$\mathcal{I}_\phi(T) = \mathbb{E}(\phi(T - \psi(\mu_0))).$$

Indeed, for any $p > 1$ (note $q = p/(p-1)$), Hölder's inequality can be written for $\phi : x \mapsto x^p$ as

$$\int fg \; d\mu \leq \phi^{-1}\left(\int \phi(f) d\mu\right) \phi^{*-1}\left(\int \phi^*(g) d\mu\right), \tag{3.7}$$

where ϕ^* denotes the convex conjugate of ϕ. If such an inequality holds for a larger family of convex functions, it would be possible to provide by this technique, efficiency bounds for a larger class of Bregman Informations. Unfortunately, it can be proved that only power functions satisfy Equation 3.7 for all functions f, g and all measure μ.

Chapter 4

Maximum entropy estimation for survey sampling

Calibration methods have been widely studied in survey sampling over the last decades. Viewing calibration as an inverse problem, we extend the calibration technique by using a maximum entropy method. Finding the optimal weights is achieved by considering random weights and looking for a discrete distribution which maximizes an entropy under the calibration constraint. This method points a new frame for the computation of such estimates and the investigation of its statistical properties.

4.1 Introduction

Calibration is a well spread method to improve estimation in survey sampling, using extra information from auxiliary variables. This method provides approximately unbiased estimators with variance smaller than that of the usual Horvitz-Thompson estimator. Calibration has been introduced by Deville and Särndal in [DS92], extending an idea of [Dev88]. For general references, we refer to [S0́7] and for an extension to variance estimation to [Sin01].
Finding the solution to a calibration equation involves minimizing a distance under some constraint. More precisely, let s be a random sample of size n drawn from a population U of size N, y be the variable of interest and $\mathbf{x}$ be a given auxiliary vector variable, for which the total $t_{\mathbf{x}}$ over the population is known. Further, let $d \in \mathbb{R}^n$ be the standard sampling weights (that is the Horvitz-Thompson ones). Calibration derives an estimator $\hat{t}_y = \sum_{i \in s} w_i y_i$ of the population total t_y of y. The weights w_i are chosen to minimize a dissimilarity (or distance) $\mathcal{D}(., d)$ on $\mathbb{R}^n$ with respect to the Horvitz-Thompson weights d_i and under the constraint

$$\sum_{i \in s} w_i \mathbf{x}_i = t_{\mathbf{x}}. \tag{4.1}$$

Following [Thé99], we will view here calibration as a linear inverse problem. In this paper, we use Maximum Entropy Method on the Mean (MEM) to build the calibration weights. Indeed, MEM is a strong machinery for solving linear inverse problems. It tackles a linear inverse problem by finding a measure maximizing an entropy under some suitable constraint. It has been extensively studied and used in many applications, see for example [BLN96], [Gzy95], [GG97], [KS02], [Gam99], [FLLn06] or [KT04].

Let us roughly explain how MEM works in our context. First we fix a *prior* probability measure ν on $\mathbb{R}^n$ with mean value equal to d. Then, the idea is to modify the standard weights d in order to get a representative sample for the auxiliary variable $\mathbf{x}$, but still being as close as possible to d, which have the desirable property of yielding an unbiased estimate for the population total. So, we will look for a *posterior* probability measure minimizing the entropy (or Kullback information) with respect to ν and satisfying a constraint related to (4.1). It appears that the MEM estimator is in fact a specific calibration estimator for which the corresponding dissimilarity $\mathcal{D}(.,d)$ is determined by the choice of the prior distribution ν. Hence, the MEM methodology provides a general Bayesian frame to fully understand calibration procedures in survey sampling where the different choices of dissimilarities appear as different choices of prior distributions.

An important problem when studying calibration methods is to understand the amount of information contained in the auxiliary variable. Indeed, the relationships between the variable to be estimated and the auxiliary variable are crucial to improve estimation (see for example [MR05], [WS01] or [WZ06]). When complete auxiliary information is available, *model calibration* introduced by Wu and Sitter [WS01] aims to increase the correlation between the variables by replacing the auxiliary variable $\mathbf{x}$ by some function of it, say $u(\mathbf{x})$. We consider efficiency issues for a collection of calibration estimators, depending on both the choice of the auxiliary variable and the dissimilarity. Finally, we provide an optimal way of building an efficient estimator using the MEM methodology.

The chapter falls into the following parts. The first section recalls the calibration method in survey sampling, while the second exposes the MEM methodology in a general framework, and its application to calibration and instrument estimation. Section 4.4 is devoted to the choice of a data driven calibration constraint in order to build an efficient calibration estimator. It is shown to be optimal under strong asymptotic assumptions on the sampling design. Proofs are postponed to Section 4.5.

4.2 Calibration Estimation of a linear parameter

Consider a large population $U = \{1, ..., N\}$ and an unknown characteristic $y = (y_1, ..., y_N) \in \mathbb{R}^N$. Our aim is to estimate its total $t_y := \sum_{i \in U} y_i$ when only a random subsample s of the whole population is available. So the observed data are $\{y_i\}_{i \in s}$. Each sample s has a probability $p(s)$ of being observed. The distribution $p(.)$ is called sampling design. We define the inclusion probabilities $\pi_i := p\ (i \in s) = \sum_{s,\ i \in s} p(s)$ which we assume to be strictly positive for all $i \in U$ so that $d_i = 1/\pi_i$ is well defined. A standard estimator of t_y is given by the Horvitz-Thompson estimator:

$$\hat{t}_y^{HT} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} d_i y_i.$$

This estimator is unbiased and is widely used for practical cases, see for instance [GFC$^+$04].

Suppose that it exists an auxiliary vector variable $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_N)$ entirely observed and set $t_{\mathbf{x}} = \sum_{i \in U} \mathbf{x}_i \in \mathbb{R}^k$. If the Horvitz-Thompson estimator of $t_{\mathbf{x}}$, $\hat{t}_{\mathbf{x}}^{HT} = \sum_{i \in s} d_i \mathbf{x}_i$ is far from the true value $t_{\mathbf{x}}$, we may reasonably assume that the sample will not adequately reflect the behavior of the variable of interest in the whole population. So, to prevent inefficient estimation due to bad sample selection, inference on the sample can be achieved by considering a modification of the weights of the individuals chosen in the sample.

One of the main methodology used to correct this effect is *calibration* (see [DS92]). The possible bad sample effect is corrected by replacing the Horvitz-Thompson weights d_i by new weights w_i close to d_i. Let $w \mapsto \mathcal{D}(w, d)$ be a dissimilarity between w and the Horvitz-Thompson weights that is minimal for $w_i = d_i$. The method consists in choosing weights $\hat{w}_i$ minimizing $\mathcal{D}(., d)$ under the constraint

$$\sum_{i \in s} \hat{w}_i \mathbf{x}_i = t_{\mathbf{x}}.$$

Then, consider the new weighted estimators $\hat{t}_y = \sum_{i \in s} \hat{w}_i y_i$.

A typical dissimilarity is the χ^2 distance $w \mapsto \sum_{i \in s} (\pi_i w_i - 1)^2 / (q_i \pi_i)$ for $\{q_i\}_{i \in s}$ some known positive sequence. In most applications, the q_i's are taken equal to 1 which generally warrants a consistent estimator. Nevertheless unequal weights can be used as treated in Example 1 in [DS92], in order to lay more or less stress on the distance between some of the weights and the original Horvitz-Thompson ones. The new estimator is defined as $\hat{t}_y = \sum_{i \in s} \hat{w}_i y_i$, where the weights $\hat{w}_i$ minimizes $\mathcal{D}(w, d) = \sum_{i \in s} (\pi_i w_i - 1)^2 / q_i \pi_i$ under the constraint $\sum_{i \in s} \hat{w}_i \mathbf{x}_i = t_{\mathbf{x}}$. The solution of this minimization problem is given by

$$\hat{t}_y = \hat{t}_y^{HT} + (t_{\mathbf{x}} - \hat{t}_{\mathbf{x}}^{HT})^t \hat{B},$$

where $\hat{B} = [\sum_{i \in s} q_i d_i \mathbf{x}_i \mathbf{x}_i^t]^{-1} \sum_{i \in s} q_i d_i y_i \mathbf{x}_i$. Note that this is a generalized regression estimator. It is natural to consider alternative dissimilarities, see for instance [DS92]. We first point out that the existence of a solution to the constrained minimization problem depends on the choice of the dissimilarities. Then, different choices can lead to weights with different behaviors, different ranges of values for the weights that may be found unacceptable by the users. We propose an approach where dissimilarities are given a probabilistic interpretation.

4.3 Maximum Entropy for Survey Sampling

4.3.1 MEM methodology

Consider the problem of recovering an unknown measure μ on a measurable space $\mathcal{X}$ under moment conditions. This issue belongs to the class of generalized moment problems with convex constraints (we refer to [EHN96] for general references). This inverse problem has been widely studied and in particular it can be solved using the maximum entropy on the mean (MEM).

Here, we aim at estimating μ from random observations $T_1, ..., T_n \sim \mu$ and knowing there exists a given function $\tilde{\mathbf{x}} : \mathcal{X} \to \mathbb{R}^k$ and a known quantity $t_{\mathbf{x}} \in \mathbb{R}^k$ such that

$$\int_{\mathcal{X}} \tilde{\mathbf{x}} d\mu = t_{\mathbf{x}}. \tag{4.2}$$

Solving this problem using the MEM framework amounts to approximate the inverse problem (4.2) by a sequence of finite dimensional problems which are obtained by a discretization of the space $\mathcal{X}$ using the random sample $T_1, \dots, T_n$. For this, first consider the empirical distribution $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{T_i}$, δ standing for the Dirac mass. The general idea is to modify μ_n in order to take into account the additional information on μ given by the moment equation (4.2). For this, we associate to each observation T_i a random weight P_i and consider the corresponding random weighted version of the empirical measure

$$\tilde{\mu}_n = \frac{1}{n}\sum_{i=1}^n P_i \delta_{T_i}.$$

Choosing properly the weights is the second step of the MEM procedure. The underlying idea is to incorporate some prior information by choosing $P = (P_1, ..., P_n)$, drawn from a finite measure ν^* close to a *prior* ν, and looking at the weighted measures satisfying the constraint in mean. This prior distribution conveys the information that $\hat{\mu}_n$ must be close, in a given sense, to the empirical distribution μ_n. Given our prior ν, we now define ν^* as the probability measure minimizing the relative entropy $\mathcal{K}(.|\nu)$ under the constraint that the linear constraint holds in mean:

$$\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{x}}_i \mathbb{E}_{\nu^*}(P_i) = t_{\mathbf{x}},$$

where we set $\tilde{\mathbf{x}}_i = \tilde{\mathbf{x}}(T_i)$. Note that, among the literature in optimization, the relative entropy is often defined as the opposite of the entropy defined above, which explains the name of maximum entropy method, while with our notations, we consider the minimum of the entropy. We then build the MEM estimator as

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \hat{p}_i \delta_{T_i} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\nu^*}(P_i)\delta_{T_i}.$$

For a fixed n, $\hat{\mu}_n$ is the maximum entropy reconstruction of μ with reference ν^*. This method provides an efficient way to estimate linear parameters of the form $t_y = \int_{\mathcal{X}} \tilde{y} d\mu$ for $\tilde{y} : \mathcal{X} \to \mathbb{R}$ a given map. The empirical mean $\int_{\mathcal{X}} \tilde{y} d\mu_n$ is an unbiased and consistent estimator of t_y but may not have the smallest variance in this model. Integrating $\tilde{y}$ against $\hat{\mu}_n$ provides an asymptotically unbiased estimate of t_y with a lower variance than the empirical mean (see [GG97]).

In many actual situations, the function $\tilde{\mathbf{x}}$ is unknown and only an approximation to it, say $\tilde{\mathbf{x}}_m$, is available. Under regularity conditions, the efficiency properties of the MEM estimator built with the approximate constraint have been studied in [LP08] and [LR09b], introducing the approximate maximum entropy on the mean method (AMEM). More precisely, the AMEM estimate of the weights is defined as the expectation of the variable P under the distribution ν_m^* minimizing $\mathcal{K}(.|\nu)$ under the approximate constraint

$$\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{x}}_m(T_i)\mathbb{E}_{\nu^*}(P_i) = t_{\mathbf{x}}. \tag{4.3}$$

It is shown that, under assumptions on $\tilde{\mathbf{x}}_m$, the AMEM estimator of t_y obtained in this way is consistent as n and m tend to infinity. This procedure enables to increase the efficiency of a calibration estimator while remaining in a Bayesian framework, as shown in Section 4.4.2. This situation occurs for instance, when dealing with inverse problem with unknown operators which still can be approximated either using another sample or directly from the data. For instance, in Econometry, when dealing with instrumental variables the operator which corresponds here to the function $\tilde{x}$ is unknown but can be estimated, see [CFR06]. The practical case of aerosol remote sensing is tackled in [LP08].

4.3.2 Maximum entropy method for calibration

Recall that our original problem is to estimate the population total $t_y = \sum_{i \in U} y_i$ based on the observations $\{y_i, i \in s\}$ and auxiliary information $\{\mathbf{x}_i, i \in U\}$. We introduce the following notations:

$$\tilde{y}_i = n d_i y_i, \ \tilde{\mathbf{x}}_i = n d_i \mathbf{x}_i, \ p_i = \pi_i w_i.$$

The variables of interest are rescaled to match the MEM framework. The collection of weights $\{p_i\}_{i \in s}$ is now identified with a discrete measure on the sample s. The Horvitz-Thompson estimator $\hat{t}_y^{HT} = \sum_{i \in s} d_i y_i = \frac{1}{n} \sum_{i \in s} \tilde{y}_i$ is the preliminary estimator we aim at improving. The calibration constraint $\frac{1}{n} \sum_{i \in s} p_i \tilde{\mathbf{x}}_i = t_{\mathbf{x}}$ stands for the linear condition satisfied by the discrete measure $\{p_i\}_{i \in s}$. In these settings, it appears that the calibration problem follows the pattern of maximum entropy on the mean. Let ν be a prior distribution on the vector of the weights $\{p_i\}_{i \in s}$. The solution $\hat{p} = \{\hat{p}_i\}_{i \in s}$ is the expectation of the random vector $P = \{\pi_i W_i\}_{i \in s}$ drawn from a *posterior* distribution ν^*, defined as the minimizer of the Kullback information $\mathcal{K}(., \nu)$ under the condition that the calibration constraint holds in mean

$$\mathbb{E}_{\nu^*} \left[\frac{1}{n} \sum_{i \in s} P_i \tilde{\mathbf{x}}_i \right] = \mathbb{E}_{\nu^*} \left[\sum_{i \in s} W_i \mathbf{x}_i \right] = t_{\mathbf{x}}. \tag{4.4}$$

We take the solution $\hat{p} = \mathbb{E}_{\nu^*}(P)$ and define the corresponding MEM estimator $\hat{t}_y$ as

$$\hat{t}_y = \frac{1}{n} \sum_{i \in s} \hat{p}_i \tilde{y}_i = \sum_{i \in s} \hat{w}_i y_i,$$

where we set $\hat{w}_i = d_i \hat{p}_i$ for all $i \in s$. Under the following assumptions, we will show in Theorem 4.3.1 that maximum entropy on the mean gives a Bayesian interpretation of calibration methods.

The random weights P_i, $i \in s$ (and therefore the W_i's) are taken independent. We denote by ν_i the prior distribution of P_i. It follows that $\nu = \otimes_{i \in s} \nu_i$. Moreover, all prior distributions ν_i are probability measures with mean 1. This last assumption conveys the information that $\hat{p}_i$ must be close to 1, equivalently, $\hat{w}_i = d_i \hat{p}_i$ must be close to d_i.

Denote by S_ν the interior of the convex hull of the support of ν and let $D(\nu)$ denote the domain of the log-Laplace transform Λ_ν, $D(\nu) = \{s \in \mathbb{R} : \ \Lambda_\nu(s) < \infty\}$. In the sequel, we assume that Λ_{ν_i} is essentially smooth (see [Roc97]) for all i, strictly convex and that ν_i is not concentrated on a single point. The last assumption means that if $D(\nu_i) = (-\infty; \alpha_i)$, $(\alpha_i \leq +\infty)$, then $\Lambda'_{\nu_i}(s)$

goes to $+\infty$ whenever $\alpha_i < +\infty$ and s goes to α_i. Under these assumptions, Λ'_{ν_i} is an increasing bijection between the interior of $D(\nu_i)$ and S_{ν_i}. So, denote by $\psi_i = {\Lambda'_{\nu_i}}^{-1}$ its inverse function, the Cramer transform $\Lambda^*_{\nu_i}$ of ν_i, which is defined as the convex conjugate of Λ_{ν_i}, satisfies

$$\Lambda^*_{\nu_i}(s) = s\psi_i(s) - \Lambda_{\nu_i}(\psi_i(s)).$$

Classical choices of priors ν_i lead to easily computable functions $\Lambda^*_{\nu_i}$ in most cases. Some examples are given at the end of the section.

Definition : We say that the optimization problem is feasible if there exists a vector $\delta = \{\delta_i\}_{i\in s} \in \otimes_{i\in s} S_{\nu_i}$ such that:

$$\sum_{i\in s} \delta_i \mathbf{x}_i = t_{\mathbf{x}}. \tag{4.5}$$

Under the last assumptions, the following proposition claims that the solutions $\{\hat{w}_i\}_{i\in s}$ are easily tractable.

Theorem 4.3.1 (survey sampling as a MEM procedure) *Assume that the optimization problem is feasible. The MEM estimator $\hat{w} = \{\hat{w}_i\}_{i\in s}$ minimizes over $\mathbb{R}^n$*

$$\{w_i\}_{i\in s} \mapsto \sum_{i\in s} \Lambda^*_{\nu_i}(\pi_i w_i)$$

under the constraint $\sum_{i\in s} \hat{w}_i \mathbf{x}_i = t_{\mathbf{x}}$.

Hence, we point out that maximum entropy on the mean method leads to calibration estimation, where the dissimilarity is determined by the Cramer transforms $\Lambda^*_{\nu_i}, i \in s$ of the prior distributions ν_i. Conditions we require on the priors in the MEM procedure correspond to regularity conditions on the dissimilarity. Indeed, taking priors ν_i with mean 1 yields $\Lambda^*_{\nu_i}(1) = {\Lambda^*_{\nu_i}}'(1) = 0$, which is a classical condition in calibration, see for instance [DS92] or Theorem 2.7.1 in [Ful09]. To see it, apply Jensen inequality to $\Lambda_\nu(t) = \log \int e^{tx} d\nu(x)$ to show that $\Lambda_\nu(t) \geq t$, which implies $\Lambda^*_\nu(1) = 0$. Since Λ^*_ν is smooth, non negative and strictly convex by construction, we also get ${\Lambda^*_\nu}'(1) = 0$.

Note that we require the feasibility condition (4.5) since we only consider here exact constraints in (4.4). An alternative would have been to consider a weakened constraint of the form

$$\left\| \mathbb{E}_{\nu^*_m} \left[\tfrac{1}{n} \textstyle\sum_{i=1}^n P_i \, \tilde{\mathbf{x}}_m(T_i) \right] - t_x \right\| \leq \epsilon$$

for a well chosen ϵ.

Remark Taking the priors ν_i in a certain class of measures may lead to specific dissimilarities known as Bregman divergences (see [KT04]). The definition of a Bregman divergence requires a strictly convex function, which in our situation, is given by the Cramer transform Λ^*_ν of some probability measure ν. Since we know that $\Lambda_\nu^*(1) = \Lambda_\nu^{*\prime}(1) = 0$, taking equal priors $\nu_i = \nu$ for all $i \in s$ leads to a dissimilarity that can be written

$$\mathcal{D}(w,d) = \sum_{i\in s} \Lambda_\nu^*(\pi_i w_i) = \sum_{i\in s} \left[\Lambda_\nu^*(\pi_i w_i) - \Lambda_\nu^*(1) - \Lambda_\nu^{*\prime}(1)(\pi_i w_i - 1)\right].$$

Here, we recognize the expression of the Bregman divergence between the weights $\pi_i w_i, i \in s$ and 1 associated to the convex function Λ_ν^*. Another possibility is to take prior distributions ν_i lying in some suitable exponential family. More precisely, define the prior distributions as

$$d\nu_i(x) = \exp(\alpha_i x + \beta_i) d\nu(d_i x), i \in s,$$

where $\beta_i = -\Lambda_\nu(\Lambda_\nu^{*\prime}(d_i))$ and $\alpha_i = d_i \Lambda_\nu^{*\prime}(d_i)$ are taken so that ν_i is a probability measure with mean 1. We recover after calculation the following dissimilarity

$$\mathcal{D}(w,d) = \sum_{i\in s} \left[\Lambda_\nu^*(w_i) - \Lambda_\nu^*(d_i) - \Lambda_\nu^{*\prime}(d_i)(w_i - d_i)\right],$$

which is the Bregman divergence between w and d associated to Λ_ν^*.

4.3.3 Bayesian interpretation of calibration using MEM

The two basic components of calibration are the set of constraint equations and the choice of a dissimilarity. Here, the latter is justified by prior measures $\{\nu_i\}_{i\in s}$ on the weights. We now see the probabilistic interpretation of some commonly used distances.

Stochastic interpretation of some usual calibrated survey sampling estimators

1. Generalized Gaussian prior. For a given positive sequence $\{q_i\}_{i\in s}$, take $\nu_i \sim \mathcal{N}(1, \pi_i q_i)$. We get

   $$\forall t \in \mathbb{R},\ \Lambda_{\nu_i}(t) = \frac{q_i \pi_i t^2}{2} + t\ ;\ \Lambda_{\nu_i}^*(t) = \frac{(t-1)^2}{2\pi_i q_i}$$

   The calibrated weights in that cases minimize the criterion

   $$\mathcal{D}_1(w,d) = \sum_{i\in s} \frac{(\pi_i w_i - 1)^2}{q_i \pi_i}.$$

   So, we recover the χ^2 distance discussed in Section 4.2. This is one of the main distance used in survey sampling. The q_i's can be seen as a smoothing sequence determined by the variance of the Gaussian prior. The larger the variance, the less stress is laid on the distance between the weights and the original Horvitz-Thompson weights.

2. Exponential prior. We take a unique prior ν with an exponential distribution with parameter 1. That is, $\nu = \nu^{\otimes n}$. We have in that case

   $$\forall t \in \mathbb{R}_+^*,\ \Lambda_\nu^*(t) = -\log t + t - 1.$$

This corresponds to the following dissimilarity

$$\mathcal{D}_2(w,d) = \sum_{i \in s} -\log(\pi_i w_i) + \pi_i w_i.$$

We here recognize the Bregman divergence between $\{\pi_i w_i\}_{i \in s}$ and 1 associated to Λ^*_ν, as explained in the previous remark. A direct calculation shows that this is also the Bregman divergence between w and d associated to Λ^*_ν. The two distances are the same in that case.

3. Poisson prior. If we choose for prior $\nu_i = \nu, \forall i \in s$, where ν is the Poisson distribution with parameter 1, then we obtain

$$\forall t \in \mathbb{R}^*_+,\ \Lambda^*_\nu(t) = t \log t - t + 1.$$

So, we have the following contrast

$$\mathcal{D}_3(w,d) = \sum_{i \in s} \pi_i w_i \log(\pi_i w_i) - \pi_i w_i.$$

We recover the Kullback information where $\{\pi_i w_i\}_{i \in s}$ is identified with a discrete measures on s.

MEM leads to a classical calibration problem where the solution is defined as a minimizer of a convex function subject to linear constraints. The following result gives another expression of the solution for which the computation may be easier in practical cases.

Proposition 4.3.2 *Assume that the optimization problem is feasible, the MEM estimator $\hat{w}$ is given by:*

$$\forall i \in s,\ \hat{w}_i = d_i \Lambda'_{\nu_i}(\hat{\lambda}^t d_i \mathbf{x}_i)$$

where $\hat{\lambda}$ minimizes over $\mathbb{R}^k$ $\lambda \mapsto \sum_{i \in s} \Lambda_{\nu_i}(\lambda^t d_i \mathbf{x}_i) - \lambda^t t_{\mathbf{x}}$.

We endow y with new weights obtaining the MEM estimator $\hat{t}_y = \sum_{i \in s} \hat{w}_i y_i$. Note that the function $\Lambda_{\nu_i}(d_i .\ )$ corresponds to the function F_i in [DS92], while taking identical priors $\nu_i = \nu$ for all $i \in s$ recovers the particular case $\Lambda_\nu = F$ according to the notations used in [DS92].

Calibration using maximum entropy framework turns into a general convex optimization program, which can be easily solved. Indeed, computing the new weights w_i only involves a two step procedure. First, we find the unique $\hat{\lambda} \in \mathbb{R}^k$ such that

$$\sum_{i \in s} d_i \Lambda'_{\nu_i}(d_i \hat{\lambda}^t \mathbf{x}_i)\mathbf{x}_i - t_{\mathbf{x}} = 0. \tag{4.6}$$

This is achieved by optimizing a convex function. Then, compute the new weights $\hat{w}_i = d_i \Lambda'_{\nu_i}(d_i \hat{\lambda}^t \mathbf{x}_i)$.

4.3.4 Extension to generalized calibration and instrument estimation

Computing a calibration estimator requires that (4.6) has a unique solution. This condition follows from the convexity of the functions $\Lambda_{\nu_i}, i \in s$. Aiming to provide wider possibilities of estimation, the method of generalized calibration (GC) considered in [Sau] consists in replacing the functions $\lambda \mapsto \Lambda'_{\nu_i}(d_i \lambda^t \mathbf{x}_i)$ by more general functions $f_i : \mathbb{R}^k \to \mathbb{R}$. Assume that the equation

$$F(\lambda) = \sum_{i \in s} d_i f_i(\lambda) \mathbf{x}_i = t_{\mathbf{x}} \tag{4.7}$$

has a unique solution $\hat{\lambda}$. Assume also that the f_i are continuously differentiable at 0, and are such that $f_i(0) = 1$ so that $F(0) = \hat{t}_{\mathbf{x}}^{HT}$. Then, take as the solution to the generalized calibration procedure, the weights:

$$\forall i \in s, \ \hat{w}_i = d_i f_i(\hat{\lambda}).$$

Calibration is of course a particular example of generalized calibration where we set $f_i : \lambda \mapsto \Lambda'_{\nu_i}(d_i \lambda^t \mathbf{x}_i)$ to recover a calibration problem seen in Section 4.3. An interesting example of GC is to take affine functions $\lambda \mapsto 1 + \mathbf{z}_i^t \lambda$, where $\{\mathbf{z}_i\}_{i \in s}$ is a sequence of vectors of $\mathbb{R}^k$. The $\mathbf{z}_i$'s are called instruments (see [Sau]). If the matrix $X_n := \frac{1}{N} \sum_{i \in s} d_i \mathbf{z}_i \mathbf{x}_i^t$ is invertible, the resulting estimator $\hat{t}_y$, referred to as the instrument estimator obtained with the instruments $\mathbf{z}_i$, is given by:

$$\hat{t}_y = \hat{t}_y^{HT} + (t_{\mathbf{x}} - \hat{t}_{\mathbf{x}}^{HT})^t X_n^{-1} \sum_{i \in s} d_i \mathbf{z}_i y_i. \tag{4.8}$$

Remark As pointed out in [Wu03] in the case $\mathbf{z}_i = \mathbf{x}_i$, the estimator of the population total is identical to the one obtained with one-dimensional auxiliary variable $\hat{B}^t \mathbf{x}$, where $\hat{B}$ is estimated by least squares. More generally, reducing the dimension of the auxiliary variable to one is always possible when using instruments. The new auxiliary variable and instruments are linear transformations $\hat{B}^t \mathbf{x}$ and $\hat{B}^t \mathbf{z}$ of the original variables $\mathbf{x}$ and $\mathbf{z}$, where

$$\hat{B} = \left[\sum_{i \in s} d_i \mathbf{z}_i \mathbf{x}_i^t \right]^{-1} \sum_{i \in s} d_i y_i \mathbf{z}_i.$$

This points out the relationship between calibration and linear regression discussed in [DS92]. The method implicitly aims at constructing a variable $\tilde{y} = y - \hat{B}^t \mathbf{x}$ with a lower variance than that of y (at least for sufficiently large samples), and for which the population total is known up to t_y. The calibrated estimator $\hat{t}_y$ can be written

$$\hat{t}_y = \sum_{i \in s} d_i \tilde{y}_i + \hat{B}^t t_{\mathbf{x}},$$

that is, $\hat{t}_y$ is the Horvitz-Thompson estimator (up to a known additive constant, here $\hat{B}^t t_{\mathbf{x}}$) of the variable $\tilde{y}$.

Instrument estimators play a crucial role when studying the asymptotic properties of generalized calibration estimation. A classical asymptotic framework in calibration is to consider that n and N simultaneously go to infinity while the Horvitz-Thompson estimators of the mean converge at a rate of convergence of $\sqrt{n}$, as described in [DS92] and [Wu03] for instance. Hence, we assume that

$$\frac{1}{N}\|\hat{t}^{HT}_{\mathbf{x}} - t_{\mathbf{x}}\| = O\left(\frac{1}{\sqrt{n}}\right) \quad \text{and} \quad \frac{1}{N}(\hat{t}^{HT}_y - t_y) = O\left(\frac{1}{\sqrt{n}}\right),$$

further assumptions on our asymptotic framework are made in Section 4.4. Besides, we say that two GC estimators $\hat{t}_y$ and $\tilde{t}_y$ are asymptotically equivalent if

$$\frac{1}{N}(\hat{t}_y - \tilde{t}_y) = o\left(\frac{1}{\sqrt{n}}\right).$$

Proposition 4.3.3 *Let $\hat{t}_y$ and $\tilde{t}_y$ be the GC estimators obtained respectively with the functions $f_i, i \in s$ and $g_i, i \in s$. If for all $i \in s$, $\nabla f_i(0) = \nabla g_i(0) = \mathbf{z}_i$, and if the matrix $X_n := \frac{1}{N}\sum_{i\in s} d_i \mathbf{z}_i \mathbf{x}_i^t$ converges toward an invertible matrix X, then $\hat{t}_y$ and $\tilde{t}_y$ are asymptotically equivalent. In particular, two MEM estimators are asymptotically equivalent as soon as their prior distributions have the same respective variances.*

This proposition is a consequence of Result 3 in [DS92]. It states that first order asymptotic behavior of GC estimators in only determined by the gradient vectors $\mathbf{z}_i = \nabla f_i(0), i \in s$, where the f_i's are the functions used in (4.7). As a result, all GC estimator can be shown to have an asymptotically equivalent instrument estimator.

The frame of calibration and MEM estimation corresponds to instruments of the form $\mathbf{z}_i = q_i \mathbf{x}_i$. This particular case is discussed in [DS92] where the authors prove that a calibration estimator can always be approximated by a regression estimator under regularity conditions. A different proof of this result is also given in Theorem 2.7.1 in [Ful09]. Thus, a MEM estimator $\hat{t}_y$ obtained with prior distributions $\nu_i, i \in s$ with respective variances $\pi_i q_i$ satisfies

$$\hat{t}_y = \hat{t}^{HT}_y + (t_{\mathbf{x}} - \hat{t}^{HT}_{\mathbf{x}})^t \hat{B} + o\left(\frac{N}{\sqrt{n}}\right)$$

where $\hat{B} = \left[\sum_{i\in s} d_i q_i \mathbf{x}_i \mathbf{x}_i^t\right]^{-1} \sum_{i\in s} d_i q_i \mathbf{x}_i y_i$. The negligible term is null for Gaussian priors, leading to a χ^2 dissimilarity in the frame of calibration (see Example 1 in Section 4.3.3) and to the instrument estimator built with instruments $\mathbf{z}_i = q_i \mathbf{x}_i$. This choice of instruments, and in particular the case $q_i = 1$ for all $i \in s$, is often used in practice since it provides a consistent estimate which can be easily computed.

4.4 Efficiency of calibration estimator with MEM method

The accuracy of the estimator heavily relies on the linear correlation between y and the auxiliary variable $\mathbf{x}$. If a relationship other than linear prevails, $\mathbf{x}$ may not be an efficient choice of calibration variable. When complete information is available, *model calibration* proposed by Wu and Sitter aims to generalize the calibration procedure by considering an auxiliary variable of the form $u(\mathbf{x})$ for $u : \mathbb{R}^k \to \mathbb{R}^d$ a given function. Their objective is to increase the linear correlation between the variables, leading to a better efficiency of the estimation. In [WS01], Wu and Sitter assume that the optimal calibration function u belongs to a known parametric class of functions for which the true value of the parameter is estimated from the data. Montanari and Ranalli [MR05] discuss the estimation of the optimal choice for the function u in a non parametric model.

With complete information, the choice of the calibration function u and the instruments are the two main aspects of the estimation of t_y. In this section, we first study the influence of the instruments $\mathbf{z}$ when the calibration function u is fixed. Then, we discuss ways of improving the estimation by allowing both the instruments and the calibration variable to vary with the observations.

4.4.1 Asymptotic efficiency

We consider the usual asymptotic framework in survey sampling where there is a sequence of sampling designs and finite populations, indexed by r. The population size and the sample size, denoted respectively by N_r and n_r, both grow to infinity as $r \to +\infty$. The asymptotic framework is to be understood in the sense that $r \to +\infty$, but, in the following, the index r will be suppressed to simplify notation. We consider the population measurements $\{(\mathbf{x}_i, y_i), i = 1, ..., N\}$ as independent realizations of a random variable (X, Y) from a superpopulation model ξ.
For $u : \mathbb{R}^k \to \mathbb{R}^d$ a given function, we note $u_i = u(\mathbf{x}_i)$ and

$$t_u = \sum_{i \in U} u_i, \qquad \hat{t}_u^{HT} = \sum_{i \in s} d_i u_i.$$

In the sequel, we assume that $\mathbb{E}|Y|^3 < \infty$ and $\mathbb{E}\|u(X)^3\| < \infty$, where $\mathbb{E}$ denotes the expectation with respect to the distribution of (X, Y). In a general setting where the auxiliary variable takes the form $u(\mathbf{x})$, instrument estimators have the following expression

$$\hat{t}_y = \hat{t}_y(u) = \hat{t}_y^{HT} + (t_u - \hat{t}_u^{HT})^t \hat{B}_u,$$

where $\hat{B}_u = \left[\sum_{i \in s} d_i \mathbf{z}_i u_i^t\right]^{-1} \sum_{i \in s} d_i y_i \mathbf{z}_i$ is assumed to be well defined. Furthermore, define the joint inclusion probabilities $\pi_{ij} = \sum_{s:\ i,j \in s} p(s)$ and set $\Delta_{ij} := \pi_{ij} d_i d_j - 1$.

The nonlinearity of $\hat{t}_y$ makes it difficult to evaluate its quadratic risk. Following [ES06], easier is to consider its linear asymptotic expansion $\hat{t}_{y,\text{lin}}(u) := \hat{t}_y^{HT} + (t_u - \hat{t}_u^{HT}) B_u$ where B_u is a vector, independent from the sample s, such that $\|\hat{B}_u - B_u\| = o(1)$. The linear expansion $\hat{t}_{y,\text{lin}}(u)$ is design unbiased and is asymptotically equivalent to $\hat{t}_y$. As proved in [Mon87], the variance of $\hat{t}_{y,\text{lin}}(u)$, which is given by

$$\text{var}_p(\hat{t}_{y,\text{lin}}(u)) = \sum_{i,j \in U} \Delta_{ij}\ (y_i - B_u^t u_i)(y_j - B_u^t u_j),$$

depends on the instruments only through the value of B_u and is minimal for $B_u = B_u^*$ given by

$$B_u^* = \left[\text{var}_p(\hat{t}_u^{HT})\right]^{-1} \text{cov}_p(\hat{t}_u^{HT}, \hat{t}_y^{HT}) = \left[\sum_{i,j \in U} \Delta_{ij}\ u_i u_j\right]^{-1} \sum_{i,j \in U} \Delta_{ij}\ u_j y_i,$$

where var_p and cov_p denote respectively the variance and covariance under the sampling design p. We make the following assumptions

A4.1. The sampling design $p(.)$ is weakly dependent of ξ in the sense that for any sequence $\{a(x_i, y_i)\}_{i\in U} = \{a_i\}_{i\in U}$ such that $\frac{1}{N}\sum_{i\in U} |a_i|^3 = O(1)$,

$$\mathbb{E} \sum_{i,j\in U} \Delta_{ij}\ a_i a_j = \sum_{i,j\in U} \Delta_{ij}\ \mathbb{E}(a_i a_j) + o\left(\frac{N^2}{n}\right).$$

A4.2. There exists $0 \leq \pi < 1$, such that $\limsup_{r\to\infty} \dfrac{n}{N} = \pi$.

A4.3. $\lim \dfrac{n}{N^2} \sum_{i\in U} \Delta_{ii} = -\lim \dfrac{n}{N^2} \sum_{i\in U} \sum_{j\neq i} \Delta_{ij} = 1 - \pi.$

The first assumption conveys the information that no design weight is disproportionately large compared to the others. It holds for instance if p and ξ are independent and if $\sum_{i\in U} \Delta_{ii}^2 = o(N^4 n^{-2})$ and $\sum_{i\in U}\sum_{j\neq i} \Delta_{ij}^2 = o(N^3 n^{-2})$. The condition A4.2 is not restrictive, it simply states that the number of unobserved data never becomes negligible compared to the population size. This is a classical assumption in survey sampling, see for instance [MR05]. The condition A4.3 is essentially made to ensure the existence of efficient estimators as shown further. It is fulfilled when the sampling design is uniform on the samples of size n, provided that the sample size and the population size remain of the same order. In that case, the Horvitz-Thompson weights are $\pi_i = n/N$, $\pi_{ij} = n(n-1)/N(N-1), \forall i \neq j$, yielding $\Delta_{ii} = N/n - 1$ and $\Delta_{ij} = -(N-n)/n(N-1)$.

Lemma 4.4.1 *Suppose that A4.1 and A4.2 hold. Then,*

$$\frac{n}{N^2}\ \mathbb{E}\ \mathbb{E}_p(t_y - \hat{t}_y(u))^2 \geq (1-\pi)\mathrm{var}\left(Y - B_u^t u(X)\right) + o(1),$$

with equality if, and only if, A4.3 also holds.

The proof is a direct consequence of Lemma 4.5.1. This result provides a natural criterion of asymptotic efficiency. Indeed, finding instruments for which the right term of the inequality is minimal appears as a natural objective, whether the sampling design satisfies A4.3 or not. So, the variance lower bound is defined as the minimum $V^*(u)$ of $(1-\pi)\mathrm{var}(Y - B^t u(X))$ as B ranges over $\mathbb{R}^d$. We say that an estimator $\hat{t}_y(u)$ is asymptotically efficient if the expectation of its design quadratic risk converges toward $V^*(u)$. This is an analog of optimal calibration in [Mon87], where in our framework, optimality requires that

$$\lim \hat{B}_u = [\mathrm{var}(u(X))]^{-1} \mathrm{cov}(Y, u(X)), \tag{4.9}$$

assuming that $\mathrm{var}(u(X))$ is invertible. In this case, we get

$$V^*(u) = (1-\pi)\mathrm{var}\left(Y - \mathrm{cov}(Y, u(X))^t \left[\mathrm{var}(u(X))\right]^{-1} u(X)\right). \tag{4.10}$$

Note that this lower bound can not be reached if A4.3 is not fulfilled.

Estevao and Särndal [ES06] propose the instruments $\mathbf{z}_i^* = \sum_{j\in U} \Delta_{ij}\ u_j$ as a natural choice by identification, noticing that the optimal value B_u^* for a fixed N verifies

$$B_u^* = \left[\sum_{i,j\in U} \Delta_{ij}\ u_i u_j\right]^{-1} \sum_{i,j\in U} \Delta_{ij}\ u_j y_i = \left[\sum_{i\in U} u_i \mathbf{z}_i^*\right]^{-1} \sum_{i\in U} y_i \mathbf{z}_i^*.$$

In our framework, these instruments satisfy condition (4.9), as a consequence of Lemma 4.5.1. However, a noticeable drawback is that the calculation of each instrument $\mathbf{z}_i^*$ involves the whole population $\{\mathbf{x}_i, i \in U\}$, yielding a computationally expensive estimate.

The simple choice $\mathbf{z}_i = u_i$ provides a good alternative. As shown in Proposition 4.3.3, the resulting estimator is asymptotically equivalent to MEM estimators built using prior distributions ν_i with variance π_i. The consistency of the Horvitz-Thomspon estimates leads to

$$\hat{B}_u = \left[\sum_{i\in s} d_i u_i u_i^t\right]^{-1} \sum_{i\in s} d_i y_i u_i \longrightarrow \left[\mathbb{E}(u(X)u(X)^t)\right]^{-1} \mathbb{E}(Yu(X)).$$

Although condition (4.9) for optimality is not fulfilled for most functions u, the problem can be easily solved by adding the constant variable 1 in the calibration constraint. We then consider the MEM estimator $\hat{t}_y(v)$ where $v = (1, u)^t : \mathbb{R}^k \to \mathbb{R}^{d+1}$, the calibrated weights now satisfy the constraints

$$\sum_{i\in s} w_i u_i = t_u, \ \sum_{i\in s} w_i = N.$$

Here, the matrix $\text{var}(v(X))$ is not invertible although it is simple algebra to see that $V^*(v) = V^*(u)$. So, the auxiliary variable is modified but the asymptotic lower bound is unchanged. As a result of the dimension reduction property of calibration, adding the constant in the calibration constraint reduces to use the instruments $\mathbf{z}_i = u_i - \hat{t}_u^{HT}$ up the a negligible term. A direct calculation shows that these instruments now satisfy condition (4.9).

4.4.2 Approximate Maximum Entropy on the Mean

We now turn to the question of the optimal auxiliary variable. By minimizing the asymptotic variance lower bound $V^*(u)$ with respect to u, the conditional expectation $\Phi(\mathbf{x}_i) = \mathbb{E}(Y|X = \mathbf{x}_i)$ appears as the optimal choice since $\Phi(.)$ is the unique (up to affine transformations) minimizer of the functional $u \mapsto V^*(u)$ in Equation (4.10) (u can be taken real-valued without loss of generality). This confirms the result stated in Theorem 1 in [Wu03], where Wu proves the variable $\Phi(\mathbf{x}_i), i \in U$ to be optimal. In that case, the asymptotic lower bound is given by:

$$V^* = (1-\pi)\mathbb{E}(Y - \mathbb{E}(Y|X))^2.$$

Note that, since this optimal choice depends on the unknown distribution of (X, Y), this result does not provide a tractable estimator. A natural solution is to replace Φ by an estimate Φ_m, and then plug it into the calibration constraint. Under regularity conditions that will be made precise later, we show that this approach enables to compute an asymptotically optimal estimator of t_y, in the sense that its asymptotic expected variance is equal to the lower bound V^* defined above.

In this section, $\hat{t}_y(u)$ will denote a MEM estimator of t_y obtained with auxiliary variable $(u(\mathbf{x}),1)^t$ and prior distributions ν_i with variance π_i. We recall that for any measurable function u, $\hat{t}_y(u)$ is $\sqrt{n}$-consistent with asymptotic variance $V^*(u)$. Assume that we observe approximations $\{\Phi_m\}_{m\in\mathbb{N}}$ of Φ, we define the AMEM estimator as $\hat{t}_y(\Phi_m)$, i.e., the MEM estimator calibrated with the variable $(\Phi_m(\mathbf{x}),1)^t$.

Theorem 4.4.2 *Suppose that Assumptions A4.1 to A4.3 hold. Let $\{\Phi_m\}_{m\in\mathbb{N}}$ be a sequence of functions satisfying*

$$i) \quad \frac{n}{N^2}\,\mathbb{E}\ \mathbb{E}_p(\hat{t}^{HT}_{\Phi-\Phi_m} - t_{\Phi-\Phi_m})^2 \longrightarrow 0$$

$$ii) \quad \hat{B}_{\Phi_m} = \left[\sum_{i\in s} d_i \Phi_m(\mathbf{x}_i)^2\right]^{-1} \sum_{i\in s} d_i y_i \Phi_m(\mathbf{x}_i) \longrightarrow 1,$$

as $(r,m)\to\infty$. Then, the AMEM estimator $\hat{t}_y(\Phi_m)$ is asymptotically optimal among all GC estimators in the sense that $\frac{n}{N^2}\mathbb{E}\ \mathbb{E}_p(t_y - \hat{t}_y(\Phi_m))^2$ converges toward V^* as $(r,m)\to\infty$.

This theorem does not rule out that the functions Φ_m are estimated using the data. Hence, it is possible to compute an asymptotically efficient estimator of t_y with a single sample, since a data driven estimator Φ_n provides an asymptotically efficient estimator of t_y, as soon as the two conditions of Theorem 4.4.2 are fulfilled.

Remark that although this natural way to extend calibration to non parametric procedures can be claimed to be asymptotically optimal, the resulting estimator may still be highly unstable for relatively small samples or under irregular sampling designs.

Many non parametric methods could be used in the frame of calibration, see for instance [MR05]. Here, we study an approach where the conditional expectation Φ is estimated by projection onto finite dimensional subspaces. Let $\phi = (1,\phi_1,\phi_2,...)$ be a sequence of linearly independent functions, total in the space of square integrable functions. This sequence is referred to as a projection basis. Typically, it can be polynomials if X takes values in a compact subset of $\mathbb{R}^k$ or wavelets but other forms may be chosen, depending on the situation.

Denote by $\phi^m = (1,\phi_1,...,\phi_m)$ the vector of the first $m+1$ components of ϕ, we build a projection estimator Φ_m of Φ by considering a suitable linear combination $\hat{B}^t_m\phi^m$ of the functions, the vector $\hat{B}_m$ being generally obtained by least squares on the variables y and $\phi^m(\mathbf{x})$. In the context of calibration, it is natural to consider design based estimates Φ_m. As a result of a reciprocal effect of the dimension reduction property, taking $\Phi_m = \hat{B}^t_m\phi^m$ with

$$\hat{B}_m = \hat{B}_{\phi^m} = \left[\sum_{i\in s} d_i \phi^m_i \phi^{mt}_i\right]^{-1} \sum_{i\in s} d_i y_i \phi^m_i,$$

leads to the estimator calibrated with the vector variable $\phi^m(\mathbf{x})$ up to a negligible term. Indeed, the auxiliary variable $\Phi_m(\mathbf{x}) = \hat{B}_m^t \phi^m(\mathbf{x})$ is obtained as the one dimensional equivalent of $\phi^m(\mathbf{x})$ discussed in Section 4.3. So, we point out that calibration is here extended to non parametric procedures by simply increasing the number of auxiliary variables. The estimator calibrated with a χ^2 dissimilarity can be expressed as

$$\hat{t}_y(\Phi_m) = \hat{t}_y^{HT} + (t_{\Phi_m} - \hat{t}_{\Phi_m}^{HT}) = \hat{t}_y^{HT} + (t_{\phi^m} - \hat{t}_{\phi^m}^{HT})^t \hat{B}_{\phi^m},$$

which illustrates the equivalence between auxiliary variables $\Phi_m(\mathbf{x})$ and $\phi^m(\mathbf{x})$. With the notations of Theorem 4.4.2, the vector $\hat{B}_{\Phi_m}$ corresponding to $\Phi_m(\mathbf{x})$ is equal to 1 for all m and therefore satisfies the condition *ii*) in the corollary. The condition *i*) can also be fulfilled with this method, although, a proper number of constraints must be chosen. If m is fixed, we know that $\hat{t}_y(\Phi_m)$ converges toward t_y with an asymptotic variance $V^*(\phi^m)$. The convergence of $V^*(\phi^m)$ toward V^* as $m \to \infty$ warrants the existence of a sequence of integers $\{m(n)\}_{n\in\mathbb{N}}$ such that $\Phi_{m(n)}$ undergoes the first condition of Theorem 4.4.2. Note however that finding such a sequence is a difficult task and belongs to the class of model selection issues.

Asymptotic results of non parametric methods are to be taken with care since it may require a large number of observations before the method becomes really effective. Here we assumed strong regularity conditions on the sampling design, allowing good consistency of the non parametric estimation with relatively small samples. AMEM procedures in survey sampling have the advantage to enable to implement non parametric procedures while remaining in a Bayesian framework.

4.4.3 Simulations

We provide in this section some numerical applications. In particular, we want to study the influence of the choice of the projection basis and the number of constraints, in an AMEM procedure where the conditional expectation is estimated by projection. For this, we consider two kinds of relationships between the variable of interest and the auxiliary variable, namely $Y = \exp(X) + \varepsilon$ and $Y = 4/X + \varepsilon$ where X is a uniform random variable on $[2;3]$ and ε is a standard Gaussian noise.

A sample s of size n is drawn from a uniform sampling design in a population $U = 2500$. Results are obtained for $n = 100$ and $n = 500$. We compute eights estimates $\hat{\rho}_1$ to $\hat{\rho}_8$ of the population mean $N^{-1}t_y$, obtained by calibration with regular χ^2 dissimilarity with different auxiliary variables.

- $\hat{\rho}_1$ is the usual Horvitz-Thompson estimator.
- $\hat{\rho}_2$ is the regular calibration estimator calibrated with $\mathbf{x}$.
- $\hat{\rho}_3$ is obtained with perfect calibration variable $\Phi(\mathbf{x})$ and plays the role of an oracle.
- $\hat{\rho}_4$ to $\hat{\rho}_8$ are obtained by increasing the set of calibration variables, adding for each step a higher power of $\mathbf{x}$.

Estimators $\hat{\rho}_4$ to $\hat{\rho}_8$ can be viewed as specific AMEM estimators, where $\Phi(\mathbf{x})$ is approximated by a projection $\Phi_m(\mathbf{x})$ of y onto a polynomial subspace of dimension $m+1$. So, the projection basis is

$(1, \mathbf{x}, \mathbf{x}^2, ..., \mathbf{x}^m, ...)$ and we consider here the cases $m = 1$ to $m = 5$, with the respective estimators $\hat{\rho}_4$ to $\hat{\rho}_8$. The construction of the estimators is detailed in the following table. Quadratic risks are estimated by Monte Carlo with 10000 replications of the procedure, and are given in the last two columns of the table for the different sample sizes $n = 100$ and $n = 500$.

First, we consider the model $Y = \exp(X) + \varepsilon$.

estimator	auxiliary variable	$n = 100$	$n = 500$
$\hat{\rho}_1$ (Horvitz-Thompson)	none	$2.64.10^{-1}$	$5.02.10^{-2}$
$\hat{\rho}_2$ (calibration)	$\mathbf{x}$	$6.90.10^{-2}$	$1.33.10^{-2}$
$\hat{\rho}_3$ (oracle)	$\exp \mathbf{x}$	$1.00.10^{-2}$	$2.00.10^{-3}$
$\hat{\rho}_4$ (AMEM $m = 1$)	$(1, \mathbf{x})$	$2.60.10^{-2}$	$5.01.10^{-3}$
$\hat{\rho}_5$ (AMEM $m = 2$)	$(1, \mathbf{x}, \mathbf{x}^2)$	$1.05.10^{-2}$	$2.30.10^{-3}$
$\hat{\rho}_6$ (AMEM $m = 3$)	$(1, \mathbf{x}, \mathbf{x}^2, \mathbf{x}^3)$	$1.02.10^{-2}$	$2.10.10^{-3}$
$\hat{\rho}_7$ (AMEM $m = 4$)	$(1, \mathbf{x}, \mathbf{x}^2, \mathbf{x}^3, \mathbf{x}^4)$	$1.03.10^{-2}$	$2.03.10^{-3}$
$\hat{\rho}_8$ (AMEM $m = 5$)	$(1, \mathbf{x}, \mathbf{x}^2, \mathbf{x}^3, \mathbf{x}^4, \mathbf{x}^5)$	$1.08.10^{-2}$	$2.24.10^{-3}$

The high quadratic risk of the Horvitz-Thompson estimate $\hat{\rho}_1$ compared to the other estimators points out the improvement of the estimation due to calibration. Note also a significant gain of efficiency between $\hat{\rho}_2$ and $\hat{\rho}_4$, simply due to the addition of the constant in the set of calibration constraints. Among AMEM estimators, $\hat{\rho}_6$ seems to be the most efficient when $n = 100$ while $\hat{\rho}_7$ has the smallest estimated quadratic risk when the sample size grows to $n = 500$. So, the optimal number of constraint increases with the sample size, going from $m = 3$ when $n = 100$ to $m = 4$ when $n = 500$. A problem of over-fitting is observed for a too large number of constraints, as we see that the quadratic risks increase when m goes past its optimal value. Increasing the number of constraint seems necessary to have an efficient estimation, although a good balance must be found between the two indexes m and n.

With an exponential relationship between Y and X, polynomial functions appear to yield a good estimation of the population mean in a AMEM procedure, as we see that AMEM estimators perform almost as well as the oracle $\hat{\rho}_3$ under a suitable number of constraint. This is not surprising, given that the exponential function can be relatively well approximated by low degree polynomials on the interval $[2; 3]$.

We now study the case $Y = 4/X + \varepsilon$. We consider the same eight estimators of $\frac{1}{N} t_y$, the quadratic risks are estimated by Monte-Carlo from 10000 simulated samples, as previously. Results are given for three values of (N, n), namely $(2500, 100)$, $(2500, 500)$ and $(5000, 1000)$.

estimator	auxiliary variable	$(2500, 100)$	$(2500, 500)$	$(5000, 1000)$
$\hat{\rho}_1$	none	$1.53.10^{-2}$	$3.1.10^{-3}$	$1.5.10^{-3}$
$\hat{\rho}_2$	$\mathbf{x}$	$2.58.10^{-2}$	$5.1.10^{-3}$	$2.4.10^{-3}$
$\hat{\rho}_3$	$4\mathbf{x}^{-1}$	$1.01.10^{-2}$	$1.9.10^{-3}$	$8.9.10^{-4}$
$\hat{\rho}_4$	$(1, \mathbf{x})$	$1.07.10^{-2}$	$2.3.10^{-3}$	$1.2.10^{-3}$
$\hat{\rho}_5$	$(1, \mathbf{x}, \mathbf{x}^2)$	$1.03.10^{-2}$	$2.2.10^{-3}$	$1.1.10^{-3}$
$\hat{\rho}_6$	$(1, \mathbf{x}, \mathbf{x}^2, \mathbf{x}^3)$	$1.07.10^{-2}$	$2.0.10^{-3}$	$9.9.10^{-4}$
$\hat{\rho}_7$	$(1, \mathbf{x}, \mathbf{x}^2, \mathbf{x}^3, \mathbf{x}^4)$	$1.08.10^{-2}$	$2.0.10^{-3}$	$9.7.10^{-4}$
$\hat{\rho}_8$	$(1, \mathbf{x}, \mathbf{x}^2, \mathbf{x}^3, \mathbf{x}^4, \mathbf{x}^5)$	$1.08.10^{-2}$	$2.1.10^{-3}$	$9.6.10^{-4}$

A first surprising result is the negative effect of calibration, pointed out by the quadratic risk of $\hat{\rho}_2$ larger than that of $\hat{\rho}_1$. This bad use of the auxiliary information is due to weak linear correlation between Y and X with $\mathrm{cov}(X, Y) \approx 0$ while the correlation coefficient $\mathbb{E}(XY)/\mathbb{E}(X^2)$ is far from zero. Nevertheless, this drawback is partly corrected by adding the constant in the set of calibration constraints, which corresponds to $\hat{\rho}_4$. The optimal number of constraints grows with the sample size as expected, as we see that the most efficient AMEM estimator is given for $m = 2$ when $(N, n) = (2500, 100)$, $m = 3$ or $m = 4$ when $N = 2500, n = 500$ and $m = 5$ when $N = 5000, n = 1000$. The quadratic risk is in each case quite close to that of the oracle $\hat{\rho}_3$. In spite of the inappropriate projection basis in this example, AMEM estimation still turns out to be rather satisfactory, provided that the number of constraint is properly chosen.

4.5 Proofs

4.5.1 Technical lemmas

Lemma 4.5.1 *Under A4.1 and A4.2, for any sequence* $\{a(x_i, y_i)\}_{i\in U} = \{a_i\}_{i\in U}$ *such that* $\frac{1}{N}\sum_{i\in U} |a_i|^3 = O(1)$,

$$\frac{n}{N^2}\,\mathbb{E}\sum_{i,j\in U} \Delta_{ij} a_i a_j \geq (1-\pi)\mathrm{var}(a(X, Y)) + o(1)$$

with equality if and only if A4.3 also holds. Moreover, under A4.1 to A4.3, the quantity $\frac{n}{N^2}\sum_{i,j\in U} \Delta_{ij} a_i b_j$ *converges in probability toward* $\mathrm{cov}(a(X, Y), b(X, Y))$ *for all sequence* $\{b_i\}_{i\in U}$ *satisfying the same conditions as* $\{a_i\}_{i\in U}$.

Proof. For such a sequence $a = \{a_i\}_{i\in U}$, A4.1 and A4.2 yield:

$$\frac{n}{N^2}\sum_{i,j\in U} \Delta_{ij}\ a_i a_j = \frac{n}{N^2}\sum_{i\in U} \Delta_{ii}\ a_i^2 + \frac{n}{N^2}\sum_{i\neq j} \Delta_{ij}\ a_i a_j$$

$$= \left(\frac{n}{N^2}\sum_{i\in U} \Delta_{ii}\right)\mathbb{E}(a(X, Y)^2) + \left(\frac{n}{N^2}\sum_{i\neq j} \Delta_{ij}\right)\mathbb{E}(a(X, Y))^2 + o(1)$$

Denote by $\mathcal{P}_n(U)$ the set of all subsamples s of U with n elements. By Jensen inequality,

$$\sum_{i,j\in U}\Delta_{ij}=\sum_{s\in\mathcal{P}_n(U)}\left(\sum_{i\in s}d_i\right)^2 p(s)-N^2\geq\left[\sum_{s\in\mathcal{P}_n(U)}\left(\sum_{i\in s}d_i\right)p(s)\right]^2-N^2\geq 0$$

which implies that $\sum_{i\neq j}\Delta_{ij}\geq-\sum_{i\in U}\Delta_{ii}$. Thus:

$$\frac{n}{N^2}\sum_{i,j\in U}\Delta_{ij}\ f_if_j\geq\left(\frac{n}{N^2}\sum_{i\in U}\Delta_{ii}\right)\mathrm{var}(f(X,Y))+o(1).$$

Since $\sum_{i\in U}\pi_i=n$, we know that $\frac{n}{N^2}\sum_{i\in U}\Delta_{ii}\geq 1-\frac{n}{N}$ by convexity of $x\mapsto 1/x$ on $\mathbb{R}_+^*$. Hence

$$\frac{n}{N^2}\sum_{i,j\in U}\Delta_{ij}\ a_ia_j\geq(1-\pi)\mathrm{var}(a(X,Y))+o(1)$$

without equality for all sequence a if A4.3 is not true. The end of the lemma follows directly by using the same guideline applied to a and b. In particular, it holds when $a=b$.

4.5.2 Proof of Theorem 4.3.1

Let $\hat{p}=\{\hat{p}_i\}_{i\in s}$ where $\hat{p}_i=\pi_i\hat{w}_i$. We have

$$\hat{p}=\arg\min_{p\in S}\sum_{i=1}^n\Lambda_\nu^*(p_i),$$

where $S=\{p\in\mathbb{R}^n:\frac{1}{n}\sum_{i=1}^n p_i\tilde{\mathbf{x}}_i=t_{\mathbf{x}}\}$. The proof is similar to that of Theorem 1.2.3. For a fixed $p\in\mathbb{R}^n$, let ν_p be the I-projection of ν_0 onto $S_p=\{\mu\in\mathcal{P}(\mathbb{R}^n):\mathbb{E}_\mu(P)=p\}$. Following the proof of Theorem 1.2.3, we know that $\mathcal{K}(\nu_p|\nu_0)=\Lambda_{\nu_0}^*(p)=\sum_{i=1}^n\Lambda_{\nu_i}^*(p_i)$, where the last equality follows from the assumption $\nu_0=\otimes_{i\in s}\nu_i$. We conclude in the same way as for Theorem 1.2.3.

4.5.3 Proof of Proposition 4.3.2

This is a classic convex optimization problem. Let $\mathcal{L}$ be the Lagrangian associated to the problem:

$$\mathcal{L}(\lambda,w)=\sum_{i\in s}\Lambda_{\nu_i}^*(w_i\pi_i)-\lambda^t\left(\sum_{i\in s}w_i\mathbf{x}_i-Nt_{\mathbf{x}}\right)$$

where $\lambda\in\mathbb{R}^k$ is the Lagrange multiplier. The solutions to the first order conditions satisfy for all $i\in s$,

$$w_i=d_i{\Lambda_{\nu_i}^*}'^{-1}(\lambda^td_i\mathbf{x}_i),$$

where we recall that the functions $\Lambda_{\nu_i}^*$ are assumed to be strictly convex, so that ${\Lambda_{\nu_i}^*}'^{-1}$ exists for all i, and is equal to Λ_{ν_i}'. Now it suffices to apply the solutions of the first order conditions to the constraint to obtain an expression of the solution $\hat{\lambda}$:

$$\frac{1}{N}\sum_{i\in s}d_i\Lambda_{\nu_i}'(\hat{\lambda}^td_i\mathbf{x}_i)\mathbf{x}_i-t_{\mathbf{x}}=0\Longleftrightarrow\hat{\lambda}=\arg\min_{\lambda\in\mathbb{R}^k}\sum_{i\in s}\Lambda_{\nu_i}(\lambda^td_i\mathbf{x}_i)-\lambda^tt_{\mathbf{x}}.$$

The equivalence is justified by the fact that Λ_{ν_i} is strictly convex, and therefore, so is $\lambda \mapsto \sum_{i\in s} \Lambda_{\nu_i}(\lambda^t d_i \mathbf{x}_i) - \lambda^t t_{\mathbf{x}}$. For that reason, $\hat{\lambda}$ is uniquely defined. We finally obtain an expression of the calibrated weights

$$\forall i \in s,\ \hat{w}_i = d_i \Lambda'_{\nu_i}(\hat{\lambda}^t d_i \mathbf{x}_i).$$

4.5.4 Proof of Proposition 4.3.3

Let $F : \lambda \mapsto \frac{1}{N}\sum_{i\in s} d_i f_i(\lambda)\mathbf{x}_i$, and $G : \lambda \mapsto \frac{1}{N}\sum_{i\in s} d_i g_i(\lambda)\mathbf{x}_i$. We call respectively $\hat{\lambda}$ and $\tilde{\lambda}$ the solutions to $F(\lambda) = t_{\mathbf{x}}$ and $G(\lambda) = t_{\mathbf{x}}$. We have

$$F(\hat{\lambda}) = F(0) + X_n\hat{\lambda} + o(\|\hat{\lambda}\|)$$

and then $(t_{\mathbf{x}} - \hat{t}_{\mathbf{x}}^{HT}) = X_n\hat{\lambda} + o(\|\hat{\lambda}\|)$. By assumption, X_n is invertible for large values of n since it converges toward an invertible matrix X. Thus, whenever $\hat{t}_{\mathbf{x}}^{HT}$ is close enough to $t_{\mathbf{x}}$, there exists λ_0 in a neighborhood of 0 such that $F(\lambda_0) = t_{\mathbf{x}}$. By uniqueness of the solution, we conclude that $\lambda_0 = \hat{\lambda}$. Hence, for large values of n,

$$\hat{\lambda} = X_n^{-1}(t_{\mathbf{x}} - \hat{t}_{\mathbf{x}}^{HT}) + o\left(\frac{1}{\sqrt{n}}\right).$$

A similar reasoning for $\tilde{\lambda}$ yields $\|\tilde{\lambda} - \hat{\lambda}\| = o(n^{-1/2})$. Thus, $\hat{\lambda}$ and $\tilde{\lambda}$ converge toward 0 and by Taylor formula:

$$f_i(\hat{\lambda}) = 1 + \mathbf{z}_i^t\hat{\lambda} + o\left(\frac{1}{\sqrt{n}}\right) = 1 + \mathbf{z}_i^t\tilde{\lambda} + o\left(\frac{1}{\sqrt{n}}\right) = g_i(\tilde{\lambda}) + o\left(\frac{1}{\sqrt{n}}\right).$$

It follows that $\hat{t}_y$ and $\tilde{t}_y$ are asymptotically equivalent.

We know that MEM estimation reduces to taking $f_i(.) = \Lambda'_{\nu_i}(d_i\mathbf{x}_i^t.)$ in a GC procedure. Hence, in that case, $\nabla f_i(0) = d_i\Lambda''_{\nu_i}(0)\mathbf{x}_i$. Since the variance of a probability measure ν_i is given by $\Lambda''_{\nu_i}(0)$, two MEM estimators with prior distributions having the same respective variances are asymptotically equivalent. Furthermore, a Gaussian prior $\nu_i \sim \mathcal{N}(1, q_i\pi_i)$ has a log-Laplace transform $\Lambda_{\nu_i} : t \mapsto \pi_i q_i t^2/2 + t$ which corresponds to $f_i(\lambda) = \Lambda'_{\nu_i}(d_i\mathbf{x}_i^t\lambda) = 1 + q_i\mathbf{x}_i^t\lambda$. The resulting MEM estimator is thus the instrument estimator obtained with instruments $\mathbf{z}_i = q_i\mathbf{x}_i, i \in s$.

4.5.5 Proof of Theorem 4.4.2

We decompose the AMEM estimator as follow

$$\hat{t}_y(\Phi_m) = \hat{t}_y^{HT} + (t_\Phi - \hat{t}_\Phi^{HT}) + (\hat{t}_\Phi^{HT} - t_\Phi - (\hat{t}_{\Phi_m}^{HT} - t_{\Phi_m})) + (\hat{B}_{\Phi_m} - 1)(t_{\Phi_m} - \hat{t}_{\Phi_m}^{HT}).$$

We have by assumption

$$\frac{n}{N^2}\ \mathbb{E}\ \mathbb{E}_p(\hat{t}_\Phi^{HT} - t_\Phi - (\hat{t}_{\Phi_m}^{HT} - t_{\Phi_m}))^2 \to 0 \ \text{ and } \ (\hat{B}_{\Phi_m} - 1) \to 0$$

Therefore, the terms $(\hat{t}_{\Phi-\Phi_m}^{HT} - t_{\Phi-\Phi_m})$ and $(\hat{B}_{\Phi_m} - 1)(t_{\Phi_m} - \hat{t}_{\Phi_m}^{HT})$ are asymptotically negligible in comparison to $(t_\Phi - \hat{t}_\Phi^{HT})$ as $m \to \infty$. We conclude using Lemma 4.5.1.

Chapter 5

Threshold regularization of inverse problems

A number of regularization methods for discrete inverse problems consist in considering weighted versions of the usual least square solution. These filter methods are generally restricted to monotonic transformations, e.g. the Tikhonov regularization or the spectral cut-off. In this paper, we point out that in several cases, non-monotonic sequences of filters may appear more appropriate. We study a regularization method that naturally extends the spectral cut-off procedure to non-monotonic sequences and provide several oracle inequalities, showing the method to be nearly optimal under mild assumptions. We extend the method to inverse problems with noisy operator and provide efficiency results in a conditional framework.

5.1 Introduction

We are interested in recovering an unobservable signal x_0, based on noisy observations of the image of x_0 through a linear operator A. The observation y satisfies the following relation

$$y(t) = Ax_0(t) + \varepsilon(t),$$

where $\varepsilon(.)$ is a random process representing the noise. This problem is studied in [CGPT00], [HO93], [Lou08] and in many applied fields such as medical imaging in [Nat01] or seismography in [SG88] for instance. When the measured signal is only available at a finite number of points $t_1, ..., t_n$, the operator A must be replaced by a discrete version $A_n : x \mapsto (Ax(t_1), ..., Ax(t_n))^t$, leading to a discrete linear model

$$y = A_n x_0 + \varepsilon,$$

with $y \in \mathbb{R}^n$. Difficulties in estimating x_0 occur when the problem is *ill-posed*, in the sense that small perturbations in the observations induce large changes in the solution. This is caused by an ill-conditioning of the operator A_n, reflected by a fast decay of its spectral values b_i. In such problems, the least square solution, although having a small bias, is generally inefficient due to a too large variance. Hence, *regularization* of the problem is required to improve the estimation. A large number of regularization methods are based on considering weighted versions of the least square estimator. The idea is to allocate low weights λ_i, or *filters*, to the least square coefficients that are highly contaminated with noise, thus reducing the variance, at the cost of increasing the bias at the same time. The most famous filter-based method is arguably the one due to Tikhonov (see [TA77]), where a collection of filters is indirectly obtained via a minimization procedure with ℓ^2 penalization. Tikhonov filters are entirely determined by a parameter τ that controls the balance between the minimization of the ℓ^2 norm of the estimator and the residual. Another well spread filter method that will be given a particular attention, is the *spectral cut-off* discussed in [BHMR07], [EHN96] and [Han87]. One simply considers a truncated version of the least square solution, where all coefficients corresponding to arbitrarily small eigenvalues are removed. Thus, spectral cut-off is associated to binary filters λ_i, equal to 1 if the corresponding eigenvalue b_i exceeds in absolute value a certain threshold τ, and 0 otherwise.

A common feature of spectral cut-off and Tikhonov regularization is the predetermined nature of the filters λ_i, defined in each case as a fixed non-decreasing function $f(\tau, .)$ of the eigenvalues b_i^2, and where only the parameter τ is allowed to depend on the observations. However, in many situations, non-monotonic sequences of filters may seem to be more appropriate. Actually, optimal values for λ_i generally depend on both the noise level, which is determined by the eigenvalue b_i, and the component, say x_i, of x_0 in the direction associated to b_i. A restriction to monotonic collections of filters may turn out to be inefficient in situations where the coefficients x_i are uncorrelated to the spectral values b_i of the operator A_n.

Regularization methods involving more general classes of filters have also been treated in the literature. For example, the *unbiased risk estimation* (URE) introduced by Stein in [Ste81] and studied in this context in [CGPT00], applies to arbitrary classes of filters, dealing in particular with non-monotonic collections. However, this approach has proven inefficient in low regularity cases. More recently, the *risk hull method*, discussed in [CG06], is shown to be an improvement of URE and this is confirmed by simulation studies. Here, we focus on a specific class of projection estimators that extends the spectral cut-off to non-monotonic collections of filters. Precisely, we consider the collection of unrestricted binary filters $\lambda_i \in \{0, 1\}$, known as *projection filters.* The computation of the estimator relies on the choice of a proper set of coefficients $m \subseteq \{1, ..., n\}$, which increases the number of possibilities compared to the spectral cut-off. We show this method to satisfy a non-asymptotic oracle inequality, when the oracle is computed in the class of projection filters. Moreover, we show our estimator to nearly achieve the rate of convergence of the best linear estimator in the maximal class of filters, i.e. when no restriction is made on λ_i.

It many actual situations, the operator A_n is not known precisely and only an approximation of it is available. Regularization of inverse problems with approximate operator is studied in [CH05], [EK01] and [HR08]. In this paper, we tackle the problem of estimating x_0 in the situation where we observe independently a noisy version $\hat{b}_i$ of each eigenvalue b_i. We consider a framework where the observations $\hat{b}_i$ are made once and for all, and are thus seen as non-random. We provide a bound on the conditional risk of the estimator, given the values of $\hat{b}_i$, in the form of a conditional oracle inequality.

The chapter is organized as follows. We introduce the problem in Section 5.2. We define our estimator in Section 5.3, and provide two kinds of oracle inequalities and numerical applications. Section 5.4 is devoted to an application of the method to inverse problems with noisy operators.

The proofs of our results are postponed to Section 5.5.

5.2 Problem setting

Let $(\mathcal{X}, \|.\|)$ be a Hilbert space and $A_n : \mathcal{X} \to \mathbb{R}^n$ $(n > 2)$ a linear operator. We want to recover an unknown signal $x_0 \in \mathcal{X}$ based on the indirect observations

$$y = A_n x_0 + \varepsilon, \tag{5.1}$$

where ε is a random noise vector. We assume that ε is centered with covariance matrix $\sigma^2 I$, where I denotes the identity matrix. We endow $\mathbb{R}^n$ with the scalar product $\langle u, v\rangle_n = \frac{1}{n}\sum_{i=1}^n u_i v_i$ and the associated norm $\|.\|_n$ and we note $A_n^* : \mathbb{R}^n \to \mathcal{X}$ the adjoint of A_n. Let $\mathcal{K}_n$ be the kernel of A_n and $\mathcal{K}_n^\perp$ its orthogonal in $\mathcal{X}$ which we assume to be of dimension n. The fact that A_n is surjective ensures that the observation y provides information in all directions. If this condition is not met, one may simply reduce the dimension of the image in order to make A_n surjective.

The efficiency of the estimator relies first of all on the accuracy of the discrete operator A_n and how "close" it is to the true value A. The convergence of the estimator toward x_0 is subject to the condition that the distance of x_0 to the set $\mathcal{K}_n^\perp$ tends to 0, which is reflected by a proper asymptotic behavior of the design $t_1, ..., t_n$. This aspect is not discussed here, we consider a framework where we have no control over the design $t_1, ..., t_n$ and we focus on the convergence of the estimator toward the projection $x^\dagger$.

Let $\{b_i; \phi_i, \psi_i\}_{i=1,...,n}$ be a singular system for the linear operator A_n, that is, $A_n\phi_i = b_i\psi_i$ and $A_n^*\psi_i = b_i\phi_i$ and $b_1^2 \geq ... \geq b_n^2 > 0$ are the ordered non-zero eigenvalues of the self-adjoint operator $A_n^*A_n$. The ϕ_i's (resp. ψ_i's) form an orthonormal system of $\mathcal{K}_n^\perp$ (resp. $\mathbb{R}^n$).

In this framework, the available information on x_0 consists in a noisy version of $A_n x_0$. As a result, estimating the part of x_0 lying in $\mathcal{K}_n$ is impossible, based only on the observations. The best approximation of x_0 one can get without prior information is the orthogonal projection of x_0 onto $\mathcal{K}_n^\perp$. This projection, noted $x^\dagger$, is called *best approximate solution* and is obtained as the image of $A_n x_0$ through the generalized Moore-Penrose inverse operator $A_n^\dagger = (A_n^*A_n)^\dagger A_n^*$, where $(A_n^*A_n)^\dagger$ denotes the inverse of $A_n^*A_n$, restricted to $\mathcal{K}_n^\perp$. By construction, the generalized Moore-Penrose inverse $A_n^\dagger$ can also be defined as the operator for which $\{b_i^{-1}; \psi_i, \phi_i\}_{i=1,...,n}$ is a singular system. We refer to [EHN96] for further details.

Searching for a solution in the subspace $\mathcal{K}_n^\perp$ allows to reduce the number of regressors to n. Then, estimating $x^\dagger$ can be made using a classical linear regression framework where the number of regressors is equal to the dimension of the observation. Decomposing the observation in the singular basis $\{\psi_i\}_{i=1,...,n}$ leads to the following model

$$y_i = b_i x_i + \varepsilon_i, i = 1, ..., n,$$

where we set $y_i = \langle y, \psi_i\rangle_n$, $\varepsilon_i = \langle \varepsilon, \psi_i\rangle_n$ and $x_i = \langle x_0, \phi_i\rangle$. It now suffices to divide each term by the known singular value b_i to observe the coefficient x_i, up to a noise term $\eta_i := b_i^{-1}\varepsilon_i$. Equivalently, this is obtained by applying the Moore-Penrose inverse $A_n^\dagger$ in the model (5.1). We thus consider the function $y^\dagger = A_n^\dagger y \in \mathcal{K}_n^\perp$, defined as the inverse image of y through A_n

with minimal norm. Identifying $y^\dagger$ with the vector of its coefficients $y_i^\dagger = b_i^{-1} y_i$ in the basis $\{\phi_i\}_{i=1,...,n}$, we obtain

$$y_i^\dagger = x_i + \eta_i, \ i = 1, ..., n. \tag{5.2}$$

The covariance matrix of the noise $\eta = (\eta_1, ..., \eta_n)^t$ is diagonal in this model, as we have $\mathbb{E}(\eta_i \eta_j) = \frac{\sigma^2}{n} b_i^{-1} b_j^{-1} \langle \psi_i, \psi_j \rangle_n$ which is null for all $i \neq j$ and equal to $\sigma_i^2 := \frac{\sigma^2}{n} b_i^{-2}$ if $i = j$. The model can be somehow interpreted as a linear regression model with heteroscedastic noises, the variances σ_i^2 being inversely proportional to the eigenvalues b_i^2. In the case where ε in the original model (5.1) is Gaussian with distribution $\mathcal{N}(0, \sigma^2 I)$, the noises η_i remain Gaussian in (5.2).

This representation points out the effect of the decay of the singular values b_i on the noise level, making the problem ill-posed. To control the noise with a too large variance σ_i^2, a solution is to consider weighted versions of $y^\dagger$. For some filter $\lambda = (\lambda_1, ..., \lambda_n)^t$, note $\hat{x}(\lambda) \in \mathcal{K}_n^\perp$ the function defined by $\langle \hat{x}(\lambda), \phi_i \rangle = \lambda_i y_i^\dagger$ for $i = 1, ..., n$. Filter-based methods aim to cancel out the high frequency noises by allocating low weights to the components $y_i^\dagger$ corresponding to small singular values. A widely used example is the Tikhonov regularization, with weights of the form $\lambda_i = (1 + \tau \sigma_i^2)^{-1}$ for some $\tau > 0$. The Tikhonov solution can be expressed as the minimizer of the functional

$$\|y - A_n x\|^2 + \frac{\tau \sigma^2}{n} \|x\|^2, \ x \in \mathcal{X},$$

which makes the method particularly convenient in cases where the SVD of $A_n^* A_n$ or the coefficients $y_i^\dagger$ are not easily computable. We refer to [Cav08] and [TA77] for further details.

Another common filter-based method is the *truncated singular value decomposition* or *spectral cut-off* studied in [BHMR07], [EHN96] and [Han87]. An estimator of x_0 is obtained as a truncated version of $y^\dagger$, where all coefficient $y_i^\dagger$ corresponding to arbitrarily small singular values are replaced by 0. This approach can be viewed as a principal component analysis, where only the highly explanatory directions are selected. The spectral cut-off estimator is associated to filter factors of the form $\lambda_i = \mathbb{1}\{i \leq k\}$, where $\mathbb{1}\{.\}$ denotes the indicator function and k is a bandwidth to be determined. Data-driven methods for selecting suitable values of k are discussed in [Cav08], [CG06], [Han87], [Var73] and [Var79].

A natural way to generalize the spectral cut-off procedure is to enlarge the class of estimators by considering non-ordered truncated versions of $y^\dagger$, as made in [Lou08], [LL08] or [LL10] (see also Examples 1 and 2 in [CGPT00]). This approach reduces to a model selection issue where each model is identified with a set of indices $m \subseteq \{1, ..., n\}$. Precisely, for m a given model, define $\hat{x}_m \in \mathcal{K}_n^\perp$ as the orthogonal projection of $y^\dagger$ onto $\mathcal{X}_m := \text{span}\{\phi_i, i \in m\}$, that is, $\hat{x}_m$ satisfies

$$\langle \hat{x}_m, \phi_i \rangle = \begin{cases} y_i^\dagger & \text{if } \ i \in m, \\ 0 & \text{otherwise.} \end{cases}$$

The objective is to find a model m that makes the expected risk $\mathbb{E}\|\hat{x}_m - x_0\|^2$ small. The computation of the estimator no longer relies on the choice of one parameter $k \in \{1, ..., n\}$ as for spectral cut-off, but on the choice of a set of indices $m \subseteq \{1, ..., n\}$, which increases the number of possibilities. In particular, this approach allows non-monotonic collections of filters that may perform better than decreasing sequences obtained by spectral cut-off. To see this, write the bias-variance decomposition of the estimator $\hat{x}_m$ for a deterministic model m:

$$\mathbb{E}\|\hat{x}_m - x_0\|^2 = \mathbb{E}\|x_0 - x^\dagger\|^2 + \sum_{i \notin m} x_i^2 + \sum_{i \in m} \sigma_i^2.$$

In these settings, it appears that in order to minimize the risk, best is to select indices i for which the component x_i^2 is larger than the noise level σ_i^2. A proper choice of filter should depend on both the variance σ_i^2 and the coefficient x_i^2. Consequently, the resulting sequence $\{\lambda_i\}_{i=1,\ldots,n}$ has no reason of being a decreasing function of σ_i^2 if some coefficients x_i^2 are large enough to compensate for a large variance.

5.3 Threshold regularization

The construction of the projection estimator reduces to finding a proper set m. An optimal value for m (minimizing the risk) is obtained by keeping small simultaneously the bias term $\sum_{i \notin m} x_i^2$ and the variance term $\sum_{i \in m} \sigma_i^2$ in the expression of the risk $\mathbb{E}\|\hat{x}_m - x_0\|^2$. Following this argument, a minimizer of the risk $\mathbb{E}\|\hat{x}_m - x_0\|^2$ is obtained by selecting only the indices i for which the coefficient x_i^2 is larger than the noise level σ_i^2. An optimal model is thus given by $m^* := \left\{i : \; x_i^2 \geq \sigma_i^2\right\}$. The coefficients x_i being unknown to the practitioner, the optimal set m^* can not be computed in practical cases. For this reason it is referred to as an *oracle*.

We shall now provide a model $\widehat{m}$ constructed from the available information, that mimics the oracle m^*. Fixing a threshold on the coefficients x_i being impossible, we propose to use a threshold on the coefficients $y_i^\dagger$. Precisely, consider the set

$$\widehat{m} = \left\{i : y_i^{\dagger 2} \geq 4\sigma_i^2 \mu_i\right\} = \left\{i : y_i^2 \geq \frac{4\sigma^2 \mu_i}{n}\right\},$$

for $\{\mu_i\}_{i=1,\ldots,n}$ a sequence of positive parameters to be chosen and where we recall that $y_i = b_i y_i^\dagger$. Obviously, the behavior of the resulting estimator $\hat{x}_{\widehat{m}}$ relies on the choice of the sequence $\{\mu_i\}_{i=1,\ldots,n}$: the larger the μ_i's, the more sparse is $\hat{x}_{\widehat{m}}$. It must be chosen so that the resulting set $\widehat{m}$ contains only the indices i for which the noise level is small compared to the actual value of x_i, but the only knowledge of the observations $y_i^\dagger$ and the variances σ_i^2 makes it a difficult task.

A number of thresholding procedures have been studied in the inverse problem literature. In [LL10], Loubes proposes a ℓ^1 penalization procedure to the inverse problem, corresponding to a soft-thresholding approach with a threshold on y_i^2 of the order $c \, \frac{\log n}{n} \, \sigma^2$, for some $c > 0$. In [AS98], Abramovich and Silverman discuss an approach based on the decomposition of the observation in a wavelet basis, for which the coefficients can be selected via a thresholding criterion. Here again, a threshold of the order $c \, \frac{\log n}{n} \, \sigma^2$ is suggested. For these two approaches, the threshold is a linear function of the variance, which with our notations, corresponds to taking a parameter $\mu_i = c \log n$ that does not depend on i. In Theorem 5.3.1, we discuss the accuracy of a non-linear threshold that involves a logarithmic term of the variance.

5.3.1 Oracle inequalities

In the definition of $\widehat{m}$, the choice of the parameters μ_i is crucial. Too large values of μ_i will result in an under-adjustment, keeping too few relevant components $y_i^\dagger$ to estimate x_0. On the contrary, a small value of μ_i increases the probability of selecting a component $y_i^\dagger$ that is highly affected with noise. Thus, it is essential to find a good balance between these two types of errors. In the next theorem, we provide a nearly optimal choice for the parameters μ_i, under the condition that ε has finite exponential moments.

For $i = 1, ..., n$, note $\gamma_i := \eta_i^2/\sigma_i^2 = n\varepsilon_i^2/\sigma^2$. We make the following assumption.

A5.1. There exist $K, \beta > 0$ such that $\forall t > 0, \forall i = 1, ..., n,\ \mathbb{P}(\gamma_i > t) \leq Ke^{-t/\beta}$.

In a Gaussian model, the γ_i's have χ^2 distribution with one degree of freedom. The condition A5.1 holds for any $\beta > 2$, taking $K = \sqrt{1 - 2/\beta}$.

Theorem 5.3.1 *Assume that* A5.1 *holds. For some* $\theta > 0$*, set* $\mu_i = \beta \log(e + \theta\sigma_i^2)$*. The estimator* $\hat{x}_{\widehat{m}}$ *satisfies*

$$\mathbb{E}\|\hat{x}_{\widehat{m}} - x^\dagger\|^2 \leq \mathbb{E}\|\hat{x}_{m^*} - x^\dagger\|^2 + \Big(6\beta \log(e + \theta\|x^\dagger\|^2) + 2\Big) \sum_{i \in m^*} \sigma_i^2 + \frac{2K\beta n}{\theta}.$$

Remark 1 This theorem establishes a non-asymptotic oracle inequality. The first residual term is comparable to that in Corollary 1 in [Lou08] for $\theta \sim n^2$. The second residual term can be controlled by an appropriate choice of θ. While choosing $\theta \sim n^2$ is sufficient to make this term negligible, a more accurate value may be chosen depending on the known sequence $\{\sigma_i^2\}_{i=1,...,n}$, in order to find a good balance between the two residual terms, making the inequality as sharp as possible.

Remark 2 The method requires knowledge of the operator A_n, the variance σ^2 and the constant β in the condition A5.1. Note however that knowing the constant K is not necessary to build the estimator.

Remark 3 In an asymptotic concern, the accuracy of the result stated in Theorem 5.3.1 relies on the convergence rate of the residual term to zero, compared to the risk of the oracle. The residual term $\sum_{i \in m^*} \sigma_i^2$ is actually the variance term in the bias-variance decomposition of $\hat{x}_{m^*}$, and therefore, it is bounded by the risk of the oracle. As a result, for θ of the order n, the estimator $\hat{x}_{\widehat{m}}$ is shown to reach at least the same rate of convergence as the oracle up to a logarithmic term, which warrants good adaptivity properties. The logarithmic term vanishes in the convergence rate if the bias term $\sum_{i \notin m^*} x_i^2$ dominates in the risk of the $\hat{x}_{m^*}$. Precisely, the oracle inequality is asymptotically exact as soon as the residual term $\log n \sum_{i \in m^*} \sigma_i^2$ is negligible compared to the bias term $\sum_{i \notin m^*} x_i^2$. In this case, it follows from Theorem 5.3.1 that

$$\mathbb{E}\|\hat{x}_{\widehat{m}} - x^\dagger\|^2 = (1 + o(1))\ \mathbb{E}\|\hat{x}_{m^*} - x^\dagger\|^2.$$

Of course, this condition is hard to verify in practice and assuming it is true reduces to make strong regularity assumptions on the asymptotic behavior of x_0 and A_n.

The estimator $\hat{x}_{\widehat{m}}$ being built using binary filters $\lambda_i \in \{0,1\}$, it is natural to measure its efficiency by comparing its risk to that of the best linear estimator in this class. Nevertheless, we see in the next corollary that a similar oracle inequality holds if we consider the oracle in the maximal class of filters, that is, allowing the λ_i's to take any real value.

Corollary 5.3.2 *Assume that the condition* A5.1 *holds, the estimator $\hat{x}_{\widehat{m}}$ of Theorem 5.3.1 satisfies*

$$\mathbb{E}\|\hat{x}_{\widehat{m}} - x^\dagger\|^2 \leq \Big(12\beta \log(e + \theta\|x^\dagger\|^2) + 4\Big) \inf_{\lambda \in \mathbb{R}^n} \mathbb{E}\|\hat{x}(\lambda) - x^\dagger\|^2 + \frac{2K\beta n}{\theta}.$$

This result is a straightforward consequence of Lemma 5.5.2 in the Appendix, where it is shown that the oracle in the class of binary filters $\lambda_i \in \{0,1\}$ achieves the same rate of convergence up to a factor 2, as the best filter estimator obtained with non-random values of λ. This results points out that the class of unrestricted binary filters only induces a slight loss of efficiency compared to the maximal class.

5.3.2 Rates of convergence and adaptation

Interest of oracles lies in the fact that the best estimator in a given class will often achieve the optimal rate of convergence. In many situations, comparing the risk of the estimator to that of an oracle might be sufficient to deduce optimality results, as well as adaptivity properties. In the literature of inverse problems, rates of convergence of oracles are obtained under regularity conditions on the map x_0 and the spectrum of A_n. In the literature, examples of commonly studied regularity spaces for x_0 are Sobolev classes of functions

$$x_0 \in \mathcal{S}(s,L) = \left\{ x \in \mathcal{X} : \sum_{i=1}^{n} \langle x, \phi_i \rangle^2 \ i^{2s} \leq L \right\}, \ s, L > 0,$$

and analytical functions

$$x_0 \in \mathcal{A}(s,L) = \left\{ x \in \mathcal{X} : \sum_{i=1}^{n} \langle x, \phi_i \rangle^2 \ e^{2is} \leq L \right\}, \ s, L > 0.$$

We refer for instance to [Cav08] and [LR09a]. To calculate rates of convergence, we also need to take into consideration the nature of ill-posedness of the inverse problem. For instance, we say the problem is mildly ill-posed if there exist constants $0 < c < C$ and $t > 0$ such that $c\, i^{-t} \leq |b_i| \leq C\, i^{-t}$ and severely ill-posed if $c\, e^{-it} \leq |b_i| \leq C\, e^{-it}$.

In this setting, one may be interested in calculating optimal rates of convergences of estimators. A way of defining these optimal rates is to consider the risk of the best possible estimator when the true function is the hardest to estimate in a given class $\mathcal{C} \subset \mathcal{X}$. Precisely, denote by X the set of estimators of x_0 (i.e. the set of measurable functions of y), the so-called *minimax* risk knowing that $x_0 \in \mathcal{C}$, is given by

$$\mathcal{R}^*(\mathcal{C}) = \inf_{\hat{x} \in X} \sup_{x \in \mathcal{C}} \mathbb{E}\|\hat{x} - x\|^2.$$

For the Sobolev and analytical classes of functions, the minimax rates of convergence are given by

$$\mathcal{R}^*(\mathcal{S}(s,L)) = O\left(n^{-\frac{2s}{2s+2t+1}}\right) \quad \text{and} \quad \mathcal{R}^*(\mathcal{A}(s,L)) = O\left(\frac{(\log n)^{2t+1}}{n}\right)$$

for mildly ill-posed problems and

$$\mathcal{R}^*(\mathcal{S}(s,L)) = O\left((\log n)^{-2s}\right) \quad \text{and} \quad \mathcal{R}^*(\mathcal{A}(s,L)) = O\left(n^{-\frac{2s}{2s+2t}}\right)$$

for severely ill-posed problems (see [Cav08]). In general, the regularity of the true function x_0 is unknown. Therefore, a desirable property for an estimator is that it achieves minimax rates of convergence for several classes of functions.

It is also possible to gather the regularity conditions on x_0 and A_n into a single *source condition*, relating the behavior of x_0 to the regularity of the operator A_n (see for instance [BHMR07], [CR07], [EHN96] or [FLn08]). Here is a simple example for which we deduce the rate of convergence of the estimator under a polynomial source condition.

Proposition 5.3.3 (Polynomial source condition) *Assume there exists* $\delta \in (0;2)$ *such that* $\sum_{i=1}^n |x_i|^\delta |b_i|^{\delta-2} = O(1)$, *then the estimator* $\hat{x}_{\widehat{m}}$ *obtained for* $\theta = n^2$ *satisfies*

$$\mathbb{E}\|\hat{x}_{\widehat{m}} - x^\dagger\|^2 = O\left(n^{\frac{\delta-2}{2}}.\log n\right).$$

For mildly ill-posed inverse problems with $|b_i| \sim i^{-t}$, it is pointed out in [Lou08] that this rate of convergence corresponds to the minimax rate in the Sobolev class $\mathcal{S}(s,L)$ if $\frac{1}{\delta} = \frac{1}{2} + \frac{s}{2t+1}$ up to a $\log n$ factor. Indeed, we have in this case

$$\frac{\delta-2}{2} = \frac{2t+1}{2t+2s+1} - 1 = -\frac{2s}{2t+2s+1}.$$

For more discussion on rates of convergence, we refer to [Cav08], [DJ98] and [LR09a].

5.3.3 Comparison with unbiased risk estimation and risk hull method

In a general point of view, the estimator $\hat{x}_{\widehat{m}}$ can be obtained via a minimization procedure, using a BIC-type criterion for heteroscedastic models,

$$\hat{x}_{\widehat{m}} = \arg\min_{x \in \mathcal{X}} \left\{ \|y^\dagger - x\|^2 + 4\sum_{i=1}^n \sigma_i^2 \mu_i \mathbb{1}\{\langle x, \phi_i \rangle \neq 0\} \right\}.$$

However, expressing the estimator as the solution to a minimization problem does not ease the computation. The method requires in any case calculation of the SVD of $A_n^* A_n$ and the coefficients $y_i^\dagger$, which may be computationally expensive. On the other hand, the computation of the estimator is simple once the decomposition of $y^\dagger$ in the SVD of $A_n^* A_n$ is known, as it suffices to compare each coefficient $y_i^{\dagger 2}$ to the threshold $4\sigma_i^2 \mu_i$.

Let us compare our approach to some other methods. First, we discuss the *unbiased risk estimation* (URE) studied in [CGPT00] in the inverse problem framework. The method constructs an estimator of x_0 via the minimization of an unbiased estimate of the risk, over an arbitrary set Λ of filters. When restricted to the class of projection filters $\lambda_i \in \{0, 1\}$, unbiased risk estimation reduces to minimizing over the collection $\mathcal{M}$ of all subsets of $\{1, ..., n\}$, the criterion

$$m \mapsto \|y^\dagger - \hat{x}_m\|^2 + 2 \sum_{i \in m} \sigma_i^2.$$

The minimum is achieved for the set $m = \{i : y_i^{\dagger 2} \geq 2\sigma_i^2\}$, which corresponds to taking $\mu_i = 1/2$. This choice is shown to be asymptotically efficient in Proposition 2 in [CGPT00] under strong regularity conditions. However, it is pointed out that this threshold is too low whenever the inverse problem has a high degree of ill-posedness.

A good alternative to URE is the *risk hull method* (RHM) discussed in [CG06]. Rather than considering an unbiased estimate of the risk, the idea of RHM is to find a function $\ell(\lambda)$ that bounds the risk from above, uniformly over the class Λ of filters. So, let $\ell(.)$ be such that

$$\mathbb{E} \sup_{\lambda \in \Lambda} \left\{ \|\hat{x}(\lambda) - x^\dagger\|^2 - \ell(\lambda) \right\} \leq 0. \tag{5.3}$$

The estimator is then defined via the minimizer $\tilde{\lambda}$ of $\lambda \mapsto \ell(\lambda)$ over Λ. By the previous inequality, we obtain an upper bound of the risk by

$$\mathbb{E}\|\hat{x}(\tilde{\lambda}) - x^\dagger\|^2 \leq \mathbb{E}\ \ell(\tilde{\lambda}).$$

The *risk hull* ℓ has to be chosen as small as possible, while still satisfying (5.3), in order to obtain a sharp bound on the risk of the estimator $\hat{x}(\tilde{\lambda})$. An analytic form of the minimal risk hull may be difficult to obtain but it can be computed by Monte-Carlo. In the class of projection filters where all filter λ can be canonically identified with a model $m \subseteq \mathcal{M}$, the objective is to find $\ell : \mathcal{M} \to \mathbb{R}$ such that

$$\mathbb{E} \sup_{m \in \mathcal{M}} \left\{ \sum_{i \notin m} x_i^2 + \sum_{i \in m} \eta_i^2 - \ell(m) \right\} \leq 0.$$

Although it is not necessarily minimal, convenient is to consider a risk hull of the form $\ell(m) = \delta + \sum_{i \notin m} x_i^2 + \sum_{i \in m} c_i$, where $\delta \geq 0$ is a tolerance term and the c_i's are such that

$$\mathbb{E} \sup_{m \in \mathcal{M}} \left\{ \sum_{i \in m} (\eta_i^2 - c_i) \right\} = \sum_{i=1}^{n} \mathbb{E}\left[(\eta_i^2 - c_i) \mathbb{1}\{\eta_i^2 \geq c_i\} \right] \leq \delta,$$

in order to recover (5.3). Of course, the true coefficients x_i^2 are unknown, but they can be replaced by their unbiased estimates $y_i^{\dagger 2} - \sigma_i^2$, as suggested in [CG06]. Under A5.1, it appears that taking $c_i \sim c \log n\ \sigma_i^2$ yields a δ of the order $n^{-\alpha} \sum_{i=1}^n \sigma_i^2$ for some $\alpha \geq 0$. On the other hand, adding a term $\log \sigma_i^2$ in the expression of c_i enables to obtain a tolerance term δ that does not involve the variances σ_i^2 (see for instance the proof of Lemma 5.5.1), which somehow justifies the choice of the threshold used in Theorem 5.3.1.

5.3.4 Simulations

We shall now see numerical applications. We consider an heteroscedastic non-parametric regression model,

$$y_i = x_i + \sigma_i \varepsilon_i,\ i = 1, ..., n,$$

with $\varepsilon_i \sim \mathcal{N}(0, 1)$. This model illustrates the discrete inverse problem, where the observation is expressed via the singular value decomposition of the operator A_n. So, the x_i's stand for the coefficients of x_0 in the singular basis $\{\phi_i\}$, i.e. $x_i = \langle x_0, \phi_i \rangle$. The noises ε_i are independently drawn from a standard Gaussian distribution. The variance of the model is determined by the non-decreasing sequence $\{\sigma_i^2\}_{i=1,...,n}$ which reflects the decay of the spectrum of $A_n A_n^*$. For now, we do not need to specify the value of basis $\{\phi_i\}$, as it is not directly involved in the model. Consequently, the function of interest x_0 is not fully determined. Nevertheless, this framework covers several possible values for x_0, depending on the underlying value of the operator A_n. Graphical examples will be given in the sequel. For sake of objectivity, the coefficients x_i are randomly drawn from independent centered Gaussian variables $x_i \sim \mathcal{N}(0, v_i^2)$, with variances v_i^2 to be made precise later. The coefficients x_i are drawn once and for all and are treated as non-random, which means that the risk of an estimator $R(\widehat{x}) = \mathbb{E}\|\widehat{x} - x^\dagger\|^2$ is to be understood as an expectation conditionally to the x_i's.

The risk of $\widehat{x}_{\widehat{m}}$ is compared to that of the following oracles

- x_{m^*} is the optimal threshold estimator defined in Section 5.3.3, obtained with the filters $\lambda_i = \mathbb{1}\{x_i^2 \geq \sigma_i^2\}$.
- x_{sco}^* is the best spectral cut-off estimator, obtained with the filters $\lambda_i = \mathbb{1}\{i \leq k^*\}$, with optimal bandwidth k^*.
- x_{lin}^* is the best filter estimator, obtained with the filters $\lambda_i = x_i^2/(x_i^2 + \sigma_i^2)$.

We calculate the risk of the estimators by Monte Carlo with 10000 replications of the procedure, for different degrees of ill-posedness. We consider a well-posed problem with $\sigma_i^2 = 1/n$, mildly ill-posed inverse problems with a polynomial growth of the variances (here $\sigma_i^2 = i/n$ and $\sigma_i^2 = i^2/n$) and a severely ill-posed problem with σ_i^2 growing exponentially (here $\sigma_i^2 = 2^i/n$). The noises ε_i being Gaussian, we compute our estimator taking a small value of β satisfying the condition A5.1, namely $\beta = 2.1$. The risks $R(.)$ are given in the following table for the estimator and for the oracles. Next to the risk of $\widehat{x}_{\widehat{m}}$, is noted between brackets the ratio $R(\widehat{x}_{\widehat{m}})/R(x_{m^*})$. We consider the estimator obtained with two values of θ, namely $\theta = 1$ and $\theta = n$. In each case, we vary two aspects which are the sample size n and the value of the sequence $\{v_i^2\}_{i=1,...,n}$ reflecting the decay of the coefficients x_i.

Case 1. $n = 50$, $x_i^2 \sim 1$.

	$\sigma_i^2 = 1/n$	$\sigma_i^2 = i/n$	$\sigma_i^2 = i^2/n$	$\sigma_i^2 = 2^i/n$
$R(\widehat{x}_{\widehat{m}})$ $(\theta = 1)$	1.00 (×1.08)	34.2 (×2.07)	61.9 (×1.09)	58.6 (×1.01)
$R(\widehat{x}_{\widehat{m}})$ $(\theta = n)$	1.00 (×1.08)	50.2 (×3.02)	61.1 (×1.08)	59.8 (×1.04)
$R(x_{m^*})$	0.93	16.6	56.7	57.8
$R(x^*_{\text{sco}})$	1.00	24.4	59.0	57.8
$R(x^*_{\text{lin}})$	0.84	12.0	44.7	56.3

Here, the signal has roughly the same intensity along all directions, as the coefficients x_i are independently drawn from the same standard Gaussian distribution. As a result, the function is easy to calculate whenever the noise is small compared to the signal (e.g. the case $\sigma_i^2 = 1/n$), but it yields a high risk, even for the best possible estimators when the variance increases. Remark that the estimator $\widehat{x}_{\widehat{m}}$ performs especially well with a high degree of ill-posedness (e.g. the cases $\sigma_i^2 = i^2/n$ and $\sigma_i^2 = 2^i/n$), where only few components of the signal are tractable. Indeed, we see that the risk of $\widehat{x}_{\widehat{m}}$ is close to that of the oracles (with a ratio smaller than 1.1). It seems that the estimator obtained with $\theta = 1$ shows good adaptivity properties.

To illustrate these results, we consider the family of functions $\{\cos(k\pi.), \sin(k\pi.), k \in \mathbb{N}\}$, forming an orthogonal system on $\mathbb{L}^2([-1;1])$. We assume that the coefficients x_i are the decomposition of the signal in this basis. The well-posed situation is not problematic as the function can be easily estimated in this case. Here, we see an example of the mildly ill-posed case with $\sigma_i^2 = i/n$. We here compare the oracle on the left graphic to the estimator obtained with tuning parameter $\theta = 1$.

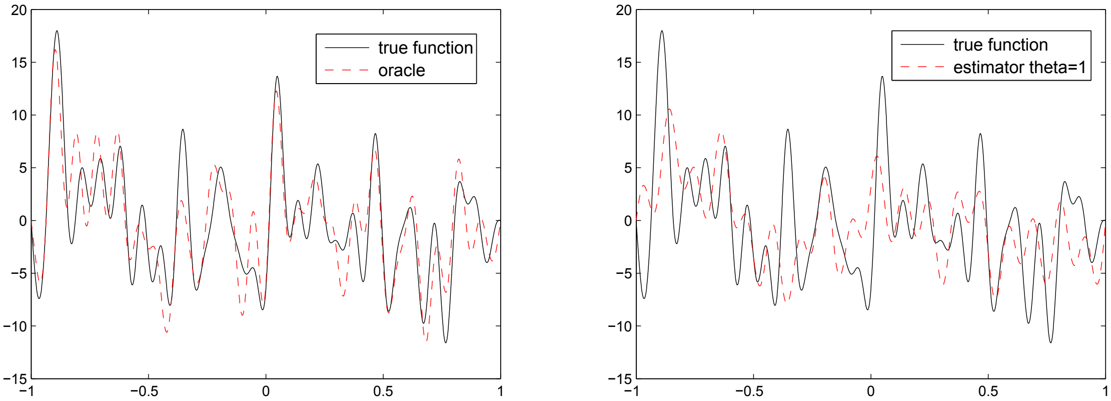

Figure 5.1: Mildly ill-posed problem $\sigma_i^2 = i/n$

Taking the value $\theta = 1$ (right graphic) clearly causes underfitting, although it is not too far from the oracle (left graphic). Naturally, the smoothness of the estimation increases with larger values of the parameter θ. The following graphic confirms that taking a larger tuning parameter (here $\theta = n$) is overly cautious.

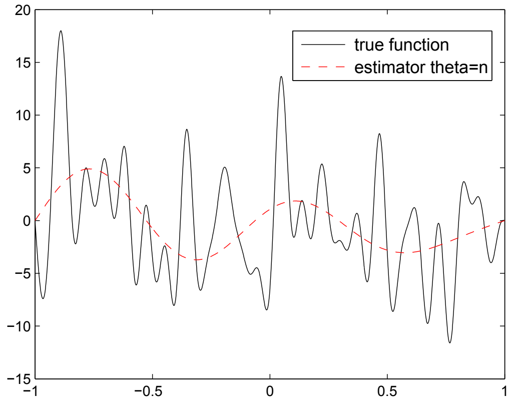

Figure 5.2: Mildly ill-posed problem $\sigma_i^2 = i/n$

With a too high degree of ill-posedness, a part of the signal can not be estimated. In particular, this situation where the coefficients x_i are roughly of the same order makes the signal hardly tractable when the noise level increases. As a result, even in mildly ill-posed problems, the oracle can be far from the true function. We now see the case $\sigma_i^2 = i^2/n$.

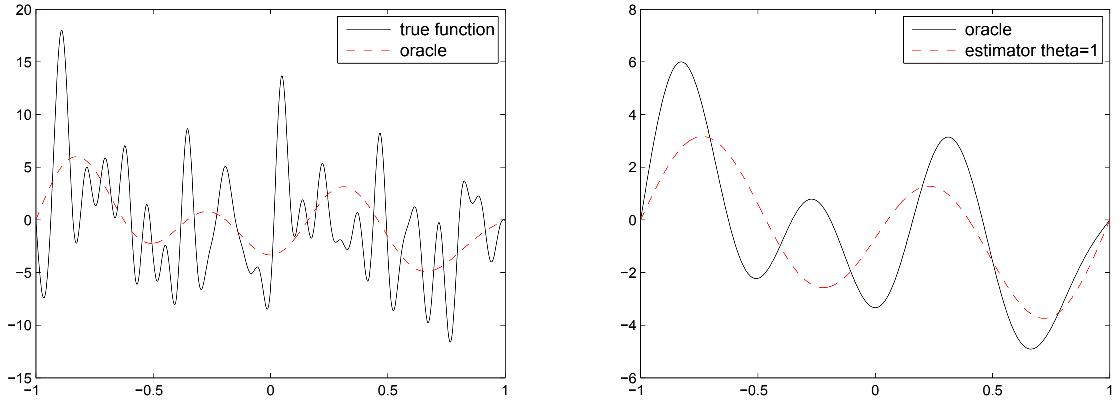

Figure 5.3: Mildly ill-posed problem $\sigma_i^2 = i^2/n$

Again, the estimator selects too few variables. Although, the difference with the oracle is negligible compared to the gap with the actual function.

Case 2. $n = 50$, $x_i^2 \sim i^{-1}$.

	$\sigma_i^2 = 1/n$	$\sigma_i^2 = i/n$	$\sigma_i^2 = i^2/n$	$\sigma_i^2 = 2^i/n$
$R(\widehat{x}_{\widehat{m}})$ $(\theta = 1)$	1.60 (×1.69)	26.9 (×1.98)	70.7 (×1.60)	55.4 (×1.02)
$R(\widehat{x}_{\widehat{m}})$ $(\theta = n)$	2.07 (×2.18)	39.6 (×2.91)	72.1 (×1.63)	56.6 (×1.05)
$R(x_{m^*})$	0.95	13.6	44.2	54.0
$R(x^*_{\text{sco}})$	1.00	18.7	48.5	54.0
$R(x^*_{\text{lin}})$	0.87	9.69	36.1	50.2

Setting the sequence v_i^2 of the order i^{-1} (precisely here $v_i^2 = 25/i$) causes an attenuation in the signal corresponding to a decreasing trend of $1/i$ in the coefficients x_i^2. As we see in this table, the method seems adapted to well-posed problems as well as severely ill-posed problems. The method seems less efficient when dealing with mildly ill-posed problems although the ratio between the risk of the estimator and that of the oracle remains satisfactory (less than 2 in the worst situation for the estimator with $\theta = 1$). It appears that the relative efficiency of the estimator is increased as the degree of ill-posedness grows.

The decay of the coefficients x_i, reflected by a decreasing sequence $\{v_i^2\}$ in the calculation of x_0 leads to a smoother function, which makes it more natural to estimate. The mildly ill-posed case with $\sigma_i^2 = i/n$ is shown in the two graphics below.

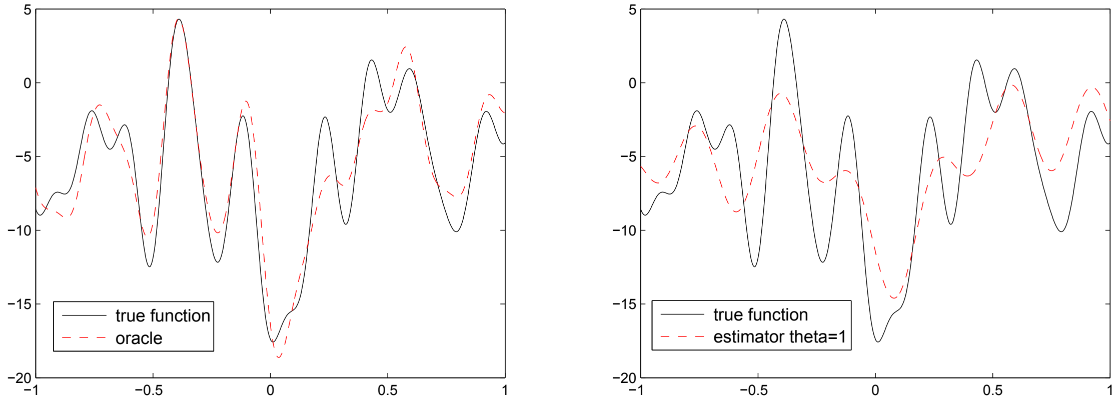

Figure 5.4: Mildly ill-posed problem $\sigma_i^2 = i/n$

With a higher degree of ill-posedness, only a small part of the signal is tractable by the oracle, which is efficiently recovered by the estimator. In the two following graphics the estimator $\widehat{x}_{\widehat{m}}$ is close to the oracle $\widehat{x}_{m^*}$.

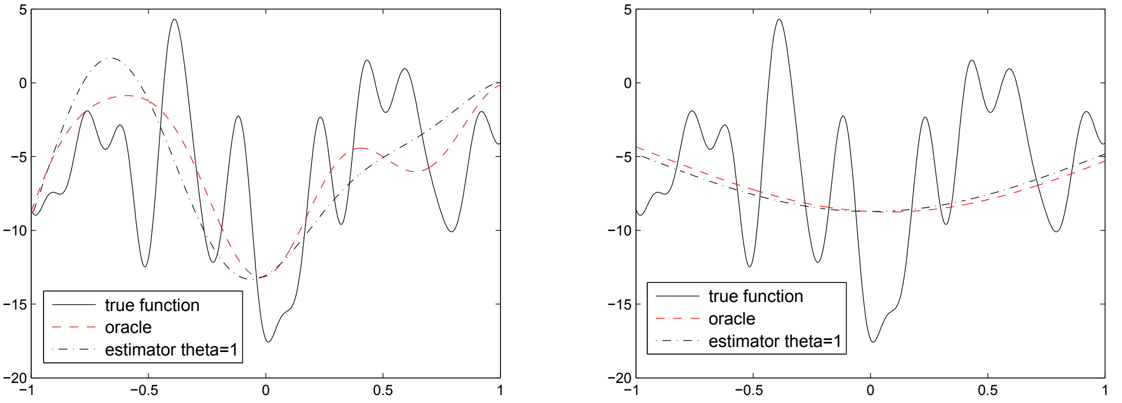

Figure 5.5: Ill-posed problems $\sigma_i^2 = i^2/n$ and $\sigma_i^2 = 2^i/n$

Case 3. $n = 50$, $x_i^2 \sim i^{-2}$.

	$\sigma_i^2 = 1/n$	$\sigma_i^2 = i/n$	$\sigma_i^2 = i^2/n$	$\sigma_i^2 = 2^i/n$
$R(\widehat{x}_{\widehat{m}})$ $(\theta = 1)$	1.96 (×2.38)	9.27 (×1.47)	30.4 (×1.94)	44.0 (×1.29)
$R(\widehat{x}_{\widehat{m}})$ $(\theta = n)$	2.32 (×2.82)	11.2 (×1.78)	46.9 (×2.98)	46.9 (×1.38)
$R(x_{m^*})$	0.82	6.29	15.7	34.1
$R(x^*_{\text{sco}})$	0.97	6.89	18.4	39.8
$R(x^*_{\text{lin}})$	0.71	4.50	13.0	26.8

Here the decrease of coefficients x_i is more important, due to smaller values of v_i^2, taken of the order $1/i^2$. As previously, we see in this table that the estimator performs well for severely ill-posed problems, as we observe a particularly small relative risk for $\sigma_i^2 = i^2/n$ and $\sigma_i^2 = 2^i/n$.

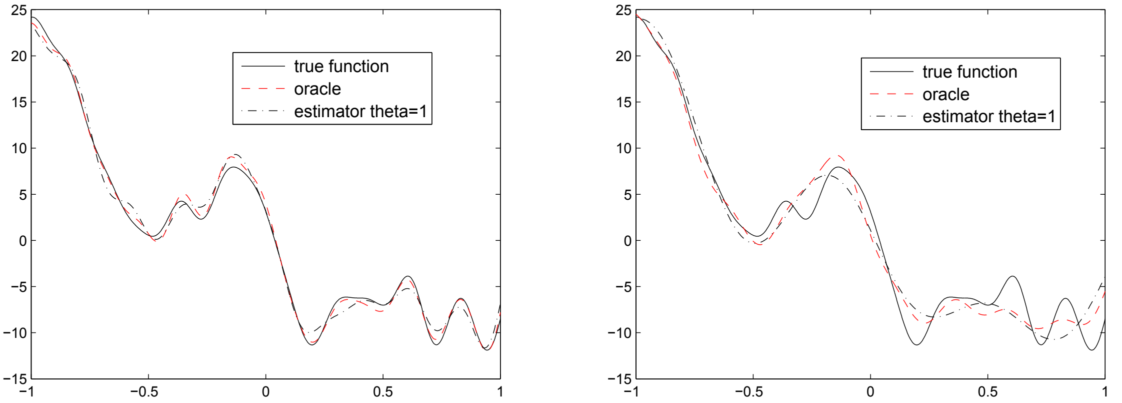

Figure 5.6: Well-posed and mildly ill-posed problems $\sigma_i^2 = 1/n$ and $\sigma_i^2 = i/n$

The estimation with a high degree of ill-posedness is also satisfactory compared to the oracle.

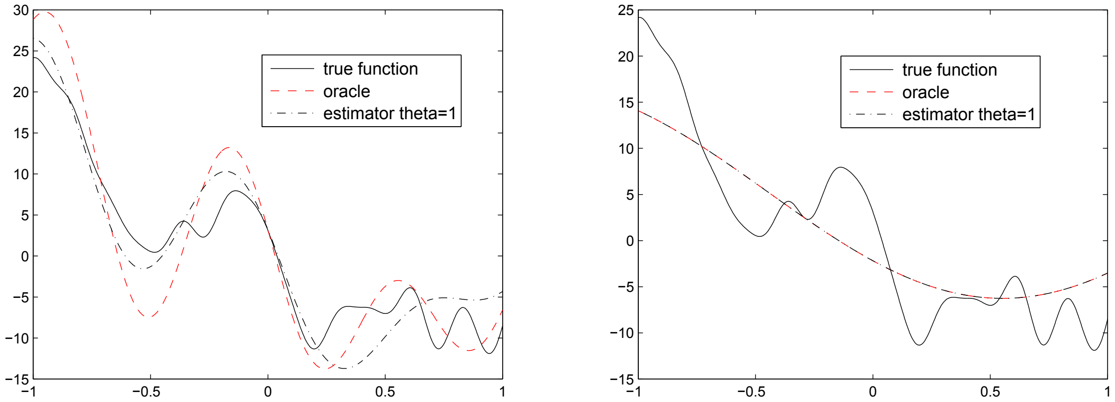

Figure 5.7: Ill-posed problems $\sigma_i^2 = i^2/n$ and $\sigma_i^2 = 2^i/n$

Case 4. $n = 100$, $x_i^2 \sim 2^{-i}$.

	$\sigma_i^2 = 1/n$	$\sigma_i^2 = i/n$	$\sigma_i^2 = i^2/n$	$\sigma_i^2 = 2^i/n$
$R(\widehat{x}_{\widehat{m}})$ $(\theta = 1)$	0.15 (×2.46)	1.19 (×7.18)	0.68 (×2.24)	0.52 (×1.70)
$R(\widehat{x}_{\widehat{m}})$ $(\theta = n)$	0.14 (×2.31)	0.23 (×1.37)	1.26 (×4.13)	1.09 (×3.59)
$R(x_{m^*})$	0.06	0.17	0.30	0.30
$R(x_{\text{sco}}^*)$	0.06	0.21	0.30	0.30
$R(x_{\text{lin}}^*)$	0.05	0.13	0.26	0.28

In the case treated here, we consider a fast decay of the coefficients x_i^2, leading to a relatively smooth function. We observe that with a low degree of ill-posedness, the parameter $\theta = n$ actually leads to a more efficient estimate than the value $\theta = 1$. What happens is that the threshold obtained with $\theta = 1$ is not sufficiently high to prevent selecting observations y_i that are strongly affected with noise. Although the events where the estimator wrongfully selects a highly noisy observation is relatively rare, it has a devastating effect on the estimation is the noise level is high. To avoid this issue, increasing the tuning parameter is effective, although it generally yields an overly smooth estimator.

On the next graphics, we see the same estimator, obtained with parameter $\theta = 1$ for different observations y. The estimation is good most of the time, the situation on the right graphic occurs with probability approximately 0.1.

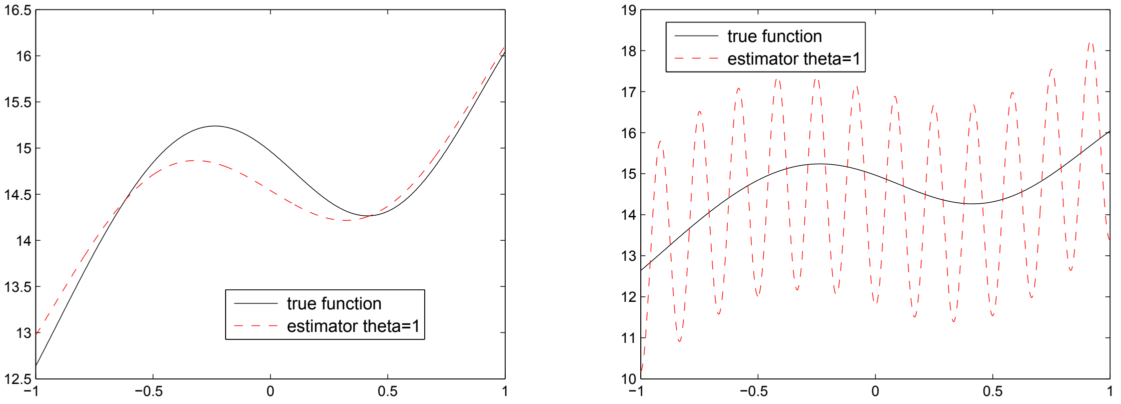

Figure 5.8: Mildly ill-posed problem $\sigma_i^2 = i/n$

As a matter of fact, the nature of the inverse problem makes the function difficult to estimate by some common methods. On the left graphic below, we show the comparison with the Tikhonov estimator obtained with optimal tuning parameter τ^*. In this situation, the Tikhonov method seems unable to achieve the approximate shape of x_0. On the right graphic, the threshold estimator obtained with tuning value $\theta = n$ is clearly well adapted to this specific problem.

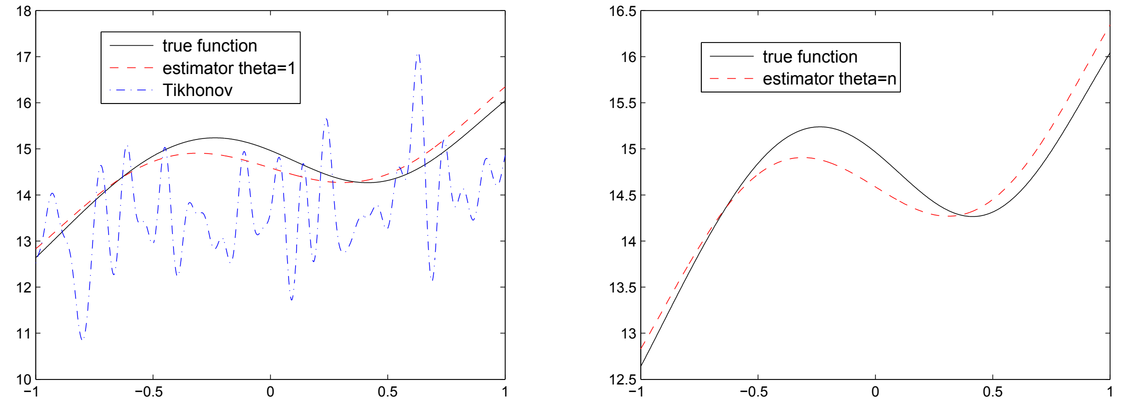

Figure 5.9: Mildly ill-posed problem $\sigma_i^2 = i/n$

The threshold is calculated to prevent selecting too noisy coefficients. For instance, assume that the signal is null in a given direction, i.e. $x_i = 0$. The observation is thus only noise $y_i = \eta_i$. For a given threshold, say $c_i\sigma_i^2$, the probability of wrongfully selecting the variable is only bounded by $\mathbb{P}(y_i^2 \geq c_i) \leq K \exp(-c_i/\beta)$, while the error in the risk caused by wrongfully selecting the index i is greater than c_i. Thus, we may choose a threshold sufficiently high to compensate the maximal loss in the risk. As a result, the estimator is robust to low probability events that have a devastating effect on the estimation, at the cost of being overly smoothed in most cases. In the previous examples, we observed that taking θ of order n, as it is somehow suggested by the theory, may lead to underfitting issues. Actually, this threshold is overly cautious as it generally selects too few variables, although it enables to control the risk. On the other hand, the value $\theta = 1$ yields an efficient estimate with high probability, but may be far from the true value in presence of outliers.

5.4 Regularization with unknown operator

We shall now discuss a situation where the operator A_n is not precisely known and is observed with a noise, independently from y. This situation is studied in [CH05], [EK01] or [HR08]. Here, the method discussed in the previous section does not apply since it required complete knowledge of the operator A_n.

As in [CH05], we assume that the eigenvectors ϕ_i and ψ_i are known. This seemingly strong assumption is actually met in many situations, for instance if the problem involves convolution or differential operators which can be decomposed in Fourier basis (see also the examples in [Cav08]). Thus, only the eigenvalues b_i are unknown and we assume they are observed independently of y, with a centered noise ξ_i with known variance $s^2 > 0$:

$$\hat{b}_i = b_i + \xi_i, \; i = 1, ..., n.$$

The method discussed in this paper is different according to whether the eigenvalues are known exactly or observed with a noise. Thus, we need to assume here that s is positive and the known operator framework can not be seen as a particular case. Moreover, we assume the ξ_i's are independent and satisfy the two following conditions.

A5.2. There exist $K', \beta' > 0$ such that $\forall t > 0, \forall i = 1, ..., n, \; \mathbb{P}(\xi_i^2/s^2 > t) \leq K' e^{-t/\beta'}$.

A5.3. There exist $C, \alpha > 0$ such that $\forall i = 1, ..., n, \; \min\{\mathbb{P}(\xi_i < -\alpha s), \mathbb{P}(\xi_i > \alpha s)\} \geq C$.

As discussed previously, the condition A5.2 means that that the ξ_i's have finite exponential moments. The condition A5.3 is hardly restrictive, and is fulfilled for instance as soon as the ξ_i's are identically distributed. As we shall see in the sequel, the method requires knowledge of the constant α (or at least an upper bound for it), but no information on the constants β', K' or C is needed to build the estimator.

Knowing the eigenvectors of $A_n^* A_n$ allows us to write the model in the form

$$y_i = b_i x_i + \varepsilon_i, i = 1, ..., n.$$

In our framework where the actual eigenvalues b_i are unknown, a natural estimator of each component x_i is obtained by $\tilde{y}_i = \hat{b}_i^{-1} y_i$, provided that $\hat{b}_i \neq 0$. However, it is clear that this estimate is not satisfactory if $\hat{b}_i$ is far from the true value (consider for instance the extreme case where $\hat{b}_i = 0$ or if $\hat{b}_i$ and b_i are of opposite signs). Actually, the naive estimator $\hat{b}_i^{-1}$ can not be used efficiently to estimate b_i^{-1} because it may have an infinite variance. In [CH05], the authors fix a threshold w the estimate can not exceed and consider an estimator of b_i^{-1} equal to

$\hat{b}_i^{-1}$ if $|\hat{b}_i| > 1/w$ and null otherwise. As we shall see below, we use the same idea here, where the threshold fixed on the $\hat{b}_i$'s is implicitly part of the variable selection process.

We can reasonably assume that null values of $\hat{b}_i$ do not provide any relevant information and can not be used to estimate x_0. Thus, to avoid considering trivial situations, we assume that all $\hat{b}_i$ are non-zero. In all generality, the $\tilde{y}_i$'s can be viewed as noisy observations of x_i by writing

$$\tilde{y}_i = x_i + \tilde{\eta}_i, \ i = 1, ..., n,$$

with $\tilde{y}_i = \hat{b}_i^{-1} \langle y, \psi_i \rangle_n$ and $\tilde{\eta}_i = \hat{b}_i^{-1}(\varepsilon_i - \xi_i x_i)$, where we recall $\varepsilon_i = \langle \varepsilon, \psi_i \rangle_n$. As in the previous section, we propose a threshold procedure to filter out the observations $\tilde{y}_i$ that are potentially highly contaminated with noise. Here, the noise $\tilde{\eta}_i$ is more difficult to deal with because it depends on the unknown coefficient x_i.

Our objective is to find an optimal variable selection criterion conditionally to the $\hat{b}_i$'s. In order to do so, we consider a framework where the $\hat{b}_i$'s are observed once and for all, and are treated as non-random. Thus, we define as an oracle, a model m_ξ^* minimizing the conditional risk $\mathbb{E}_\xi \|\hat{x}_m - x^\dagger\|^2$, where $\mathbb{E}_\xi(.)$ denotes the expectation knowing $\xi = (\xi_1, ..., \xi_n)^t$. Following a similar argument as in the previous section, a model minimizing the conditional risk contains only the indices i for which the coefficient x_i^2 is larger than the noise level. Hence, we may define $m_\xi^* = \{i : x_i^2 > \mathbb{E}_\xi(\tilde{\eta}_i^2)\}$. A notable difference here is that the noise $\tilde{\eta}_i$ actually depends on the value x_i. We can calculate the conditional expectation of $\tilde{\eta}_i^2$, given by

$$\mathbb{E}_\xi(\tilde{\eta}_i^2) = \hat{\sigma}_i^2 + \hat{b}_i^{-2} \xi_i^2 x_i^2,$$

where we set $\hat{\sigma}_i^2 = n^{-1} \hat{b}_i^{-2} \sigma^2$. After simplifications, it appears that the optimal model conditionally to the ξ_i's can be expressed in the two following equivalent forms

$$m_\xi^* = \left\{ i : 2|\hat{b}_i| > \frac{\sigma^2}{n|b_i|x_i^2} + |b_i| \right\} = \left\{ i : x_i^2 > \frac{\sigma^2}{n(\hat{b}_i^2 - \xi_i^2)}, \ |\hat{b}_i| > \frac{|b_i|}{2} \right\}.$$

In the first expression, we see that the oracle selects indices i for which the observation $\hat{b}_i$ exceeds a certain value depending on both x_i and b_i. Interestingly, components $\tilde{y}_i$ corresponding to observations $\hat{b}_i$ smaller than half the true eigenvalue b_i are not selected in the oracle, regardless of the coefficient x_i. Here again, the optimal model m_ξ^* can not be used in practical cases since it involves the unknown values x_i and ξ_i. We can only try to mimic the optimal threshold, based on the observations $\tilde{y}_i$ and $\hat{b}_i$. Consider the set

$$\widehat{m}_\xi = \left\{ i : \tilde{y}_i^2 > 8\hat{\sigma}_i^2 \nu_i, \ |\hat{b}_i| > \alpha s \right\},$$

where $\{\nu_i\}_{i=1,...,n}$ are parameters to be chosen and α is the constant defined in A5.3. With this definition, only the indices for which the observation $\hat{b}_i$ is larger than a certain value, namely αs, are selected. This conveys the idea discussed in [CH05], that when b_i is small compared to the noise level, the observation $\hat{b}_i$ is potentially mainly noise. Remark however that in [CH05], the lower limit for the observed eigenvalues is $s \log^2(1/s)$, while in our method, it is chosen of the same order as the standard deviation s.

Define the set $M = \{i : |b_i| < 2\alpha s\}$.

Theorem 5.4.1 *Assume that the condition* A5.1 *holds. The threshold estimator obtained with* $\nu_i = \beta \log(n^2 \hat{\sigma}_i^2)$ *satisfies,*

$$\mathbb{E}_\xi \|\hat{x}_{\widehat{m}_\xi} - x^\dagger\|^2 \leq \left(K_1' \log n + K_2'\right) \mathbb{E}_\xi \|\hat{x}_{m_\xi^*} - x^\dagger\|^2 + \sum_{i \in M} x_i^2 + \kappa(\xi),$$

with $K_1' = \max\{18\beta, 4\alpha^{-2}\beta'\}$, $K_2' = \max\{9(\beta \log \|x^\dagger\|^2 + 1), 1\}$, *and*

$$\kappa(\xi) = \frac{4K\beta}{n} + 4 \sum_{i \notin m_\xi^*} \frac{\xi_i^2 x_i^2}{\alpha^2 s^2} \mathbb{1}\{\xi_i^2 > s^2 \beta' \log n\}.$$

Moreover, if A5.2 *holds,* $\mathbb{E}(\kappa(\xi)) = O\left(\frac{\log n}{n}\right)$.

The main interest of this result lies in the fact that it provides an oracle inequality, conditionally to the $\hat{b}_i$'s. In particular, the conditional oracle $\hat{x}_{m_\xi^*}$ is more efficient than the estimator obtained by minimizing the expected risk $m \mapsto \mathbb{E}\|\hat{x}_m - x^\dagger\|^2$, since the optimal set m_ξ^* is allowed to depend on the ξ_i's. We see that the estimator $\hat{x}_{\widehat{m}_\xi}$ performs almost as well as the conditional oracle. Indeed, the residual term $\kappa(\xi)$ is independent from ξ with high probability, and its expectation is negligible under A5.2 as pointed out in the theorem. The non-random term $\sum_{i \in M} x_i^2$ is small if the eigenvalues b_i are observed with a good precision, i.e. if the variance s^2 is small. Moreover, this term can be shown to be of the same order as the risk under the condition A5.3.

Corollary 5.4.2 *If the conditions* A5.1, A5.2 *and* A5.3 *hold, the threshold estimator defined in Theorem 5.4.1 satisfies*

$$\mathbb{E}\|\hat{x}_{\widehat{m}_\xi} - x^\dagger\|^2 \leq K_4' \log n \; \mathbb{E}\|\hat{x}_{m_\xi^*} - x^\dagger\|^2 + \frac{K_5' \log n}{n},$$

for some constants K_4' *and* K_5' *independent from* n *and* s^2.

With a noisy operator, we manage to provide an estimator that achieves the rate of convergence of the conditional oracle, regardless of the precision of the approximation of the spectrum of A_n. Indeed, the constants K_4' and K_5' in Corollary 5.4.2 do not involve the variance s^2 of ξ. Actually, the variance only plays a role in the accuracy of the oracle. The result is non-asymptotic and requires no assumption on s^2.

5.5 Proofs

5.5.1 Technical lemmas

Lemma 5.5.1 *Assume the condition* A5.1 *holds. We have*

- $\mathbb{E}\left[(\eta_i^2 - x_i^2)\mathbb{1}\{i \in \widehat{m}\}\right] \leq 2K\beta\sigma_i^2 e^{-\mu_i/\beta}$.
- $\mathbb{E}\left[(x_i^2 - \eta_i^2)\mathbb{1}\{i \notin \widehat{m}\}\right] \leq \sigma_i^2(6\mu_i + 2)$.

Proof. Using the inequality $(a+b)^2 \leq 2a^2+2b^2$, we find that $\eta_i^2 - x_i^2 \leq 2\eta_i^2 - y_i^{\dagger 2}/2$. By definition of $\widehat{m}$, we get

$$(\eta_i^2 - x_i^2)\mathbb{1}\{i \in \widehat{m}\} \leq 2\sigma_i^2(\gamma_i - \mu_i)\mathbb{1}\{i \in \widehat{m}\} \leq 2\sigma_i^2(\gamma_i - \mu_i)\mathbb{1}\{\gamma_i \geq \mu_i\},$$

where we used that $X \leq X\mathbb{1}\{X \geq 0\}$. We finally obtain for all $i \notin m^*$,

$$\mathbb{E}\left[(\eta_i^2 - x_i^2)\mathbb{1}\{i \in \widehat{m}\}\right] \leq 2\sigma_i^2 \int_0^\infty \mathbb{P}(\gamma_i \geq t + \mu_i)\ dt \leq 2K\beta\sigma_i^2 e^{-\mu_i/\beta},$$

as a consequence of A5.1. For the second part of the lemma, write $x_i^2 - \eta_i^2 = y_i^{\dagger 2} - 2\eta_i y_i^\dagger$ which is bounded by $3y_i^{\dagger 2}/2 + 2\eta_i^2$, using the inequality $2ab \leq 2a^2 + b^2/2$. This leads to

$$\mathbb{E}\left[(x_i^2 - \eta_i^2)\mathbb{1}\{i \notin \widehat{m}\}\right] \leq \sigma_i^2(6\mu_i + 2).$$

Lemma 5.5.2

$$\inf_{m\in\mathcal{M}} \mathbb{E}\|\hat{x}_m - x^\dagger\|^2 \leq 2 \inf_{\lambda\in\mathbb{R}^n} \mathbb{E}\|\hat{x}(\lambda) - x^\dagger\|^2.$$

Proof. The minimal values of the expected risks can be calculated explicitly in the two classes considered here. Minimizing over $\mathbb{R}^n$ the function $\lambda \mapsto \mathbb{E}\|\hat{x}(\lambda) - x^\dagger\|^2$, we find that the optimal value of λ_i is reached for $\lambda_i^* = x_i^2/(x_i^2+\sigma_i^2)$. On the other hand, we know that $m \mapsto \mathbb{E}\|\hat{x}_m - x^\dagger\|^2$ reaches its minimum at $m^* = \{i : x_i^2 \geq \sigma_i^2\}$, yielding

$$\inf_{\lambda\in\mathbb{R}^n} \mathbb{E}\|\hat{x}(\lambda) - x^\dagger\|^2 = \sum_{i=1}^n \frac{x_i^2\sigma_i^2}{x_i^2 + \sigma_i^2} \quad \text{and} \quad \inf_{m\in\mathcal{M}} \mathbb{E}\|\hat{x}_m - x^\dagger\|^2 = \sum_{i\in m^*} \sigma_i^2 + \sum_{i\notin m^*} x_i^2.$$

By definition, if $i \in m^*$, $2x_i^2/(x_i^2 + \sigma_i^2) \geq 1$. In the same way, $2\sigma_i^2/(x_i^2 + \sigma_i^2) \geq 1$, for all $i \notin m^*$. We conclude by summing all the terms.

Lemma 5.5.3 *Assume the condition* A5.1 *holds. We have, for all* $i = 1, ..., n$,

- $\mathbb{E}_\xi\left[(\tilde{\eta}_i^2 - x_i^2)\mathbb{1}\{i \in \widehat{m}_\xi\}\right] \leq 4K\beta\ \hat{\sigma}_i^2 e^{-\nu_i/\beta} + \dfrac{4\xi_i^2 x_i^2}{\alpha^2 s^2}$.
- $\mathbb{E}_\xi\left[(x_i^2 - \tilde{\eta}_i^2)\mathbb{1}\{i \notin \widehat{m}_\xi\}\right] \leq 9\hat{\sigma}_i^2\nu_i + 8\mathbb{E}_\xi(\tilde{\eta}_i^2) + x_i^2\mathbb{1}\{|\hat{b}_i| \leq \alpha s\}$.

Proof. Remark that $\tilde{\eta}_i^2 = \hat{b}_i^{-2}(\varepsilon_i - \xi_i x_i)^2 \leq 2\hat{b}_i^{-2}\varepsilon_i^2 + 2\hat{b}_i^{-2}\xi_i^2 x_i^2$. Using that $x_i^2 \geq \tilde{y}_i^2/2 - \tilde{\eta}_i^2$, we deduce

$$\tilde{\eta}_i^2 - x_i^2 \leq 4\hat{b}_i^{-2}\varepsilon_i^2 + 4\hat{b}_i^{-2}\xi_i^2 x_i^2 - \frac{\tilde{y}_i^2}{2}.$$

Writing $\widehat{m}_\xi = \{\tilde{y}_i^2 > 8\hat{\sigma}_i^2\nu_i\} \cap \{|\hat{b}_i| > \alpha s\}$, we find

$$(\tilde{\eta}_i^2 - x_i^2)\mathbb{1}\{i \in \widehat{m}_\xi\} \leq 4\hat{\sigma}_i^2(\gamma_i - \nu_i)\mathbb{1}\{\gamma_i \geq \nu_i\} + 4\hat{b}_i^{-2}\xi_i^2 x_i^2\mathbb{1}\{|\hat{b}_i| > \alpha s\},$$

where we recall that $\gamma_i = n\varepsilon_i^2/\sigma^2$. Clearly, $\hat{b}_i^{-2}\mathbb{1}\{|\hat{b}_i| > \alpha s\} < \alpha^{-2}s^{-2}$ and the result follows using the condition A5.1. For the second part of the lemma, remark that the complement of $\widehat{m}_\xi$ is $\{\tilde{y}_i^2 \leq 8\hat{\sigma}_i^2\nu_i,\ |\hat{b}_i| > \alpha s\} \cup \{|\hat{b}_i| \leq \alpha s\}$. Using the inequality $x_i^2 - \tilde{\eta}_i^2 \leq (1+\theta^{-1})\tilde{y}_i^2 + \theta\tilde{\eta}_i^2$ for $\theta = 8$, we get

$$(x_i^2 - \tilde{\eta}_i^2)\mathbb{1}\{i \notin \widehat{m}_\xi\} \leq 9\hat{\sigma}_i^2\nu_i + 8\tilde{\eta}_i^2 + x_i^2\mathbb{1}\{|\hat{b}_i| \leq \alpha s\}.$$

Lemma 5.5.4 *If* A5.2 *holds, we have*

$$\xi_i^2 \leq s^2\beta' \log n + \xi_i^2 \mathbb{1}\{\xi_i^2 > s^2\beta' \log n\},$$

with $\mathbb{E}\left(\xi_i^2 \mathbb{1}\{\xi_i^2 > s^2\beta' \log n\}\right) = O(n^{-1}\log n)$.

Proof. Write $\xi_i^2 \leq s^2\beta' \log n \ \mathbb{1}\{\xi_i^2 \leq s^2\beta' \log n\} + \xi_i^2 \mathbb{1}\{\xi_i^2 > s^2\beta' \log n\}$. To bound the first term, we use the crude inequality $\mathbb{1}\{\xi_i^2 \leq s^2\beta' \log n\} \leq 1$. For the second term, we have as a consequence of A5.2,

$$\begin{aligned}\mathbb{E}\left[\xi_i^2 \mathbb{1}\{\xi_i^2 > s^2\beta' \log n\}\right] &= \int_0^\infty \mathbb{P}\left(\xi_i^2 \mathbb{1}\{\xi_i^2/s^2 > \beta' \log n\} > t\right)\ dt \\ &= s^2\beta' \log n \ \mathbb{P}(\xi_i^2/s^2 > \beta' \log n) + s^2 \int_{\beta' \log n}^\infty \mathbb{P}(\xi_i^2/s^2 > t)\ dt \\ &\leq \frac{K'\beta' s^2(1+\log n)}{n}.\end{aligned}$$

5.5.2 Proof of Theorem 5.3.1

Write

$$\|\hat{x}_{\widehat{m}} - x_0\|^2 = \|\hat{x}_{m^*} - x_0\|^2 + \sum_{i \notin m^*} (\eta_i^2 - x_i^2)\mathbb{1}\{i \in \widehat{m}\} + \sum_{i \in m^*} (x_i^2 - \eta_i^2)\mathbb{1}\{i \notin \widehat{m}\}.$$

The objective is to bound the terms $\mathbb{E}[(\eta_i^2 - x_i^2)\mathbb{1}\{i \in \widehat{m}\}]$ and $\mathbb{E}[(x_i^2 - \eta_i^2)\mathbb{1}\{i \notin \widehat{m}\}]$ separately. By Lemma 5.5.1, we know that $\mathbb{E}\left[(\eta_i^2 - x_i^2)\mathbb{1}\{i \in \widehat{m}\}\right] \leq 2K\beta\sigma_i^2 e^{-\mu_i/\beta}$, which gives for $\mu_i = \beta \log\left(e + \theta\sigma_i^2\right)$,

$$\mathbb{E}\left[(\eta_i^2 - x_i^2)\mathbb{1}\{i \in \widehat{m}\}\right] \leq 2K\beta\ \frac{\sigma_i^2}{e + \theta\sigma_i^2} \leq \frac{2K\beta}{\theta}.$$

On the other hand, if $i \notin \widehat{m}$, Lemma 5.5.1 warrants

$$\mathbb{E}\left[(x_i^2 - \eta_i^2)\mathbb{1}\{i \notin \widehat{m}\}\right] \leq \sigma_i^2\left(6\beta \log(e + \theta\sigma_i^2) + 2\right).$$

Since $i \in m^*$, $\log(e + \theta\sigma_i^2) \leq \log(e + \theta\|x^\dagger\|^2)$. We conclude by summing all the terms.

5.5.3 Proof of Proposition 5.3.3

It suffices to show that the oracle $\widehat{x}_{m^*}$ achieves the rate of convergence $\mathbb{E}\|\widehat{x}_{m^*} - x^\dagger\|^2 = O\left(n^{\frac{\delta-2}{2}}\right)$. For this, write

$$\mathbb{E}\|\widehat{x}_{m^*} - x^\dagger\|^2 = \sum_{i \in m^*} \sigma_i^2 + \sum_{i \notin m^*} x_i^2 \leq \sum_{i=1}^n |x_i|^\delta \sigma_i^{2-\delta},$$

by definition of m^*. We deduce

$$\mathbb{E}\|\widehat{x}_{m^*} - x^\dagger\|^2 \leq n^{-\frac{2-\delta}{2}}\ \sigma^{2-\delta} \sum_{i=1}^n |x_i|^\delta |b_i|^{2-\delta},$$

proving the result.

5.5.4 Proof of Theorem 5.4.1

The proof starts as in Theorem 5.3.1. We have

$$\|\hat{x}_{\widehat{m}_\xi} - x^\dagger\|^2 = \|\hat{x}_{m^*_\xi} - x^\dagger\|^2 + \sum_{i\notin m^*_\xi} (\tilde{\eta}_i^2 - x_i^2)\mathbb{1}\{i \in \widehat{m}_\xi\} + \sum_{i\in m^*_\xi} (x_i^2 - \tilde{\eta}_i^2)\mathbb{1}\{i \notin \widehat{m}_\xi\},$$

and the objective is to bound the conditional expectation of each term separately. Using successively Lemma 5.5.3 and Lemma 5.5.4, we get

$$\mathbb{E}_\xi\left[(\tilde{\eta}_i^2 - x_i^2)\mathbb{1}\{i \in \widehat{m}_\xi\}\right] \le \frac{4K\beta}{n^2} + 4\alpha^{-2}s^{-2}\xi_i^2x_i^2 \le \frac{4\beta'\log n}{\alpha^2}\, x_i^2 + \kappa_i(\xi),$$

with

$$\kappa_i(\xi) = \frac{4K\beta}{n^2} + \frac{4\xi_i^2x_i^2}{\alpha^2s^2}\mathbb{1}\{\xi_i^2 > s^2\beta'\log n\}.$$

By Lemma 5.5.4, we know that $\kappa(\xi) = \sum_{i\notin m^*_\xi}\kappa_i(\xi)$ is such that

$$\mathbb{E}(\kappa(\xi)) \le \frac{4(K\beta + 2\alpha^{-2}K'\beta'\|x^\dagger\|^2\log n)}{n} = O\left(\frac{\log n}{n}\right).$$

On the other hand, Lemma 5.5.3 gives, for $\theta = 8$,

$$\mathbb{E}_\xi\left[(x_i^2 - \tilde{\eta}_i^2)\mathbb{1}\{i \notin \widehat{m}_\xi\}\right] \le 9\hat{\sigma}_i^2\nu_i + 8\mathbb{E}_\xi(\tilde{\eta}_i^2) + x_i^2\mathbb{1}\{|\hat{b}_i| \le \alpha s\}.$$

For all $i \in m^*_\xi$, we know that $|\hat{b}_i| \ge |b_i|/2$. Thus, if $i \in m^*_\xi$, $\mathbb{1}\{|\hat{b}_i| \le \alpha s\} \le \mathbb{1}\{i \in M\}$, where we recall $M = \{i : |b_i| < 2\alpha s\}$. We know also that, if $i \in m^*_\xi$, then $\hat{\sigma}_i^2 \le x_i^2$. Thus, $\nu_i = \beta\log(n^2\hat{\sigma}_i^2) \le 2\beta\log n + \beta\log\|x^\dagger\|^2$. Noticing that $\hat{\sigma}_i^2 \le \mathbb{E}_\xi(\tilde{\eta}_i^2)$, we find

$$\mathbb{E}_\xi\left[(x_i^2 - \tilde{\eta}_i^2)\mathbb{1}\{i \notin \widehat{m}_\xi\}\right] \le (18\beta\log n + 9\beta\log\|x^\dagger\|^2 + 8)\mathbb{E}_\xi(\tilde{\eta}_i^2) + x_i^2\mathbb{1}\{i \in M\}.$$

The result follows by summing all the term, using that the risk of the oracle $\hat{x}_{\widehat{m}_\xi}$ is

$$\mathbb{E}_\xi\|\hat{x}_{m^*_\xi} - x^\dagger\|^2 = \sum_{i\notin m^*_\xi} x_i^2 + \sum_{i\in m^*_\xi}\mathbb{E}_\xi(\tilde{\eta}_i^2).$$

5.5.5 Proof of Corollary 5.4.2

It suffices to show that the term $\sum_{i\in M} x_i^2$ is of the same order as the risk of the oracle. Write

$$\mathbb{E}\|\hat{x}_{m^*_\xi} - x^\dagger\|^2 \ge \sum_{i=1}^n x_i^2\mathbb{P}(i \notin m^*_\xi) \ge \sum_{i=1}^n x_i^2\mathbb{P}(|\hat{b}_i| \le |b_i|/2).$$

For all $i \in M$, the probability $\mathbb{P}(|\hat{b}_i| \le |b_i|/2)$ is greater than C as a consequence of A5.3. We deduce $\sum_{i\in M} x_i^2 \le C^{-1}\mathbb{E}\|\hat{x}_{m^*_\xi} - x^\dagger\|^2$.

5.6 Appendix: Regularization by aggregation

In the problem of recovering the function x_0, several competing estimation procedures can be used. Rather than searching for the best estimator among all considered solutions, one may be interested in considering a solution expressed as a combination of the existing estimates. This approach, known as *aggregation*, has been recently studied in the frame of non-parametric regression models in [BTW06], [BTW07], [JN00] and [Tsy03]. Assume we have a collection $\mathbf{x} = (\widehat{x}_1, ..., \widehat{x}_M)$ (with $2 \leq M \leq n$) of preliminary estimators of x_0, independent from the observations. The $\widehat{x}_j$'s can be viewed as preliminary estimators of x_0, constructed from a training sample. Aggregation procedures aim to build an estimator of x_0 by combining in a suitable way the functions $\widehat{x}_1, ..., \widehat{x}_M$. The purpose is to filter out irrelevant elements in the collection $\widehat{x}_1, ..., \widehat{x}_M$ as well as to combine several possibly competing estimators. Thus, an estimator is sought as a linear combination of the $\widehat{x}_j$'s, called *aggregate*, and noted

$$x_\lambda = \mathbf{x}\lambda = \sum_{j=1}^{M} \lambda_j \, \widehat{x}_j,$$

for $\lambda = (\lambda_1, ..., \lambda_M)^t$ lying in some subset Λ of $\mathbb{R}^M$. Several kinds of aggregation frameworks are studied in the literature, depending on restrictions made on the possible values of λ. This restrictions are reflected through the choice of the set Λ. Most common examples are arguably convex aggregation (C), linear aggregation (L) and model selection aggregation (MS) studied in [BTW07], [BTW06], [Tsy03], [JN00] in a homoscedastic model. The objective of convex aggregation is to select the optimal estimator lying in the convex hull of $\widehat{x}_1, ..., \widehat{x}_M$, which corresponds to $\Lambda = \{\lambda \in [0;1]^M, \sum_{j=1}^M \lambda_j \leq 1\}$. Linear aggregation aims at finding the best linear combination of the $\widehat{x}_j$'s, allowing λ to take any value in $\mathbb{R}^M$. Finally, model selection aggregation aims to select the best estimator among the collection $\widehat{x}_1, ..., \widehat{x}_M$, which corresponds to considering λ in the usual basis of $\mathbb{R}^M$.

For the regularization of the discrete inverse problem (5.1) treated in Section 5.2, most methods restrict the choice of the solution in a set of linear estimators $\{\widehat{x}_\alpha = R_\alpha y, \ \alpha \in S\}$, where $\{R_\alpha, \alpha \in S\}$ is a collection of linear operators and α is a tuning parameter. This is the case for instance of the Tikhonov regularization as well as the spectral cut-off or more general projection estimators, where the solution can be expressed using a smooth version of the pseudo inverse of A_n,

$$\widehat{x}_\alpha = \Phi_\alpha(A_n^* A_n) A_n^* y.$$

Here, Φ_α is a bounded approximation of the inverse function and the application of Φ_α to the diagonalizable operator $A_n^* A_n$ is to be understood as an operation on the spectrum (see Section 1.3.1). In general, we may restrict without loss of efficiency the number of possible values of α to a finite set $S = \{\alpha_j, j = 1, ..., M\}$, leading to a finite collection of candidate estimators $\{\widehat{x}_j = \widehat{x}_{\alpha_j}, \ j = 1, ..., M\}$. The main question in the estimation problem is then to choose a value $\widehat{\alpha}$ of the tuning parameter leading to an efficient estimate $\widehat{x} = \widehat{x}_{\widehat{\alpha}}$. Equivalently, we search for the best estimator in the finite class $\{\widehat{x}_j, \ j = 1, ..., M\}$ given an observation of y. A natural objective is to minimize the expected quadratic risk $\mathbb{E}\|x_0 - \widehat{x}_{\widehat{\alpha}}\|^2$, however this quantity is hard to evaluate because the resulting estimator $\widehat{x}_{\widehat{\alpha}}$ is no longer linear, as $\widehat{\alpha}$ depends on the observation y. So, a usual compromise in such situations is to compare the risk of the estimator to that of the best estimator in the class, or to the best linear combination

$$\mathbb{E}\|x_0 - \widehat{x}\|^2 \leq \inf_{\lambda \in \mathbb{R}^M} \mathbb{E}\|x_\lambda - x_0\|^2 + \Delta_n,$$

where Δ_n is a residual term we want as small as possible (ideally of the same order as the minimal risk). While some methods to select a suitable value of α have been treated in the literature (see for instance [Cav08] or [FLn08]), we propose an aggregation approach that constructs an estimator as a linear combination of the candidate estimators $\widehat{x}_1, ..., \widehat{x}_M$.

5.6.1 The heteroscedastic case

As discussed in Section 5.2, a discrete inverse problem with known operator can be treated as an heteroscedastic model with known variance. Thus, in order to extend the aggregation framework to inverse problems, we study the aggregation process in a heteroscedastic model. We consider the usual non-parametric model where the function of interest x_0 is observed at a finite design $t_1, ..., t_n$. For sake of simplicity, we shall identify a map u with the vector of its coordinates $u = (u(t_1), ..., u(t_n))^t \in \mathbb{R}^n$. Thus, in the sequel, we use the notation $x_0 = (x_0(t_1), ..., x_0(t_n))^t \in \mathbb{R}^n$. We consider the following model

$$y = x_0 + \varepsilon,$$

with $y \in \mathbb{R}^n$. Here, the noise ε is assumed centered with known covariance matrix Σ, which we assume positive definite. Typically, we have $\Sigma = \sigma^2(A_n^* A_n)^\dagger/n$ in the inverse problem framework treated in Section 5. We note $S = \text{span}\{\widehat{x}_1, ..., \widehat{x}_M\}$ in $\mathbb{R}^n$ and Π_S the orthogonal projector onto S in $\mathbb{R}^n$. Moreover, for a set of indices $m \subseteq \{1, ..., M\}$, we define in the same way $S_m = \text{span}\{\widehat{x}_j, j \in m\}$ and the associated projector Π_{S_m}.

Viewing the preliminary estimators $\widehat{x}_1, ..., \widehat{x}_M$ as a collection of regressors, the aggregation problem can be treated as a classical linear model. It is well known that the least square solution $\hat{x} = \arg\min_{x \in S} \|y - x\|_n^2$ may lead to overfitting issues, especially with a large number of regressors. So, rather than to minimize the quadratic loss over the linear span S, it is generally more efficient to consider a solution in a subspace $E \subseteq S$. In this way, the construction of the estimator involves two distinct aspects which are finding a proper model $E \subseteq S$ (variable selection) and choosing the best candidate in this model (regression). When the regression step is made via minimum least square, the estimator can be simply expressed as the orthogonal projection of y onto E and the accuracy of such estimates relies on the choice of the projection subspace E.

In a homoscedastic model (i.e. when $\Sigma = \sigma^2 I$), the question of the best projection space E (which is to be understood as the minimizer of the quadratic risk $\mathbb{E}\|\Pi_E y - x_0\|^2$) can be given a simple and natural answer. Remark that the risk can be expressed as

$$\mathbb{E}\|\Pi_E y - x_0\|^2 = \|(I - \Pi_E)x_0\|^2 + \mathbb{E}\|\Pi_E \varepsilon\|^2 = \|x_0\|^2 + \text{Tr}\left(\Pi_E(\sigma^2 I - x_0 x_0^t)\right),$$

where $\text{Tr}(M)$ stands for the trace of M. Thus, it appears that the best projection space contained in S is given by $E^* = \text{span}\{\Pi_S x_0\}$ if $\|\Pi_S x_0\|^2 > \sigma^2$ and $E^* = \{0\}$ otherwise. However, this information is irrelevant since E^* is unknown and estimating E^* basically reduces to estimating $\Pi_S x_0$. In the heteroscedastic case, determining the best projection space E^* is not as direct. The same calculation shows that minimizing the quadratic risk reduces to minimizing the criterion

$$E \mapsto \operatorname{Tr}\left(\Pi_E(\Sigma - x_0 x_0^t)\right),\ E \subseteq S.$$

Actually, determining the optimal projection space E^* involves calculating the spectral decomposition of the symmetric operator $\Pi_S \Sigma \Pi_S - (\Pi_S x_0)(\Pi_S x_0)^t$. More precisely, it appears that the minimizer E^* is the linear span of the eigenvectors of $\Pi_S \Sigma \Pi_S - (\Pi_S x_0)(\Pi_S x_0)^t$ associated to negative eigenvalues. Once again, E^* is unknown in practice since it depends on x_0.

It is presumably hopeless to intend to estimate the best projection model over all subspaces of S. For computational feasibility, it is generally necessary to restrict the choice of E to a finite collection of submodels. In the Gaussian case, classical penalized procedures such as Mallows Cp, Akaike information criterion (AIC) or Bayesian information criterion (BIC) lead to projection estimators where the projection space is estimated in the class $\{S_m, m \subseteq \{1, ..., M\}\}$. This class of submodels is quite natural although one drawback of these methods remains their computational cost. Indeed, the correlation between the regressors $\widehat{x}_1, ..., \widehat{x}_M$ makes it difficult to evaluate the accuracy of selecting a variable $\widehat{x}_j$ independently from the other variables. To overcome this issue, a solution is to consider a collection of submodels generated by orthogonal variables. As we shall see in the heteroscedastic case, the family of eigenvectors of $\Pi_S \Sigma \Pi_S$ turns out to be a particularly convenient orthogonal basis. The idea is to restrict the observation to the linear span of $\widehat{x}_1, ..., \widehat{x}_M$ and then to write the data in an appropriate basis in order to have uncorrelated noises (but still with possibly unequal variances). Precisely, let $\{\sigma_i^2, v_i\}_{i=1,...,M}$ be an orthogonal system of $\Pi_S \Sigma \Pi_S$, that is $\langle v_i, v_j \rangle = \mathbb{1}\{i = j\}$ and

$$\forall u \in \mathbb{R}^n,\ \Pi_S \Sigma \Pi_S\ u = \sum_{i=1}^{M} \sigma_i^2 \langle u, v_i \rangle v_i.$$

Let $y_i = \langle y, v_i \rangle$, $x_i = \langle x_0, v_i \rangle$ and $\varepsilon_i = \langle \varepsilon, v_i \rangle$, we have the following relation

$$y_j = x_j + \varepsilon_j,\ j = 1, ..., M,$$

where the noises ε_j satisfy $\mathbb{E}(\varepsilon_i \varepsilon_j) = v_i^t \Pi_S \Sigma \Pi_S v_j = \sigma_i^2 \mathbb{1}\{i = j\}$. The objective is to estimate x_0 by a linear combination of the $\widehat{x}_j$'s. Equivalently, we aim to estimate $\Pi_S x_0 = \sum_{i=1}^M x_i v_i$, which can be made using a threshold procedure on the coefficients y_i. Denote by $\mathcal{M}$ the set of all subsets of $\{1, ..., M\}$ and for all $m \in \mathcal{M}$, note $V_m = \operatorname{span}\{v_j, j \in m\}$. We consider a projection estimator of the form $\widehat{x}_m = \Pi_{V_m} y$. We know that the model minimizing the risk is given by $m^* = \{j : x_j^2 \geq \sigma_j^2\}$ (see Section 5.3). So, in order to mimic the oracle $\widehat{x}_{m^*}$, we define the estimator $\widehat{x} = \widehat{x}_{\widehat{m}}$ where

$$\widehat{m} = \{j : y_j^2 \geq 4\sigma_j^2 \mu_j\},$$

with $\{\mu_j\}_{j=1,...,M}$ a sequence of tuning parameters to be determined. Although the construction of the estimator requires the calculation of the singular system $\{\sigma_i^2, v_i\}_{i=1,...,M}$, the computational cost is low compared to penalized methods such as AIC or BIC because we do not need to compare each submodel one at a time. Here, we simply choose to select or not the variable

v_j in the model in function of the value of y_j. By construction, the estimator $\widehat{x}$ belongs to S and thus, it can be written as an aggregate $\widehat{x} = x_{\widehat{\lambda}}$ for some $\widehat{\lambda} \in \mathbb{R}^M$. Define the function

$$p(\lambda) = \sum_{i=1}^{M} \sigma_j^2 \, \mathbb{1}\{\langle x_\lambda, v_j \rangle \neq 0\}, \ \lambda \in \mathbb{R}^M.$$

Theorem 5.6.1 *Assume there exist positive constants K, β such that $\mathbb{E}[\exp(\varepsilon_j^2/\beta\sigma_j^2)] \leq K$, for all $j = 1, ..., M$. For some $\theta > 0$, set $\mu_j = \beta \log(e + \theta\sigma_j^2)$, then the estimator $\widehat{x}$ satisfies*

$$\mathbb{E}\|\widehat{x} - x_0\|^2 \ \leq \ \inf_{\lambda \in \mathbb{R}^M} \left\{ \|x_\lambda - x_0\|^2 + p(\lambda) \right\} + (6\beta \log(e + \theta\|x^\dagger\|^2) + 2) \sum_{j \in m^*} \sigma_j^2 + \frac{2K\beta n}{\theta}.$$

This theorem provides an oracle inequality for the estimator $\widehat{x}$. Remark that the first residual term can be expressed in function of the minimizer $\lambda^* = \arg\min_{\lambda \in \mathbb{R}^M} \|x_\lambda - x_0\|^2 + p(\lambda)$ as we have $\sum_{j \in m^*} \sigma_j^2 = p(\lambda^*)$.

Proof. A direct application of Theorem 5.3.1 yields

$$\mathbb{E}\|\widehat{x} - x_0\|^2 \leq \inf_{m \in \mathcal{M}} \mathbb{E}\|\widehat{x}_m - x_0\|^2 + (6\beta \log(e + \theta\|x^\dagger\|^2) + 2) \sum_{j \in m^*} \sigma_j^2 + \frac{2K\beta n}{\theta}.$$

Note $K(m) = \{\lambda \in \mathbb{R}^M : \langle x_\lambda, v_j \rangle \neq 0 \Leftrightarrow j \in m\}$ and remark that $p(.)$ is constant over $K(m)$. We conclude using that $\mathbb{R}^M = \cup_{m \in \mathcal{M}} K(m)$ and writing for all $m \in \mathcal{M}$,

$$\mathbb{E}\|\widehat{x}_m - x_0\|^2 = \|\Pi_{V_m} x_0 - x_0\|^2 + \sum_{j \in m} \sigma_j^2 = \inf_{\lambda \in K(m)} \{\|x_\lambda - x_0\|^2 + p(\lambda)\}.$$

The aggregate $\widehat{x}$ obtained with this approach requires the calculation of the singular value decomposition of the matrix $\Pi_S \Sigma \Pi_S$. Writing the observations in this particular basis leads to a model with uncorrelated noises, which makes it easier to handle. Thus, while the method can be interpreted as a penalized procedure, the computation of the estimator is simple and does not involve the calculation of a penalized criterion for all 2^M models.

One desirable property of most penalized procedures is the sparsity of the solution. Here, the method induces the sparsity of the solution in the orthogonal basis $\{v_i\}_{i=1,...,M}$ and therefore, it has no reason of being associated to a sparse parameter λ. If we are concerned with the sparsity of λ, penalized procedures for aggregation as studied in [BTW07] and [BTW06] are more appropriate. The methods proposed in these papers rely on penalizations on the number of non-zero components of λ or on the ℓ^1-norm of λ, both known to favor low-dimensional values of the parameter. While the properties of the aggregate are obtained in a homoscedastic Gaussian regression framework, a generalization of these results to heteroscedastic models could provide an efficient regularization method for inverse problems via aggregation.

Bibliography

[AC03] C. Ai and X. Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

[AC09] C. Ai and X. Chen. Semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. Cowles Foundation Discussion Papers 1731, Cowles Foundation for Research in Economics, Yale University, October 2009.

[ACH11] D. Ackerberg, X. Chen, and J. Hahn. A practical asymptotic variance estimator for two-step semiparametric estimators. CeMMAP working papers CWP22/11, Centre for Microdata Methods and Practice, Institute for Fiscal Studies, June 2011.

[AS98] F. Abramovich and B. W. Silverman. Wavelet decomposition approaches to statistical inverse problems. *Biometrika*, 85(1):115–129, 1998.

[BHMR07] N. Bissantz, T. Hohage, A. Munk, and F. Ruymgaart. Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J. Numer. Anal.*, 45(6):2610–2636 (electronic), 2007.

[BKRW93] P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and adaptive estimation for semiparametric models.* Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, 1993.

[BL91] J. M. Borwein and A. S. Lewis. Duality relationships for entropy-like minimization problems. *SIAM J. Control Optim.*, 29(2):325–338, 1991.

[BL93] J. M. Borwein and A. S. Lewis. Partially-finite programming in L_1 and the existence of maximum entropy estimates. *SIAM J. Optim.*, 3(2):248–267, 1993.

[BLN96] J. M. Borwein, A. S. Lewis, and D. Noll. Maximum entropy reconstruction using derivative information. I. Fisher information and convex duality. *Math. Oper. Res.*, 21(2):442–468, 1996.

[BTW06] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation and sparsity via l1 penalized least squares. In *Learning theory*, volume 4005 of *Lecture Notes in Comput. Sci.*, pages 379–391. Springer, Berlin, 2006.

[BTW07] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.

[Cav08] L. Cavalier. Nonparametric statistical inverse problems. *Inverse Problems*, 24(3):034004, 19, 2008.

[CF00] M. Carrasco and J. P. Florens. Generalization of GMM to a continuum of moment conditions. *Econometric Theory*, 16(6):797–834, 2000.

[CFR06] M. Carrasco, J. P. Florens, and E. Renault. *Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization*, volume 6. North Holland, 2006.

[CG06] L. Cavalier and G. K. Golubev. Risk hull method and regularization by projections of ill-posed inverse problems. *Ann. Statist.*, 34(1):1653–1677, 2006.

[CGPT00] L. Cavalier, G. K. Golubev, D. Picard, and A. B. Tsybakov. Oracle inequalities for inverse problems. *Ann. Statist.*, 30(3):843–874, 2000.

[CH05] L. Cavalier and N. W. Hengartner. Adaptive estimation for inverse problems with noisy operators. *Inverse Problems*, 21(4):1345–1361, 2005.

[Cha87] G. Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econometrics*, 34(3):305–334, 1987.

[Cha92a] G. Chamberlain. Efficiency bounds for semiparametric regression. *Econometrica*, 60(3):567–96, May 1992.

[Cha92b] G. Chamberlain. Sequential moment restrictions in panel data: Comment. *Journal of Business & Economic Statistics*, 10(1):20–26, January 1992.

[CHT08] X. Chen, H. Hong, and A. Tarozzi. Semiparametric efficiency in GMM models with auxiliary data. *Ann. Statist.*, 36(2):808–843, 2008.

[CP08] X. Chen and D. Pouzo. Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. Cowles Foundation Discussion Papers 1640R, Cowles Foundation for Research in Economics, Yale University, February 2008.

[CR07] X. Chen and M. Reiss. On rate optimality for ill-posed inverse problems in econometrics. Cowles Foundation Discussion Papers 1626, Cowles Foundation for Research in Economics, Yale University, September 2007.

[Csi75] I. Csiszár. *I*-divergence geometry of probability distributions and minimization problems. *Ann. Probability*, 3:146–158, 1975.

[Dev88] J. C. Deville. Estimation linéaire et redressement sur informations auxiliaires d'enquête par sondages. *Economica*, 1988.

[DHLN92] A. Decarreau, D. Hilhorst, C. Lemaréchal, and J. Navaza. Dual methods in entropy maximization. Application to some problems in crystallography. *SIAM J. Optim.*, 2(2):173–197, 1992.

[DIN09] S. G. Donald, G. W. Imbens, and W. K. Newey. Choosing instrumental variables in conditional moment restriction models. *J. Econometrics*, 152(1):28–36, 2009.

[DJ98] D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 26(3):879–921, 1998.

[DRM96] A. K. Dey, F. H. Ruymgaart, and B. A. Mair. Cross-validation for parameter selection in inverse estimation problems. *Scand. J. Statist.*, 23(4):609–620, 1996.

[DS92] J. C. Deville and C. E. Sárndal. Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*, 87(418):376–382, 1992.

[EHN96] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

[EK01] S. Efromovich and V. Koltchinskii. On inverse problems with unknown operators. *IEEE Trans. Inform. Theory*, 47(7):2876–2894, 2001.

[ES02] S. N. Evans and P. B. Stark. Inverse problems as statistics. *Inverse Problems*, 18(4):R55–R97, 2002.

[ES06] V. M. Estevao and C. E. Sárndal. Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, 74(2):127–147, 2006.

[FLLn06] A. K. Fermin, J. M. Loubes, and C. Ludeña. Bayesian methods for a particular inverse problem seismic tomography. *Int. J. Tomogr. Stat.*, 4(W06):1–19, 2006.

[FLn08] A. K. Fermin and C. Ludeña. A statistical view of iterative methods for linear inverse problems. *TEST*, 17(2):381–400, 2008.

[Flo03] J. P. Florens. Inverse problems and structural econometrics: The example of instrumental variables. *Advances in Economics and Econometrics: Theory and Applications - Eight World Congress*, 36, 2003.

[Ful09] W. A. Fuller. *Sampling Statistics.* Wiley Series in Survey Methodology. John Wiley & Sons Inc, NJ, 2009.

[Gam99] F. Gamboa. New Bayesian methods for ill posed problems. *Statist. Decisions*, 17(4):315–337, 1999.

[GFC$^+$04] R. M. Groves, F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. *Survey methodology.* Wiley Series in Survey Methodology. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2004.

[GG91] F. Gamboa and E. Gassiat. Maximum d'entropie et problème des moments: cas multidimensionnel. *Probab. Math. Statist.*, 12(1):67–83 (1992), 1991.

[GG97] F. Gamboa and E. Gassiat. Bayesian methods and maximum entropy for ill-posed inverse problems. *Ann. Statist.*, 25(1):328–350, 1997.

[GKP99] M. Gabella, V. Kisselev, and G. Perona. Retrieval of aerosol profile variations from reflected radiation in the oxygen absorption a band. *Applied Optics*, 38(15):1390–1395, 1999.

[GLR11] F. Gamboa, J. M. Loubes, and P. Rochet. Maximum entropy estimation for survey sampling. *J. Statist. Plann. Inference*, 141(1):305–317, 2011.

[Goz05] N. Gozlan. Principe conditionnel de gibbs pour des contraintes fines approchées et inégalité de transport. *Thèse*, 2005.

[GZ02] H. Gzyl and N. Zeev. Probabilistic approach to an image reconstruction problem. *Methodol. Comput. Appl. Probab.*, 4(3):279–290 (2003), 2002.

[Gzy95] H. Gzyl. *The method of maximum entropy*, volume 29 of *Series on Advances in Mathematics for Applied Sciences.* World Scientific Publishing Co. Inc., River Edge, NJ, 1995. Sections (6.19)–(6.21) by Aldo Tagliani.

[Gzy02] H. Gzyl. Tomographic reconstruction by maximum entropy in the mean: unconstrained reconstructions. *Appl. Math. Comput.*, 129(2-3):157–169, 2002.

[Han82] L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

[Han87] P. C. Hansen. The truncated SVD as a method for regularization. *BIT*, 27(4):534–553, 1987.

[HHY96] L. P. Hansen, J. Heaton, and A. Yaron. Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*, 14(3):262–80, July 1996.

[HN00] U. Hermann and D. Noll. Adaptive image reconstruction using information measures. *SIAM J. Control Optim.*, 38(4):1223–1240 (electronic), 2000.

[HO93] P. C. Hansen and D. P. O'Leary. The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM J. Sci. Comput.*, 14(6):1487–1503, 1993.

[HR08] M. Hoffmann and M. Reiss. Nonlinear estimation for linear inverse problems with error in the operator. *Ann. Statist.*, 36(1):310–336, 2008.

[HS82] L. P. Hansen and K. J. Singleton. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica*, 50(5):1269–1286, 1982.

[JN00] A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric regression. *Ann. Statist.*, 28(3):681–712, 2000.

[Kit06] Y. Kitamura. Empirical likelihood methods in econometrics: Theory and practice. Cowles Foundation Discussion Papers 1569, Cowles Foundation for Research in Economics, Yale University, 2006.

[KS97] Y. Kitamura and M. Stutzer. An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65(4):861–874, 1997.

[KS02] Y. Kitamura and M. Stutzer. Connections between entropic and linear projections in asset pricing estimation. *J. Econometrics*, 107(1-2):159–174, 2002. Information and entropy econometrics.

[KT04] A. Kaplan and R. Tichatschke. Extended auxiliary problem principle using Bregman distances. *Optimization*, 53(5-6):603–623, 2004.

[LL08] J. M. Loubes and C. Ludeña. Adaptive complexity regularization for linear inverse problems. *Electron. J. Stat.*, 2:661–677, 2008.

[LL10] J. M. Loubes and C. Ludeña. Penalized estimators for non linear inverse problems. *ESAIM Probab. Stat.*, 14:173–191, 2010.

[Lou08] J. M. Loubes. l1 penalty for ill-posed inverse problems. *Comm. Statist. Theory Methods*, 37(8-10):1399–1411, 2008.

[LP08] J. M. Loubes and B. Pelletier. Maximum entropy solution to ill-posed inverse problems with approximately known operator. *J. Math. Anal. Appl.*, 344(1):260–273, 2008.

[LR09a] J. M. Loubes and V. Rivoirard. Review of rates of convergence and regularity conditions for inverse problems. *Int. J. Tomogr. Stat.*, 11(S09):61–82, 2009.

[LR09b] J. M. Loubes and P. Rochet. Regularization with approximated l^2 maximum entropy method. In *submitted, Electronic version HAL 00389698.* 2009.

[McF89] D. McFadden. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57(5):995–1026, 1989.

[Mon87] G. E. Montanari. Post-sampling efficient QR-prediction in large-sample surveys. *Internat. Statist. Rev.*, 55(2):191–202, 1987.

[MR05] G. E. Montanari and M. G. Ranalli. Nonparametric model calibration estimation in survey sampling. *J. Amer. Statist. Assoc.*, 100(472):1429–1442, 2005.

[MRVP97] Gabella M., Guzzi R., Kisselev V., and G. Perona. Retrieval of aerosol profile variations in the visible and near infrared: theory and application of the single-scattering approach. *Applied Optics*, 36(6):1328–1336, 1997.

[Nat01] F. Natterer. *The mathematics of computerized tomography*, volume 32 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001. Reprint of the 1986 original.

[New90] W. K. Newey. Efficient instrumental variables estimation of nonlinear models. *Econometrica*, 58(4):809–837, 1990.

[NS04] W. K. Newey and R. J. Smith. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

[Owe91] A. Owen. Empirical likelihood for linear models. *Ann. Statist.*, 19(4):1725–1747, 1991.

[PR94] Florens J. P. and J.M. Rolin. Bayes, bootstrap, moments. *Discussion Paper 9413, Institut de Statistique, Université catholique de Louvain, Belgium*, 1994.

[QL94] J. Qin and J. Lawless. Empirical likelihood and general estimating equations. *Ann. Statist.*, 22(1):300–325, 1994.

[RB90] Y. Ritov and P. J. Bickel. Achieving information bounds in non and semiparametric models. *Ann. Statist.*, 18(2):925–938, 1990.

[Roc97] R. T. Rockafellar. *Convex analysis*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997. Reprint of the 1970 original, Princeton Paperbacks.

[S0́7] C. E. Sárndal. The calibration approach in survey theory and practice. *Statistics Canada*, 33(2):33–119, 2007.

[Sau] O. Sautory. A new version for the calmar calibration adjustment program. In *Statistics Canada International Symposium Series*.

[SG88] J. A. Scales and A. Gersztenkorn. Robust methods in inverse theory. *Inverse Problems*, 4(4):1071–1091, 1988.

[Sin01] S. Singh. Generalized calibration approach for estimating variance in survey sampling. *Ann. Inst. Statist. Math.*, 53(2):404–417, 2001.

[Ski88] J. Skilling. Maximum entropy spectroscopy—DIMES and MESA. In *Maximum-entropy and Bayesian methods in science and engineering, Vol. 2 (Laramie, WY, 1985 and Seattle, WA, 1986/1987)*, Fund. Theories Phys., pages 127–145. Kluwer Acad. Publ., Dordrecht, 1988.

[Ste81] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 1981.

[TA77] A. N. Tikhonov and V. Y. Arsenin. *Solutions of ill-posed problems*. V. H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York, 1977. Translated from the Russian, Preface by translation editor Fritz John, Scripta Series in Mathematics.

[Thé99] A. Théberge. Extensions of calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*, 94(446):635–644, 1999.

[Tsy03] A. B. Tsybakov. Optimal rates of aggregation. *Proceedings of 16th Annual Conference on Learning Theory (COLT) and 7th Annual Workshop on Kernel Machines. Lecture Notes in Artificial Intelligence*, 2777:303–313, 2003.

[Var73] J. M. Varah. On the numerical solution of ill-conditioned linear systems with applications to ill-posed problems. *SIAM J. Numer. Anal.*, 10:257–267, 1973. Collection of articles dedicated to the memory of George E. Forsythe.

[Var79] J. M. Varah. A practical examination of some numerical methods for linear discrete ill-posed problems. *SIAM Rev.*, 21(1):100–111, 1979.

[vdV98] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[WS01] C. Wu and R. R. Sitter. A model-calibration approach to using complete auxiliary information from survey data. *J. Amer. Statist. Assoc.*, 96(453):185–193, 2001.

[Wu03] C. Wu. Optimal calibration estimators in survey sampling. *Biometrika*, 90(4):937–951, 2003.

[WZ06] C. Wu and R. Zhang. A model-calibration approach to using complete auxiliary information from stratified sampling survey data. *Chinese Quart. J. Math.*, 21(2):309–316, 2006.