École Doctorale de Mathématiques de la région Paris-Sud

Thèse de doctorat

Discipline : Mathématiques

préparée à l'Université Paris Sud par

Lucie Montuelle

Inégalités d'oracle et mélanges

Soutenue le 4 décembre 2014 devant le jury composé de :

M.	Olivier	Catoni	cnrs et ens Paris	Examinateur
M.	Gilles	Celeux	Université Paris-Sud	Examinateur
M.	Serge	Cohen	cnrs et Synchrotron Soleil	Invité
M.	Arnak	Dalalyan	ensae et Université Paris-Est	Rapporteur
M.	Gérard	Govaert	Université de Technologie de Compiègne	Rapporteur
M.	Erwan	Le Pennec	École Polytechnique	Directeur
M.	Pascal	Massart	Université Paris-Sud	Président

Thèse préparée au
Département de Mathématiques d'Orsay
Laboratoire de Mathématiques (UMR 8628), Bât. 425
Université Paris-Sud
91405 Orsay CEDEX

Inégalités d'oracle et mélanges

Ce manuscrit se concentre sur deux problèmes d'estimation de fonction. Pour chacun, une garantie non asymptotique des performances de l'estimateur proposé est fournie par une inégalité d'oracle.

Pour l'estimation de densité conditionnelle, des mélanges de régressions gaussiennes à poids exponentiels dépendant de la covariable sont utilisés. Le principe de sélection de modèle par maximum de vraisemblance pénalisé est appliqué et une condition sur la pénalité est établie. Celle-ci est satisfaite pour une pénalité proportionnelle à la dimension du modèle. Cette procédure s'accompagne d'un algorithme mêlant EM et algorithme de Newton, éprouvé sur données synthétiques et réelles.

Dans le cadre de la régression à bruit sous-gaussien, l'agrégation à poids exponentiels d'estimateurs linéaires permet d'obtenir une inégalité d'oracle en déviation, au moyen de techniques PAC-bayésiennes. Le principal avantage de l'estimateur proposé est d'être aisément calculable. De plus, la prise en compte de la norme infinie de la fonction de régression permet d'établir un continuum entre inégalité exacte et inexacte.

Mots-clefs : Inégalité d'oracle, sélection de modèle, pénalisation, poids exponentiels, apprentissage, agrégation, modèles de mélange, maximum de vraisemblance.

Oracle inequalities and mixtures

This manuscript focuses on two functional estimation problems. A non asymptotic guarantee of the proposed estimator's performances is provided for each problem through an oracle inequality.

In the conditional density estimation setting, mixtures of Gaussian regressions with exponential weights depending on the covariate are used. Model selection principle through penalized maximum likelihood estimation is applied and a condition on the penalty is derived. If the chosen penalty is proportional to the model dimension, then the condition is satisfied. This procedure is accompanied by an algorithm mixing EM and Newton algorithm, tested on synthetic and real data sets.

In the regression with sub-Gaussian noise framework, aggregating linear estimators using exponential weights allows to obtain an oracle inequality in deviation, thanks to PAC-bayesian technics. The main advantage of the proposed estimator is to be easily calculable. Furthermore, taking the infinity norm of the regression function into account allows to establish a continuum between sharp and weak oracle inequalities.

Keywords : Oracle inequality, model selection, penalization, exponential weight, learning, aggregation, mixture model, maximum likelihood.

Remerciements

La reconnaissance est la mémoire du cœur,
Andersen.

Mes premiers remerciements vont à Erwan et Serge, qui m'ont accompagnée pendant le stage et la thèse. L'enthousiasme de Serge et de l'équipe IPANEMA m'a donné envie de poursuivre ma petite aventure dans la recherche pour trois années supplémentaires, sous les conseils avisés d'Erwan. Erwan, merci pour ta disponibilité, ton écoute, ton optimisme féroce et la grande liberté que tu m'as laissée durant ces années, me permettant d'explorer mes envies scientifiques.

Je suis honorée qu'Arnak Dalalyan et Gérard Govaert aient accepté de rapporter ma thèse. Les conseils prodigués ont permis d'améliorer ce manuscrit. Je remercie Pascal Massart d'avoir assumé la présidence de mon jury, et Olivier Catoni et Gilles Celeux d'y avoir gentiment pris part. L'intérêt porté à mes travaux au travers des questions et réflexions est source de motivation supplémentaire pour poursuivre dans cette voie.

Merci Pascal de m'avoir orientée vers Erwan et de m'avoir conseillée quand le besoin s'en faisait sentir. Je tiens aussi à remercier Élisabeth Gassiat et Jean-François Le Gall pour l'attention qu'ils m'ont accordée durant le M2. Étudier et travailler à Orsay a été un plaisir ces dernières années grâce aux membres du laboratoire avec qui j'ai pu interagir. Merci à Nathalie Castelle, Christine Keribin, Claire Lacour, Patrick Pamphile, Thanh Mai Pham Ngoc, et tous ceux qui ont pris de leur temps pour assister à ma soutenance et participer avec enthousiasme au pot.

Qu'auraient été ces années sans la bonne ambiance de bureau assurée par Giancarlo et Simon puis Laure, Lionel, Thierry, Émilien et Zheng ? Merci aussi aux fidèles participants des pauses thé : Tristan, Olivier, Cagri, Élodie, Morzi, Arthur, Vincent(s), Céline, Valérie, Mélina, Alba, Pierre-Antoine, et ceux qui me pardonneront d'avoir oublié de les nommer. Je suis heureuse d'avoir participé à des conférences et écoles d'été mémorables comme Saint-Flour qui m'ont permis d'avoir des compagnons de fortune sympathiques comme Cécile, Benjamin, Baptiste et Erwan, Mélisande et Carole, Sarah et Émilie, Andrés, Sébastien ou Christophe. Ces voyages n'auraient pu être accomplis si facilement sans l'aide de Katia Evrat, Valérie Lavigne et Catherine Ardin, qui m'ont guidée dans les méandres administratifs. Je tiens aussi à souligner l'efficacité et la gentillesse de Christine Bailleul, Nathalie Carrière, Olivier Chaudet et Pascale Starck.

Merci à l'équipe IPANEMA du Synchrotron Soleil : Alessandra, Laurianne, Marie-Angélique, Régina, Loïc, Phil, Sebastian et Serge, pour votre accueil chaleureux.

L'accomplissement de cette thèse a été grandement facilité par les sas de décompression offerts par les soirées parisiennes avec Nicolas, Vincent, Ugo ou, dans un registre plus lyrique, Sébastien. Je souhaite exprimer ma gratitude au noyau dur d'(ex-)orcéens : Yann, Tony, Adrien, Henri et Jérémie pour leur appui en toute circonstance et les soirées entre matheux mais loin des maths. Merci aussi aux amis de prépa Gilles, Rémi, Pécu, Schul et à Solenne toujours présents depuis 10 ans et j'espère pour longtemps encore. Enfin je souhaite exprimer toute ma reconnaissance à François pour son soutien indéfectible, les merveilleux moments passés et tout ce que nous continuons à partager.

Merci à ma famille et particulièrement à mes parents pour leurs prouesses culinaires à l'occasion du pot (et aussi au quotidien !). Vous avez notamment enduré les deux semaines de « vacances » estivales en pleine rédaction sans craquer. Pour cette raison, pour l'éducation que vous vous êtes efforcés de m'offrir, pour votre soutien sans faille, je souhaite vous rendre hommage. Les « hiéroglyphes » qui suivent sont un peu les vôtres. Pour conclure, merci à Clément pour ses petits plats, et tout le reste...

Table des matières

Chapitre 1

État de l'art et contributions

Sommaire

1.1 Problème statistique étudié

Le fil conducteur de cette thèse est l'estimation d'une fonction θ_0 à partir de couples indépendants de données $((X_i, Y_i))_{1 \leq i \leq n}$. Diverses méthodes peuvent être utilisées pour résoudre ce problème statistique. Dans ce panel, il est souhaitable de choisir celle qui fournit la *meilleure* estimation. Cependant, ce choix dépend de la loi des données, rendant l'optimum inconnu. Le but devient alors l'imitation des performances de la meilleure méthode, en s'adaptant aux données. Cela est garanti par un résultat théorique non asymptotique, appelé inégalité d'oracle.

Le problème envisagé ici a été décliné sous deux formes. La première est l'estimation de densité conditionnelle. Y a pour densité θ_0 conditionnellement à X par rapport à la mesure de Lebesgue. La question posée est celle du choix du meilleur estimateur parmi des mélanges de régressions gaussiennes dont les poids exponentiels dépendent de la covariable X. Elle fait l'objet du chapitre 2. Le second cadre

considéré est l'estimation de fonction de régression, dans un modèle sous-gaussien à design fixe : $Y = \theta_0(X) + W$. Ce problème est lié au précédent lorsque la loi du bruit W est connue : un estimateur de la densité conditionnelle peut alors être facilement obtenu à partir de l'estimateur de la fonction de régression. Dans le cas du bruit sous-gaussien, l'estimateur étudié est une moyenne pondérée d'estimateurs par projection des données. Se pose alors la question du choix de ces poids pour l'obtention d'un estimateur aussi performant que la meilleure des projections. Une stratégie est proposée dans le chapitre 3.

En préambule à ces deux objectifs, ce chapitre présente quelques méthodes d'estimation classiques dans les deux cadres considérés, précise le sens de *meilleure* estimation ainsi que le type de résultat espéré, une inégalité d'oracle. Deux stratégies courantes d'adaptation au meilleur estimateur sont exposées, la sélection et l'agrégation, avant d'être détaillées dans les cadres étudiés. Enfin, mon apport est précisé pour chacun de ces deux problèmes.

1.2 Méthodes d'estimation

De nombreux estimateurs peuvent être employés pour estimer une fonction de régression ou une densité conditionnelle [voir Wasserman, 2004]. Le problème de l'estimation de densité conditionnelle a été introduit à la fin des années 60 par Rosenblatt [1969], qui estime alors la loi jointe du couple (X, Y) ainsi que la loi marginale en X à l'aide de noyaux avant d'en faire le quotient. A la même époque, Nadaraya [1965] et Watson [1964] proposent l'estimation de la fonction de régression à l'aide de noyaux. Une revue de la littérature sur cet estimateur est disponible notamment dans le livre de Györfi et al. [2002], qui couvre l'estimation non paramétrique de la fonction de régression. Les estimateurs à noyaux ont été depuis largement étudiés [voir Tsybakov, 2009]. Pour l'estimation de densité conditionnelle, peu de références semblent exister avant le milieu des années 90 et la correction par Hyndman et al. [1996] du biais de l'estimateur proposé par Rosenblatt. Des propriétés de convergence ponctuelle des estimateurs localement polynomiaux (Fan et al. [1996], De Gooijer and Zerom [2003] pour la densité conditionnelle et Stone [1982] pour la régression qui montre aussi la convergence en moyenne quadratique), ou localement logistiques [Hall et al., 1999, Hyndman and Yao, 2002] ont été obtenues et souvent étendues à des données corrélées. Le livre de Fan and Gijbels [1996] passe en revue les propriétés des estimateurs par polynômes locaux et propose des critères de choix de fenêtre, notamment dans le cadre de la régression.

Cependant, pour l'obtention de résultats, le choix de la largeur de fenêtre est crucial mais dépend de la régularité de la fonction à estimer. En pratique, ce choix est rarement discuté, aux exceptions notables de Bashtannyk and Hyndman [2001], Fan and Yim [2004], Hall et al. [2004]. Li and Racine [2007] proposent une explication détaillée des méthodes possibles. van Keilegom and Veraverbeke [2002] ont considéré des extensions au cas de données censurées. Enfin Tsybakov [2009] passe en revue l'ensemble des méthodes d'estimation non paramétrique.

Parmi les méthodes non-paramétriques, les projections dans des bases de Fourier, trigonométriques, d'ondelettes [Donoho et al., 1995], ou de splines [Wahba, 1990] sont couramment utilisées. Ces approches ont les mêmes propriétés que les estimateurs à

noyaux, où l'ordre de la série joue le rôle de fenêtre. Ces estimateurs atteignent les vitesses minimax en régression pour le risque quadratique sur les espaces de Sobolev [Tsybakov, 2009].

Délaissant les noyaux, Stone [1994] propose une modélisation paramétrique de la densité conditionnelle. Il étudie l'estimateur du maximum de vraisemblance basé sur des splines. Cette idée a été reprise par Györfi and Kohler [2007] avec une approche par histogramme, Efromovich [2010] par base de Fourier, et Brunel et al. [2007] et Akakpo and Lacour [2011] par une représentation polynomiale par morceaux. Ces auteurs contrôlent une erreur intégrée : une perte intégrée en variation totale pour le premier et une perte quadratique pour les autres. Brunel et al. [2007] proposent une extension aux données censurées et Akakpo and Lacour [2011] aux données faiblement dépendantes. Blanchard et al. [2007] reprennent l'idée du maximum de vraisemblance dans un cadre de classification à l'aide d'estimateurs par histogrammes, alors que Cohen and Le Pennec [2013] l'appliquent à l'estimation par des polynômes par morceaux.

Stone [1994] applique aussi sa méthode au cas de la régression avec l'estimateur des moindres carrés pour lequel il obtient des résultats similaires à ceux obtenus pour l'estimation de densité conditionnelle, pour la perte quadratique intégrée. L'estimateur des moindres carrés a été le premier introduit pour l'estimation de régression par Gauss et Legendre [1805] pour la détermination des orbites des comètes. Il a depuis été largement utilisé et étudié, notamment dans le cadre des modèles linéaires généralisés [McCullagh and Nelder, 1983]. Lorsque les erreurs sont gaussiennes, c'est un cas particulier d'estimateur du maximum de vraisemblance. Des versions régularisées de cet estimateur ont été proposées comme le lasso [Tibshirani, 1994], ou l'estimateur ridge [Hastie et al., 2009]. Il a été prouvé dans différents cadres que le minimiseur du risque empirique (moins log-vraisemblance ou moindres carrés) a un risque optimal à constante près au sens minimax pour un choix adéquat de modèle [voir Massart and Nédélec, 2006].

Qualité de l'estimation La qualité de ces estimateurs peut être mesurée par une mesure de dissimilarité, comme la norme ℓ_2 ou la divergence de Kullback-Leibler. Pour simplifier, considérons le cas de l'erreur quadratique intégrée, le même phénomène se produisant avec la divergence de Kullback-Leibler. Soit θ un estimateur de la fonction θ_0. Son erreur quadratique intégrée se décompose comme suit :

$$\begin{aligned}\mathbb{E}\left[\|\theta-\theta_0\|^2\right] &= \int \left(\mathbb{E}\left[\theta(x)\right]-\theta_0(x)\right)^2 d\mu(x) + \int \mathbb{E}\left[|\theta(x)-\mathbb{E}\left[\theta(x)\right]|^2\right] d\mu(x)\\ &= \int biais(\theta(x))^2 d\mu(x) + \int Var(\theta(x)) d\mu(x).\end{aligned}$$

Un compromis entre le biais et la variance doit donc être établi afin de minimiser le risque de l'estimateur. Celui-ci peut être illustré par l'exemple de la régression gaussienne homoscédastique à design fixe : $Y = \theta_0 + W$, dans lequel les calculs se font aisément. Considérons les estimateurs $\theta_m = A_m Y$ où A_m est la matrice de projection orthogonale sur l'espace vectoriel S_m. Alors

$$\begin{aligned}\mathbb{E}\left[\|\theta_m-\theta_0\|^2\right] = \|A_m\theta_0-\theta_0\|^2 + \mathbb{E}\left[\|A_m W\|^2\right] &= \|A_m\theta_0-\theta_0\|^2 + \sigma^2 tr(A_m^\top A_m)\\ &= \|A_m\theta_0-\theta_0\|^2 + \sigma^2 \dim(S_m).\end{aligned}$$

Si l'espace S_m est de grande dimension, le biais sera faible mais la variance importante. Inversement, si notre choix se porte sur un estimateur par projection sur un espace de petite dimension, la variance sera faible mais l'erreur d'estimation sera grande. Un équilibre doit être trouvé. Cependant, ce compromis dépend de la fonction à estimer. Il est donc nécessaire de construire une procédure adaptative.

1.3 Oracle, inégalité d'oracle et adaptation

Les estimateurs présentés ci-dessus ont chacun des qualités qui permettent d'espérer une bonne estimation à partir de cette collection sans rien supposer sur la fonction à estimer. L'un des estimateurs va réaliser le risque minimal sur la collection. Or la connaissance de ce risque suppose l'accès à la fonction θ_0 inconnue.

La qualité de l'estimation est évaluée par une perte ℓ dont l'espérance est le risque R. La perte et le risque choisis sont adaptés au contexte. Ainsi, pour l'estimation de densité conditionnelle, la divergence de Kullback-Leibler mesure l'erreur alors que l'erreur quadratique intégrée, portée par la norme ℓ_2, lui a été préférée en régression. De cette façon, selon le cadre étudié, $\ell(\theta_0, \theta)$ peut être $\ln \theta_0 - \ln \theta$ ou $\|\theta_0 - \theta\|_2$. Remarquons que si θ est un estimateur basé sur les données, $R(\theta_0, \theta)$ est une variable aléatoire. Pour souligner cette dépendance, Θ_n désignera l'ensemble des estimateurs candidats, notés θ_n.

Oracle Le *meilleur* estimateur $\theta_n^\star$ de θ_0 parmi l'ensemble des candidats Θ_n est celui qui minimise le risque $R(\theta_0, \theta_n)$:

$$\theta_n^\star \in \arg\min_{\theta_n \in \Theta_n} R(\theta_0, \theta_n).$$

Cependant, il dépend de la loi des données, qui est inconnue. Il est donc inaccessible, mais son risque peut servir de référence. Cet estimateur idéal est appelé oracle, selon la terminologie introduite par Donoho and Johnstone [1998].

Inégalité d'oracle Bien que l'oracle soit inconnu, il est possible de construire un estimateur $\hat{\theta}_n$ qui mime ses performances en terme de risque. Cela peut être garanti dans le cadre asymptotique par l'optimalité asymptotique :

$$\frac{R(\theta_0, \hat{\theta}_n)}{\inf_{\theta_n \in \Theta_n} R(\theta_0, \theta_n)} \xrightarrow[n \to +\infty]{p.s} 1.$$

Nous nous sommes concentrés sur le cadre non-asymptotique, où la garantie théorique est une inégalité d'oracle :

$$R(\theta_0, \hat{\theta}_n) \leq C_n \inf_{\theta_n \in \Theta_n} R(\theta_0, \theta_n) + \epsilon_n,$$

où $C_n \geq 1$ est une quantité déterministe bornée et $\epsilon_n > 0$ est un terme résiduel, souvent négligeable devant $\inf_{\theta_n \in \Theta_n} R(\theta_0, \theta_n)$. Lorsque C_n vaut 1, l'inégalité est dite exacte et $\hat{\theta}_n$ imite l'oracle sur Θ_n en terme de risque. Cette situation est particulièrement intéressante puisqu'elle permet d'évaluer la vitesse minimax. Si $C_n > 1$ pour

tout n, l'estimateur imite seulement la vitesse de convergence de l'oracle. L'inégalité est alors appelée inégalité d'oracle ou inégalité d'oracle approximative. Ces types d'inégalités peuvent être satisfaits en espérance ou avec grande probabilité. Notons que $\hat{\theta}_n$ n'est pas forcément un estimateur de la collection Θ_n. Les inégalités d'oracles établissent des propriétés non asymptotiques d'adaptation à l'oracle et permettent ainsi de transférer les propriétés intéressantes de l'oracle.

Adaptation Le problème que nous nous proposons de résoudre peut être énoncé comme suit. Soit $((X_i, Y_i))_{1 \leq i \leq n}$ un ensemble de couples de données indépendants, de loi inconnue liée à la fonction θ_0. Soit Θ_n une collection d'estimateurs construits à partir des données et $\theta_n \mapsto R(\theta_0, \theta_n)$ un risque pour l'estimation de θ_0. Comment construire un estimateur $\hat{\theta}_n$ tel que $R(\theta_0, \hat{\theta}_n) \approx \inf_{\theta_n \in \Theta_n} R(\theta_0, \theta_n)$ ou $\mathbb{E}\left[R(\theta_0, \hat{\theta}_n)\right] \approx \inf_{\theta_n \in \Theta_n} \mathbb{E}[R(\theta_0, \theta_n)]$?

Une première idée consiste à prendre $\hat{\theta}_n$ dans la collection Θ_n. C'est la sélection. Ce choix dur parmi les estimateurs de la collection a l'avantage d'être facilement interprétable. Cependant, la sélection est peu robuste aux fluctuations, dues au bruit des données. Une légère variation de celles-ci peut mener à la sélection d'un estimateur totalement différent [Yang, 2001]. Vient alors l'idée d'opérer un choix mou, en combinant plusieurs estimateurs de la collection en fonction de la confiance qu'il leur est accordée. Cette stratégie s'appelle l'agrégation.

1.4 Sélection et agrégation

1.4.1 Sélection

Quelques procédures de sélection Plusieurs techniques ont été développées pour sélectionner un modèle, et donc un estimateur. Un panorama de méthodes est disponible dans l'article de Rao and Wu [2001]. Parmi les plus classiques, citons la validation croisée [Stone, 1974, Allen, 1974, Geisser, 1975], qui consiste à découper le jeu de données en un échantillon d'apprentissage sur lequel les estimateurs sont calculés et un échantillon de test permettant d'estimer le risque de chaque estimateur. Cette technique s'adapte facilement à des contextes variés, mais peu de résultats théoriques sont disponibles (voir Arlot and Celisse [2010] pour un état de l'art, Celisse [2008] pour l'estimation de densité et Arlot [2008] pour la régression). Deux autres méthodes usuelles imitent la décomposition biais-variance du risque de l'estimateur : la méthode de Goldenshluger and Lepski [2011] et la sélection par contraste pénalisé. Mentionnons aussi la sélection par test d'hypothèse [Kass and Raftery, 1995, Berger and Pericchi, 1996], qui est intimement liée à l'agrégation bayésienne de modèles [voir Raftery et al., 1997, pour les modèles de régression], avant de détailler la sélection par contraste pénalisé.

Sélection par contraste pénalisé Cette approche a été développée par Barron et al. [1999] (voir aussi Massart [2007]), mais remonte aux années 70, lorsque Akaike [1973] a proposé un critère pénalisé pour la log-vraisemblance dans le cadre de l'estimation de densité, l'AIC, et Mallows [1973] le critère C_p pour la régression avec

les moindres carrés, lorsque la variance des erreurs est supposée connue. Dans les deux cas, la pénalité est proportionnelle au nombre de paramètres du modèle. Par la suite, Schwarz [1978] a proposé le BIC (pénalisation par le nombre de paramètres du modèle avec une approche bayésienne), Tibshirani [1994] le LASSO (pénalisation par la norme ℓ_1), Tikhonov la pénalisation ridge [Hastie et al., 2009] (pénalisation par la norme ℓ_2), puis est apparu l'elastic net [Zou and Hastie, 2005], mélangeant pénalisation ℓ_1 et ℓ_2. La pénalisation ℓ_1 est arrivée naturellement comme relaxation convexe de la norme ℓ_0, permettant de résoudre le problème de minimisation de façon efficace d'un point de vue algorithmique.

Pour la sélection par contraste pénalisé, une collection dénombrable $\mathcal{M}$ de modèles et un contraste γ sont fixés. Le risque R associé à γ est l'espérance du contraste selon la loi des données. Plus précisément, si (X, Y) est un nouveau couple de données indépendantes des précédentes, tiré selon la même loi, alors $R(\theta_0, \theta_n) = \mathbb{E}[\gamma(\theta_n, (X, Y)) - \gamma(\theta_0, (X, Y))]$. Par exemple, le contraste associé à la divergence de Kullback-Leibler est l'opposé du logarithme. Sur chaque modèle m, un minimiseur θ_m du risque empirique est construit. L'idée sous-jacente est d'estimer sans biais le risque de chaque estimateur. L'estimateur choisi $\hat{\theta}$ est celui associé au modèle qui minimise le contraste empirique pénalisé. En notant γ_n le contraste empirique et $\Theta = \{\theta_m, m \in \mathcal{M}\}$, l'estimateur sélectionné $\hat{\theta}$ est $\theta_{\hat{m}}$ tel que

$$\hat{m} \in \arg \inf_{m \in \mathcal{M}} \{\gamma_n(\theta_m) + pen(m)\},$$

où $pen : \mathcal{M} \to \mathbb{R}^+$ est la pénalité. Par exemple, $pen(m)$ vaut $\frac{\ln(n)}{2} D_m$ pour le BIC où D_m est la dimension du modèle m, $\lambda \|\theta_m\|_1$ pour le LASSO, $\lambda \|\theta_m\|_2^2$ pour le ridge et $\lambda_1 \|\theta_m\|_1 + \lambda_2 \|\theta_m\|_2^2$ pour l'elastic net, avec λ, λ_1 et λ_2 des paramètres de régularisation, réels positifs, à calibrer. Remarquons que la pénalisation par la norme ℓ_0 dans une base orthonormée est équivalente au seuillage. Les estimateurs par seuillage sont plus aisément calculables. Des inégalités d'oracle existent pour ces estimateurs ainsi que les autres estimateurs pénalisés précédemment cités [voir entre autres Donoho and Johnstone, 1994, Donoho et al., 1995, Massart, 2007, Massart and Meynet, 2011]

1.4.2 Agrégation

Une alternative à la sélection est l'agrégation. L'estimateur produit n'est plus nécessairement un des estimateurs de la collection Θ, mais une combinaison linéaire de ceux-ci. Le problème n'est plus de choisir un estimateur de la collection, mais suite de coefficients. Cette technique générale de génération d'estimateurs adaptatifs est plus puissante que les techniques classiques, puisqu'elle permet de combiner des estimateurs de natures différentes. Cette idée a été étudiée entre autres par Vovk [1990], Littlestone and Warmuth [1994], Cesa-Bianchi et al. [1997], Cesa-Bianchi and Lugosi [1999]. Elle est au cœur de procédures telles que le bagging [Breiman, 1996], le boosting [Freund, 1995, Schapire, 1990] ou les forêts aléatoires (Amit and Geman [1997] ou Breiman [2001] ; voir plus récemment Biau et al. [2008], Biau and Devroye [2010], Biau [2012], Genuer [2011]) dont les performances expérimentales sont reconnues.

Le cadre général de l'agrégation est détaillé par Nemirovski [2000] et étendu dans les livres de Catoni [2004, 2007] par une approche PAC-Bayésienne, ainsi que les travaux de Yang [2000c,b,a, 2001, 2003, 2004a,b]. Enfin Tsybakov [2008] en dresse un panorama. Dans le cadre de la régression, Tsybakov [2003] a introduit la notion de vitesse optimale d'agrégation pour les familles finies d'estimateurs, liée à des inégalités d'oracle en espérance, et proposé des procédures les atteignant pour l'agrégation linéaire, convexe et l'agrégation de type sélection de modèle. Lounici [2007], Rigollet and Tsybakov [2007], Rigollet [2006] ainsi que Lecué [2007] se sont penchés sur ces problèmes en régression ou pour l'estimation de densité. Remarquons que le choix des coefficients d'agrégation peut être fait en utilisant les méthodes de sélection par contraste pénalisé précédentes avec la perte quadratique. Par exemple, l'estimateur BIC atteint les vitesses optimales pour les problèmes précédents [voir Bunea et al., 2007] tout comme le LASSO pour le problème de sélection [Massart and Meynet, 2011].

Cependant, les estimateurs candidats sont « gelés » dans ces travaux : soit ils sont déterministes, soit ils sont construits à partir d'un échantillon indépendant de celui utilisé pour l'agrégation. Cette hypothèse a été levée par Leung and Barron [2006] qui ont obtenu la première inégalité d'oracle exacte en espérance en agrégeant des projections à l'aide de poids exponentiels, dans le cadre de la régression gaussienne. Ces résultats ont été étendus à un bruit plus général et une famille non dénombrable d'estimateurs potentiellement gelés par Dalalyan and Tsybakov [2007, 2008, 2012]. Ces derniers résultats ont été obtenus à l'aide de poids exponentiels, qui sont un exemple important d'agrégation (voir Rigollet and Tsybakov [2012] pour un historique récent).

Agrégation à poids exponentiels L'agrégation à poids exponentiels se justifie d'un point de vue quasi-bayésien dans le cadre de la régression à design déterministe

$$Y_i = \theta_0(x_i) + \xi_i,$$

avec les ξ_i indépendants de loi normale centrée de variance σ^2. Supposons la variance connue et qu'un dictionnaire de fonctions déterministes $\{\theta_1, \dots, \theta_M\}$ est à notre disposition. L'estimateur agrégé à poids exponentiels peut alors être vu comme l'estimateur de Bayes dans le modèle fantôme

$$Y_i = \sum_{j=1}^{M} w_j \theta_j(x_i) + \xi_i',$$

où $w_j \geq 0$, les ξ_i' sont indépendants de loi normale centrée de variance $\frac{\beta}{2}$, et la loi a priori porte sur les vecteurs de la base canonique de $\mathbb{R}^M$. Ce n'est que *quasi* bayésien car en général, θ_0 n'est pas une combinaison linéaire des θ_j et le réel β doit être supérieur à $4\sigma^2$ pour l'obtention d'inégalités oracles.

L'agrégation à poids exponentiels peut aussi être vue comme une approximation de la sélection des coefficients par minimisation de risque empirique pénalisé. La pénalité employée est la divergence de Kullback-Leibler, favorisant ainsi les distributions proches de la loi a priori, et le contraste est associé à la perte quadratique. Ce problème de minimisation est en général dur à résoudre, c'est pourquoi le risque quadratique de l'estimateur agrégé est majoré par la combinaison convexe des risques de

chaque élément du dictionnaire. Les poids exponentiels sont solution de ce nouveau problème de minimisation sur tous les poids dans le simplexe.

Cette propriété leur a valu d'être largement utilisés pour l'obtention d'inégalités d'oracle via les résultats PAC-bayésiens. L'approche Probably Approximately Correct est employée dans ce contexte pour contrôler en probabilité le risque théorique par le risque empirique pour toute mesure. La technique PAC a été initiée par Shawe-Taylor and Williamson [1997] et McAllester [1998], et améliorée par Seeger [2003] dans le cas des processus gaussiens. Elle a été étendue par Catoni [2004] en classification et en régression avec la perte quadratique où les vitesses atteintes sont optimales au sens minimax dans certains cas. Audibert [2004] a obtenu les premiers résultats adaptatifs, suivi de Catoni [2007]. La technique PAC-bayésienne a aussi été élargie à d'autres contextes.

Lorsque des résultats en espérance sont souhaités, l'estimation sans biais du risque via l'estimateur SURE de Stein [1981] se substitue aux contrôles de déviations. Pour les estimateurs affines, il est possible sous certaines conditions de majorer l'estimateur SURE de l'estimateur agrégé par l'agrégat des estimateurs SURE des estimateurs de la collection [voir Leung, 2004, Dalalyan and Salmon, 2012]. La pénalisation par la divergence de Kullback-Leibler entre la mesure d'agrégation et la loi a priori sur la collection d'estimateurs rend les poids exponentiels optimaux et permet d'avoir une solution explicite de ce problème. La divergence de Kullback-Leibler joue le rôle de borne d'union pour la minimisation sur l'ensemble des mesures d'agrégation.

Voyons maintenant comment la sélection et l'agrégation ont été mises en œuvre dans deux contextes particuliers : l'estimation de densité conditionnelle par des mélanges gaussiens et l'estimation de fonction de régression à l'aide d'estimateurs linéaires respectivement.

1.5 Sélection par contraste pénalisé pour les mélanges gaussiens

1.5.1 Estimation par des mélanges gaussiens

Un modèle couramment utilisé pour l'estimation de densité ou la classification est le mélange gaussien. Il se présente sous la forme d'une moyenne pondérée de densités gaussiennes dont les paramètres sont le nombre de composantes, les poids ainsi que les moyennes et variances de chaque composante. Le contraste employé est alors l'opposé du logarithme et la perte associée est la distance de Kullback-Leibler.

La classification a souvent recourt aux pénalisations AIC et BIC pour le choix du nombre de composantes du mélange [voir Burnham and Anderson, 2002]. Cependant, les heuristiques associées à ces pénalités supposent l'appartenance de la fonction estimée à la collection de modèles. S'affranchissant de cette hypothèse et s'appuyant sur le travail de Massart [2007], Maugis and Michel [2011] ont fourni une pénalité permettant d'obtenir une inégalité d'oracle aussi bien pour la sélection du nombre de composantes que des autres paramètres. Ce résultat repose sur le contrôle de la complexité de chaque modèle par son entropie à crochets, ainsi que de la collection

de modèles par une hypothèse du type inégalité de Kraft.

L'entropie à crochets a aussi été employée dans un cadre bayésien pour obtenir des propriétés asymptotiques de consistance des distributions a posteriori pour les mélanges gaussiens par Choi [2008], et des vitesses de convergence par Genovese and Wasserman [2000] ou van der Vaart and Wellner [1996] lorsque la densité cible est un mélange gaussien.

En présence d'une covariable, il est tentant de complexifier le modèle de mélange gaussien en transformant les moyennes constantes en fonctions. Ce modèle est bien documenté (voir McLachlan and Peel [2000]). En particulier, dans le contexte bayésien, Viele and Tong [2002] ont majoré des entropies à crochets pour prouver la consistance de la loi a posteriori pour les mélanges de régressions gaussiennes. Dans le cadre paramétrique, Young [2014] utilise des mélanges à poids constants de gaussiennes dont les moyennes sont des fonctions continues affines par morceaux (nommées *regression with changepoints*). Il use du maximum de vraisemblance pénalisé par le BIC pour construire un estimateur calculable par un algorithme semblable à l'EM. Une alternative consiste à faire varier les poids en fonction de la covariable. En considérant des poids constants par morceaux et en se basant sur une idée de Kolaczyk et al. [2005], Antoniadis et al. [2009] se sont intéressés à la vitesse de convergence de leur estimateur basé sur la log-vraisemblance pénalisée. Ils ont supposé les composantes gaussiennes connues. Cette hypothèse peut être levée pour l'obtention d'une inégalité d'oracle, comme l'ont montré Cohen and Le Pennec [2013]. Ces modèles sont fréquemment utilisés en économétrie [voir Li and Racine, 2007].

Des mélanges dans lesquels à la fois les moyennes et les poids dépendent d'une covariable sont présentés par Ge and Jiang [2006] mais uniquement pour les mélanges de régressions logistiques. Ils donnent des conditions sur le nombre de composantes du mélange à poids logistiques qui assurent la consistance de la loi a posteriori. Lee [2000] a étudié des propriétés similaires pour les réseaux de neurones.

1.5.2 Estimation par des mélanges de régressions gaussiennes à poids variables

Les mélanges de régressions gaussiennes à poids logistiques ne semblent apparaître dans la littérature qu'au milieu des années 90. Jordan and Jacobs [1994] proposent un algorithme basé sur l'EM et l'algorithme des moindres carrés pondérés (en anglais, Iteratively Reweighted Least Squares) pour estimer les paramètres du modèle de mélange hiérarchisé d'experts, mais ne font pas d'analyse théorique. Dans ce modèle, les composantes et les poids sont des modèles linéaires généralisés. Young and Hunter [2010] considèrent un mélange de régressions gaussiennes à poids variables, pas nécessairement logistiques, estimés par une approche non paramétrique mêlant noyaux et validation croisée. Cette procédure est soutenue par des simulations à l'aide d'un algorithme proche de l'EM, dont les auteurs comparent les performances. Ce travail trouve une extension dans [Hunter and Young, 2012, Huang and Yao, 2012], où sont considérés des mélanges semi-paramétriques de régressions, pour lesquels des conditions pour l'identifiabilité et un algorithme similaire sont donnés. Huang et al. [2013] proposent une estimation non-paramétrique

des moyennes, variances et poids. Ils établissent la normalité asymptotique de l'estimateur construit à l'aide de noyaux et de validation croisée, et l'accompagnent d'un algorithme. Chamroukhi et al. [2010] proposent des mélanges de régressions polynomiales par morceaux à poids logistiques pour l'estimation fonctionnelle. Ils sélectionnent un estimateur par maximum de vraisemblance pénalisé par le critère BIC et fournissent un algorithme, mais pas de justification théorique.

1.5.3 Ma contribution

Les simulations encourageantes de Chamroukhi et al. [2009] sur l'estimation de fonction par des mélanges à poids logistiques de régressions polynomiales par morceaux, nous ont poussé à chercher une garantie théorique. C'est pourquoi nous avons considéré des mélanges de régressions gaussiennes à poids logistiques. Les travaux de Cohen and Le Pennec [2011] esquissaient la possibilité d'estimer des densités conditionnelles par de tels mélanges, y compris avec des régressions gaussiennes par morceaux. Tout comme eux, nous avons obtenu une condition sur la pénalité assurant une inégalité d'oracle en espérance pour l'estimateur sélectionné par maximum de vraisemblance pénalisé. Nous montrons qu'une pénalité proportionnelle à la dimension du modèle convient, ce qui n'avait pas encore été vérifié dans ce cadre.

De plus, la procédure d'estimation est facilement implémentable en combinant les algorithmes EM et de Newton. L'usage de poids logistiques permet une partition souple des données selon la covariable, en autorisant plus d'une régression pour une même valeur de la covariable. Cela se traduit en dimension 1 par des frontières non parallèles aux axes entre les différentes classes lors de la représentation graphique des données. La question cruciale de l'initialisation de l'algorithme EM a été traitée dans le cas de moyennes affines avec des données unidimensionnelles. Elle devient beaucoup plus délicate lorsque la dimension des données augmente, la faute incombant au fléau de la dimension. La procédure que nous proposons a été utilisée pour illustration sur les données de Brinkman [1981] et sur des données génétiques étudiées par Martin-Magniette et al. [2008]. Les données de Brinkman présentent une mesure du mélange air-éthanol utilisé dans le test d'un moteur mono-cylindre et la concentration en monoxyde d'azote des émissions du moteur. L'estimateur sélectionné par notre procédure retrouve des phases connues et interprétables du fonctionnement d'un moteur. Dans le jeu de données génétiques, le but est de trouver des groupes de protéines accrochées aux brins d'ADN. Ces exemples illustrent l'intérêt des mélanges de régressions à poids logistiques puisqu'ils améliorent l'estimation en prenant en compte l'information apportée par la covariable à plusieurs niveaux.

1.6 Agrégation PAC-bayésienne à poids exponentiels

1.6.1 Agrégation d'estimateurs

Dans la plupart des travaux précédemment cités traitant d'agrégation, les résultats portent sur l'agrégation d'estimateurs « gelés » car l'analyse devient trop

complexe, et des astuces de division ou de clonage de l'échantillon [Tsybakov, 2014] sont utilisées. Actuellement, seuls les estimateurs par projection orthogonale [Leung, 2004] et les estimateurs affines [Dalalyan and Salmon, 2012] permettent d'estimer et d'agréger à partir du même échantillon sans sur-apprentissage. Cependant, les inégalités d'oracle obtenues sont en espérance, aux exceptions notables de Audibert [2008] pour les *progressive mixture rules*, Lecué and Mendelson [2009], Gaïffas and Lecué [2011] avec un procédure basée sur la découpe de l'échantillon dans le cas du design aléatoire, et Rigollet [2012] pour le design fixe.

Dai et al. [2012] montrent que les poids exponentiels sont sous-optimaux en déviation : l'espérance du risque quadratique intégré est de l'ordre optimal mais pas les déviations autour de l'espérance. Ils corrigent cela pour la régression gaussienne à design fixe en changeant le problème de minimisation dont les poids exponentiels sont solution. Ils remplacent le risque empirique de l'estimateur agrégé par une combinaison convexe du risque de l'estimateur agrégé et l'agrégat des risques des estimateurs du dictionnaire. Pour la régression gaussienne homoscédastique, Dai et al. [2014] ont obtenu une inégalité d'oracle pour les poids exponentiels et une inégalité d'oracle exacte pour leur nouvelle procédure avec grande probabilité en agrégeant des estimateurs affines.

Les travaux cités en régression supposent la variance connue, aux exceptions de Dalalyan and Salmon [2012] et Dalalyan et al. [2013] dans le cadre hétéroscédastique ; et Belloni et al. [2011], Dalalyan [2012], Giraud [2008], Giraud et al. [2012], Sun and Zhang [2012] qui se placent dans le cadre homoscédastique.

1.6.2 Ma contribution

Parallèlement à ces travaux, nous avons obtenu une inégalité d'oracle en probabilité pour la régression à bruit sous-gaussien en agrégeant des estimateurs linéaires à l'aide de poids exponentiels. L'idée principale consiste à remplacer l'estimateur sans biais du risque traditionnellement utilisé dans les poids par une version pénalisée. Si la norme infinie de la fonction de régression est connue, l'inégalité oracle peut être rendue exacte en prenant en compte le rapport signal sur bruit dans la pénalité. Le lien entre la prise en compte de ce rapport dans la pénalité et le coefficient devant le risque de l'oracle est mis en évidence, établissant un continuum entre inégalité exacte et inexacte. Remarquons que les poids exponentiels proposés par Dai et al. [2014] utilisent un estimateur biaisé du risque et que nous obtenons la même borne sur la température dans le cas gaussien.

Chapter 2

Mixture of Gaussian regressions model with logistic weights, a penalized maximum likelihood approach

This chapter is an extended version of the article [Montuelle and Le Pennec, 2014], published in *Electronic Journal of Statistics*.

Sommaire

Abstract In the framework of conditional density estimation, we use candidates taking the form of mixtures of Gaussian regressions with logistic weights and means depending on the covariate. We aim at estimating the number of components of this mixture, as well as the other parameters, by a penalized maximum likelihood approach. We provide a lower bound on the penalty that ensures an oracle inequality for our estimator. We perform some numerical experiments that support our theoretical analysis.

Keywords : Mixture of Gaussian regressions models, Mixture of regressions models, Penalized likelihood, Model selection.

2.1 Framework

In classical Gaussian mixture models, the density is modeled by

$$s_{K,\upsilon,\Sigma,w}(y) = \sum_{k=1}^{K} \pi_{w,k} \Phi_{\upsilon_k,\Sigma_k}(y),$$

where $K \in \mathbb{N} \setminus \{0\}$ is the number of mixture components, $\Phi_{\upsilon,\Sigma}$ is the Gaussian density with mean υ and covariance matrix Σ,

$$\Phi_{\upsilon,\Sigma}(y) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} e^{\frac{-1}{2}(y-\upsilon)'\Sigma^{-1}(y-\upsilon)}$$

and $\pi_{w,k}$ are the mixture weights, that can always be defined from a K-tuple $w = (w_1, \ldots, w_K)$ with a logistic scheme:

$$\pi_{w,k} = \frac{e^{w_k}}{\sum_{k'=1}^{K} e^{w_{k'}}}.$$

In this article, we consider such a model in which the mixture weights as well as the means can depend on a, possibly multivariate, covariate.

More precisely, we observe n pairs of random variables $((X_i, Y_i))_{1 \leq i \leq n}$ where the covariates X_is are independent while the Y_is are conditionally independent given the X_is. We assume that the covariates are in some subset $\mathcal{X}$ of $\mathbb{R}^d$ and the Y_is

are in $\mathbb{R}^p$. We want to estimate the conditional density $s_0(\cdot|x)$ with respect to the Lebesgue measure of Y given X. We model this conditional density by a mixture of Gaussian regressions with varying logistic weights

$$s_{K,\upsilon,\Sigma,w}(y|x) = \sum_{k=1}^{K} \pi_{w(x),k} \Phi_{\upsilon_k(x),\Sigma_k}(y),$$

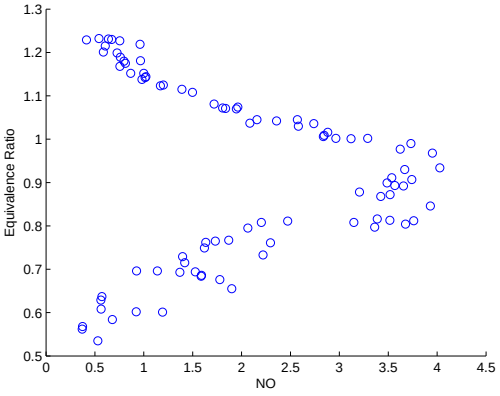
where $\upsilon = (\upsilon_1, \ldots, \upsilon_K)$ and $w = (w_1, \ldots, w_K)$ are now K-tuples of functions chosen, respectively, in a set Υ_K and W_K. Our aim is then to estimate those functions υ_k and w_k, the covariance matrices Σ_k as well as the number of classes K so that the *error* between the estimated conditional density and the true conditional density is *as small as possible*.

The classical Gaussian mixture case has been extensively studied [McLachlan and Peel, 2000]. Nevertheless, theoretical properties of such model have been less considered. In a Bayesian framework, asymptotic properties of the posterior distribution are obtained by Choi [2008], Genovese and Wasserman [2000], van der Vaart and Wellner [1996] when the true density is assumed to be a Gaussian mixture. AIC/BIC penalization scheme are often used to select a number of clusters (see Burnham and Anderson [2002] for instance). Non asymptotic bounds are obtained by Maugis and Michel [2011] even when the true density is not a Gaussian mixture. All these works rely heavily on a *bracketing* entropy analysis of the models, that will also be central in our analysis.

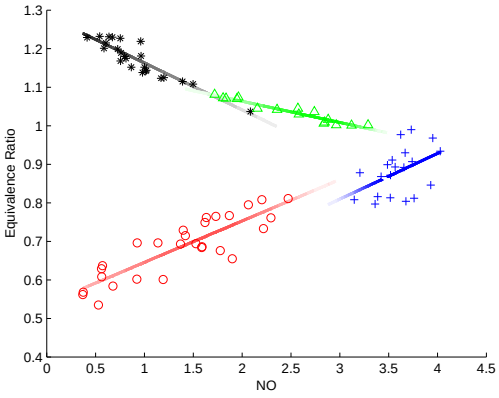
When there is a covariate, the most classical extension of this model is a mixture of Gaussian regressions, in which the means υ_k are now functions. It is well studied as described in McLachlan and Peel [2000]. In particular, in a Bayesian framework, Viele and Tong [2002] have used bracketing entropy bounds to prove the consistency of the posterior distribution. Models in which the proportions vary have been considered by Antoniadis et al. [2009]. Using an idea of Kolaczyk et al. [2005], they have considered a model in which only proportions depend in a piecewise constant manner from the covariate. Their theoretical results are nevertheless obtained under the strong assumption they exactly know the Gaussian components. This assumption can be removed as shown by Cohen and Le Pennec [2013]. Models in which both mixture weights and means depend on the covariate are considered by Ge and Jiang [2006], but in a mixture of logistic regressions framework. They give conditions on the number of components (experts) to obtain consistency of the posterior with logistic weights. Note that similar properties are studied by Lee [2000] for neural networks.

Although natural, mixture of Gaussian regressions with varying logistic weights seems to be mentioned first by Jordan and Jacobs [1994]. They provide an algorithm similar to ours, based on EM and Iteratively Reweighted Least Squares, for hierarchical mixtures of experts but no theoretical analysis. Young and Hunter [2010] choose a non-parametric approach to estimate the weights, which are not supposed logistic anymore, using kernels and cross-validation. They also provide an EM-like algorithm and some convincing simulations. This work has an extension in a series of papers [Hunter and Young, 2012], [Huang and Yao, 2012]. Young [2014] considers mixture of regressions with changepoints but constant proportions. More recently, Huang et al. [2013] have considered a non-parametric modeling for the means, the

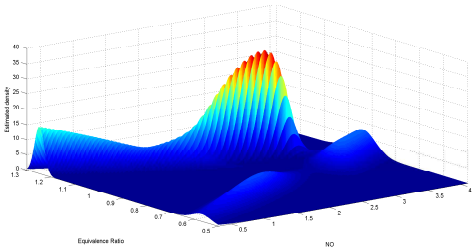
proportions as well as the variance for which they give asymptotic properties as well as a numerical algorithm. Closer to our work, Chamroukhi et al. [2010] consider the case of piecewise polynomial regression model with affine logistic weights. In our setting, this corresponds to a specific choice for Υ_K and W_K: a collection of piecewise polynomials and a set of affine functions. They use a variation of the EM algorithm and a BIC criterion and provide numerical experiments to support the efficiency of their scheme.



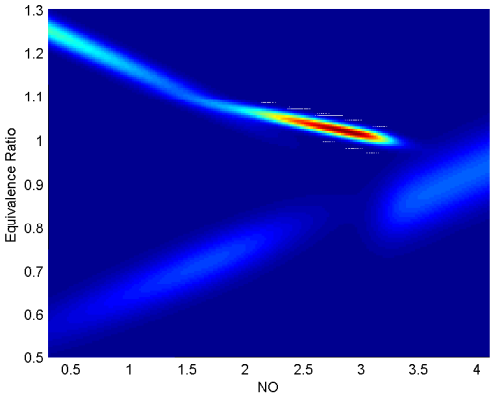
(a) Raw Ethanol data set



(b) Clustering deduced from the estimated conditional density by a MAP principle.



(c) 3D view of the resulting conditional density showing the 4 regression components.



(d) 2D view of the same conditional density. The different variances are visible as well as the connectedness of the two topmost clusters.

Figure 2.1 – Estimated density with 4 components based upon the NO data set

Young [2014] provides a relevant example for our analysis. The ethanol data set of Brinkman [1981] (Figure 2.1a) shows the relationship between the equivalence ratio, a measure of the air-ethanol mix used as a spark-ignition engine fuel in a single-cylinder automobile test, and the engine's concentration of nitrogen oxide (NO) emissions for 88 tests. Using the methodology described in this paper, we obtain a conditional density modeled by a mixture of four Gaussian regressions. Using a classical maximum likelihood approach, each point of the data set can be assigned to one the four class yielding the clustering of Figure 2.1b. The use of logistic

weight allows a soft partitioning along the NO axis while still allowing more than one regression for the same NO value. The two topmost classes seem to correspond to a single population whose behavior changes around 1.7 while the two bottommost classes appear to correspond to two different populations with a gap around $2.6 - 2.9$. Such a result could not have been obtained with non varying weights.

The main contribution of our paper is a theoretical result: an oracle inequality, a non asymptotic bound on the risk, that holds for penalty slightly different from the one used by Chamroukhi et al. [2010].

In Section 2.2, we recall the penalized maximum likelihood framework, introduce the losses considered and explain the meaning of such an oracle inequality. In Section 2.3, we specify the models considered and their collections, state our theorem under mild assumptions on the sets Υ_K and W_K and apply this result to polynomial sets. Those results are then illustrated by some numerical experiments in Section 2.4. Our analysis is based on an abstract theoretical analysis of penalized maximum likelihood approach for conditional densities conducted in Cohen and Le Pennec [2011] that relies on bracketing entropy bounds. Appendix 2.6 summarizes those results while Appendix 2.7 contains the proofs specific to this paper, the ones concerning bracketing entropies.

2.2 A model selection approach

2.2.1 Penalized maximum likelihood estimator

We will use a model selection approach and define some conditional density models S_m by specifying sets of conditional densities, taking the shape of mixtures of Gaussian regressions, through their number of classes K, a structure on the covariance matrices Σ_k and two function sets Υ_K and W_K to which belong respectively the K-tuple of means $(\upsilon_1, \ldots, \upsilon_K)$ and the K-tuple of logistic weights $(w_1, \ldots, w_K)$. Typically those sets are compact subsets of polynomials of low degree. Within such a conditional density set S_m, we estimate s_0 by the maximizer $\widehat{s}_m$ of the likelihood

$$\widehat{s}_m = \operatorname*{argmax}_{s_{K,\upsilon,\Sigma,w} \in S_m} \sum_{i=1}^{n} \ln s_{K,\upsilon,\Sigma,w}(Y_i|X_i),$$

or more precisely, to avoid any existence issue since the infimum may not be unique or even not be reached, by any η-minimizer of the negative log-likelihood:

$$\sum_{i=1}^{n} -\ln \widehat{s}_m(Y_i|X_i) \leq \inf_{s_{K,\upsilon,\Sigma,w} \in S_m} \sum_{i=1}^{n} -\ln s_{K,\upsilon,\Sigma,w}(Y_i|X_i) + \eta.$$

Assume now we have a collection $\{S_m\}_{m \in \mathcal{M}}$ of models, for instance with different number of classes K or different maximum degree for the polynomials defining Υ_K and W_K, we should choose the best model within this collection. Using only the log-likelihood is not sufficient since this favors models with large complexity. To balance this issue, we will define a penalty $\mathrm{pen}(m)$ and select the model $\widehat{m}$ that minimizes (or rather η'-almost minimizes) the sum of the negative log-likelihood and this penalty:

$$\sum_{k=1}^{K} -\ln \widehat{s}_{\widehat{m}}(Y_i|X_i) + \mathrm{pen}(\widehat{m}) \leq \inf_{m \in \mathcal{M}} \sum_{k=1}^{K} -\ln \widehat{s}_m(Y_i|X_i) + \mathrm{pen}(m) + \eta'.$$

2.2.2 Losses

Classically in maximum likelihood context, the estimator loss is measured with the Kullback-Leibler divergence KL. Since we work in a conditional density framework, we use a *tensorized* version of it. We define the tensorized Kullback-Leibler divergence $\mathrm{KL}^{\otimes n}$ by

$$\mathrm{KL}^{\otimes n}(s,t) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}(s(.|X_i), t(.|X_i))\right]$$

which appears naturally in this setting. Replacing t by a convex combination between s and t and dividing by ρ yields the so-called tensorized Jensen-Kullback-Leibler divergence, denoted $\mathrm{JKL}_\rho^{\otimes n}$,

$$\mathrm{JKL}_\rho^{\otimes n}(s,t) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\rho}\mathrm{KL}(s(.|X_i), (1-\rho)s(.|X_i)+\rho t(.|X_i))\right]$$

with $\rho \in (0,1)$. This loss is always bounded by $\frac{1}{\rho}\ln\frac{1}{1-\rho}$ but behaves as KL when t is close to s. This boundedness turns out to be crucial to control the loss of the penalized maximum likelihood estimate under mild assumptions on the complexity of the model and their collection.

Furthermore $\mathrm{JKL}_\rho^{\otimes n}(s,t) \leq \mathrm{KL}_\rho^{\otimes n}(s,t)$. If we let $d^{2\otimes n}$ be the tensorized extension of the squared Hellinger distance d^2, Cohen and Le Pennec [2011] prove that there is a constant C_ρ such that $C_\rho d^{2\otimes n}(s,t) \leq \mathrm{JKL}_\rho^{\otimes n}(s,t)$. Moreover, if we assume that for any $m \in \mathcal{M}$ and any $s_m \in S_m$, $s_0 d\lambda \ll s_m d\lambda$, then

$$\frac{C_\rho}{2+\ln\|s_0/s_m\|_\infty}\mathrm{KL}^{\otimes n}(s_0,s_m) \leq \mathrm{JKL}_\rho^{\otimes n}(s_0,s_m)$$

with $C_\rho = \frac{1}{\rho}\min\left(\frac{1-\rho}{\rho},1\right)\left(\ln\left(1+\frac{\rho}{1-\rho}\right)-\rho\right)$ (see Cohen and Le Pennec [2011]).

2.2.3 Oracle inequality

Our goal is now to define a penalty $\mathrm{pen}(m)$ which ensures that the maximum likelihood estimate in the selected model performs almost as well as the maximum likelihood estimate in the best model. More precisely, we will prove an oracle type inequality

$$\mathbb{E}\left[\mathrm{JKL}_\rho^{\otimes n}(s_0,\widehat{s}_{\widehat{m}})\right] \leq C_1 \inf_{m\in\mathcal{M}}\left(\inf_{s_m\in S_m}\mathrm{KL}^{\otimes n}(s_0,s_m)+\frac{\mathrm{pen}(m)}{n}+\frac{\eta+\eta'}{n}\right)+\frac{C_2}{n}$$

with a $\mathrm{pen}(m)$ chosen of the same order as the variance of the corresponding single model maximum likelihood estimate.

The name oracle type inequality means that the right-hand side is a proxy for the estimation risk of the best model within the collection. The Kullback-Leibler term $\inf_{s_m\in S_m}\mathrm{KL}_\lambda^{\otimes n}(s_0,s_m)$ is a typical bias term while $\frac{pen(m)}{n}$ plays the role of the variance term. We have three sources of loss here: the constant C_1 can not be taken equal to 1, we use a different divergence on the left and on the right and $\frac{pen(m)}{n}$ is not

directly related to the variance. Under a strong assumption, namely a finite upper bound on $\sup_{m\in\mathcal{M}}\sup_{s_m\in S_m}\|s_0/s_m\|_\infty$, the two divergences are *equivalent* for the conditional densities considered and thus the second issue disappears.

The first issue has a consequence as soon as s_0 does not belong to the best model, i.e. when the model is misspecified. Indeed, in that case, the corresponding modeling bias $\inf_{s_m\in S_m}\mathrm{KL}^{\otimes n}(s_0,s_m)$ may be large and the error bound does not converge to this bias when n goes to infinity but to C_1 times this bias. Proving such an oracle inequality with $C_1=1$ would thus be a real improvement.

To our knowledge, those two first issues have not been solved in penalized density estimation with Kullback-Leibler loss but only with L^2 norm or aggregation of a finite number of densities as in Rigollet [2012].

Concerning the third issue, if S_m is parametric, whenever $pen(m)$ can be chosen approximately proportional to the dimension $\dim(S_m)$ of the model, which will be the case in our setting, $\frac{pen(m)}{n}$ is approximately proportional to $\frac{\dim(S_m)}{n}$, which is the asymptotic variance in the parametric case. The right-hand side matches nevertheless the best known bound obtained for a single model within such a general framework.

2.3 Mixtures of Gaussian regressions and penalized conditional density estimation

2.3.1 Models of mixtures of Gaussian regressions

As explained in introduction, we are using candidate conditional densities of type

$$s_{K,\upsilon,\Sigma,w}(y|x)=\sum_{k=1}^{K}\pi_{w,k}(x)\Phi_{\upsilon_k(x),\Sigma_k}(y),$$

to estimate s_0, where $K\in\mathbb{N}\setminus\{0\}$ is the number of mixture components, $\Phi_{\upsilon,\Sigma}$ is the density of a Gaussian of mean υ and covariance matrix Σ, υ_k is a function specifying the mean given x of the k-th component while Σ_k is its covariance matrix and the mixture weights $\pi_{w,k}$ are defined from a collection of K functions $w_1,\dots,w_K$ by a logistic scheme:

$$\pi_{w,k}(x)=\frac{e^{w_k(x)}}{\sum_{k'=1}^{K}e^{w_{k'}(x)}}.$$

We will estimate s_0 by conditional densities belonging to some model S_m defined by

$$S_m=\left\{(x,y)\mapsto\sum_{k=1}^{K}\pi_{w,k}(x)\Phi_{\upsilon_k(x),\Sigma_k}(y)\middle|(w_1,\dots,w_K)\in W_K,\right.$$
$$\left.(\upsilon_1,\dots,\upsilon_K)\in\Upsilon_K,(\Sigma_1,\dots,\Sigma_K)\in V_K\right\}$$

where W_K is a compact set of K-tuples of functions from $\mathcal{X}$ to $\mathbb{R}$, Υ_K a compact set of K-tuples of functions from $\mathcal{X}$ to $\mathbb{R}^p$ and V_K a compact set of K-tuples of

covariance matrices of size $p \times p$. From now on, we will assume that those sets are parametric subsets of dimensions respectively $\dim(W_K)$, $\dim(\Upsilon_K)$ and $\dim(V_K)$. The dimension $\dim(S_m)$ of the now parametric model S_m is thus nothing but $\dim(S_m) = \dim(W_K) + \dim(\Upsilon_K) + \dim(V_K)$.

Before describing more precisely those sets, we recall that S_m will be taken in a model collection $\mathcal{S} = (S_m)_{m \in \mathcal{M}}$, where $m \in \mathcal{M}$ specifies a choice for each of those parameters. Within this collection, the number of components K will be chosen smaller than an arbitrary $K_{\max}$, which may depend on the sample size n. The sets W_K and Υ_K will be typically chosen as a tensor product of a same compact set of moderate dimension, for instance a set of polynomial of degree smaller than respectively d'_W and d'_Υ whose coefficients are smaller in absolute values than respectively T_W and T_Υ.

The structure of the set V_K depends on the *noise* model chosen: we can assume, for instance, it is common to all regressions, that they share a similar volume or diagonalization matrix or they are all different. More precisely, we decompose any covariance matrix Σ into $LPAP'$, where $L = |\Sigma|^{1/p}$ is a positive scalar corresponding to the volume, P is the matrix of eigenvectors of Σ and A the diagonal matrix of normalized eigenvalues of Σ. Let L_-, L_+ be positive values and λ_-, λ_+ real values. We define the set $\mathcal{A}(\lambda_-, \lambda_+)$ of diagonal matrices A such that $|A| = 1$ and $\forall i \in \{1, \dots, p\}, \lambda_- \leq A_{i,i} \leq \lambda_+$. A set V_K is defined by

$$V_K = \{(L_1 P_1 A_1 P_1', \dots, L_K P_K A_K P_K') | \forall k, L_- \leq L_k \leq L_+, P_k \in SO(p), \\ A_k \in \mathcal{A}(\lambda_-, \lambda_+)\},$$

where $SO(p)$ is the special orthogonal group. Those sets V_K correspond to the classical covariance matrix sets described by Celeux and Govaert [1995].

2.3.2 A conditional density model selection theorem

The penalty should be chosen of the same order as the estimator's complexity, which depends on an intrinsic model complexity and, also, a collection complexity.

We will bound the model complexity term using the *dimension* of S_m: we prove that those two terms are roughly proportional under some structural assumptions on the sets W_K and Υ_K. To obtain this result, we rely on an entropy measure of the complexity of those sets. More precisely, for any K-tuples of functions $(s_1, \dots, s_K)$ and $(t_1, \dots, t_K)$, we let

$$d_{\|\sup\|_\infty}((s_1, \dots, s_K), (t_1, \dots, t_K)) = \sup_{x \in \mathcal{X}} \sup_{1 \leq k \leq K} \|s_k(x) - t_k(x)\|_2,$$

and define the metric entropy of a set F_K, $H_{d_{\|\sup\|_\infty}}(\sigma, F_K)$, as the logarithm of the minimal number of balls of radius at most σ, in the sense of $d_{\|\sup\|_\infty}$, needed to cover F_K. We will assume that the parametric dimension D of the set considered coincides with an entropy based definition, namely there exists a constant C such that for $\sigma \in (0, \sqrt{2}]$

$$H_{d_{\|\sup\|_\infty}}(\sigma, F_K) \leq D\left(C + \ln \frac{1}{\sigma}\right).$$

Assumption (DIM) There exist two constants C_W and C_Υ such that, for every sets W_K and Υ_K of the models S_m in the collection $\mathcal{S}$, $\forall \sigma \in (0, \sqrt{2}]$,

$$H_{d_{\|\sup\|_\infty}}(\sigma, W_K) \leq \dim(W_K)\left(C_W + \ln\frac{1}{\sigma}\right)$$

and

$$H_{d_{\|\sup\|_\infty}}(\sigma, \Upsilon_K) \leq \dim(\Upsilon_K)\left(C_\Upsilon + \ln\frac{1}{\sigma}\right)$$

Note that one can extend our result to any compact sets for which those assumptions hold for *dimensions* that could be different from the usual ones.

The complexity of the estimator depends also on the complexity of the collection. That is why one needs further to control the complexity of the collection as a whole through a coding type (Kraft) assumption [Barron et al., 2008].

Assumption (K) There is a family $(x_m)_{m\in\mathcal{M}}$ of non-negative numbers and a real number Ξ such that
$\sum_{m\in\mathcal{M}} e^{-x_m} \leq \Xi < +\infty$.

We can now state our main result, a weak oracle inequality:

Theorem 1. *For any collection of mixtures of Gaussian regressions model $\mathcal{S} = (S_m)_{m\in\mathcal{M}}$ satisfying (K) and (DIM), there is a constant C such that for any $\rho \in (0,1)$ and any $C_1 > 1$, there is a constant κ_0 depending only on ρ and C_1 such that, as soon as for every index $m \in \mathcal{M}$, $pen(m) = \kappa((C + \ln n)\dim(S_m) + x_m)$ with $\kappa > \kappa_0$, the penalized likelihood estimate $\widehat{s}_{\widehat{m}}$ with $\widehat{m}$ such that*

$$\sum_{i=1}^n -\ln(\widehat{s}_{\widehat{m}}(Y_i|X_i)) + pen(\widehat{m}) \leq \inf_{m\in\mathcal{M}}\left(\sum_{i=1}^n -\ln(\widehat{s}_m(Y_i|X_i)) + pen(m)\right) + \eta'$$

satisfies

$$\mathbb{E}\left[\mathrm{JKL}_\rho^{\otimes n}(s_0, \widehat{s}_{\widehat{m}})\right] \leq C_1 \inf_{m\in\mathcal{M}}\left(\inf_{s_m\in S_m} \mathrm{KL}^{\otimes n}(s_0, s_m) + \frac{pen(m)}{n} + \frac{\kappa_0\Xi + \eta + \eta'}{n}\right).$$

Remind that under the assumption that $\sup_{m\in\mathcal{M}}\sup_{s_m\in S_m}\|s_0/s_m\|_\infty$ is finite, $\mathrm{JKL}_\rho^{\otimes n}$ can be replaced by $\mathrm{KL}^{\otimes n}$ up to a multiplication by a constant depending on ρ and the upper bound. Note that this strong assumption is nevertheless satisfied if we assume that $\mathcal{X}$ is compact, s_0 is compactly supported, the regression functions are uniformly bounded and there is a uniform lower bound on the eigenvalues of the covariance matrices.

As shown in the proof, in the previous theorem, the assumption on $\mathrm{pen}(m)$ could be replaced by the milder one

$$\mathrm{pen}(m) \geq \kappa\left(2\dim(S_m)C^2 + \dim(S_m)\left(\ln\frac{n}{C^2\dim(S_m)}\right)_+ + x_m\right).$$

It may be noticed that if $(x_m)_m$ satisfies Assumption (K), then for any permutation τ $(x_{\tau(m)})_m$ satisfies this assumption too. In practice, x_m should be chosen such

that $\frac{2\kappa x_m}{\text{pen}(m)}$ is as small as possible so that the penalty can be seen as proportional to the two first terms. Notice that the constant C only depends on the model collection parameters, in particular on the maximal number of components $K_{\max}$. As often in model selection, the collection may depends on the sample size n. If the constant C grows no faster than $\ln(n)$, the penalty shape can be kept intact and a similar result holds uniformly in n up to a slightly larger κ_0. In particular, the apparent dependency in $K_{\max}$ is not an issue: $K_{\max}$ only appears in C through a logarithmic term and $K_{\max}$ should be taken smaller than n for identifiability issues. Finally, it should be noted that the $\ln n$ term in the penalty of Theorem 1 may not be necessary as hinted by a result of Gassiat and van Handel [2014] for one dimensional mixtures of Gaussian distribution with the same variance.

2.3.3 Linear combination of bounded functions for the means and the weights

We postpone the proof of this theorem to the Appendix and focus on Assumption (DIM). This assumption is easily verified when the function sets W_K and Υ_K are defined as the linear combination of a finite set of bounded functions whose coefficients belong to a compact set. This quite general setting includes the polynomial basis when the covariable are bounded, the Fourier basis on an interval as well as suitably renormalized wavelet dictionaries. Let d_W and d_Υ be two positive integers, let $(\psi_{W,i})_{1\leq i\leq d_W}$ and $(\psi_{\Upsilon,i})_{1\leq i\leq d_\Upsilon}$ two collections of functions bounded functions from $\mathcal{X}\to[-1,1]$ and define

$$W=\left\{w:[0,1]^d\to\mathbb{R}|w(x)=\sum_{i=0}^{d_W}\alpha_i\psi_{W,i}(x)\text{ and }\|\alpha\|_\infty\leq T_W\right\}$$

$$\Upsilon=\left\{\upsilon:[0,1]^d\to\mathbb{R}^p\middle|\forall j\in\{1,\dots,p\},\forall x,\upsilon_j(x)=\sum_{i=0}^{d_\Upsilon}\alpha_i^{(j)}\psi_{\Upsilon,i}(x)\text{ and }\|\alpha\|_\infty\leq T_\Upsilon\right\}$$

where the (j) in $\alpha_r^{(j)}$ is a notation to indicate the link with υ_j. We will be interested in tensorial construction from those sets, namely $W_K=\{0\}\times W^{K-1}$ and $\Upsilon_K=\Upsilon^K$, for which we prove in Appendix that

Lemma 1. *W_K and Υ_K satisfy Assumption (DIM), with $C_W=\ln\left(\sqrt{2}+T_W d_W\right)$ and $C_\Upsilon=\ln\left(\sqrt{2}+\sqrt{p}d_\Upsilon T_\Upsilon\right)$, not depending on K.*

Note that in this general case, only the functions $\psi_{W,i}$ and $\psi_{\Upsilon,i}$ need to be bounded and not the covariate X itself.

For sake of simplicity, we focus on the bounded case and assume $\mathcal{X}=[0,1]^d$. In that case, we can use a polynomial modeling: $\psi_{W,i}$ and $\psi_{\Upsilon,i}$ can be chosen as monomials $x^r=x_1^{r_1}\dots x_d^{r_d}$. If we let d'_W and d'_Υ be two maximum (non negative) degrees for those monomials and define the sets of W_K and Υ_K accordingly, the previous Lemma becomes

Lemma 2. *W_K and Υ_K satisfy Assumption (DIM), with $C_W=\ln\left(\sqrt{2}+T_W\binom{d'_W+d}{d}\right)$ and $C_\Upsilon=\ln\left(\sqrt{2}+\sqrt{p}\binom{d'_\Upsilon+d}{d}T_\Upsilon\right)$, not depending on K.*

To apply Theorem 1, it remains to describe a collection $\mathcal{S} = (S_m)_{m\in\mathcal{M}}$ and a suitable choice for $(x_m)_{m\in\mathcal{M}}$. Assume, for instance, that the models in our collection are defined by an arbitrary maximal number of components $K_{\max}$, a common free structure for the covariance matrix K-tuple and a common maximal degree for the sets W_K and Υ_K. Then one can verify that $\dim(S_m) = (K - 1 + Kp)\binom{d'_W+d}{d} + Kp\frac{p+1}{2}$ and that the weight family $(x_m = K)_{m\in\mathcal{M}}$ satisfy Assumption (K) with $\Xi \leq 1/(e-1)$. Theorem 1 yields then an oracle inequality with $\mathrm{pen}(m) = \kappa\left((C + \ln(n))\dim(S_m) + x_m\right)$. Note that as $x_m \ll (C + \ln(n))\dim(S_m)$, one can obtain a similar oracle inequality with $\mathrm{pen}(m) = \kappa(C+\ln(n))\dim(S_m)$ for a slightly larger κ. Finally, as explained in the proof, choosing a covariance structure from the finite collection of Celeux and Govaert [1995] or choosing the maximal degree for the sets W_K and Υ_K among a finite family can be obtained with the same penalty but with a larger constant Ξ in Assumption (K).

2.4 Numerical scheme and numerical experiment

We illustrate our theoretical result in a setting similar to the one considered by Chamroukhi et al. [2010] and on two real data sets. We observe n pairs (X_i, Y_i) with X_i in a compact interval, namely $[0, 1]$ for simulated data and respectively $[0, 5]$ and $[0, 17]$ for the first and second real data set, and $Y_i \in \mathbb{R}$ and look for the best estimate of the conditional density $s_0(y|x)$ that can be written

$$s_{K,\upsilon,\Sigma,w}(y|x) = \sum_{k=1}^{K} \pi_{w,k}(x)\Phi_{\upsilon_k(x),\Sigma_k}(y),$$

with $w \in W_K$ and $\upsilon \in \Upsilon_K$. We consider the simple case where W_K and Υ_K contain linear functions. We do not impose any structure on the covariance matrices. Our aim is to estimate the *best* number of components K as well as the model parameters. As described with more details later, we use an EM type algorithm to estimate the model parameters for each K and select one using the penalized approach described previously.

2.4.1 The procedure

As often in model selection approach, the first step is to compute the maximum likelihood estimate for each number of components K. To this purpose, we use a numerical scheme based on the EM algorithm [Dempster et al., 1977] similar to the one used by Chamroukhi et al. [2010]. The only difference with a classical EM is in the Maximization step since there is no closed formula for the weights optimization. We use instead a Newton type algorithm. Note that we only perform a few Newton steps (5 at most were enough in our experiments) and ensure that the likelihood does not decrease. We have noticed that there is no need to fully optimize at each step: we did not observe a better convergence and the algorithmic cost is high. We denote from now on this algorithm *Newton-EM*. Notice that the lower bound on the variance required in our theorem appears to be necessary in practice. It avoids the spurious local maximizer issue of EM algorithm, in which a class degenerates

to a minimal number of points allowing a perfect Gaussian regression fit. We use a lower bound shape of $\frac{C}{n}$. Biernacki and Castellan [2011] provide a precise data-driven bound for mixture of Gaussian regressions: $\frac{\min_{1\leq i<j\leq n}(Y_i-Y_j)^2}{2\chi^2_{n-2K+1}((1-\alpha)^{1/K})}$, with χ^2_{n-2K+1} the chi-squared quantile function, which is of the same order as $\frac{1}{n}$ in our case. In practice, the constant 10 gave good results for the simulated data.

An even more important issue with EM algorithms is initialization, since the local minimizer obtained depends heavily on it. We observe that, while the weights w do not require a special care and can be simply initialized uniformly equal to 0, the means require much more attention in order to obtain a good minimizer. We propose an initialization strategy based on short runs of *Newton-EM* with random initialization.

We draw randomly K lines, each defined as the line going through two points (X_i, Y_i) drawn at random among the observations. We perform then a K-means clustering using the distance along the Y axis. Our *Newton-EM* algorithm is initialized by the regression parameters as well as the empirical variance on each of the K clusters. We perform then 3 steps of our minimization algorithm and keep among 50 trials the one with the largest likelihood. This winner is used as the initialization of a final *Newton-EM* algorithm using 10 steps.

We consider two other strategies: a *naive* one in which the initial lines chosen at random and a common variance are used directly to initialize the *Newton-EM* algorithm and a *clever* one in which observations are first normalized in order to have a similar variance along both the X and the Y axis, a K-means on both X and Y with 5 times the number of components is then performed and the initial lines are drawn among the regression lines of the resulting cluster containing more than 2 points.

The complexity of those procedures differs and as stressed by Celeux and Govaert [1995] the fairest comparison is to perform them for the same amount of time (5 seconds, 30 seconds, 1 minute...) and compare the obtained likelihoods. The difference between the 3 strategies is not dramatic: they yield very similar likelihoods. We nevertheless observe that the *naive* strategy has an important dispersion and fails sometime to give a satisfactory answer. Comparison between the *clever* strategy and the regular one is more complex since the difference is much smaller. Following Celeux and Govaert [1995], we have chosen the regular one which corresponds to more random initializations and thus may explore more local maxima.

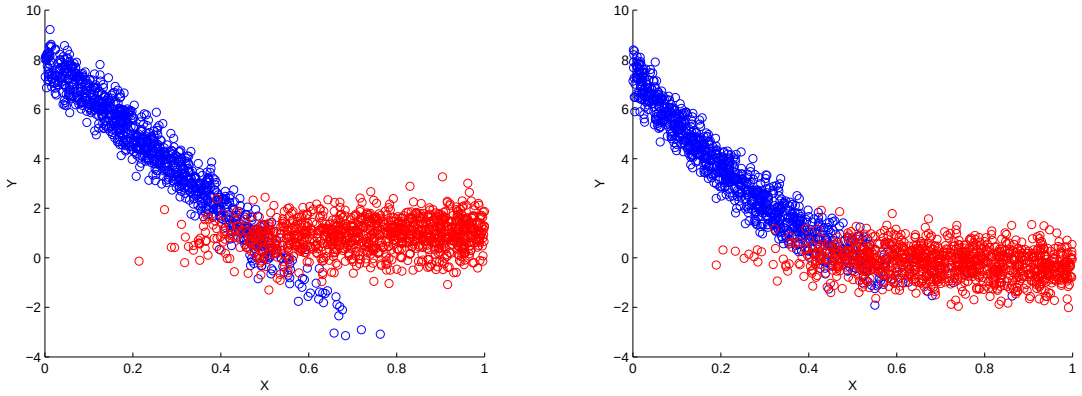
Once the parameters' estimates have been computed for each K, we select the model that minimizes

$$\sum_{i=1}^{n} -\ln(\widehat{s}_m(Y_i|X_i)) + \text{pen}(m)$$

with $\text{pen}(m) = \kappa \dim(S_m)$. Note that our theorem ensures that there exists a κ large enough for which the estimate has good properties, but does not give an explicit value for κ. In practice, κ has to be chosen. The two most classical choices are $\kappa = 1$ and $\kappa = \frac{\ln n}{2}$ which correspond to the AIC and BIC approach, motivated by asymptotic arguments. We have used here the slope heuristic proposed by Birgé and Massart [2007] and described for instance in Baudry et al. [2011]. This heuristic comes with two possible criterions: the jump criterion and the slope criterion. The

first one consists in representing the dimension of the selected model according to κ (Fig 2.3), and finding $\hat{\kappa}$ such that if $\kappa < \hat{\kappa}$, the dimension of the selected model is large, and reasonable otherwise. The slope heuristic prescribes then the use of $\kappa = 2\hat{\kappa}$. In the second one, one computes the *asymptotic* slope of the log-likelihood drawn according to the model dimension, and penalizes the log-likelihood by twice the slope times the model dimension. With our simulated data sets, we are in the not so common situation in which the jump is strong enough so that the first heuristic can be used.

2.4.2 Simulated data sets



(a) 2 000 data points of example WS

(b) 2 000 data points of example MS

Figure 2.2 – Typical realizations

The previous procedure has been applied to two simulated data sets: one in which true conditional density belongs to one of our models, a *well-specified* case, and one in which this is not true, a *misspecified* case. In the first situation, we expect to perform almost as well as the maximum likelihood estimation in the true model. In the second situation, we expect our algorithm to automatically balance the model bias and its variance. More precisely, we let

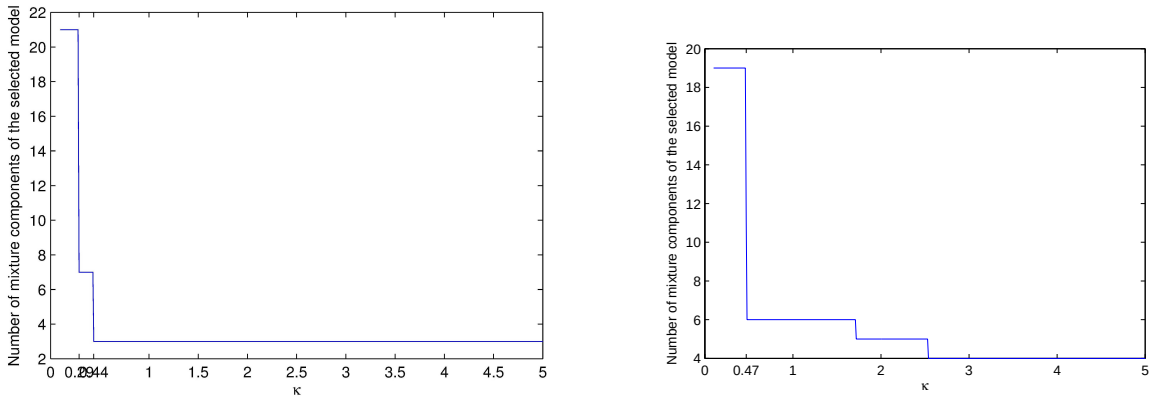
$$s_0(y|x) = \frac{1}{1+\exp(15x-7)}\Phi_{-15x+8,0.3}(y) + \frac{\exp(15x-7)}{1+\exp(15x-7)}\Phi_{0.4x+0.6,0.4}(y)$$

in the first example, denoted example WS, and

$$s_0(y|x) = \frac{1}{1+\exp(15x-7)}\Phi_{15x^2-22x+7.4,0.3}(y) + \frac{\exp(15x-7)}{1+\exp(15x-7)}\Phi_{-0.4x^2,0.4}(y)$$

in the second example, denoted example MS. For both experiments, we let X be uniformly distributed over $[0,1]$. Figure 2.2 shows a typical realization.

In both examples, we have noticed that the sample's size had no significant influence on the choice of κ, and that very often 1 was in the range of possible values indicated by the jump criterion of the slope heuristic. According to this observation, we have chosen in both examples $\kappa = 1$.



(a) Example WS with 2 000 points

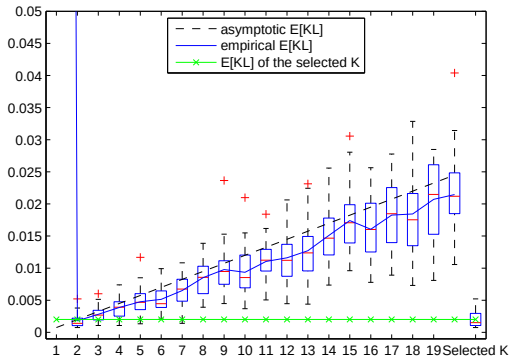
(b) Example MS with 2 000 points

Figure 2.3 – Slope heuristic: plot of the selected model dimension with respect to the penalty coefficient κ. In both examples, $\widehat{\kappa}$ is of order 1/2.

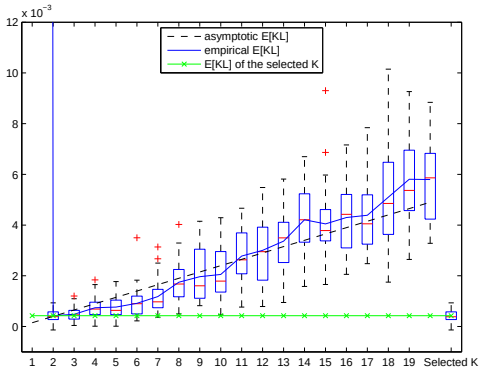
We measure performances in term of tensorized Kullback-Leibler divergence. Since there is no known formula for tensorized Kullback-Leibler divergence in the case of Gaussian mixtures, and since we know the true density, we evaluate the divergence using Monte Carlo method. The variability of this randomized approximation has been verified to be negligible in practice.

For several numbers of mixture components and for the selected K, we draw in Figure 2.4 the box plots and the mean of tensorized Kullback-Leibler divergence over 55 trials. The first observation is that the mean of tensorized Kullback-Leibler divergence between the penalized estimator $\hat{s}_{\hat{K}}$ and s_0 is smaller than the mean of tensorized Kullback-Leibler divergence between $\hat{s}_K$ and s_0 over $K \in \{1, \ldots, 20\}$. This is in line with the oracle type inequality of Theorem 1. Our numerical results hint that our theoretical analysis may be pessimistic. A close inspection shows that the bias-variance trade-off differs between the two examples. Indeed, since in the first one the true density belongs to the model, the best choice is $K = 2$ even for large n. As shown on the histogram of Figure 2.5, this is almost always the model chosen by our algorithm. Observe also that the mean of Kullback-Leibler divergence seems to behave like $\frac{\dim(S_m)}{2n}$ (shown by a dotted line). This is indeed the expected behavior when the true model belongs to a nested collection and corresponds to the classical AIC heuristic. In the second example, the misspecified one, the true model does not belong to the collection. The best choice for K should thus balance a model approximation error term and a variance one. We observe in Figure 2.5 such a behavior: the larger n the more complex the model and thus K. Note that the slope of the mean error seems also to grow like $\frac{\dim(S_m)}{2n}$ even though there is no theoretical guarantee of such a behavior.

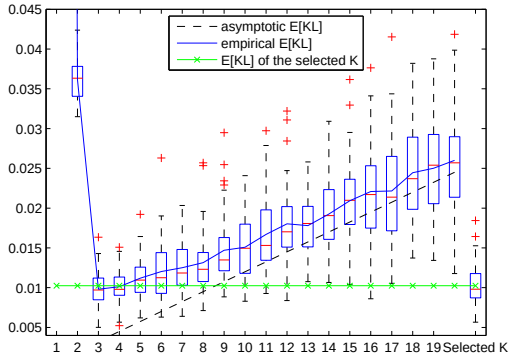
Figure 2.6 shows the error decay when the sample size n grows. As expected in the well-specified case, example W, we observe the decay in t/n predicted in the theory, with t some constant. The rate in the second case appears to be slower. Indeed, as the true conditional density does not belong to any model, the selected models are more and more complex when n grows which slows the error decay. In our theoretical analysis, this can already be seen in the decay of the *variance* term



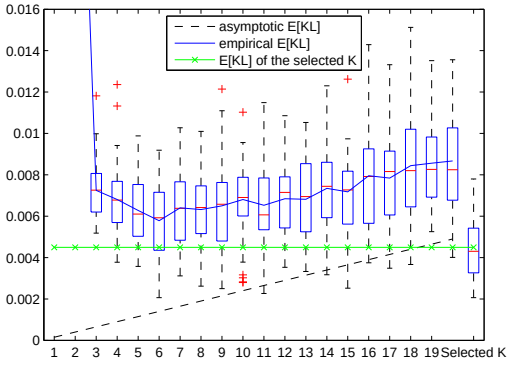
(a) Example WS with 2 000 data points



(b) Example WS with 10 000 data points



(c) Example MS with 2 000 data points



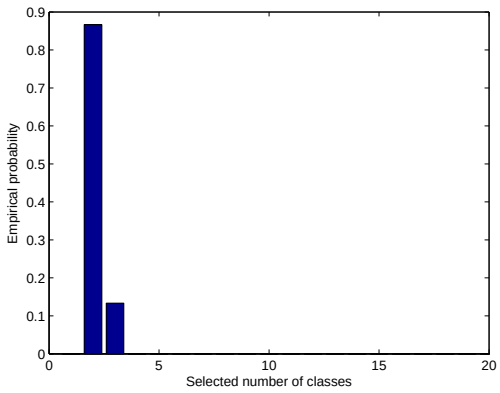
(d) Example MS with 10 000 data points

Figure 2.4 – Box-plot of the Kullback-Leibler divergence according to the number of mixture components. On each graph, the right-most box-plot shows this Kullback-Leibler divergence for the penalized estimator $\hat{s}_{\widehat{K}}$

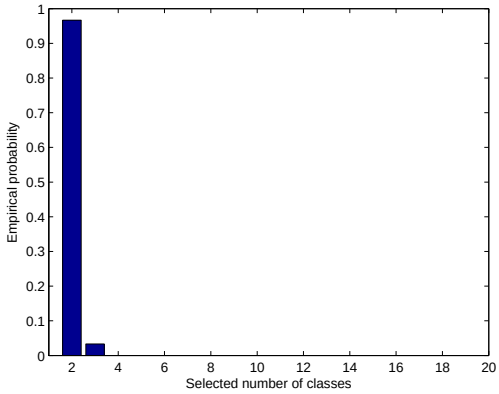
of the oracle inequality. Indeed, if we let $m_0(n)$ be the optimal oracle model, the one minimizing the right-hand side of the oracle inequality, the variance term is of order $\frac{\dim(S_{m_0(n)})}{n}$ which is larger than $\frac{1}{n}$ as soon as $\dim(S_{m_0(n)}) \to +\infty$. It is well known that the decay depends on the regularity of the true conditional density. Providing a minimax analysis of the proposed estimator, as have done Maugis and Michel [2012], would be interesting but is beyond the scope of this paper.

2.4.3 Ethanol data set

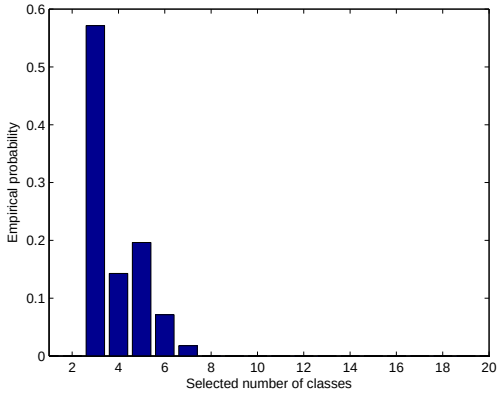
We explain now with more details the result of Figure 2.1 for the 88 data point Ethanol data set of Brinkman [1981]. Young [2014] proposes to estimate the density of the equivalence ratio R conditioned to the concentration in NO and to use this conditional density to do a clustering of the data set. In our framework, this amounts



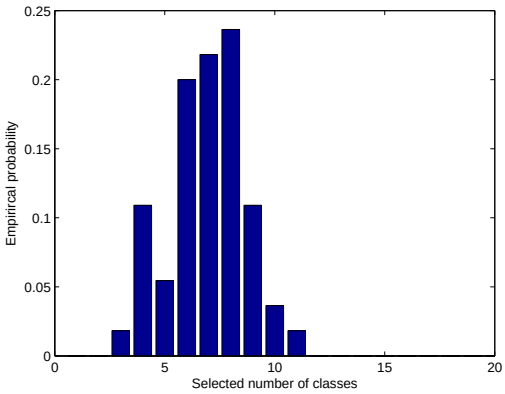
(a) Example WS with 2 000 data points



(b) Example WS with 10 000 data points



(c) Example MS with 2 000 data points



(d) Example MS with 10 000 data points

Figure 2.5 – Histograms of the selected K

to estimate the conditional density by

$$\sum_{k=1}^{\widehat{K}} \pi_{\widehat{w_k}(NO)} \Phi_{\widehat{\upsilon_k}(NO),\widehat{\Sigma_k}}(R)$$

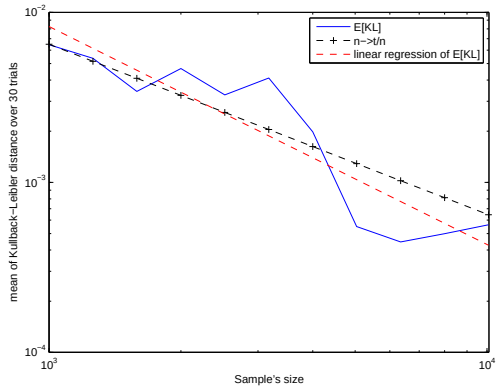
with our proposed penalized estimator and to use the classical maximum likelihood approach that associates (NO, R) to the class

$$\arg \max_{1 \leq k \leq \widehat{K}} \pi_{\widehat{w_k}(NO)} \Phi_{\widehat{\upsilon_k}(NO),\widehat{\Sigma_k}}(R)$$

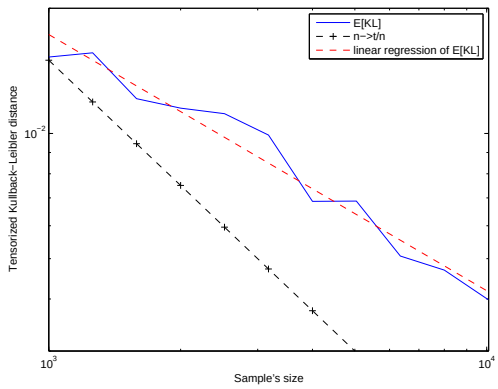
to perform the clustering.

An important parameter of the method is the lower bound of the variance used in the estimation for a given number of class. This is required to avoid spurious maximizers of the likelihood. Here, the value 10^{-4} chosen *by hand* yields satisfactory results.

Since we only have 88 points and roughly 5 parameters per class, the random initialization may yield classes with too few points to have a good estimation. We have slightly modified our K-means procedure in order to ensure than at least 10



(a) Example WS. The slope of the free regression line is $\simeq -1,3$



(b) Example MS. The slope of the regression line is $\simeq -0,6$.

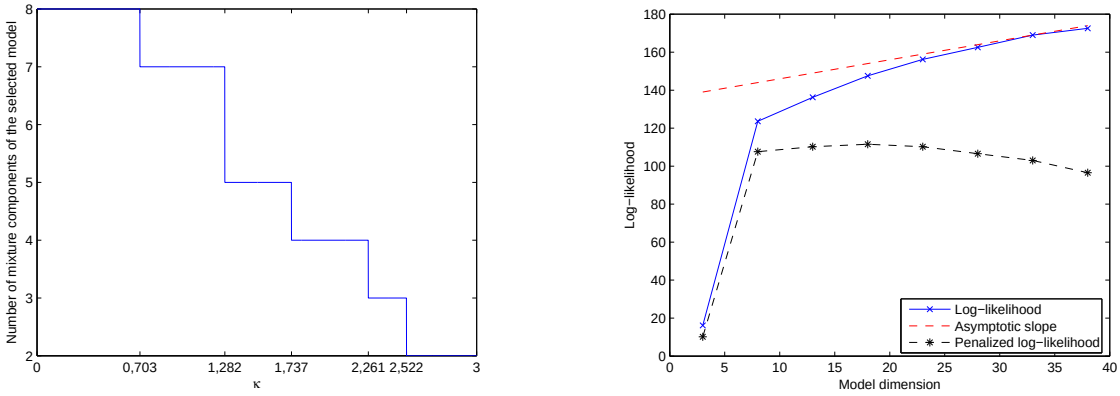
Figure 2.6 – Kullback-Leibler divergence between the true density and the computed density using $(X_i, Y_i)_{i\leq N}$ with respect to the sample size, represented in a log-log scale. For each graph, we added a free linear least-square regression and one with slope -1 to stress the two different behavior.

points are assigned to each class. In that case, we have verified that the estimated parameters of the conditional density were very stable.

Note that with this strategy, no more than 8 classes can be considered. This prevents the use of the jump criterion to calibrate the penalty because the *big* jump is hard to define. We use instead the slope heuristic. Figure 2.7 shows that this slope is of order 1 and thus the slope heuristic prescribes a penalty of $2\dim(S_K)$, providing an estimate with 4 components.

It is worth pointing out that the maximum of the penalized likelihood is not sharp, just like in the example MS of simulated data (see figure 2.5). Indeed, it is quite unlikely that the true density belongs to our model collection. So, there may be an uncertainty on the selected number of components between 4, 3 and 5. Note that AIC penalization would have lead to 7 classes while BIC would also have lead to 4 classes. Our estimated penalty is nevertheless in the middle of the zone corresponding to 4 while BIC is nearby the boundary with 3 and thus we expect this choice to be more stable. In Figure 2.1b of the introduction we have shown only this clustering with 4 classes. Figure 2.8 shows that the choices of 3 or 5 may make sense, even though the choice 5 may seem slightly too complex. A common feature among all those clusterings is the change of slope in the topmost part around 1.7. This phenomena is also visible in Young [2014] in which an explicit change point model is used, ours is only implicit and thus more versatile

To complete our study, in Figure 2.9, we have considered the more natural regression of NO with respect to the equivalence ratio that has not been studied by Young [2014]. Using the same methodology, we have recovered also 4 clusters corresponding to a soft partitioning of the equivalence ratio value. Note that this clustering, which is easily interpretable, is very similar to the one obtained with the previous parameterization.



(a) Jump criterion

(b) Slope criterion

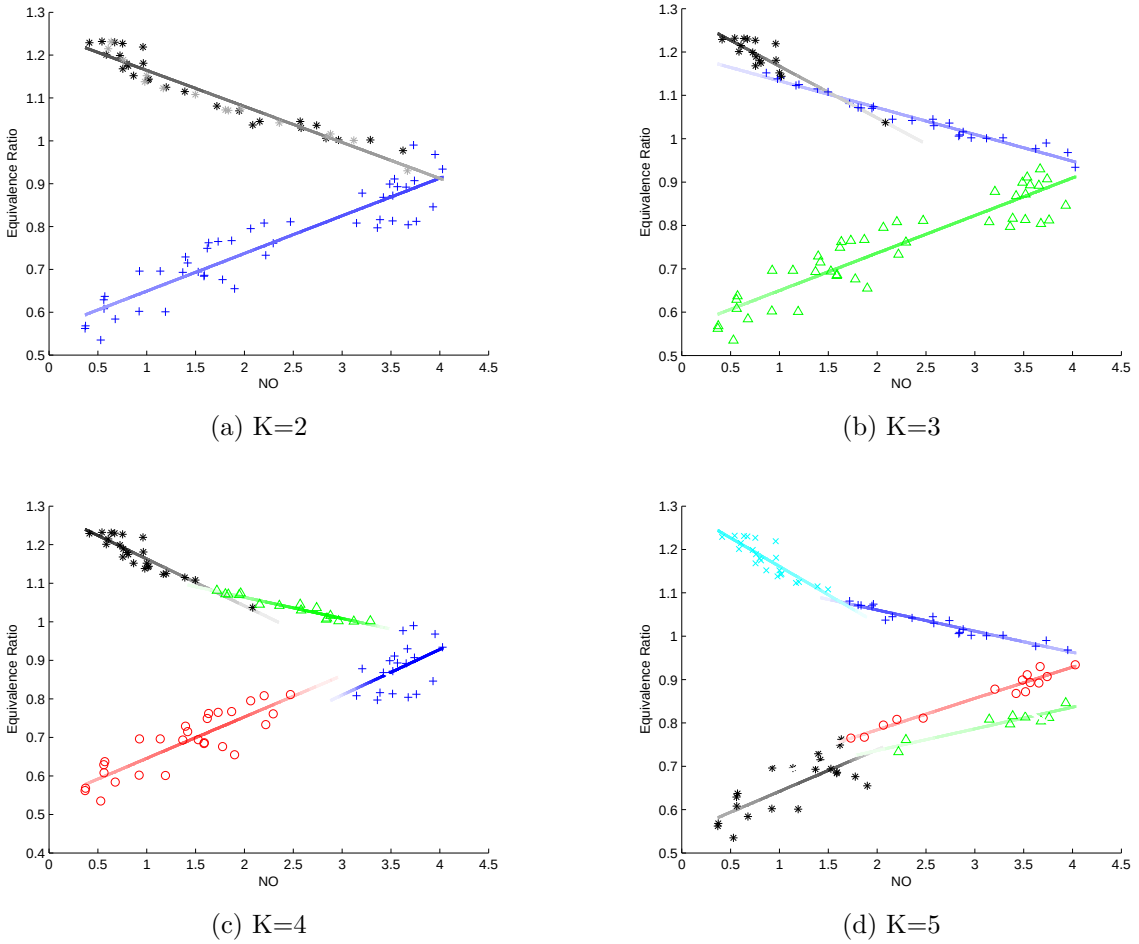
Figure 2.7 – Slope heuristic for the ethanol data set

2.4.4 ChIP-chip data set

We consider here a second real data set: a Chromatin immunoprecipitation (ChIP) on chip genomic data set. Chromatin immunoprecipitation (ChIP) is a procedure used to investigate proteins associated with DNA. The data set considered is the one used by Martin-Magniette et al. [2008]. In this experiment, two variables are studied: DNA fragments crosslinked to a protein of interest (IP) and genomic DNA (Input). Martin-Magniette et al. [2008] model the density of log-IP conditioned to log-Input by a mixture of two Gaussian regressions with the same variance. One component corresponds to an enriched one, in which there is more proteins than expected, and the other to a normal one. They use classical proportions that do not depend on the Input. The parameters are estimated using the EM algorithm initialized by values derived from a Principal Component Analysis of the whole data set. The best model between one and two components is selected according to the BIC criterion. For the histone modification in Arabidopsis thaliana data set, they select a two components model similar to the one obtained with logistic weights (Figure 2.10).

We have first compared the constant proportions model with $K = 2$ to the one proposed in their conclusion in which the proportions depend on the Input. We have used our affine logistic weights model and observed that this model greatly improves the log-likelihood. The dimension of this new model is 8 while the dimension of the original model is 7 so that the log-likelihood increase does not seem to be due to overfitting. We have also compared our solution to the one obtained with a constant weights model with $K = 3$, of dimension 11. The BIC criterion selects the $K = 2$ with affine weights solution.

We have then tested more complex models with K up to 20 with a penalty obtained with the slope heuristic. The models chosen are quite complex ($K = 10$ for constant proportions models and $K = 7$ for affine logistic weight models, the later being the overall winner). Although they better explain the data from the statistical point of view, those models become hard to interpret from the biological point of view. We think this is due to the too simple affine models used. Although



(a) K=2 (b) K=3

(c) K=4 (d) K=5

Figure 2.8 – Clustering of NO data set into K classes. The strength of the color of the regression lines corresponds to the mixture proportion.

no conceptual difficulties occur by using more complex function familie (or going to the multivariate setting), the *curse of dimensionality* makes everything more complicated in practice. In particular, initialization becomes harder and harder as the dimension grows and requires probably a more clever treatment than the one proposed here. In the spirit of Cohen and Le Pennec [2013], we are currently working on a first extension: a numerical algorithm for a bivariate piecewise linear logistic weights model applied to hyperspectral image segmentation.

2.5 Discussion

We have studied a penalized maximum likelihood estimate for mixtures of Gaussian regressions with logistic weights. Our main contribution is the proof that a penalty proportional, up to a logarithmic factor of the sample size, to the dimension of the model is sufficient to obtain a non asymptotic theoretical control on the estimator loss. This result is illustrated in the simple univariate case in which both the means and the logistic weights are linear. We study a toy model which exhibits the

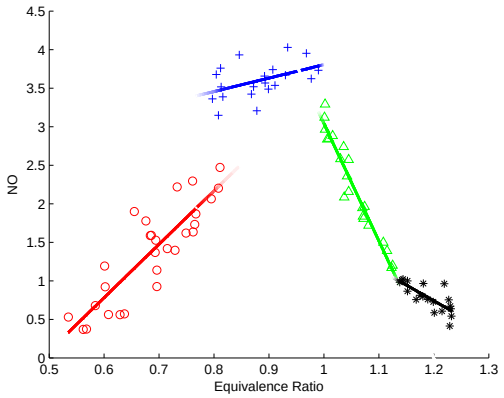


Figure 2.9 – Clustering of NO data set into 4 classes, considering the regression of NO with respect to the equivalence ratio

behavior predicted by our theoretical analysis and proposes two simple applications of our methodology. We hope that our contribution helps to popularize those mixtures of Gaussian regressions by giving a theoretical foundation for model selection technique in this area and showing some possible interesting uses even for simple models.

Besides some important theoretical issues on the loss used and the tightness of the bounds, the major future challenge is the extension of the numerical scheme to more complex cases than univariate linear models.

2.6 Appendix : A general conditional density model selection theorem

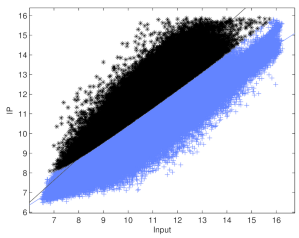
We summarize in this section the main result of Cohen and Le Pennec [2011] that will be our main tool to obtain the previous oracle inequality.

To any model S_m, a set of conditional densities, we associate a complexity defined in term of a specific entropy, the bracketing entropy with respect to the square root of the tensorized square of the Hellinger distance $d^{2\otimes n}$. Recall that a bracket $[t^-, t^+]$ is a pair of real functions such that $\forall (x,y) \in \mathcal{X}\times\mathcal{Y}, t^-(x,y) \leq t^+(x,y)$ and a function s is said to belong to the bracket $[t^-, t^+]$ if $\forall (x,y) \in \mathcal{X}\times\mathcal{Y}, t^-(x,y) \leq s(x,y) \leq t^+(x,y)$. The bracketing entropy $H_{[],d}(\delta, S)$ of a set S is defined as the logarithm of the minimal number $N_{[],d}(\delta, S)$ of brackets $[t^-, t^+]$ covering S, such that $d(t^-, t^+) \leq \delta$. The main assumption on models is a property that should satisfies the bracketing entropy:

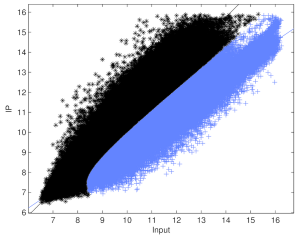
Assumption (H) For every model S_m in the collection $\mathcal{S}$, there is a non-decreasing function ϕ_m such that $\delta \mapsto \frac{1}{\delta}\phi_m(\delta)$ is non-increasing on $(0,+\infty)$ and for every $\sigma \in \mathbb{R}^+$,

$$\int_0^\sigma \sqrt{H_{[.],d^{\otimes n}}(\delta, S_m)}d\delta \leq \phi_m(\sigma).$$

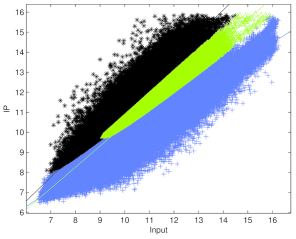
Such an integral is ofter called a Dudley type integral of these bracketing entropies and is commonly used in empirical process theory [van der Vaart and Wellner, 1996].



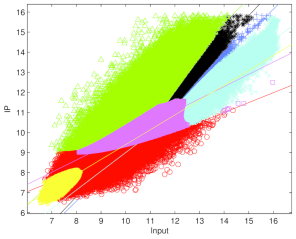
(a) K=2, constant proportions, dimension= 7



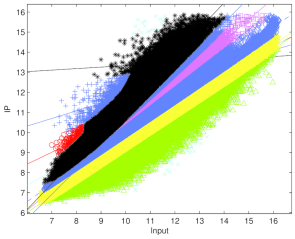
(b) K=2, affine logistic weights, dimension= 8



(c) K=3, constant proportions, dimension= 9



(d) K=7, affine logistic weights, dimension= 33



(e) K=10, constant proportions, dimension= 39

Figure 2.10 – Clustering of ChIP-chip data set into K classes.

The complexity of S_m is then defined as $n\sigma_m^2$ where σ_m is the unique square root of $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{n}\sigma$.

For technical reason, a separability assumption, always satisfied in the setting of this paper, is also required. It is a mild condition, classical in empirical process theory (see for instance van der Vaart and Wellner [1996]).

Assumption (Sep) For every model S_m in the collection $\mathcal{S}$, there exists some countable subset S'_m of S_m and a set $\mathcal{Y}'_m$ with $\lambda(\mathcal{Y}\backslash\mathcal{Y}'_m) = 0$ such that for every t in S_m, there exists some sequence $(t_k)_{k\geq 1}$ of elements of S'_m such that for every $x \in \mathcal{X}$ and every $y \in \mathcal{Y}'_m, \ln(t_k(y|x)) \xrightarrow[k\to+\infty]{} \ln(t(y|x))$.

The main result of Cohen and Le Pennec [2011] is a condition on the penalty pen(m) which ensures an oracle type inequality:

Theorem 2. *Assume we observe (X_i, Y_i) with unknown conditional density s_0. Let $\mathcal{S} = (S_m)_{m\in\mathcal{M}}$ an at most countable conditional density model collection. Assume Assumptions (H), (Sep) and (K) hold. Let $\widehat{s}_m$ be a η minimizer of the negative log-likelihood in S_m*

$$\sum_{i=1}^{n} -\ln(\widehat{s}_m(Y_i|X_i)) \leq \inf_{s_m\in S_m}\left(\sum_{i=1}^{n} -\ln(s_m(Y_i|X_i))\right) + \eta$$

Then for any $\rho \in (0,1)$ and any $C_1 > 1$, there is a constant κ_0 depending only on ρ and C_1 such that, as soon as for every index $m \in \mathcal{M}$,

$$pen(m) \geq \kappa(n\sigma_m^2 + x_m)$$

with $\kappa > \kappa_0$ and σ_m the unique square root of $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{n}\sigma$, the penalized likelihood estimate $\widehat{s}_{\widehat{m}}$ with $\widehat{m}$ such that

$$\sum_{i=1}^{n} -\ln(\widehat{s}_{\widehat{m}}(Y_i|X_i)) + pen(\widehat{m}) \leq \inf_{m\in\mathcal{M}} \left(\sum_{i=1}^{n} -\ln(\widehat{s}_m(Y_i|X_i)) + pen(m) \right) + \eta'$$

satisfies

$$\mathbb{E}\left[\mathrm{JKL}_{\rho}^{\otimes n}(s_0, \widehat{s}_{\widehat{m}})\right] \leq C_1 \inf_{m\in\mathcal{M}} \left(\inf_{s_m\in S_m} \mathrm{KL}^{\otimes n}(s_0, s_m) + \frac{pen(m)}{n} \right) + C_1 \frac{\kappa_0 \Xi + \eta + \eta'}{n}.$$

In the next section, we show how to apply this result in our mixture of Gaussian regressions setting and prove that the penalty can be chosen roughly proportional to the intrinsic dimension of the model, and thus of the order of the variance.

2.7 Appendix : Proofs

In Appendix 2.7.1, we give a proof of Theorem 1 relying on several bracketing entropy controls proved in Appendix 2.7.2.

2.7.1 Proof of Theorem 1

We will show that Assumption (DIM) ensures that for all $\delta \in (0, \sqrt{2}]$, $H_{[.],d^{\otimes n}}(\delta, S_m) \leq \dim(S_m)(\mathfrak{C} + \ln(\frac{1}{\delta}))$ with a common $\mathfrak{C}$.

We show in Appendix that if

Assumption (DIM) There exist two constants C_W and C_Υ such that, for every model S_m in the collection $\mathcal{S}$,

$$H_{d_{\|\sup\|_\infty}}(\sigma, W_K) \leq \dim(W_K)\left(C_W + \ln\frac{1}{\sigma}\right)$$

and

$$H_{d_{\|\sup\|_\infty}}(\sigma, \Upsilon_K) \leq \dim(\Upsilon_K)\left(C_\Upsilon + \ln\frac{1}{\sigma}\right)$$

then, if $n \geq 1$, the complexity of the corresponding model S_m satisfies for any $\delta \in (0, \sqrt{2}]$,

$$H_{[.],d^{\otimes n}}(\delta, S_m) \leq \dim(S_m)(\mathfrak{C} + \ln(\frac{1}{\delta}))$$

with $\dim(S_m) = \dim(W_K) + \dim(\Upsilon_K) + \dim(V_K)$ and $\mathfrak{C}$ that depends only on the constants defining V_K and the constants C_W and C_Υ.

If this happens, Proposition 1 yields the results.

Proposition 1. *If for any $\delta \in (0, \sqrt{2}]$, $H_{[.],d^{\otimes n}}(\delta, S_m) \leq \dim(S_m)(C_m + \ln(\frac{1}{\delta}))$, then the function $\phi_m(\sigma) = \sigma\sqrt{\dim(S_m)}\left(\sqrt{C_m} + \sqrt{\pi} + \sqrt{\ln\left(\frac{1}{\min(\sigma,1)}\right)}\right)$ satisfies Assumption (H). Furthermore, the unique square root σ_m of $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{n}\sigma$ satisfies*

$$n\sigma_m^2 \leq \dim(S_m)\left(2(\sqrt{C_m} + \sqrt{\pi})^2 + \left(\ln\frac{n}{(\sqrt{C_m} + \sqrt{\pi})^2 \dim(S_m)}\right)_+\right).$$

In other words, if we can control models' bracketing entropy with a uniform constant $\mathfrak{C}$, we get a suitable bound on the complexity. This result will be obtain by first decomposing the entropy term between the weights and the Gaussian components. Therefore we use the following distance over conditional densities:

$$\sup_x d_y(s,t) = \sup_{x\in\mathcal{X}} \left(\int_y \left(\sqrt{s(y|x)} - \sqrt{t(y|x)} \right)^2 dy \right)^{\frac{1}{2}}.$$

Notice that $d^{2\otimes n}(s,t) \leq \sup_x d_y^2(s,t)$.

For all weights π and π', we define

$$\sup_x d_k(\pi,\pi') = \sup_{x\in\mathcal{X}} \left(\sum_{k=1}^{K} \left(\sqrt{\pi_k(x)} - \sqrt{\pi'_k(x)} \right)^2 \right)^{\frac{1}{2}}.$$

Finally, for all densities s and t over $\mathcal{Y}$, depending on x, we set

$$\begin{aligned}\sup_x \max_k d_y(s,t) =& \sup_{x\in\mathcal{X}} \max_{1\leq k\leq K} d_y(s_k(x,.),t_k(x,.)) \\ =& \sup_{x\in\mathcal{X}} \max_{1\leq k\leq K} \left(\int_y \left(\sqrt{s_k(x,y)} - \sqrt{t_k(x,y)} \right)^2 dy \right)^{\frac{1}{2}}.\end{aligned}$$

Lemma 3. *Let* $\mathcal{P} = \left\{ (\pi_{w,k})_{1\leq k\leq K} \middle| w \in W_K, \text{ and } \forall (k,x), \pi_{w,k}(x) = \frac{e^{w_k(x)}}{\sum_{l=1}^{K} e^{w_l(x)}} \right\}$ *and* $\mathcal{G} = \left\{ (\Phi_{\upsilon_k,\Sigma_k})_{1\leq k\leq K} \,|\upsilon \in \Upsilon_K, \Sigma \in V_K \right\}$. *Then for all* δ *in (0,*$\sqrt{2}$*], for all* m *in* $\mathcal{M}$*,*

$$H_{[.],\sup_x d_y}(\delta, S_m) \leq H_{[.],\sup_x d_k}\left(\frac{\delta}{5},\mathcal{P}\right) + H_{[.],\sup_x \max_k d_y}\left(\frac{\delta}{5},\mathcal{G}\right).$$

One can then relate the bracketing entropy of $\mathcal{P}$ to the entropy of W_K

Lemma 4. *For all* $\delta \in (0,\sqrt{2}]$*,*

$$H_{[.],\sup_x d_k}\left(\frac{\delta}{5},\mathcal{P}\right) \leq H_{d_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}\delta}{20\sqrt{K}}, W_K\right)$$

Since $\mathcal{P}$ is a set of weights, $\frac{3\sqrt{3}\delta}{20\sqrt{K}}$ could be replaced by $\frac{3\sqrt{3}\delta}{20\sqrt{K-1}}$ with an identifiability condition. For example, $W'_K = \{(0, w_2 - w_1, \ldots, w_K - w_1) | w \in W_K\}$ can be covered using brackets of null size on the first coordinate, lowering squared Hellinger distance between the brackets' bounds to a sum of $K-1$ terms. Therefore, $H_{[.],\sup_x d_k}\left(\frac{\delta}{5},\mathcal{P}\right) \leq H_{d_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}\delta}{20\sqrt{K-1}}, W'_K\right)$.

Since we have assumed that $\exists C_W$ s.t $\forall \delta \in (0,\sqrt{2}]$,

$$H_{d_{\|\sup\|_\infty}}(\delta, W_K) \leq \dim(W_K)\left(C_W + \ln\left(\frac{1}{\delta}\right)\right)$$

Then

$$H_{[.],\sup_x d_k}\left(\frac{\delta}{5},\mathcal{P}\right) \leq \dim(W_K)\left(C_W + \ln\left(\frac{20\sqrt{K}}{3\sqrt{3}\delta}\right)\right)$$

To tackle the Gaussian regression part, we rely heavily on the following proposition,

Proposition 2. *Let* $\kappa \geq \frac{17}{29}$, $\gamma_\kappa = \dfrac{25(\kappa - \frac{1}{2})}{49(1 + \frac{2\kappa}{5})}$. *For any* $0 < \delta \leq \sqrt{2}$ *and any* $\delta_\Sigma \leq \frac{1}{5\sqrt{\kappa^2 \cosh(\frac{2\kappa}{5}) + \frac{1}{2}}} \frac{\delta}{p}$, $(\upsilon, L, A, P) \in \Upsilon \times [L_-, L_+] \times \mathcal{A}(\lambda_-, \lambda_+) \times SO(p)$ *and* $(\tilde{\upsilon}, \tilde{L}, \tilde{A}, \tilde{P}) \in \Upsilon \times [L_-, L_+] \times \mathcal{A}(\lambda_-, +\infty) \times SO(p), \Sigma = LPAP'$ *and* $\tilde{\Sigma} = \tilde{L}\tilde{P}\tilde{A}\tilde{P}'$, *assume that* $t^-(x,y) = (1 + \kappa\delta_\Sigma)^{-p} \Phi_{\tilde{\upsilon}(x),(1+\delta_\Sigma)^{-1}\tilde{\Sigma}}(y)$ *and* $t^+(x,y) = (1 + \kappa\delta_\Sigma)^{p} \Phi_{\tilde{\upsilon}(x),(1+\delta_\Sigma)\tilde{\Sigma}}(y)$.

If

$$\begin{cases} \forall x \in \mathbb{R}^d, \|\upsilon(x) - \tilde{\upsilon}(x)\|^2 \leq p\gamma_\kappa L_- \lambda_- \frac{\lambda_-}{\lambda_+} {\delta_\Sigma}^2 \\ (1 + \frac{2}{25}\delta_\Sigma)^{-1}\tilde{L} \leq L \leq \tilde{L} \\ \forall 1 \leq i \leq p, |A_{i,i}^{-1} - \tilde{A}_{i,i}^{-1}| \leq \frac{1}{10}\frac{\delta_\Sigma}{\lambda_+} \\ \forall y \in \mathbb{R}^p, \|Py - \tilde{P}y\| \leq \frac{1}{10}\frac{\lambda_-}{\lambda_+}\delta_\Sigma \|y\| \end{cases}$$

then $[t^-, t^+]$ *is a* $\frac{\delta}{5}$ *Hellinger bracket such that* $t^-(x,y) \leq \Phi_{\upsilon(x),\Sigma}(y) \leq t^+(x,y)$.

We consider three cases: the parameter (mean, volume, matrix) is known ($\star = 0$), unknown but common to all classes ($\star = c$), unknown and possibly different for every class ($\star = K$). For example, $[\nu_K, L_0, P_c, A_0]$ denotes a model in which only means are free and eigenvector matrices are assumed to be equal and unknown. Under our assumption that $\exists C_\Upsilon$ s.t $\forall \delta \in (0, \sqrt{2}]$,

$$H_{d_{\|\sup\|_\infty}}(\delta, \Upsilon_K) \leq \dim(\Upsilon_K) \left(C_\Upsilon + \ln\left(\frac{1}{\delta}\right)\right)$$

we deduce:

$$H_{[.],\max_k \sup_x d_y}\left(\frac{\delta}{5}, \mathcal{G}\right) \leq \mathcal{D}\left(\mathcal{C} + \ln\left(\frac{1}{\delta}\right)\right) \tag{2.1}$$

where $\mathcal{D} = Z_{\upsilon,\star} + Z_{L,\star} + \dfrac{p(p-1)}{2} Z_{P,\star} + (p-1) Z_{A,\star}$ and

$$\begin{aligned} \mathcal{C} = \ln\left(5p\sqrt{\kappa^2 \cosh\left(\frac{2\kappa}{5}\right) + \frac{1}{2}}\right) + \frac{Z_{\upsilon,\star} C_\Upsilon}{\mathcal{D}} + \frac{Z_{\upsilon,\star}}{2\mathcal{D}} \ln\left(\frac{\lambda_+}{p\gamma_\kappa L_- \lambda_-^2}\right) \\ + \frac{Z_{L,\star}}{\mathcal{D}} \ln\left(\frac{4 + 129 \ln\left(\frac{L_+}{L_-}\right)}{10}\right) + \frac{Z_{P,\star}}{\mathcal{D}}\left(\ln(c_U) + \frac{p(p-1)}{2} \ln\left(\frac{10\lambda_+}{\lambda_-}\right)\right) \\ + \frac{Z_{A,\star}(p-1)}{\mathcal{D}} \ln\left(\frac{4}{5} + \frac{52\lambda_+}{5\lambda_-} \ln\left(\frac{\lambda_+}{\lambda_-}\right)\right)\right) \end{aligned}$$

$$\begin{aligned} &Z_{\upsilon,K} = \dim(\Upsilon_K), Z_{\upsilon,c} = \dim(\Upsilon_1), Z_{\upsilon,0} = 0 && Z_{L,0} = Z_{P,0} = Z_{A,0} = 0, \\ &Z_{L,c} = Z_{P,c} = Z_{A,c} = 1, && Z_{L,K} = Z_{P,K} = Z_{A,K} = K. \end{aligned}$$

We notice that the following upper-bound of $\mathcal{C}$ is independent from the model

of the collection, because we have made this hypothesis on C_Υ.

$$\mathcal{C} \leq \ln\left(5p\sqrt{\kappa^2\cosh\left(\frac{2\kappa}{5}\right)+\frac{1}{2}}\right) + C_\Upsilon + \frac{1}{2}\ln\left(\frac{\lambda_+}{p\gamma_\kappa L_-\lambda_-^2}\right) + \ln\left(\frac{4+129\ln\left(\frac{L_+}{L_-}\right)}{10}\right) + \frac{2}{p(p-1)}\ln(c_U) + \ln\left(\frac{10\lambda_+}{\lambda_-}\right) + \ln\left(\frac{4}{5}+\frac{52\lambda_+}{5\lambda_-}\ln\left(\frac{\lambda_+}{\lambda_-}\right)\right) := \mathcal{C}_1.$$

We conclude that $H_{[.],\sup_x d_y}(\delta, S_m) \leq \dim(S_m)\left(C_m + \ln\left(\frac{1}{\delta}\right)\right)$, with

$$\begin{aligned}\dim(S_m) &= \dim(W_K) + \mathcal{D}\\ C_m &= \frac{\dim(W_K)}{\dim(S_m)}\left(C_W + \ln\left(\frac{20\sqrt{K}}{3\sqrt{3}}\right)\right) + \frac{\mathcal{D}\mathcal{C}_1}{\dim(S_m)}\\ &\leq C_W + \ln\left(\frac{20\sqrt{K_{\max}}}{3\sqrt{3}}\right) + \mathcal{C}_1 := \mathfrak{C}\end{aligned}$$

Note that the constant $\mathfrak{C}$ does not depend on the dimension $\dim(S_m)$ of the model, thanks to the hypothesis that C_W is common for every model S_m in the collection. Using Proposition 1, we deduce thus that

$$n\sigma_m^2 \leq \dim(S_m)\left(2\left(\sqrt{\mathfrak{C}}+\sqrt{\pi}\right)^2 + \left(\ln\frac{n}{\left(\sqrt{\mathfrak{C}}+\sqrt{\pi}\right)^2\dim(S_m)}\right)_+\right).$$

Theorem 2 yields then, for a collection $\mathcal{S} = (S_m)_{m\in\mathcal{M}}$, with $\mathcal{M} = \{(K, W_K, \Upsilon_K, V_K) | K \in \mathbb{N}\backslash\{0\}, W_K, \Upsilon_K, V_K$ as previously defined $\}$ for which Assumption (K) holds, the oracle inequality of Theorem 1 as soon as

$$\mathrm{pen}(m) \geq \kappa\left(\dim(S_m)\left(2\left(\sqrt{\mathfrak{C}}+\sqrt{\pi}\right)^2 + \left(\ln\frac{n}{\left(\sqrt{\mathfrak{C}}+\sqrt{\pi}\right)^2\dim(S_m)}\right)_+\right) + x_m\right).$$

2.7.2 Lemma Proofs

Bracketing entropy's decomposition

We prove here a slightly more general Lemma than Lemma 3.

Lemma 5. *Let*

$$\begin{aligned}\mathcal{P} &= \left\{\pi = (\pi_k)_{1\leq k\leq K} \middle| \forall k, \pi_k : \mathcal{X} \to \mathbb{R}^+ \text{ and } \forall x \in \mathcal{X}, \sum_{k=1}^K \pi_k(x) = 1\right\},\\ \Psi &= \left\{(\psi_1, \ldots, \psi_K) \middle| \forall k, \psi_k : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}^+, \text{ and } \forall x, \forall k, \int \psi_k(x,y)dy = 1\right\},\\ \mathcal{C} &= \left\{(x,y) \mapsto \sum_{k=1}^K \pi_k(x)\psi_k(x,y) \middle| \pi \in \mathcal{P}, \psi \in \Psi\right\}.\end{aligned}$$

Then for all δ in $(0, \sqrt{2}]$,

$$H_{[.],\sup_x d_y}(\delta, \mathcal{C}) \leq H_{[.],\sup_x d_k}\left(\frac{\delta}{5}, \mathcal{P}\right) + H_{[.],\sup_x \max_k d_y}\left(\frac{\delta}{5}, \Psi\right).$$

The proof mimics the one of Lemma 7 from Cohen and Le Pennec [2011]. It is possible to obtain such an inequality if the covariate X is not bounded, using the smaller distance $d^{\otimes n}$ for the entropy with bracketing of $\mathcal{C}$. More precisely,

Lemma 6. *For all δ in $(0, \sqrt{2}]$, $H_{[.],d^{\otimes n}}(\delta, \mathcal{C}) \leq H_{[.],d_{\mathcal{P}}}\left(\frac{\delta}{2}, \mathcal{P}\right) + H_{[.],d_\Psi}\left(\frac{\delta}{2}, \Psi\right)$,*

with $d^2_{\mathcal{P}}(\pi^+, \pi^-) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n d_k^2(\pi^+(X_i), \pi^-(X_i))\right]$ and

$d^2_\Psi(\psi^+, \psi^-) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \sum_{k=1}^K d_y^2(\psi_k^+(X_i), \psi_k^-(X_i))\right]$. But bounding such bracketing entropies for $\mathcal{P}$ and Ψ becomes much more challenging.

Proof. First we will exhibit a covering of bracket of $\mathcal{C}$.

Let $([\pi^{i,-}, \pi^{i,+}])_{1\leq i\leq N_{\mathcal{P}}}$ be a minimal covering of δ bracket for $\sup_x d_k$ of $\mathcal{P}$:

$$\forall i \in \{1, \ldots, N_{\mathcal{P}}\}, \forall x \in \mathcal{X}, d_k(\pi^{i,-}(x), \pi^{i,+}(x)) \leq \delta.$$

Let $([\psi^{i,-}, \psi^{i,+}])_{1\leq i\leq N_\Psi}$ be a minimal covering of δ bracket for $\sup_x \max_k d_y$ of Ψ:

$$\forall i \in \{1, \ldots, N_\Psi\}, \forall x \in \mathcal{X}, \forall k \in \{1, \ldots, K\}, d_y(\psi_k^{i,-}(x, .), \psi_k^{i,+}(x, .)) \leq \delta.$$

Let s be a density in $\mathcal{C}$. By definition, there is π in $\mathcal{P}$ and ψ in Ψ such that for all (x, y) in $\mathcal{X} \times \mathcal{Y}$, $s(y|x) = \sum_{k=1}^K \pi_k(x)\psi_k(x, y)$.

Due to the covering, there is i in $\{1, \ldots, N_{\mathcal{P}}\}$ such that

$$\forall x \in \mathcal{X}, \forall k \in \{1, \ldots, K\}, \pi_k^{i,-}(x) \leq \pi_k(x) \leq \pi_k^{i,+}(x).$$

There is also j in $\{1, \ldots, N_\Psi\}$ such that

$$\forall x \in \mathcal{X}, \forall k \in \{1, \ldots, K\}, \forall y \in \mathcal{Y}, \psi_k^{j,-}(x, y) \leq \psi_k(x, y) \leq \psi_k^{j,+}(x, y).$$

Since for all x, for all k and for all y, $\pi_k(x)$ and $\psi_k(x, y)$ are non-negatives, we may multiply term-by-term and sum these inequalities over k to obtain:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \sum_{k=1}^K \left(\pi_k^{i,-}(x)\right)_+ \left(\psi_k^{j,-}(x, y)\right)_+ \leq s(y|x) \leq \sum_{k=1}^K \pi_k^{i,+}(x)\psi_k^{j,+}(x, y).$$

$\left(\left[\sum_{k=1}^K \left(\pi_k^{i,-}\right)_+ \left(\psi_k^{j,-}\right)_+, \sum_{k=1}^K \pi_k^{i,+}\psi_k^{j,+}\right]\right)_{\substack{1\leq i\leq N_{\mathcal{P}} \\ 1\leq j\leq N_\Psi}}$ is thus a bracket covering of $\mathcal{C}$.

Now, we focus on brackets' size using lemmas from Cohen and Le Pennec [2011] (namely Lemma 11, 12, 13), To lighten the notations, π_k^- and ψ_k^- are supposed non-negatives for all k. Following their Lemma 12, only using Cauchy-Schwarz inequality, we prove that

$$\sup_x d_y^2\left(\sum_{k=1}^K \pi_k^-(x)\psi_k^-(x, .), \sum_{k=1}^K \pi_k^+(x)\psi_k^+(x, .)\right) \leq \sup_x d_{y,k}^2(\pi^-(x)\psi^-(x, .), \pi^+(x)\psi^+(x, .)).$$

Then, using Cauchy-Schwarz inequality again, we get by their Lemma 11:

$$\begin{aligned}\sup_x d^2_{y,k}(\pi^-(x)\psi^-(x,.),\pi^+(x)\psi^+(x,.)) \\ \leq \sup_x \left(\max_k d_y(\psi_k^+(x,.),\psi_k^-(x,.)) \sqrt{\sum_{k=1}^K \pi_k^+(x)} \right. \\ \left. + d_k(\pi^+(x),\pi^-(x)) \max_k \sqrt{\int \psi_k^-(x,y)dy} \right)^2\end{aligned}$$

According to their Lemma 13, $\forall x, \sum_{k=1}^K \pi_k^+(x) \leq 1 + 2(\sqrt{2}+\sqrt{3})\delta$.

$$\begin{aligned}&\sup_x \left(\max_k d_y(\psi_k^+(x,.),\psi_k^-(x,.)) \sqrt{\sum_{k=1}^K \pi_k^+(x)} \right. \\ &\qquad \left. + d_k(\pi^+(x),\pi^-(x)) \max_k \sqrt{\int \psi_k^-(x,y)dy} \right)^2 \\ &\leq \left(\sqrt{1+2(\sqrt{2}+\sqrt{3})\delta} + 1 \right)^2 \delta^2 \leq (5\delta)^2\end{aligned}$$

The result follows from the fact we exhibited a 5δ covering of brackets of $\mathcal{C}$, with cardinality $N_{\mathcal{P}} N_\Psi$. □

Bracketing entropy of weight's families

General case We prove

Lemma 4. *For any $\delta \in (0, \sqrt{2}]$,*

$$H_{[.],\sup_x d_k}\left(\frac{\delta}{5}, \mathcal{P}\right) \leq H_{d_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}\delta}{20\sqrt{K}}, W_K\right).$$

Proof. We show that $\forall (w,z) \in (W_K)^2, \forall k \in \{1,\dots,K\}, \forall x \in \mathcal{X}, |\sqrt{\pi_{w,k}(x)} - \sqrt{\pi_{z,k}(x)}| \leq F(k,x)d(w,z)$, with F a function and d some distance. We define $\forall k, \forall u \in \mathbb{R}^K, A_k(u) = \frac{\exp(u_k)}{\sum_{k=1}^K \exp(u_k)}$, so $\pi_{w,k}(x) = A_k(w(x))$.

$\forall (u,v) \in (\mathbb{R}^K)^2$,

$$\left|\sqrt{A_k(v)} - \sqrt{A_k(u)}\right| = \left| \int_0^1 \nabla\left(\sqrt{A_k}\right)(u + t(v-u)).(v-u)dt \right|$$

Besides,

$$\begin{aligned}\nabla\left(\sqrt{A_k}\right)(u) &= \left(\frac{1}{2}\sqrt{A_k(u)} \frac{\partial}{\partial u_l}(\ln(A_k(u))) \right)_{1\leq l\leq K} \\ &= \left(\frac{1}{2}\sqrt{A_k(u)}\left(\delta_{k,l} - A_l(u)\right) \right)_{1\leq l\leq K}\end{aligned}$$

$$\begin{aligned}
&\left|\sqrt{A_k(v)}-\sqrt{A_k(u)}\right| \\
&\qquad = \frac{1}{2}\left|\int_0^1 \sqrt{A_k(u+t(v-u))}\sum_{l=1}^{K}\left(\delta_{k,l}-A_l(u+t(v-u))\right)(v_l-u_l)dt\right| \\
&\qquad \leq \frac{\|v-u\|_\infty}{2}\int_0^1 \sqrt{A_k(u+t(v-u))}\sum_{l=1}^{K}\left|\delta_{k,l}-A_l(u+t(v-u))\right|dt
\end{aligned}$$

Since $\forall u \in \mathbb{R}^K, \sum_{k=1}^K A_k(u)=1$, $\sum_{l=1}^K |\delta_{k,l}-A_l(u)| = 2(1-A_k(u))$

$$\begin{aligned}
\left|\sqrt{A_k(v)}-\sqrt{A_k(u)}\right| &\leq \|v-u\|_\infty \int_0^1 \sqrt{A_k(u+t(v-u))}\left(1-A_k(u+t(v-u))\right)dt \\
&\leq \frac{2}{3\sqrt{3}}\|v-u\|_\infty
\end{aligned}$$

since $x \mapsto \sqrt{x}(1-x)$ is maximal over [0,1] for $x=\frac{1}{3}$. We deduce that for any (w,z) in $(W_K)^2$, for all k in $\{1,\dots,K\}$, for any x in $\mathcal{X}$, $|\sqrt{\pi_{w,k}(x)}-\sqrt{\pi_{z,k}(x)}| \leq \frac{2}{3\sqrt{3}}\max_l \|w_l-z_l\|_\infty$.

By hypothesis, for any positive ϵ, an ϵ-net $\mathcal{N}$ of W_K may be exhibited. Let w be an element of W_K. There is a z belonging to the ϵ-net $\mathcal{N}$ such that $\max_l \|z_l-w_l\|_\infty \leq \epsilon$. Since for all k in $\{1,\dots,K\}$, for any x in $\mathcal{X}$,

$$|\sqrt{\pi_{w,k}(x)}-\sqrt{\pi_{z,k}(x)}| \leq \frac{2}{3\sqrt{3}}\max_l \|w_l-z_l\|_\infty \leq \frac{2}{3\sqrt{3}}\epsilon,$$

and

$$\sum_{k=1}^{K}\left(\sqrt{\pi_{z,k}(x)}+\frac{2}{3\sqrt{3}}\epsilon-\sqrt{\pi_{z,k}(x)}+\frac{2}{3\sqrt{3}}\epsilon\right)^2 = K\left(\frac{4\epsilon}{3\sqrt{3}}\right)^2,$$

$\left(\left[\left(\sqrt{\pi_z}-\frac{2}{3\sqrt{3}}\epsilon\right)^2,\left(\sqrt{\pi_z}+\frac{2}{3\sqrt{3}}\epsilon\right)^2\right]\right)_{z\in\mathcal{N}}$ is a $\frac{4\epsilon\sqrt{K}}{3\sqrt{3}}$-bracketing cover of $\mathcal{P}$. As a result, $H_{[],\sup_x d_k}\left(\frac{\delta}{5},\mathcal{P}\right) \leq H_{d_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}}{20\sqrt{K}}\delta, W_K\right)$. □

Case: $W_K = \{0\}\otimes W^{K-1}$ with W constructed from bounded functions We remind that

$$W = \left\{ w : \mathcal{X} \to \mathbb{R} / w(x) = \sum_{i=0}^{d_W} \alpha_i \psi_{W,i} \text{ and } \|\alpha\|_\infty \leq T_W \right\}$$

with $\|\psi_{W,i}\|_\infty \leq 1$.

Proof of Part 1 of Lemma 1. W_K is a finite dimensional compact set. Thanks to the result in the general case, we get

$$H_{[.],\sup_x d_k}\left(\frac{\delta}{5},\mathcal{P}\right) \leq H_{d_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}\delta}{20\sqrt{K-1}}, W_K\right)$$

now as for all w, v in W_K, $\max_k \|w_k - v_k\|_\infty \leq \max_k \sum_{i=0}^{d_W} |\alpha_{k,i}^w - \alpha_{k,i}^v| \leq d_W \max_{k,i} |\alpha_{k,i}^w - \alpha_{k,i}^v|$

$$\begin{aligned}
&\leq H_{\|.\|_\infty}\left(\frac{3\sqrt{3}\delta}{20\sqrt{K-1}d_W}, \left\{\alpha \in \mathbb{R}^{(K-1)d_W} / \|\alpha\|_\infty \leq T_W\right\}\right) \\
&\leq (K-1)d_W \ln\left(1 + \frac{20\sqrt{K-1}T_W d_W}{3\sqrt{3}\delta}\right) \\
&\leq (K-1)d_W \left[\ln\left(\sqrt{2} + \frac{20}{3\sqrt{3}}T_W\sqrt{K-1}d_W\right) + \ln\left(\frac{1}{\delta}\right)\right]
\end{aligned}$$

□

The second Lemma is just a consequence of $d_W = \binom{d'_W + d}{d}$

Bracketing entropy of Gaussian families

General case We rely on a general construction of Gaussian brackets:

Proposition. 2. *Let $\kappa \geq \frac{17}{29}$, $\gamma_\kappa = \dfrac{25(\kappa - \frac{1}{2})}{49(1 + \frac{2\kappa}{5})}$. For any $0 < \delta \leq \sqrt{2}$, any $p \geq 1$ and any $\delta_\Sigma \leq \frac{1}{5\sqrt{\kappa^2 \cosh(\frac{2\kappa}{5}) + \frac{1}{2}}} \frac{\delta}{p}$, let $(\upsilon, L, A, P) \in \Upsilon \times [L_-, L_+] \times \mathcal{A}(\lambda_-, \lambda_+) \times SO(p)$ and $(\tilde{\upsilon}, \tilde{L}, \tilde{A}, \tilde{P}) \in \Upsilon \times [L_-, L_+] \times \mathcal{A}(\lambda_-, +\infty) \times SO(p)$, define $\Sigma = LPAP'$ and $\tilde{\Sigma} = \tilde{L}\tilde{P}\tilde{A}\tilde{P}'$,*

$$t^-(x,y) = (1+\kappa\delta_\Sigma)^{-p}\Phi_{\tilde{\upsilon}(x),(1+\delta_\Sigma)^{-1}\tilde{\Sigma}}(y) \text{ and } t^+(x,y) = (1+\kappa\delta_\Sigma)^{p}\Phi_{\tilde{\upsilon}(x),(1+\delta_\Sigma)\tilde{\Sigma}}(y).$$

If

$$\begin{cases}
\forall x \in \mathcal{X}, \|\upsilon(x) - \tilde{\upsilon}(x)\|^2 \leq p\gamma_\kappa L_- \lambda_- \frac{\lambda_-}{\lambda_+}\delta_\Sigma^2 \\
\left(1 + \frac{2}{25}\delta_\Sigma\right)^{-1} \tilde{L} \leq L \leq \tilde{L} \\
\forall 1 \leq i \leq p, |A_{i,i}^{-1} - \tilde{A}_{i,i}^{-1}| \leq \frac{1}{10}\frac{\delta_\Sigma}{\lambda_+} \\
\forall y \in \mathbb{R}^p, \|Py - \tilde{P}y\| \leq \frac{1}{10}\frac{\lambda_-}{\lambda_+}\delta_\Sigma\|y\|
\end{cases}$$

then $[t^-, t^+]$ is a $\delta/5$ Hellinger bracket such that $t^-(x,y) \leq \Phi_{\upsilon(x),\Sigma}(y) \leq t^+(x,y)$.

This statement is similar to Lemma 10 in Cohen and Le Pennec [2011]. Admitting this proposition, we are brought to construct nets over the spaces of the means, the volumes, the eigenvector matrices and the normalized eigenvalue matrices. We consider three cases: the parameter (mean, volume, matrix) is known ($\star = 0$), unknown but common to all classes ($\star = c$), unknown and possibly different for every class ($\star = K$). For example, $[\nu_K, L_0, P_c, A_0]$ denotes a model in which only means are free and eigenvector matrices are assumed to be equal and unknown.

If the means are free ($\star = K$), we construct a grid G_{Υ_K} over Υ_K, which is compact. Since

$$H_{d_{\|\sup\|_\infty}}\left(\sqrt{p\gamma_\kappa L_- \lambda_- \frac{\lambda_-}{\lambda_+}}\delta_\Sigma, \Upsilon_K\right) \leq \dim(\Upsilon_K)\left(C_\Upsilon + \ln\left(\frac{1}{\sqrt{p\gamma_\kappa L_- \lambda_- \frac{\lambda_-}{\lambda_+}}\delta_\Sigma}\right)\right),$$

$$\left|G_{\Upsilon_K}\left(\sqrt{p\gamma_\kappa L_- \lambda_- \frac{\lambda_-}{\lambda_+}}\delta_\Sigma\right)\right| \leq \left(C_\Upsilon + \ln\left(\frac{1}{\sqrt{p\gamma_\kappa L_- \lambda_- \frac{\lambda_-}{\lambda_+}}\delta_\Sigma}\right)\right)^{\dim(\Upsilon_K)}.$$

If the means are common and unknown ($\star = c$), belonging to Υ_1 , we construct a grid $G_{\Upsilon_c}\left(\sqrt{p\gamma_\kappa L_- \lambda_- \frac{\lambda_-}{\lambda_+}}\delta_\Sigma\right)$ over Υ_1 with cardinality at most

$$\left(C_\Upsilon + \ln\left(\frac{1}{\sqrt{p\gamma_\kappa L_- \lambda_- \frac{\lambda_-}{\lambda_+}}\delta_\Sigma}\right)\right)^{D_{\Upsilon_1}}.$$

Finally, if the means are known ($\star = 0$), we do not need to construct a grid. In the end, $\left|G_{\Upsilon_\star}\left(\sqrt{p\gamma_\kappa L_- \lambda_- \frac{\lambda_-}{\lambda_+}}\delta_\Sigma\right)\right| \leq \left(C_\Upsilon + \ln\left(\frac{1}{\sqrt{p\gamma_\kappa L_- \lambda_- \frac{\lambda_-}{\lambda_+}}\delta_\Sigma}\right)\right)^{Z_{\upsilon,\star}}$, with $Z_{\upsilon,K} = \dim(\Upsilon_K)$, $Z_{\upsilon,c} = D_{\Upsilon_1}$ and $Z_{\upsilon,0} = 0$.

Then, we consider the grid G_L over $[L_-, L_+]$:

$$G_L\left(\frac{2}{25}\delta_\Sigma\right) = \left\{L_-\left(1+\frac{2}{25}\delta_\Sigma\right)^g / g \in \mathbb{N}, L_-\left(1+\frac{2}{25}\delta_\Sigma\right)^g \leq L_+\right\}$$

$$\left|G_L\left(\frac{2}{25}\delta_\Sigma\right)\right| \leq 1 + \frac{\ln\left(\frac{L_+}{L_-}\right)}{\ln\left(1+\frac{2}{25}\delta_\Sigma\right)}$$

Since $\delta_\Sigma \leq \frac{2}{5}$, $\ln\left(1+\frac{2}{25}\delta_\Sigma\right) \geq \frac{10}{129}\delta_\Sigma$.

$$\left|G_L\left(\frac{2}{25}\delta_\Sigma\right)\right| \leq 1 + \frac{129\ln\left(\frac{L_+}{L_-}\right)}{10\delta_\Sigma} \leq \frac{4 + 129\ln\left(\frac{L_+}{L_-}\right)}{10\delta_\Sigma}$$

By definition of a net, for any $P \in SO(p)$ there is a $\tilde{P} \in G_P\left(\frac{1}{10}\frac{\lambda_-}{\lambda_+}\delta_\Sigma\right)$ such that $\forall y \in \mathbb{R}^p, \|Py - \tilde{P}y\| \leq \frac{1}{10}\frac{\lambda_-}{\lambda_+}\delta_\Sigma\|y\|$. There exists a universal constant c_U such that $\left|G_P\left(\frac{1}{10}\frac{\lambda_-}{\lambda_+}\delta_\Sigma\right)\right| \leq c_U\left(\frac{10\lambda_+}{\lambda_-\delta_\Sigma}\right)^{\frac{p(p-1)}{2}}$.

For the grid G_A, we look at the condition on the $p-1$ first diagonal values and obtain:

$$\left|G_A\left(\frac{1}{10}\frac{\lambda_-}{\lambda_+}\delta_\Sigma\right)\right| \leq \left(2 + \frac{\ln\left(\frac{\lambda_+}{\lambda_-}\right)}{\ln\left(1+\frac{1}{10}\frac{\lambda_-}{\lambda_+}\delta_\Sigma\right)}\right)^{p-1}$$

Since $\delta_\Sigma \leq \frac{2}{5}$, $\ln\left(1+\frac{1}{10}\frac{\lambda_-}{\lambda_+}\delta_\Sigma\right) \geq \frac{5}{52}\frac{\lambda_-}{\lambda_+}\delta_\Sigma$, then

$$\left|G_A\left(\frac{1}{10}\frac{\lambda_-}{\lambda_+}\delta_\Sigma\right)\right| \leq \left(2 + \frac{52}{5\delta_\Sigma}\frac{\lambda_+}{\lambda_-}\ln\left(\frac{\lambda_+}{\lambda_-}\right)\right)^{p-1} \leq \left(4 + 52\frac{\lambda_+}{\lambda_-}\ln\left(\frac{\lambda_+}{\lambda_-}\right)\right)^{p-1}\left(\frac{1}{5\delta_\Sigma}\right)^{p-1}$$

Let $Z_{L,0} = Z_{P,0} = Z_{A,0} = 0$, $Z_{L,c} = Z_{P,c} = Z_{A,c} = 1$, $Z_{L,K} = Z_{P,K} = Z_{A,K} = K$. We define $f_{\upsilon,\star}$ from $\Upsilon_\star$ to Υ_K by $\begin{cases} 0 \mapsto (\upsilon_{0,1}, \dots, \upsilon_{0,1}) \text{ if } \star = 0 \\ \upsilon \mapsto (\upsilon, \dots, \upsilon) \text{ if } \star = c \\ (\upsilon_1, \dots, \upsilon_K) \mapsto (\upsilon_1, \dots, \upsilon_K) \text{ if } \star = K \end{cases}$ and similarly $f_{L,\star}$, $f_{P,\star}$ and $f_{A,\star}$, respectively from $(\mathbb{R}_+)^{Z_{L,\star}}$ into $(\mathbb{R}_+)^K$, from $(SO(p))^{Z_{P,\star}}$ into $(SO(p))^K$ and from $\mathcal{A}(\lambda_-, \lambda_+)^{Z_{A,\star}}$ into $\mathcal{A}(\lambda_-, \lambda_+)^K$.

We define

$$\Gamma : (\upsilon_1, \dots, \upsilon_K, L_1, \dots, L_K, P_1, \dots, P_K, A_1, \dots, A_K) \mapsto (\upsilon_k, L_k P_k A_k P_k')_{1 \le k \le K}$$

and $\Psi : (\upsilon_k, \Sigma_k)_{1 \le k \le K} \mapsto (\Phi_{\upsilon_k, \Sigma_k})_{1 \le k \le K}$. The image of $\Upsilon_\star \times [L_-, L_+]^{Z_{L,\star}} \times SO(p)^{Z_{P,\star}} \times \mathcal{A}(\lambda_-, \lambda_+)^{Z_{A,\star}}$ by $\Psi \circ \Gamma \circ (f_{\upsilon,\star} \otimes f_{L,\star} \otimes f_{P,\star} \otimes f_{A,\star})$ is the set $\mathcal{G}$ of all K-tuples of Gaussian densities of type $[\upsilon_\star, L_\star, P_\star, A_\star]$.

Now, we define B:

$$(\upsilon_k, \Sigma_k)_{1 \le k \le K} \mapsto \left((1 + \kappa\delta_\Sigma)^{-p} \Phi_{\upsilon_k, (1+\delta_\Sigma)^{-1}\Sigma_k}, (1 + \kappa\delta_\Sigma)^p \Phi_{\upsilon_k, (1+\delta_\Sigma)\Sigma_k} \right)_{1 \le k \le K}.$$

The image of $G_{\Upsilon_\star} \times G_L^{Z_{L,\star}} \times G_P^{Z_{P,\star}} \times G_A^{Z_{A,\star}}$ by $B \circ \Gamma \circ (f_{\upsilon,\star} \otimes f_{L,\star} \otimes f_{P,\star} \otimes f_{A,\star})$ is a $\delta/5$-bracket covering of $\mathcal{G}$, with cardinality bounded by

$$\left(\frac{\sqrt{\lambda_+} \exp(C_\Upsilon)}{\sqrt{p\gamma_\kappa L_- \lambda_-^2} \delta_\Sigma} \right)^{Z_{\Upsilon,\star}} \times \left(\frac{4 + 129 \ln\left(\frac{L_+}{L_-}\right)}{10\delta_\Sigma} \right)^{Z_{L,\star}} \times c_U^{Z_{P,\star}} \left(\frac{10\lambda_+}{\lambda_-\delta_\Sigma} \right)^{\frac{p(p-1)}{2} Z_{P,\star}}$$
$$\times \left(4 + 52 \frac{\lambda_+}{\lambda_-} \ln\left(\frac{\lambda_+}{\lambda_-}\right) \right)^{(p-1)Z_{A,\star}} \left(\frac{1}{5\delta_\Sigma} \right)^{(p-1)Z_{A,\star}}.$$

Taking $\delta_\Sigma = \frac{1}{5\sqrt{\kappa^2 \cosh\left(\frac{2\kappa}{5}\right) + \frac{1}{2}}} \frac{\delta}{p}$, we obtain

$$H_{[.], \sup_x \max_k d_y}\left(\frac{\delta}{5}, \mathcal{G}\right) \le \mathcal{D}\left(\mathcal{C} + \ln\left(\frac{1}{\delta}\right)\right)$$

with $\mathcal{D} = Z_{\upsilon,\star} + Z_{L,\star} + \frac{p(p-1)}{2} Z_{P,\star} + (p-1) Z_{A,\star}$ and

$$\mathcal{C} = \ln\left(5p \sqrt{\kappa^2 \cosh\left(\frac{2\kappa}{5}\right) + \frac{1}{2}} \right) + \frac{Z_{\upsilon,\star} C_\Upsilon}{\mathcal{D}} + \frac{Z_{\upsilon,\star}}{2\mathcal{D}} \ln\left(\frac{\lambda_+}{p\gamma_\kappa L_- \lambda_-^2} \right)$$
$$+ \frac{Z_{L,\star}}{\mathcal{D}} \ln\left(\frac{4 + 129 \ln\left(\frac{L_+}{L_-}\right)}{10} \right) + \frac{Z_{P,\star}}{\mathcal{D}} \left(\ln(c_U) + \frac{p(p-1)}{2} \ln\left(\frac{10\lambda_+}{\lambda_-} \right) \right)$$
$$+ \frac{Z_{A,\star}(p-1)}{\mathcal{D}} \ln\left(\frac{4}{5} + \frac{52\lambda_+}{5\lambda_-} \ln\left(\frac{\lambda_+}{\lambda_-} \right) \right)$$

Case: Υ_K generated from bounded functions Using previous work, we only have to handle Υ_K's bracketing entropy. Just like for W_K, we aim at bounding the bracketing entropy by the entropy of the parameters' space

We focus on the case of Lemma 1 where $\Upsilon_K = \Upsilon^K$ and

$$\Upsilon = \left\{ \upsilon : \mathcal{X} \to \mathbb{R}^p \middle| \forall j \in \{1, \dots, p\}, \forall x, \upsilon_j(x) = \sum_{i=0}^{d_\Upsilon} \alpha_i^{(j)} \psi_{\Upsilon,i}, \text{ and } \|\alpha\|_\infty \leq T_\Upsilon \right\}$$

We consider for any υ, ν in Υ and any x in $[0,1]^d$,

$$\begin{aligned}
\|\upsilon(x) - \nu(x)\|_2^2 &= \sum_{j=1}^{p} \left(\sum_{i=0}^{d_\Upsilon} \left(\alpha_i^{\upsilon,(j)} - \alpha_i^{\nu,(j)} \right) \psi_{\Upsilon,j}(x) \right)^2 \\
&\leq \sum_{j=1}^{p} \left(\sum_{i=0}^{d_\Upsilon} \left(\alpha_i^{\upsilon,(j)} - \alpha_i^{\nu,(j)} \right)^2 \right) \left(\sum_{i=0}^{d_\Upsilon} |\psi_{\Upsilon,j}(x)|^2 \right) \\
&\leq d_\Upsilon \sum_{j=1}^{p} \sum_{i=0}^{d_\Upsilon} \left(\alpha_r^{\upsilon,(j)} - \alpha_r^{\nu,(j)} \right)^2 \\
&\leq p d_\Upsilon^2 \max_{j,i} \left(\alpha_i^{\upsilon,(j)} - \alpha_r^{\nu,(j)} \right)^2
\end{aligned}$$

So,

$$\begin{aligned}
H_{\max_k \sup_x \|\|_2}(\delta, \Upsilon_K) &\leq H_{\max_{k,j,r} |.|} \left(\frac{\delta}{\sqrt{p} d_\Upsilon}, \left\{ \left(\alpha_r^{(j,k)} \right)_{\substack{1 \leq j \leq p \\ |r| \leq d'_\Upsilon \\ 1 \leq k \leq K}} \middle| \|\alpha\|_\infty \leq T_\Upsilon \right\} \right) \\
&\leq pK d_\Upsilon \ln \left(1 + \frac{\sqrt{p} d_\Upsilon T_\Upsilon}{\delta} \right) \\
&\leq pK d_\Upsilon \left[\ln \left(\sqrt{2} + \sqrt{p} d_\Upsilon T_\Upsilon \right) + \ln \left(\frac{1}{\delta} \right) \right] \\
&\leq \dim(\Upsilon_K) \left(C_\Upsilon + \ln \left(\frac{1}{\delta} \right) \right)
\end{aligned}$$

with $\dim(\Upsilon_K) = pK \binom{d'_\Upsilon + d}{d}$ and $C_\Upsilon = \ln \left(\sqrt{2} + \sqrt{p} \binom{d'_\Upsilon + d}{d} T_\Upsilon \right)$.

The second part of Lemma 2 is deduced from the fact that if $\mathcal{X} = [0,1]^d$ and Υ is the set of linear combination of monomials of degree less that d'_Υ then $d_\Upsilon = \binom{d'_\Upsilon + d}{d}$.

2.7.3 Proof of the key proposition to handle bracketing entropy of Gaussian families

Proof of Proposition 2

Proof. $[t^-, t^+]$ is a $\delta/5$ bracket.
Since $(1+\delta_\Sigma)\tilde{\Sigma}^{-1} - (1+\delta_\Sigma)^{-1}\tilde{\Sigma}^{-1} = ((1+\delta_\Sigma) - (1+\delta_\Sigma)^{-1})\tilde{\Sigma}^{-1}$ is a positive-definite matrix, Maugis and Michel's lemma can be applied.

Lemma 7. *[Maugis and Michel, 2011] Let $\Phi_{\upsilon_1,\Sigma_1}$ and $\Phi_{\upsilon_2,\Sigma_2}$ be two Gaussian densities with full rank covariance matrix in dimension p such that $\Sigma_1^{-1} - \Sigma_2^{-1}$ is a positive definite matrix. For any $y \in \mathbb{R}^p$,*

$$\frac{\Phi_{\upsilon_1,\Sigma_1}(y)}{\Phi_{\upsilon_2,\Sigma_2}(y)} \leq \sqrt{\frac{|\Sigma_2|}{|\Sigma_1|}} \exp \left(\frac{1}{2} (\upsilon_1 - \upsilon_2)' (\Sigma_2 - \Sigma_1)^{-1} (\upsilon_1 - \upsilon_2) \right).$$

Thus, $\forall x \in \mathcal{X}, \forall y \in \mathbb{R}^p$,

$$\begin{aligned}\frac{t^-(x,y)}{t^+(x,y)} =& \frac{(1+\kappa\delta_\Sigma)^{-p}}{(1+\kappa\delta_\Sigma)^{p}} \frac{\Phi_{\upsilon(x),(1+\delta_\Sigma)^{-1}\tilde{\Sigma}}(y)}{\Phi_{\upsilon(x),(1+\delta_\Sigma)\tilde{\Sigma}}(y)} \leq \frac{1}{(1+\kappa\delta_\Sigma)^{2p}} \sqrt{\frac{(1+\delta_\Sigma)^p}{(1+\delta_\Sigma)^{-p}}} \\ =& \left(\frac{1+\delta_\Sigma}{(1+\kappa\delta_\Sigma)^2}\right)^p = \left(\frac{1+\delta_\Sigma}{1+2\kappa\delta_\Sigma+\kappa^2\delta_\Sigma^2}\right)^p \leq 1\end{aligned}$$

For all x in $\mathcal{X}$,

$$\begin{aligned}d_y^2(t^-,t^+) &= \int t^-(x,y)\mathrm{d}y + \int t^+(x,y)\,\mathrm{d}y - 2\int \sqrt{t^-(x,y)}\sqrt{t^+(x,y)}\mathrm{d}y \\ &= (1+\kappa\delta_\Sigma)^{-p} + (1+\kappa\delta_\Sigma)^{p} - 2(1+\kappa\delta_\Sigma)^{-p/2}(1+\kappa\delta_\Sigma)^{p/2} \\ &\qquad \times \int \sqrt{\Phi_{\upsilon(x),(1+\delta_\Sigma)^{-1}\tilde{\Sigma}}(y)}\sqrt{\Phi_{\upsilon(x),(1+\delta_\Sigma)\tilde{\Sigma}}(y)}\,\mathrm{d}y \\ &= (1+\kappa\delta_\Sigma)^{-p} + (1+\kappa\delta_\Sigma)^{p} - (2 \\ &\qquad\qquad -d_y^2\left(\Phi_{\upsilon(x),(1+\delta_\Sigma)^{-1}\tilde{\Sigma}}(y), \Phi_{\upsilon(x),(1+\delta_\Sigma)\tilde{\Sigma}}(y)\right)\Big).\end{aligned}$$

Using the following lemma,

Lemma 8. *Let $\Phi_{\upsilon_1,\Sigma_1}$ and $\Phi_{\upsilon_2,\Sigma_2}$ be two Gaussian densities with full rank covariance matrix in dimension p, then*

$$\begin{aligned}d^2\left(\Phi_{\upsilon_1,\Sigma_1}, \Phi_{\upsilon_2,\Sigma_2}\right) = 2\Big(1 - 2^{p/2}|\Sigma_1\Sigma_2|^{-1/4}|\Sigma_1^{-1}+\Sigma_2^{-1}|^{-1/2} \\ \times \exp\left(-\frac{1}{4}(\upsilon_1-\upsilon_2)'(\Sigma_1+\Sigma_2)^{-1}(\upsilon_1-\upsilon_2)\right)\Big).\end{aligned}$$

we obtain

$$\begin{aligned}d_y^2(t^-,t^+) &= (1+\kappa\delta_\Sigma)^{-p} + (1+\kappa\delta_\Sigma)^{p} - 2\,2^{p/2}\left((1+\delta_\Sigma)+(1+\delta_\Sigma)^{-1}\right)^{-p/2} \\ &= 2 - 2\,2^{p/2}\left((1+\delta_\Sigma)+(1+\delta_\Sigma)^{-1}\right)^{-p/2} + (1+\kappa\delta_\Sigma)^{-p} - 2 \\ &\qquad + (1+\kappa\delta_\Sigma)^{p}\end{aligned}$$

Applying Lemma 9

Lemma 9. *For any $0 < \delta \leq \sqrt{2}$ and any $p \geq 1$, let $\kappa \geq \frac{1}{2}$ and $\delta_\Sigma \leq \frac{1}{5\sqrt{\kappa^2\cosh(\frac{2\kappa}{5})+\frac{1}{2}}}\frac{\delta}{p}$, then*

$$\delta_\Sigma \leq \frac{2}{5p} \leq \frac{2}{5}.$$

and

Lemma 10. *For any $p \in \mathbb{N}\setminus\{0\}$, for any $\delta_\Sigma > 0$,*

$$2 - 2^{p/2+1}\left((1+\delta_\Sigma)+(1+\delta_\Sigma)^{-1}\right)^{-p/2} \leq \frac{p{\delta_\Sigma}^2}{2} \leq \frac{p^2{\delta_\Sigma}^2}{2}$$

Furthermore, if $p\delta_\Sigma \leq c$, then

$$(1+\kappa\delta_\Sigma)^{p} + (1+\kappa\delta_\Sigma)^{-p} - 2 \leq \kappa^2\cosh(\kappa c)p^2{\delta_\Sigma}^2.$$

with $c = \frac{2}{5}$, it comes out that:

$$\sup_x d_y^2(t^-(x,y), t^+(x,y)) \leq \left(\frac{\delta}{5}\right)^2.$$

Now, we show that for all x in $\mathcal{X}$, for all y in $\mathbb{R}^p$, $t^-(x,y) \leq \Phi_{\upsilon(x),\Sigma}(y) \leq t^+(x,y)$. We use therefore Lemma 11, thanks to the hypothesis made on covariance matrices.

Lemma 11. *Let* $(L, A, P) \in [L_-, L_+] \times \mathcal{A}(\lambda_-, \lambda_+) \times SO(p)$ *and* $(\tilde{L}, \tilde{A}, \tilde{P}) \in [L_-, L_+] \times \mathcal{A}(\lambda_-, \infty) \times SO(p)$, *define* $\Sigma = LPAP'$ *and* $\tilde{\Sigma} = \tilde{L}\tilde{P}\tilde{A}\tilde{P}'$. *If*

$$\begin{cases} (1+\delta_L)^{-1}\tilde{L} \leq L \leq \tilde{L} \\ \forall 1 \leq i \leq p, |A_{i,i}^{-1} - \tilde{A}_{i,i}^{-1}| \leq \delta_A \lambda_-^{-1} \\ \forall y \in \mathbb{R}^p, \|Py - \tilde{P}y\| \leq \delta_P \|y\| \end{cases}$$

then $(1+\delta_\Sigma)\tilde{\Sigma}^{-1} - \Sigma^{-1}$ *and* $\Sigma^{-1} - (1+\delta_\Sigma)^{-1}\tilde{\Sigma}^{-1}$ *satisfy*

$$\forall y \in \mathbb{R}^p, y'\left((1+\delta_\Sigma)\tilde{\Sigma}^{-1} - \Sigma^{-1}\right)y \geq \tilde{L}^{-1}\left((\delta_\Sigma - \delta_L)\lambda_+^{-1} - (1+\delta_\Sigma)\lambda_-^{-1}(2\delta_P + \delta_A)\right)\|y\|^2$$

$$\forall y \in \mathbb{R}^p, y'\left(\Sigma^{-1} - (1+\delta_\Sigma)^{-1}\tilde{\Sigma}^{-1}\right)y \geq \frac{\tilde{L}^{-1}}{1+\delta_\Sigma}\left(\delta_\Sigma\lambda_+^{-1} - \lambda_-^{-1}(2\delta_P + \delta_A)\right)\|y\|^2$$

Using $\begin{cases} \delta_L = \frac{2}{25}\delta_\Sigma \\ \delta_P = \delta_A = \frac{1}{10}\frac{\lambda_-}{\lambda_+}\delta_\Sigma \end{cases}$
we get lower bounds of the same order:

$$\forall y \in \mathbb{R}^p, y'\left((1+\delta_\Sigma)\tilde{\Sigma}^{-1} - \Sigma^{-1}\right)y \geq \frac{\tilde{L}^{-1}}{2\lambda_+}\delta_\Sigma\|y\|^2$$

$$\forall y \in \mathbb{R}^p, y'\left(\Sigma^{-1} - (1+\delta_\Sigma)^{-1}\tilde{\Sigma}^{-1}\right)y \geq \frac{\tilde{L}^{-1}}{1+\delta_\Sigma}\frac{7}{10\lambda_+}\delta_\Sigma\|y\|^2$$

Let's compare $\Phi_{\upsilon,\Sigma}$ and t^+.

$$\begin{aligned} &\frac{\Phi_{\upsilon(x),\Sigma}(y)}{(1+\kappa\delta_\Sigma)^p \Phi_{\tilde{\upsilon}(x),(1+\delta_\Sigma)\tilde{\Sigma}}(y)} \\ &\leq (1+\kappa\delta_\Sigma)^{-p}\left(\sqrt{\frac{|(1+\delta_\Sigma)\tilde{\Sigma}|}{|\Sigma|}}\exp\left(\frac{1}{2}(\upsilon(x)-\tilde{\upsilon}(x))'\left((1+\delta_\Sigma)\tilde{\Sigma}-\Sigma\right)^{-1}(\upsilon(x)-\tilde{\upsilon}(x))\right)\right) \\ &\leq \frac{(1+\delta_\Sigma)^{p/2}}{(1+\kappa\delta_\Sigma)^p}\left(\sqrt{\frac{|\tilde{\Sigma}|}{|\Sigma|}}\exp\left(\frac{1}{2}(\upsilon(x)-\tilde{\upsilon}(x))'\left((1+\delta_\Sigma)\tilde{\Sigma}-\Sigma\right)^{-1}(\upsilon(x)-\tilde{\upsilon}(x))\right)\right). \end{aligned}$$

But,

$$\begin{aligned} \left((1+\delta_\Sigma)\tilde{\Sigma} - \Sigma\right)^{-1} &= \left((1+\delta_\Sigma)\tilde{\Sigma}(\Sigma^{-1} - (1+\delta_\Sigma)^{-1}\tilde{\Sigma}^{-1})\Sigma\right)^{-1} \\ &= (1+\delta_\Sigma)^{-1}\Sigma^{-1}(\Sigma^{-1} - (1+\delta_\Sigma)^{-1}\tilde{\Sigma}^{-1})^{-1}\tilde{\Sigma}^{-1} \end{aligned}$$

Thus by Lemma 11,

$$\begin{aligned}
&(\upsilon(x)-\tilde{\upsilon}(x))'\left((1+\delta_\Sigma)\tilde{\Sigma}-\Sigma\right)^{-1}(\upsilon(x)-\tilde{\upsilon}(x))\\
&\leq (1+\delta_\Sigma)^{-1}L_-^{-1}\lambda_-^{-1}(1+\delta_\Sigma)\tilde{L}\frac{10}{7}\lambda_+\delta_\Sigma^{-1}\tilde{L}^{-1}\lambda_-^{-1}\|\upsilon(x)-\tilde{\upsilon}(x)\|^2\\
&\leq \frac{10}{7}L_-^{-1}\lambda_-^{-2}\lambda_+\delta_\Sigma^{-1}\|\upsilon(x)-\tilde{\upsilon}(x)\|^2\\
&\leq \frac{10}{7}L_-^{-1}\lambda_-^{-2}\lambda_+\delta_\Sigma^{-1}p\gamma_\kappa L_-\lambda_-^2\lambda_+^{-1}\delta_\Sigma^2\\
&\leq \frac{10}{7}p\gamma_\kappa\delta_\Sigma
\end{aligned}$$

Since $\sqrt{\frac{|\tilde{\Sigma}|}{|\Sigma|}}=\left(\frac{\tilde{L}}{L}\right)^{\frac{p}{2}}\leq\left(1+\frac{2}{25}\delta_\Sigma\right)^{p/2}$,

$$\frac{\Phi_{\upsilon(x),\Sigma}(y)}{(1+\kappa\delta_\Sigma)^p\Phi_{\tilde{\upsilon}(x),(1+\delta_\Sigma)\tilde{\Sigma}}(y)}\leq\frac{(1+\delta_\Sigma)^{p/2}(1+\frac{2}{25}\delta_\Sigma)^{p/2}}{(1+\kappa\delta_\Sigma)^p}\exp\left(\frac{5\gamma_\kappa}{7}p\delta_\Sigma\right).$$

It suffices that

$$\frac{5\gamma_\kappa}{7}\delta_\Sigma\leq\ln\left(\frac{1+\kappa\delta_\Sigma}{\sqrt{1+\delta_\Sigma}\sqrt{1+\frac{2}{25}\delta_\Sigma}}\right)$$

Now let

$$\begin{aligned}
f(\delta_\Sigma)&=\ln(1+\kappa\delta_\Sigma)-\frac{1}{2}\ln(1+\delta_\Sigma)-\frac{1}{2}\ln\left(1+\frac{2}{25}\delta_\Sigma\right)\\
f'(\delta_\Sigma)&=\frac{\kappa}{1+\kappa\delta_\Sigma}-\frac{1}{2(1+\delta_\Sigma)}-\frac{1}{25\left(1+\frac{2}{25}\delta_\Sigma\right)}=\frac{(27k-4)\delta_\Sigma+50k-27}{2(1+\kappa\delta_\Sigma)(1+\delta_\Sigma)(25+2\delta_\Sigma)}
\end{aligned}$$

Since $\kappa>\frac{17}{29}$,

$$f'(\delta_\Sigma)>\frac{k-\frac{27}{50}}{(1+\kappa\delta_\Sigma)(1+\delta_\Sigma)\left(1+\frac{2}{25}\delta_\Sigma\right)}$$

Finally, since $f(0)=0$ and $\delta_\Sigma\leq\frac{2}{5}$, one deduces

$$\begin{aligned}
f(\delta_\Sigma)&>\frac{k-\frac{27}{50}}{(1+\kappa\delta_\Sigma)(1+\delta_\Sigma)\left(1+\frac{2}{25}\delta_\Sigma\right)}\delta_\Sigma\\
&\geq\frac{k-\frac{27}{50}}{\left(1+\frac{2}{5}\kappa\right)\left(1+\frac{2}{5}\right)\left(1+\frac{2}{25}\frac{2}{5}\right)}\delta_\Sigma=\frac{5}{7}\frac{125(k-\frac{27}{50})}{129\left(1+\frac{2}{5}\kappa\right)}\delta_\Sigma\\
&\geq\frac{5}{7}\gamma_\kappa\delta_\Sigma
\end{aligned}$$

So $\Phi_{\upsilon,\Sigma} \leq t^+$. $\frac{t^-}{\Phi_{\upsilon,\Sigma}}$ is handled the same way.

$$\begin{aligned}
&\frac{(1+\kappa\delta_\Sigma)^{-p}\Phi_{\tilde{\upsilon}(x),(1+\delta_\Sigma)^{-1}\tilde{\Sigma}}(y)}{\Phi_{\upsilon(x),\Sigma}(y)} \\
&\leq (1+\kappa\delta_\Sigma)^{-p}\left(\sqrt{\frac{|\Sigma|}{|(1+\delta_\Sigma)^{-1}\tilde{\Sigma}|}}\exp\left(\frac{1}{2}(\upsilon(x)-\tilde{\upsilon}(x))'\left(\Sigma-(1+\delta_\Sigma)^{-1}\tilde{\Sigma}\right)^{-1}(\upsilon(x)-\tilde{\upsilon}(x))\right)\right) \\
&\leq \frac{(1+\delta_\Sigma)^{p/2}}{(1+\kappa\delta_\Sigma)^p}\exp\left(\frac{1}{2}(\upsilon(x)-\tilde{\upsilon}(x))'\left(\Sigma-(1+\delta_\Sigma)^{-1}\tilde{\Sigma}\right)^{-1}(\upsilon(x)-\tilde{\upsilon}(x))\right)
\end{aligned}$$

Now

$$\begin{aligned}
\left(\Sigma-(1+\delta_\Sigma)^{-1}\tilde{\Sigma}\right)^{-1} &= \left(\Sigma\left((1+\delta_\Sigma)\tilde{\Sigma}^{-1}-\Sigma^{-1}\right)(1+\delta_\Sigma)^{-1}\tilde{\Sigma}\right)^{-1} \\
&= (1+\delta_\Sigma)\tilde{\Sigma}^{-1}\left((1+\delta_\Sigma)\tilde{\Sigma}^{-1}-\Sigma^{-1}\right)^{-1}\Sigma^{-1}
\end{aligned}$$

and

$$\begin{aligned}
&(\upsilon(x)-\tilde{\upsilon}(x))'\left(\Sigma-(1+\delta_\Sigma)^{-1}\tilde{\Sigma}\right)^{-1}(\upsilon(x)-\tilde{\upsilon}(x)) \\
&\qquad\leq (1+\delta_\Sigma)\tilde{L}^{-1}\lambda_-^{-1}2\tilde{L}\lambda_+\delta_\Sigma^{-1}L_-^{-1}\lambda_-^{-1}p\gamma_\kappa L_-\lambda_-^2\lambda_+^{-1}\delta_\Sigma^2 \\
&\qquad\leq 2p\gamma_\kappa(1+\delta_\Sigma)\delta_\Sigma
\end{aligned}$$

We only need to prove that

$$\gamma_\kappa(1+\delta_\Sigma)\delta_\Sigma \leq \ln\left(\frac{1+\kappa\delta_\Sigma}{\sqrt{1+\delta_\Sigma}}\right)$$

Let

$$\begin{aligned}
g(\delta_\Sigma) &= \ln\left(\frac{1+\kappa\delta_\Sigma}{\sqrt{1+\delta_\Sigma}}\right) \\
g'(\delta_\Sigma) &= \frac{\kappa}{1+\kappa\delta_\Sigma}-\frac{1}{2(1+\delta_\Sigma)} = \frac{\kappa\delta_\Sigma+2\kappa-1}{2(1+\delta_\Sigma)(1+\kappa\delta_\Sigma)}
\end{aligned}$$

Provided that $\kappa \geq \frac{1}{2}$ and $\delta_\Sigma \leq \frac{2}{5}$,

$$g'(\delta_\Sigma) > \frac{2\kappa-1}{2(1+\frac{2}{5})(1+\frac{2}{5}\kappa)}.$$

Finally, since $g(0)=0$,

$$g(\delta_\Sigma) > \frac{2\kappa-1}{2(1+\frac{2}{5})(1+\frac{2}{5}\kappa)}\delta_\Sigma = \frac{5(2\kappa-1)}{14(1+\frac{2\kappa}{5})}\delta_\Sigma \geq \frac{7}{5}\gamma_\kappa\delta_\Sigma \geq (1+\delta_\Sigma)\,\gamma_\kappa\delta_\Sigma.$$

One deduces $(1+\kappa\delta_\Sigma)^{-p}\Phi_{\tilde{\upsilon}(x),(1+\delta_\Sigma)^{-1}\tilde{\Sigma}}(y) \leq \Phi_{\upsilon(x),\Sigma}(y)$. □

2.7.4 Proofs of inequalities used for bracketing entropy's decomposition

For sake of completness, we repeat here the proof of the inequalities of Lemmas 11, 12 and 13 of Cohen and Le Pennec [2011].

Proof. of inequality of Lemma 11 of Cohen and Le Pennec [2011]

For all x in $\mathcal{X}$,

$$
\begin{aligned}
&d^2_{y,k}(\pi^-(x)\psi^-(x,.),\pi^+(x)\psi^+(x,.))\\
&=\int\sum_{k=1}^{K}\left(\sqrt{\pi_k^+(x)}\left(\sqrt{\psi_k^+(x,y)}-\sqrt{\psi_k^-(x,y)}\right)\right.\\
&\qquad\left.+\sqrt{\psi_k^-(x,y)}\left(\sqrt{\pi_k^+(x)}-\sqrt{\pi_k^-(x)}\right)\right)^2 dy\\
&=\int\sum_{k=1}^{K}\pi_k^+(x)\left(\sqrt{\psi_k^+(x,y)}-\sqrt{\psi_k^-(x,y)}\right)^2 dy\\
&+\int\sum_{k=1}^{K}\psi_k^-(x,y)\left(\sqrt{\pi_k^+(x)}-\sqrt{\pi_k^-(x)}\right)^2 dy\\
&+2\sum_{k=1}^{K}\sqrt{\pi_k^+(x)}\left(\sqrt{\pi_k^+(x)}-\sqrt{\pi_k^-(x)}\right)\int\sqrt{\psi_k^-(x,y)}\left(\sqrt{\psi_k^+(x,y)}-\sqrt{\psi_k^-(x,y)}\right)dy\\
&\leq\left(\sum_{k=1}^{K}\pi_k^+(x)\right)\max_k d_y^2(\psi_k^+(x,.),\psi_k^-(x,.))+d_k^2(\pi^+(x),\pi^-(x))\max_k\int\psi_k^-(x,y)dy\\
&+2\sum_{k=1}^{K}\sqrt{\pi_k^+(x)}\left(\sqrt{\pi_k^+(x)}-\sqrt{\pi_k^-(x)}\right)d_y(\psi_k^+(x,.),\psi_k^-(x,.))\sqrt{\int\psi_k^-(x,y)dy}\\
&\leq\left(\sum_{k=1}^{K}\pi_k^+(x)\right)\max_k d_y^2(\psi_k^+(x,.),\psi_k^-(x,.))+d_k^2(\pi^+(x),\pi^-(x))\max_k\int\psi_k^-(x,y)dy\\
&+2\max_k\sqrt{\int\psi_k^-(x,y)dy}\max_k d_y(\psi_k^+(x,.),\psi_k^-(x,.))\left(\sum_{k=1}^{K}\pi_k^+(x)\right)^{1/2}d_k(\pi^+(x),\pi^-(x))\\
&\leq\left(\max_k d_y(\psi_k^+(x,.),\psi_k^-(x,.))\sqrt{\sum_{k=1}^{K}\pi_k^+(x)}\right.\\
&\qquad\left.+d_k(\pi^+(x),\pi^-(x))\max_k\sqrt{\int\psi_k^-(x,y)dy}\right)^2
\end{aligned}
$$

□

Proof. of inequality of Lemma 12 of Cohen and Le Pennec [2011]. For all x in $\mathcal{X}$,

$$
\begin{aligned}
&d_y^2\left(\sum_{k=1}^K \pi_k^-(x)\psi_k^-(x,.), \sum_{k=1}^K \pi_k^+(x)\psi_k^+(x,.)\right) = \int \sum_{k=1}^K \pi_k^+(x)\psi_k^+(x,y)dy \\
&+ \int \sum_{k=1}^K \pi_k^-(x)\psi_k^-(x,y)dy - 2\int \sqrt{\sum_{k=1}^K \pi_k^+(x)\psi_k^+(x,y)}\sqrt{\sum_{k=1}^K \pi_k^-(x)\psi_k^-(x,y)}dy \\
&\leq \int \sum_{k=1}^K \pi_k^+(x)\psi_k^+(x,y)dy + \int \sum_{k=1}^K \pi_k^-(x)\psi_k^-(x,y)dy \\
&\qquad\qquad - 2\int \sum_{k=1}^K \sqrt{\pi_k^+(x)\psi_k^+(x,y)}\sqrt{\pi_k^-(x)\psi_k^-(x,y)}dy \\
&\leq d_{y,k}^2(\pi^-(x)\psi^-(x,.), \pi^+(x)\psi^+(x,.))
\end{aligned}
$$

□

Proof. of inequality of Lemma 13 of Cohen and Le Pennec [2011].

We need to prove that for any x and any δ-Hellinger bracket $[t^-(x,y), t^+(x,y)]$, $\int t^-(x,y)dy \leq 1$ and $\int t^+(x,y)dy \leq \left(\delta + \sqrt{1+\delta^2}\right)^2$.

The first point is straightforward as t^- is upper-bounded by a density.

For the second point,

$$
\begin{aligned}
\int t^+ dy &= \int \left(t^+ - t^-\right) dy + \int t^- dy \leq \int \left(\sqrt{t^+} - \sqrt{t^-}\right)\left(\sqrt{t^+} + \sqrt{t^-}\right) dy + 1 \\
&\leq 2\int \left(\sqrt{t^+} - \sqrt{t^-}\right)\sqrt{t^+}\, dy + 1 \leq 2\left(\int \left(\sqrt{t^+} - \sqrt{t^-}\right)^2 dy\right)^{1/2}\left(\int t^+ dy\right)^{1/2} + 1 \\
\int t^+ dy &\leq 2\delta\left(\int t^+ dy\right)^{1/2} + 1
\end{aligned}
$$

Solving the corresponding inequality yields

$$\int t^+ dy \leq \left(\delta + \sqrt{1+\delta^2}\right)^2.$$

□

2.7.5 Proofs of lemmas used for Gaussian's bracketing entropy

Proof of Lemma 9

Proof.

$$\delta_\Sigma \leq \frac{1}{5\sqrt{\kappa^2\cosh(\frac{2\kappa}{5}) + \frac{1}{2}}}\frac{\delta}{p} \leq \frac{1}{5\sqrt{\kappa^2 + \frac{1}{2}}}\frac{\delta}{p} \leq \frac{1}{5\sqrt{\left(\frac{1}{2}\right)^2 + \frac{1}{2}}}\frac{\delta}{p} \leq \frac{2\sqrt{2}}{5\sqrt{3}p} \leq \frac{2}{5p}$$

□

Proof of Lemma 10

Proof.

$$2 - 2\,2^{d/2}\left((1+\delta_\Sigma) + (1+\delta_\Sigma)^{-1}\right)^{-d/2} = 2\left(1 - \left(\frac{e^{\ln(1+\delta_\Sigma)} + e^{-\ln(1+\delta_\Sigma)}}{2}\right)^{-d/2}\right)$$
$$= 2\left(1 - (\cosh(\ln(1+\delta_\Sigma)))^{-d/2}\right)$$
$$= 2f(\ln(1+\delta_\Sigma))$$

where $f(x) = 1 - \cosh(x)^{-d/2}$. Studying this function yields

$$f'(x) = \frac{d}{2}\sinh(x)\cosh(x)^{-d/2-1}$$
$$f''(x) = \frac{d}{2}\cosh(x)^{-d/2} - \frac{d}{2}\left(\frac{d}{2}+1\right)\sinh(x)^2\cosh(x)^{-d/2-2}$$
$$= \frac{d}{2}\left(1 - \left(\frac{d}{2}+1\right)\left(\frac{\sinh(x)}{\cosh(x)}\right)^2\right)\cosh(x)^{-d/2}$$

as $\cosh(x) \geq 1$, we have thus

$$f''(x) \leq \frac{d}{2}.$$

Now since $f(0) = 0$ and $f'(0) = 0$, this implies for any $x \geq 0$

$$f(x) \leq \frac{d}{2}\frac{x^2}{2} \leq \frac{d^2}{2}\frac{x^2}{2}.$$

We deduce thus that

$$2 - 2\,2^{d/2}\left((1+\delta_\Sigma) + (1+\delta_\Sigma)^{-1}\right)^{-d/2} \leq \frac{1}{2}d^2(\ln(1+\delta_\Sigma))^2$$

and using $\ln(1+\delta_\Sigma) \leq \delta_\Sigma$

$$2 - 2\,2^{d/2}\left((1+\delta_\Sigma) + (1+\delta_\Sigma)^{-1}\right)^{-d/2} \leq \frac{1}{2}d^2\delta_\Sigma^2.$$

Now,

$$(1+\kappa\delta_\Sigma)^d + (1+\kappa\delta_\Sigma)^{-d} - 2 = 2(\cosh(d\ln(1+\kappa\delta_\Sigma)) - 1) = 2g(d\ln(1+\kappa\delta_\Sigma))$$

with $g(x) = \cosh(x) - 1$. Studying this function yields

$$g'(x) = \sinh(x) \quad \text{and} \quad g''(x) = \cosh(x)$$

and thus, since $g(0) = 0$ and $g'(0) = 0$, for any $0 \leq x \leq c$

$$g(x) \leq \cosh(c)\frac{x^2}{2}.$$

Since $\ln(1+\kappa\delta_\Sigma) \leq \kappa\delta_\Sigma$, $d\delta_\Sigma \leq c$ implies $d\ln(1+\kappa\delta_\Sigma) \leq \kappa c$, we obtain thus

$$(1+\kappa\delta_\Sigma)^d + (1+\kappa\delta_\Sigma)^{-d} - 2 \leq \cosh(\kappa c)d^2(\ln(1+\kappa\delta_\Sigma))^2 \leq \kappa^2\cosh(\kappa c)d^2\delta_\Sigma^2.$$

□

Proof of Lemma 11

Proof. By definition,

$$\begin{aligned}
x'\left((1+\delta_\Sigma)\tilde{\Sigma}^{-1}-\Sigma^{-1}\right)x &= (1+\delta_\Sigma)\tilde{L}^{-1}\sum_{i=1}^{p}\tilde{A}_{i,i}^{-1}|\tilde{D}_i'x|^2 - L^{-1}\sum_{i=1}^{p}A_{i,i}^{-1}|D_i'x|^2 \\
&= (1+\delta_\Sigma)\tilde{L}^{-1}\sum_{i=1}^{p}\tilde{A}_{i,i}^{-1}|\tilde{D}_i'x|^2 - (1+\delta_\Sigma)\tilde{L}^{-1}\sum_{i=1}^{p}\tilde{A}_{i,i}^{-1}|D_i'x|^2 \\
&\quad + (1+\delta_\Sigma)\tilde{L}^{-1}\sum_{i=1}^{p}\tilde{A}_{i,i}^{-1}|D_i'x|^2 - (1+\delta_\Sigma)\tilde{L}^{-1}\sum_{i=1}^{p}A_{i,i}^{-1}|D_i'x|^2 \\
&\quad + (1+\delta_\Sigma)\tilde{L}^{-1}\sum_{i=1}^{p}A_{i,i}^{-1}|D_i'x|^2 - L^{-1}\sum_{i=1}^{p}A_{i,i}^{-1}|D_i'x|^2
\end{aligned}$$

Along the same lines,

$$\begin{aligned}
x'\left(\Sigma^{-1}-(1+\delta_\Sigma)^{-1}\tilde{\Sigma}^{-1}\right)x &= L^{-1}\sum_{i=1}^{p}A_{i,i}^{-1}|D_i'x|^2 - (1+\delta_\Sigma)^{-1}\tilde{L}^{-1}\sum_{i=1}^{p}\tilde{A}_{i,i}^{-1}|\tilde{D}_i'x|^2 \\
&= L^{-1}\sum_{i=1}^{p}A_{i,i}^{-1}|D_i'x|^2 - (1+\delta_\Sigma)^{-1}\tilde{L}^{-1}\sum_{i=1}^{p}A_{i,i}^{-1}|D_i'x|^2 \\
&\quad + (1+\delta_\Sigma)^{-1}\tilde{L}^{-1}\sum_{i=1}^{p}A_{i,i}^{-1}|D_i'x|^2 - (1+\delta_\Sigma)^{-1}\tilde{L}^{-1}\sum_{i=1}^{p}\tilde{A}_{i,i}^{-1}|D_i'x|^2 \\
&\quad + (1+\delta_\Sigma)^{-1}\tilde{L}^{-1}\sum_{i=1}^{p}\tilde{A}_{i,i}^{-1}|D_i'x|^2 - (1+\delta_\Sigma)^{-1}\tilde{L}^{-1}\sum_{i=1}^{p}\tilde{A}_{i,i}^{-1}|D_i'x|^2
\end{aligned}$$

Now

$$\begin{aligned}
\left|\sum_{i=1}^{p}\tilde{A}_{i,i}^{-1}|\tilde{D}_i'x|^2 - \sum_{i=1}^{p}\tilde{A}_{i,i}^{-1}|D_i'x|^2\right| &\le \sum_{i=1}^{p}\tilde{A}_{i,i}^{-1}\left||\tilde{D}_i'x|^2-|D_i'x|^2\right| \\
&\le \lambda_-^{-1}\sum_{i=1}^{p}\left||\tilde{D}_i'x|^2-|D_i'x|^2\right| \\
&\le \lambda_-^{-1}\sum_{i=1}^{p}\left||\tilde{D}_i'x|-|D_i'x|\right|\left||\tilde{D}_i'x|+|D_i'x|\right| \\
&\le \lambda_-^{-1}\left(\sum_{i=1}^{p}\left|(\tilde{D}_i-D_i)'x\right|^2\right)^{1/2}\left(\sum_{i=1}^{p}\left|(\tilde{D}_i+D_i)'x\right|^2\right)^{1/2} \\
&\le \lambda_-^{-1}\delta_D\|x\|2\|x\| = \lambda_-^{-1}2\delta_D\|x\|^2.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\left|\sum_{i=1}^{p}\tilde{A}_{i,i}^{-1}|D_i'x|^2 - \sum_{i=1}^{p}A_{i,i}^{-1}|D_i'x|^2\right| &\le \sum_{i=1}^{p}\left|\tilde{A}_{i,i}^{-1}-A_{i,i}^{-1}\right||D_i'x|^2 \\
&\le \delta_A\lambda_-^{-1}\sum_{i=1}^{p}|D_i'x|^2 = \delta_A\lambda_-^{-1}\|x\|^2.
\end{aligned}$$

We notice then that

$$(1+\delta_\Sigma)\tilde{L}^{-1}\sum_{i=1}^{p}A_{i,i}^{-1}|D_i'x|^2 - L^{-1}\sum_{i=1}^{p}A_{i,i}^{-1}|D_i'x|^2 = \left((1+\delta_\Sigma)\tilde{L}^{-1}-L^{-1}\right)\sum_{i=1}^{p}A_{i,i}^{-1}|D_i'x|^2$$
$$\geq (\delta_\Sigma-\delta_L)\tilde{L}^{-1}\lambda_+^{-1}\|x\|^2$$

while

$$L^{-1}\sum_{i=1}^{p}A_{i,i}^{-1}|D_i'x|^2 - (1+\delta_\Sigma)^{-1}\tilde{L}^{-1}\sum_{i=1}^{p}A_{i,i}^{-1}|D_i'x|^2 = \left(L^{-1}-(1+\delta_\Sigma)^{-1}\tilde{L}^{-1}\right)\sum_{i=1}^{p}A_{i,i}^{-1}|D_i'x|^2$$
$$\geq \left(1-(1+\delta_\Sigma)^{-1}\right)\tilde{L}^{-1}\lambda_+^{-1}\|x\|^2$$
$$\geq \frac{\delta_\Sigma}{1+\delta_\Sigma}\lambda_+^{-1}\tilde{L}^{-1}\|x\|^2$$

We deduce thus that

$$x'\left((1+\delta_\Sigma)\tilde{\Sigma}^{-1}-\Sigma^{-1}\right)x \geq (\delta_\Sigma-\delta_L)\tilde{L}^{-1}\lambda_+^{-1}\|x\|^2 - (1+\delta_\Sigma)\tilde{L}^{-1}\lambda_-^{-1}\left(2\delta_D+2\delta_A\right)\|x\|^2$$
$$\geq \tilde{L}^{-1}\left((\delta_\Sigma-\delta_L)\lambda_+^{-1}-(1+\delta_\Sigma)\lambda_-^{-1}\left(2\delta_D+\delta_A\right)\right)\|x\|^2$$

and

$$x'\left(\Sigma^{-1}-(1+\delta_\Sigma)^{-1}\tilde{\Sigma}^{-1}\right)x \geq \frac{\delta_\Sigma}{1+\delta_\Sigma}\tilde{L}^{-1}\lambda_+^{-1}\|x\|^2 - (1+\delta_\Sigma)^{-1}\tilde{L}^{-1}\lambda_-^{-1}\left(2\delta_D+\delta_A\right)\|x\|^2$$
$$\geq \frac{\tilde{L}^{-1}}{1+\delta_\Sigma}\left(\delta_\Sigma\lambda_+^{-1}-\lambda_-^{-1}\left(2\delta_D+\delta_A\right)\right)\|x\|^2$$

□

2.8 Appendix : Description of Newton-EM algorithm

In this section, Newton-EM algorithm is detailed. It consists in the classical EM algorithm in which the update of the weights has been replaced by some Newton steps. For further details on EM algorithm, refer to the technical report related to Young and Hunter [2010].

Newton-EM

Initialization Parameters for w, υ and Σ are given.

Newton steps for w Perform at most 5 steps Newton steps for w only while the like likelihood increases.

Maximization Update of υ and Σ with usual formulas in EM algorithm.

Initialization of Newton-EM

1. Draw K couples of points (X_i, Y_i) among data, defining K lines υ_l.
2. Classify the data: $k = \arg\min_l |Y_i - \upsilon_l(X_i)|$.
3. Proceed 3 steps of Newton-EM initialized with $w = 0$ and empirical covariance matrices and means.
4. Repeat 50 times the previous steps and choose the set of parameters with the greatest likelihood among the 50.

Chapter 3

PAC-Bayesian aggregation of linear estimators

Sommaire

Abstract : Aggregating estimators using exponential weights depending on their risk performs well in expectation, but sadly not in probability. Considering exponential weights of a penalized risk is a way to overcome this issue. In this case, an oracle inequality can be obtained in probability, but is not sharp. Taking into account the estimated function's norm in the penalty offers a sharp inequality.

Keywords : Exponentially weighted aggregation, Regression, Oracle inequality

3.1 Introduction

We consider here a classical fixed design regression model

$$\forall i \in \{1, \dots, n\},\ Y_i = f_0(x_i) + W_i$$

with f_0 an unknown function, x_i the fixed design points and $W = (W_i)_{i \leq n}$ a centered sub-Gaussian noise. Our aim is to estimate the function f_0 at the grid points.

Many regression estimators are available in the literature. For non parametric estimation, Nadaraya-Watson estimator [Nadaraya, 1965, Watson, 1964] and its fixed

design counterpart [Gasser and Müller, 1984] are widely used, just like projection estimators using trigonometric, wavelet [Donoho et al., 1995] or spline [Wahba, 1990] basis for example. In the parametric framework, least squares or maximum likelihood estimators are commonly employed, sometimes with minimization constraints, leading to LASSO [Tibshirani, 1994], ridge [Hastie et al., 2009], elastic net [Zou and Hastie, 2005], AIC [Akaike, 1973] or BIC [Schwarz, 1978] estimates.

Facing this variety, the statistician may wonder which procedure provides the best estimation. Unfortunately, the answer depends on the data. For instance, a rectangular function is well approximated by wavelets but not by trigonometric functions. Since the best estimator is not known in advance, our aim is to mimic its performances in term of risk. This is theoretically guaranteed by an oracle inequality:

$$R(f_0, \tilde{f}) \leq C_n \inf_{t \in \mathcal{T}} R(f_0, \hat{f}_t) + \epsilon_n$$

comparing the risk of the constructed estimator $\tilde{f}$ to the risk of the best available procedure in the collection $\{\hat{f}_t, t \in \mathcal{T}\}$. Our strategy is based on convex combination of these preliminary estimators and relies on PAC-Bayesian aggregation to obtain a single adaptive estimator. We focus on a wide family, commonly used in practice : linear estimators $\{\hat{f}_t(Y) = P_t Y | P_t \in \mathcal{S}_n^+(\mathbb{R}), t \in \mathcal{T}\}$.

Aggregation procedures have been introduced by Vovk [1990], Littlestone and Warmuth [1994], Cesa-Bianchi et al. [1997], Cesa-Bianchi and Lugosi [1999]. They are a central ingredient of bagging [Breiman, 1996], boosting [Freund, 1995, Schapire, 1990] or random forest (Amit and Geman [1997] or Breiman [2001]; or more recently Biau et al. [2008], Biau and Devroye [2010], Biau [2012], Genuer [2011]).

The general aggregation framework is detailed in Nemirovski [2000] and studied in Catoni [2004, 2007] through a PAC-Baysian framework as well as in Yang [2000a,b,c, 2001, 2003, 2004a,b]. See for instance Tsybakov [2008] for a survey. Optimal rates of aggregation in regression and density estimation are studied by Tsybakov [2003], Lounici [2007], Rigollet and Tsybakov [2007], Rigollet [2006] and Lecué [2007].

A way to traduce the confidence in each preliminary estimate is to aggregate according to a measure exponentially decreasing when the estimate's risk rises. This widely used strategy is called exponentially weighted aggregation. More precisely, the weight of each element $\hat{f}_t$ in the collection is proportional to $\exp\left(-\frac{\widetilde{r}_t}{\beta}\right)\pi(t)$ where $\widetilde{r}_t$ is a, possibly penalized, estimate of the risk of $\hat{f}_t$, β is a positive parameter, called the temperature, that has to be calibrated and π is a prior measure over $\mathcal{T}$. The main interest of exponential weights resides in Lemma 12 [Catoni, 2007] since they explicitly minimize the aggregated risk penalized by the Kullback-Leibler divergence to the a priori measure π. Our aim is to give sufficient conditions on the risk estimate $\widetilde{r}_t$ and the temperature β to obtain an oracle inequality for the risk of the aggregate.

This procedure has shown its efficiency, offering lower risk than model selection because we bet on several estimators. Aggregation of projections has already been addressed by Leung and Barron [2006]. They have proved by the mean of an oracle inequality, that in expectation, the aggregate performs almost as well as the best projection in the collection. Those results have been extended to several settings and noise conditions [Dalalyan and Tsybakov, 2007, 2008, 2012, Giraud, 2008, Dalalyan

et al., 2013, Belloni et al., 2011, Dalalyan, 2012, Giraud et al., 2012, Sun and Zhang, 2012, Rigollet and Tsybakov, 2012] under a *frozen* estimator assumption: they should not depend on the observed sample. This restriction, not present in the work by Leung and Barron [2006], has been removed by Dalalyan and Salmon [2012] within the context of affine estimator and exponentially weighted aggregation.

However, Dai et al. [2012] have shown the sub-optimality in deviation of exponential weighting, not allowing to obtain a sharp oracle inequality in probability. Nevertheless, with Gaussian white noise, penalizing the risk in the weights and taking a temperature at least 20 times greater than the noise variance allows to upper bound the risk of the aggregate, based on affine estimators, in probability [Dai et al., 2014]. Furthermore, the corresponding oracle inequality is not sharp.

Our contribution is twofold. First, we propose an extension to general sub-Gaussian noise. Second, we conduct a fine analysis of the relationship between the choice of the penalty and the temperature. In particular, we are able to take into account the signal to noise ratio to provide sharp oracle inequalities for bounded functions.

Note that our results are similar to the one obtained for a slightly different aggregation scheme by Bellec [2014] in a preprint written while the authors were working independently on this one. In this work, the weights are implicitly defined and the noise variables are assumed independent, which is not the case here.

3.2 Framework and estimate

Recall that we observe

$$\forall i \in \{1, \dots, n\},\ Y_i = f_0(x_i) + W_i$$

with f_0 an unknown function and x_i the fixed grid points. Our main assumption on the noise is that $W = (W_i)_{i \leq n} \in \mathbb{R}^n$ is a centered sub-Gaussian variable, i.e. $\mathbb{E}(W) = 0$ and there exists $\sigma^2 \in \mathbb{R}^+$ such that

$$\forall \alpha \in \mathbb{R}^n,\ \mathbb{E}\left[\exp\left(\alpha^\top W\right)\right] \leq \exp\left(\frac{\sigma^2}{2}\|\alpha\|_2^2\right),$$

where $\|.\|_2$ is the usual euclidean norm in $\mathbb{R}^n$. If W is a centered Gaussian vector with covariance matrix Σ then σ^2 is nothing but the largest eigenvalue of Σ. Note that the components of W are not supposed independent.

The quality of our estimate will be measured through its error at the design points. More precisely, we will consider the classical euclidean loss, related to the squared norm

$$\|g\|_2^2 = \sum_{i=1}^{n} g(x_i)^2.$$

Thus, our unknown is the vector $(f_0(x_i))_{i=1}^n$ rather than the function f_0.

Assume that we have at hand a collection of data dependent smoothed projection estimates

$$\hat{f}_t(Y) = \sum_{i=1}^{n} \rho_{t,i} \langle Y, b_{t,i} \rangle b_{t,i}$$

where $(b_{t,i})_{i=1}^n$ is an orthonormal basis and $(\rho_{t,i})_{i=1}^n$ a sequence of non-negative real numbers. For such an estimate, it exists a symmetric positive semi-definite real matrix of size n, P_t such that $\hat{f}_t(Y) = P_t Y$. For the sake of simplicity, we use this representation of our estimators. Note that we depart from the affine estimator studied by Dalalyan and Salmon [2012] and Dai et al. [2014] because we consider only linear estimate. This choice was made to simplify our exposition but similar results as the one we obtain for linear estimates hold for affine ones.

To define our estimate from the collection $\{\hat{f}_t(Y) = P_t Y | P_t \in \mathcal{S}_n^+(\mathbb{R}), t \in \mathcal{T}\}$, we specify the estimate $\widetilde{r}_t$ of the risk of the estimator $\hat{f}_t(Y)$, choose a prior probability measure π over $\mathcal{T}$ and a temperature $\beta > 0$. We define f_{EWA} by $f_{EWA} = \int \hat{f}_t \, d\rho(t)$, with

$$d\rho(t) = \frac{\exp\left(-\frac{1}{\beta}\widetilde{r}_t\right)}{\int \exp\left(-\frac{1}{\beta}\widetilde{r_{t'}}\right) d\pi(t')} d\pi(t)$$

a probability measure over $\mathcal{T}$. The intuition behind this construction to favor low risk estimates.

When the temperature goes to zero this estimator becomes very similar to the one minimizing the risk estimate while it becomes an indiscriminate average when β grows to infinity. The choice of the temperature appears thus to be crucial and a low temperature seems to be desirable.

Our choice for the risk estimate $\widetilde{r}_t$ is to use the classical Stein unbiased estimate

$$r_t = \|Y - \hat{f}_t(Y)\|_2^2 + 2\sigma^2 tr(P_t) - n\sigma^2$$

to which a penalty $\mathrm{pen}(t)$ is added. Thus $\widetilde{r}_t$ is a modified version of Stein unbiased estimate, used to obtain optimal oracle inequalities in expectation. We will consider simultaneously the case of a penalty that depends on f_0 through an upper bound of a kind of sup norm and the case of a penalty that does not depend on f_0.

More precisely, we allow the use, at least in the analysis, of an upper bound $\|\widetilde{f_0}\|_\infty$ which can be thought as the supremum of the sup norm of the coefficients of f_0 in any basis appearing in $\mathcal{T}$. Indeed, we define $\|\widetilde{f_0}\|_\infty$ as the smallest non-negative real number C such that for any $t \in \mathcal{T}$,

$$\|P_t f_0\|_2^2 \leq C^2 tr(P_t^2).$$

By construction, $\|\widetilde{f_0}\|_\infty$ is smaller than the sup norm of any coefficients of f_0 in any basis appearing in the $\mathcal{T}$. Note that $\|\widetilde{f_0}\|_\infty$ can also be upper bounded by $\|f_0\|_1$, $\|f_0\|_2$ or $\sqrt{n}\|f_0\|_\infty$ where the ℓ_1 and sup norm can be taken in any basis.

We assume furthermore that the matrix collection $\{P_t\}_{t\in\mathcal{T}}$ satisfies the following condition : there exists a finite $V > 0$ such that $\sup_{t\in\mathcal{T}} \|P_t\|_2 \leq V$. Our aim is to obtain sufficient conditions on the penalty $\mathrm{pen}(t)$ and the temperature β so that an oracle inequality of type

$$\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\mu\in\mathcal{M}_+^1(\mathcal{T})} (1+\epsilon) \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t) + (1+\epsilon') \int (\mathrm{pen}(t) + \mathrm{price}(t)) d\mu(t) + (1+\epsilon')\beta \left(2KL(\mu,\pi) + \ln\frac{1}{\eta}\right)$$

holds, with ϵ and ϵ' small non-negative numbers possibly equal to 0 and price(t) a loss depending on the choice of pen(t) and β. Such an oracle proves that the risk of our aggregate estimate is of the same order as the one of the best estimate in the collection up to some controlled cost.

3.3 Penalization strategies and preliminary results

A first approach may be to consider a penalty proportional to the noise level σ^2, as in Dai et al. [2014] for the Gaussian white noise with known variance :

Proposition 3 (Dai et al. [2014])**.** *If pen$(t) = 2\sigma^2 tr(P_t)$, and $\beta \geq 4\sigma^2 \max(16, 5V)$, then for all $\eta > 0$, with probability at least $1 - \eta$,*

$$\|f_0 - f_{EWA}\|^2 \leq \min_t \left\{ \left(1 + \frac{128\sigma^2}{3\beta}\right) \|f_0 - \hat{f}_t\|^2 + 8\sigma^2 tr(P_t) + 3\beta \ln\left(\frac{1}{\eta\pi_t}\right) \right\}.$$

We obtained a similar result for sub-Gaussian noise with parameter σ^2, which may be simply stated in the case of orthogonal projections.

Proposition 4. *If $\beta > 8\sigma^2$, and pen$(t) \geq \frac{2\sigma^4}{\beta - 4\sigma^2} tr(P_t)$, let $\epsilon = \frac{4\sigma^2}{\beta - 8\sigma^2}$. Then for any $\eta > 0$, with probability at least $1 - \eta$,*

$$\|f_0 - f_{EWA}\|^2 \leq \min_t \Bigg\{ (1 + 2\epsilon) \|f_0 - \hat{f}_t\|^2 + 2(1 + \epsilon) \Big(pen(t) + \sigma^2 tr(P_t) \Big) + \beta(1 + \epsilon) \ln\left(\frac{1}{\eta\pi_t^2}\right) \Bigg\}.$$

In general, if $P_t \in \mathcal{S}_n^+(\mathbb{R})$, a technical parameter $\gamma \geq 0$ is introduced to get the same kind of inequality. We temporarily assume that $V > 1/2$ to simplify the statement.

Proposition 5. *If $\beta \geq 20\sigma^2 V$, there exists $\gamma \geq 0$, such that if pen$(t) \geq \frac{4\sigma^2}{\beta - 4\sigma^2 V} tr(P_t^2)\sigma^2$, for any $\eta > 0$, with probability at least $1 - \eta$,*

$$\|f_0 - f_{EWA}\|^2 \leq \inf_t \left\{ \left(1 + \frac{4V^2\gamma}{(2V - 1)(1 - 2V\gamma)}\right) \|f_0 - \hat{f}_t\|^2 + \frac{1}{1 - 2V\gamma} \Big(pen(t) + 2\sigma^2 tr(A_t) \Big) + \frac{\beta}{1 - 2V\gamma} \ln\left(\frac{1}{\eta\pi_t^2}\right) \right\}.$$

The parameter γ allows to link $\|P_t f_0 - P_u f_0\|_2^2$ to $\|f_0 - P_t Y\|_2^2$ and $\|f_0 - P_u Y\|_2^2$, which is natural in the case of orthogonal projections, since $\|f_0 - P_t f_0\|_2^2 \leq \|f_0 - P_t Y\|_2^2$.

A weak oracle inequality is satisfactory but sharp ones are also useful. Adding the knowledge of a sup norm of f_0, such a result can be obtained. Note that here the parameter γ is not necessary.

Proposition 6. *If* $\beta > 4\sigma^2 V$, *and* $pen(t) \geq \frac{4\sigma^2}{\beta - 4\sigma^2 V}\left(1 + \frac{\widetilde{\|f\|}_\infty^2}{\sigma^2}\right) tr(P_t^2)\sigma^2$, *then for any* $\eta > 0$, *with probability at least* $1-\eta$,

$$\|f_0 - f_{EWA}\|^2 \leq \inf_t \left\{ \|f_0 - \hat{f}_t\|^2 + pen(t) + 2\sigma^2 \left(tr(P_t) + \frac{2\sigma^2}{\beta - 4\sigma^2 V} \frac{\widetilde{\|f_0\|}_\infty^2}{\sigma^2} tr(P_t^2) \right) + \beta \ln\left(\frac{1}{\eta \pi_t^2}\right) \right\}.$$

For the sake of clarity, here is the simplified version in the case of orthogonal projections.

Corollary 1. *If* $\beta > 4\sigma^2$ *and* $pen(t) = \kappa tr(P_t)\sigma^2$, *with* $\kappa \geq \frac{2\sigma^2}{\beta - 4\sigma^2}\left(1 + 2\frac{\widetilde{\|f_0\|}_\infty^2}{\sigma^2}\right)$, *then for any* $\eta > 0$, *with probability at least* $1-\eta$,

$$\|f_0 - f_{EWA}\|^2 \leq \inf_t \left\{ \|f_0 - \hat{f}_t\|^2 + \left(\kappa + 2 + \frac{4\sigma^2}{\beta - 4\sigma^2} \frac{\widetilde{\|f_0\|}_\infty^2}{\sigma^2} \right) \sigma^2 tr(A_t) + \beta \ln\left(\frac{1}{\eta \pi_t^2}\right) \right\}.$$

To obtain a weak oracle inequality, we use the following type of inequality :

$$\|(P_t - P_u) f_0\|_2^2 \leq C_1 \|P_t Y - f_0\|_2^2 + C_2 \|P_u Y - f_0\|_2^2, \tag{3.1}$$

with some constants C_1 and C_2 depending on γ, whereas sharp oracle inequality is provided by

$$\|(P_t - P_u) f_0\|_2^2 \leq 2(\|P_t f_0\|_2^2 + \|P_u f_0\|_2^2) \leq 2\widetilde{\|f_0\|}_\infty^2 (tr(P_t^2) + tr(P_u^2)). \tag{3.2}$$

Combining these two upper bounds produce weak oracle inequalities for a wider range of temperatures than Proposition 5, drawing a continuum between Proposition 5 and Proposition 6. More precisely, using a convex combination of 3.1 and 3.2, yields

Proposition 7. *For any* $\delta \in [0,1]$, *if* $\beta \geq 4\sigma^2 V(1+4\delta)$ *and* $\beta > 4\sigma^2 V$, *there exists* $\gamma \geq 0$, *such that if* $pen(t) \geq \frac{4\sigma^4}{\beta - 4\sigma^2 V}\left(1 + (1-\delta)(1+2\gamma V)^2 \frac{\widetilde{\|f_0\|}_\infty^2}{\sigma^2}\right) tr(P_t^2)$, *then for any* $\eta > 0$, *with probability at least* $1-\eta$,

$$\|f_0 - f_{EWA}\|^2 \leq \inf_t \left\{ \left(1 + \frac{4V^2\gamma}{(2V-1)(1-2V\gamma)}\right) \|f_0 - \hat{f}_t\|^2 + \frac{\beta}{1-2V\gamma} \ln\left(\frac{1}{\eta \pi_t^2}\right) + \frac{1}{1-2V\gamma} \left(pen(t) + 2\sigma^2 tr(P_t) + \frac{4\sigma^4(1-\delta)(1+2\gamma V)^2}{\beta - 4\sigma^2 V} \frac{\widetilde{\|f_0\|}_\infty^2}{\sigma^2} tr(P_t^2) \right) \right\}.$$

The convex combination parameter δ measures the account for signal to noise ratio in the penalty. We are now ready to state the central result of this paper, which gives an explicit expression for γ and introduce an optimization parameter $\nu > 0$.

3.4 A general oracle inequality

Our main result is the following:

Theorem 3. *Assume W is a centered sub-Gaussian noise with parameter σ^2, and $\{\hat{f}_t(Y) = P_tY | P_t \in \mathcal{S}_n^+(\mathbb{R}), t \in \mathcal{T}\}$ is such that there exists $V > 0$ satisfying $\sup_{t\in\mathcal{T}} \|P_t\|_2 \leq V$.*

Let π be a arbitrary prior measure on $\mathcal{T}$, $\beta > 4\sigma^2 V$ an arbitrary temperature and $pen(t)$ a penalty so that $f_{EWA} = \int \hat{f}_t \, d\rho(t)$ with

$$d\rho(t) = \frac{\exp\left(-\frac{1}{\beta}\left[r_t + pen(t)\right]\right)}{\int \exp\left(-\frac{1}{\beta}\left[r_{t'} + pen(t')\right]\right) d\pi(t')} d\pi(t).$$

For any $\delta \in [0,1]$, if $\beta \geq 4\sigma^2 V(1+4\delta)$, let

$$\gamma = \frac{\beta - 4\sigma^2 V(1+2\delta) - \sqrt{\beta - 4\sigma^2 V}\sqrt{\beta - 4\sigma^2 V(1+4\delta)}}{16\sigma^2 \delta V^2} \mathbb{1}_{\delta>0}.$$

If for any $t \in \mathcal{T}$,

$$pen(t) \geq \frac{4\sigma^2}{\beta - 4\sigma^2 V}\left(1 + (1-\delta)(1+2\gamma V)^2 \frac{\|\widetilde{f_0}\|_\infty^2}{\sigma^2}\right) tr(P_t^2)\sigma^2,$$

then

— *for any $\eta \in (0,1]$, with probability at least $1-\eta$,*

$$\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\nu\in N} \inf_{\mu\in\mathcal{M}_+^1(\mathcal{T})} (1+\epsilon(\nu)) \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t)$$
$$+ (1+\epsilon'(\nu)) \int (pen(t) + \text{price}(t)) \, d\mu(t) + \beta(1+\epsilon'(\nu)) \left(2KL(\mu,\pi) + \ln\frac{1}{\eta}\right)$$

— *Furthermore*

$$\mathbb{E}\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\nu\in N} \inf_{\mu\in\mathcal{M}_+^1(\mathcal{T})} (1+\epsilon(\nu)) \int \mathbb{E}\|f_0 - \hat{f}_t\|_2^2 d\mu(t)$$
$$+ (1+\epsilon'(\nu)) \int (pen(t) + \text{price}(t)) \, d\mu(t) + 2\beta(1+\epsilon'(\nu)) KL(\mu,\pi)$$

with

$$\text{price}(t) = 2\sigma^2 \left(tr(P_t) + \frac{2\sigma^2(1-\delta)(1+2\gamma V)^2}{\beta - 4\sigma^2 V} \frac{\|\widetilde{f_0}\|_\infty^2}{\sigma^2} tr(P_t^2)\right)$$
$$\epsilon'(\nu) = \frac{1}{1-(1+\nu)\gamma} - 1$$
$$\epsilon(\nu) = \frac{(1+\nu)^2\gamma}{\nu(1-(1+\nu)\gamma)} = \frac{(1+\nu)^2}{\nu}\gamma(1+\epsilon'(\nu))$$

and $N = \{\nu > 0 | (1+\nu)\gamma < 1\}$.

This theorem is similar to the one obtained by Dai et al. [2014] (Prop. 3). It yields a sufficient condition on the penalty for oracle inequalities to hold both in probability and in expectation. It holds however under a milder sub-Gaussianity assumption on the noise and allows to take into account a sup norm information in the penalty as used for instance in Guedj and Alquier [2013] . Note that the result in expectation requires a penalty that is not necessary, at least in the Gaussian case, as shown by Dalalyan and Salmon [2012].

If we are authorized to use the upper bound of the sup norm, we may ensure that the penalty satisfies the lower bound condition with $\delta = 0$. In that case, $\gamma = 0$ and $\epsilon'(\nu) = 0, \epsilon(\nu) = 0$. Thus, there is no need to optimize ν and it suffices to notice that $N = (0, +\infty)$ to obtain that

Corollary 2. *Under the assumptions of Theorem 3, if $\beta > 4\sigma^2 V$, if for any $t \in \mathcal{T}$,*

$$pen(t) \geq \frac{4\sigma^4}{\beta - 4\sigma^2 V}\left(1 + \frac{\widetilde{\|f_0\|}_\infty^2}{\sigma^2}\right) tr(P_t^2),$$

then

— *for any $\eta \in (0, 1]$, with probability at least $1 - \eta$,*

$$\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\mu \in \mathcal{M}_+^1(\mathcal{T})} \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t) + \int \left(pen(t) + \text{price}(t)\right) d\mu(t) + \beta\left(2KL(\mu, \pi) + \ln\frac{1}{\eta}\right)$$

— *Furthermore*

$$\mathbb{E}\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\mu \in \mathcal{M}_+^1(\mathcal{T})} \int \mathbb{E}\|f_0 - \hat{f}_t\|_2^2 d\mu(t) + \int \left(pen(t) + \text{price}(t)\right) d\mu(t) + 2\beta KL(\mu, \pi)$$

with

$$\text{price}(t) = 2\sigma^2\left(tr(P_t) + \frac{2}{\beta - 4\sigma^2 V}\widetilde{\|f_0\|}_\infty^2 tr(P_t^2)\right).$$

As soon as $\delta > 0$, a simple calculation yields that for any $\beta \geq 4\sigma^2 V(1 + 4\delta)$, $0 < 2\gamma V \leq 1$. As a result, if $V > 0.5$, $(0, 2V - 1) \subseteq N$. Furthermore if $\beta > 4\sigma^2 V(1 + 4\delta)$ and $V > 0.5$, then $2V - 1 \in N$. Else, if $\delta > 0$ and $0 < V \leq 0.5$, N is non-empty if and only if $\beta > 4\sigma^2 V + 2\sigma^2\delta(1 + 2V)^2$, which is a stronger condition than $\beta \geq 4\sigma^2 V(1 + 4\delta)$. Since V is an upper bound, it may be chosen greater than 0.5.

If $\delta = 1$, we obtain weak oracle inequalities that do not require the use of any side information:

Corollary 3. *Under the assumptions of Theorem 3, if $\beta \geq 20\sigma^2 V$, let*

$$\gamma = \frac{\beta - 12\sigma^2 V - \sqrt{\beta - 4\sigma^2 V}\sqrt{\beta - 20\sigma^2 V}}{16\sigma^2 V^2}.$$

If for any $t \in \mathcal{T}$,

$$pen(t) \geq \frac{4\sigma^4}{\beta - 4\sigma^2 V} tr(P_t^2),$$

then

— *for any* $\eta \in (0,1]$, *with probability at least* $1-\eta$,

$$\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\nu \in N} \inf_{\mu \in \mathcal{M}_+^1(\mathcal{T})} (1+\epsilon(\nu)) \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t)$$
$$+ (1+\epsilon'(\nu)) \int (pen(t) + \text{price}(t))\, d\mu(t) + \beta(1+\epsilon'(\nu)) \left(2KL(\mu,\pi) + \ln \frac{1}{\eta}\right)$$

— *Furthermore*

$$\mathbb{E}\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\nu \in N} \inf_{\mu \in \mathcal{M}_+^1(\mathcal{T})} (1+\epsilon(\nu)) \int \mathbb{E}\|f_0 - \hat{f}_t\|_2^2 d\mu(t)$$
$$+ (1+\epsilon'(\nu)) \int (pen(t) + \text{price}(t))\, d\mu(t) + 2\beta(1+\epsilon'(\nu)) KL(\mu,\pi)$$

with

$$\text{price}(t) = 2\sigma^2 tr(P_t)$$
$$\epsilon'(\nu) = \frac{1}{1-(1+\nu)\gamma} - 1$$
$$\epsilon(\nu) = \frac{(1+\nu)^2\gamma}{\nu(1-(1+\nu)\gamma)} = \frac{(1+\nu)^2}{\nu}\gamma(1+\epsilon'(\nu))$$

and $N = \{\nu > 0 | (1+\nu)\gamma < 1\}$.

Finally, assume that we let

$$\text{pen}(t) \geq \kappa tr(P_t^2)\sigma^2.$$

The previous corollary implies that a weak oracle inequality holds for any temperature greater than $20\sigma^2 V$ as soon as $\kappa \geq \frac{4\sigma^2}{\beta - 4\sigma^2 V}$. Corollary 2 implies that an exact oracle inequality holds for any vector f_0 and any temperature β greater than $4\sigma^2 V$ as soon as

$$\frac{\beta - 4\sigma^2 V}{4\sigma^2}\kappa - 1 \geq (1+2\gamma V)^2 \frac{\|\widetilde{f_0}\|_\infty^2}{\sigma^2}.$$

For fixed κ and β, this corresponds to a low peak signal to noise ratio $\frac{\|\widetilde{f_0}\|_\infty^2}{\sigma^2}$. Theorem 3 shows that there is a continuum between those two cases as weak oracle inequalities, with smaller leading constant than the one of Corollary 3, hold as soon as there exists $\delta \in [0,1]$ such that $\beta \geq 4\sigma^2(1+4\delta)V$ and

$$\frac{\beta - 4\sigma^2 V}{4\sigma^2}\kappa - 1 \geq (1-\delta)(1+2\gamma V)^2 \frac{\|\widetilde{f_0}\|_\infty^2}{\sigma^2},$$

where the signal to noise ratio guides the transition. The temperature required remains nevertheless always above $4\sigma^2 V$.

The minimal temperature of $4\sigma^2 V(1+4\delta)$ can be replaced by some smaller values if one further restrict the smoothed projections used. As it appears in the proof, the temperature can be replaced by $4\sigma^2(1+\delta)$ or even $2\sigma^2(2+\delta)$ when the smoothed projections are respectively classical projections (see Theorem 4) and projections in the same basis. The question of the minimality of such temperature is still open. Note that in this proof, there is no loss due to the sub-Gaussianity assumption, since the same upper bound on the exponential moment of the deviation as in the Gaussian case are found, providing the same penalty and bound on temperature.

The proof of this result is quite long and thus postponed in Appendix 3.7. We provide first the generic proof of the oracle inequalities, highlighting the role of Gibbs measure and of some control in deviation. Then, we focus on the aggregation of projection estimators in the Gaussian model. This example already conveys all the ideas used in the complete proof of the deviation lemma : exponential moments inequalities for Gaussian quadratic form and the control of the bias $\|f_0 - P_t f\|_2^2$ by $\|\widetilde{f_0}\|_\infty^2$ on the one hand, to obtain an exact oracle inequality, and by $\|f_0 - P_t Y\|_2^2$ on the other hand, giving a weak inequality.

The extension to the general case is obtained by showing that similar exponential moments inequalities can be obtained for quadratic form of sub-Gaussian random variables, working along the fact that the systematic bias $\|f_0 - P_t f\|_2^2$ is no longer always smaller than $\|f_0 - P_t Y\|_2^2$ and providing a fine tuning optimization allowing the equality in the constraint on β and an optimization on the parameters ϵ.

3.5 Proof of the oracle inequalities

Theorem 3 relies on the characterization of Gibbs measure (Lemma 12) and a control of deviation of the empirical risk of any aggregate around its true risk, allowed by Lemma 13 or Lemma 14.

ρ is a Gibbs measure. Therefore it maximizes the entropy for a given expected energy. That is the subject of Lemma 1.1.3 in Catoni [2007]:

Lemma 12. *For any bounded measurable function $h : \mathcal{T} \to \mathbb{R}$, and any probability distribution $\rho \in \mathcal{M}_+^1(\mathcal{T})$ such that $KL(\rho, \pi) < \infty$,*

$$\log\left(\int \exp(h) d\pi\right) = \int h d\rho - KL(\rho, \pi) + KL(\rho, \pi_{\exp(h)}),$$

where by definition $\frac{d\pi_{\exp(h)}}{d\pi} = \frac{\exp[h(t)]}{\int \exp(h) d\pi}$. Consequently,

$$\log\left(\int \exp(h) d\pi\right) = \sup_{\rho \in \mathcal{M}_+^1(\mathcal{T})} \int h d\rho - KL(\rho, \pi).$$

With $h(t) = -\frac{1}{\beta}[r_t + \mathrm{pen}(t)]$, this lemma states that for any probability distribution $\mu \in \mathcal{M}_+^1(\mathcal{T})$ such that $KL(\mu, \pi) < \infty$,

$$\int h d\rho - KL(\rho, \pi) \geq \int h d\mu - KL(\mu, \pi).$$

Equivalently,

$$\begin{aligned}
&\int \|f_0 - \hat{f}_t\|_2^2 d\rho(t) + \int \left(r_t - \|f_0 - \hat{f}_t\|_2^2 + \mathrm{pen}(t)\right) d\rho(t) + \beta KL(\rho, \pi) \\
&\quad \leq \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t) + \int \left(r_t - \|f_0 - \hat{f}_t\|_2^2 + \mathrm{pen}(t)\right) d\mu(t) + \beta KL(\mu, \pi) \\
\Leftrightarrow &\int \|f_0 - \hat{f}_t\|_2^2 d\rho(t) - \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t) \leq \int \left(\|f_0 - \hat{f}_t\|_2^2 - r_t\right) d\rho(t) - \beta KL(\rho, \pi) \\
&\quad - \int \left(\|f_0 - \hat{f}_t\|_2^2 - r_t\right) d\mu(t) - \int \mathrm{pen}(t) d\rho(t) + \int \mathrm{pen}(t) d\mu(t) + \beta \mathrm{KL}(\mu, \pi).
\end{aligned}$$

The key is to upper bound the right-hand side with terms that may depend on ρ, but only through $\int \|f_0 - \hat{f}_t\|_2^2 d\rho(t)$ and Kullback-Leibler distance. This is the purpose of Lemma 13 in the case of Gaussian noise with projections estimators and Lemma 14 in the sub-Gaussian case. Under mild assumptions, they provide upper bounds in probability (and in expectation) of type:

$$\begin{aligned}
&\int \left(\|f_0 - \hat{f}_t\|_2^2 - r_t\right) d\rho(t) - \int \left(\|f_0 - \hat{f}_u\|_2^2 - r_u\right) d\mu(u) \\
&\qquad \leq C_1 \int \|f_0 - \hat{f}_t\|_2^2 d\rho(t) + C_2 \int \|f_0 - \hat{f}_u\|_2^2 d\mu(u) \\
&\quad + C_3 \int tr(P_t^2) d\rho(t) + C_4 \int tr(P_u) d\mu(u) + C_5 \int tr(P_u^2) d\mu(u) \\
&\qquad + \beta \left(KL(\rho, \pi) + KL(\mu, \pi) + \ln \frac{1}{\eta} \right)
\end{aligned}$$

where C_1 to C_6 are known functions. Combining with the previous inequality and taking $\mathrm{pen}(t) \geq C_3\, tr(P_t^2)$ gives

$$\begin{aligned}
&(1 - C_1) \int \|f_0 - \hat{f}_t\|_2^2 d\rho(t) - (1 + C_2) \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t) \\
&\qquad \leq C_4 \int tr(P_u) d\mu(u) + C_5 \int tr(P_u^2) d\mu(u) + \int \mathrm{pen}(t) d\mu(t) \\
&\qquad + \beta \left(2KL(\mu, \pi) + \ln \frac{1}{\eta} \right).
\end{aligned}$$

The additional condition $C_1 < 1$ allows to conclude. It is now clear that the whole work lies in the obtention of the lemma.

3.6 The expository case of Gaussian noise and projection estimates

In this section, to provide a simplified proof, we assume that P_t are the matrices of orthogonal projections and the noise W is a centered Gaussian random variable with variance $\sigma^2 I$. The previous theorem becomes:

Theorem 4. *Let π be an arbitrary prior measure over $\mathcal{T}$. For any $\delta \in [0, 1]$, any $\beta > 4\sigma^2(\delta + 1)$, the aggregate estimator f_{EWA} defined with*

$$pen(t) \geq \frac{2\sigma^4}{\beta - 4\sigma^2} \left(1 + 2(1 - \delta) \frac{\|\widetilde{f_0}\|_\infty^2}{\sigma^2} \right) tr(P_t)$$

satisfies

— *for any* $\eta \in (0,1]$, *with probability at least* $1-\eta$,

$$\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\mu \in \mathcal{M}_+^1(\mathcal{T})} (1+2\epsilon) \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t) + (1+\epsilon) \int (pen(t) + \text{price}(t))\, d\mu(t) + \beta(1+\epsilon) \left(2KL(\mu,\pi) + \ln \frac{1}{\eta} \right).$$

— *Furthermore,*

$$\mathbb{E}\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\mu \in \mathcal{M}_+^1(\mathcal{T})} (1+2\epsilon) \int \mathbb{E}\|f_0 - \hat{f}_t\|_2^2 d\mu(t) + (1+\epsilon) \int (pen(t) + \text{price}(t))\, d\mu(t) + 2\beta(1+\epsilon) 2KL(\mu,\pi),$$

with

$$\text{price}(t) = 2\left(1 + \frac{2(1-\delta)\sigma^2}{\beta - 4\sigma^2} \frac{\|\widetilde{f_0}\|_\infty^2}{\sigma^2}\right) tr(P_t)\sigma^2 \quad \text{and} \quad \epsilon = \frac{4\sigma^2\delta}{\beta - 4\sigma^2(\delta+1)}.$$

Note that $\text{pen}(t) \geq \text{price}(t) + 2\sigma^2 \left(\frac{\sigma^2}{\beta - 4\sigma^2} - 1\right) tr(P_t)$, and the result may be further simplified using $\text{pen}(t) + \text{price}(t) \leq 2\left(\text{pen}(t) + \sigma^2 tr(P_t)\right)$.

As announced in the scheme of proof of the oracle inequalities (section 3.5), the key is a control of the deviation of the empirical risk of any aggregate around its true risk. It is allowed by Lemma 13 in this case.

Lemma 13. *For any prior probability distribution* π, *any* $\delta \in [0,1]$ *and any* $\beta > 4\sigma^2$, *for any probability distributions* ρ *and* μ,

— *For any* $\eta > 0$, *with probability at least* $1-\eta$,

$$\begin{aligned}
&\int \left(\|f_0 - \hat{f}_t\|_2^2 - r_t\right) d\rho(t) - \int \left(\|f_0 - \hat{f}_u\|_2^2 - r_u\right) d\mu(u) \\
&\quad \leq \frac{4\delta\sigma^2}{\beta - 4\sigma^2} \left(\int \|f_0 - \hat{f}_t\|_2^2 d\rho(t) + \int \|f_0 - \hat{f}_u\|_2^2 d\mu(u) \right) \\
&\qquad + \frac{4\sigma^2}{\beta - 4\sigma^2} \left(\sigma^2 + (1-\delta)\|\widetilde{f_0}\|_\infty^2 \right) \int tr(P_t) d\rho(t) \\
&\qquad + 2\sigma^2 \left(1 + \frac{2(1-\delta)\|\widetilde{f_0}\|_\infty^2}{\beta - 4\sigma^2} \right) \int tr(P_u) d\mu(u) \\
&\qquad + \beta \left(KL(\rho,\pi) + KL(\mu,\pi) + \ln \frac{1}{\eta} \right)
\end{aligned}$$

— *Moreover,*

$$\mathbb{E}\left[\int \left(\|f_0 - \hat{f}_t\|_2^2 - r_t\right) d\rho(t) - \int \left(\|f_0 - \hat{f}_u\|_2^2 - r_u\right) d\mu(u)\right]$$
$$\leq \mathbb{E}\left[\frac{4\delta\sigma^2}{\beta - 4\sigma^2}\left(\int \|f_0 - \hat{f}_t\|_2^2 d\rho(t) + \int \|f_0 - \hat{f}_u\|_2^2 d\mu(u)\right)\right.$$
$$+ \frac{4\sigma^2}{\beta - 4\sigma^2}\left(\sigma^2 + (1-\delta)\|\widetilde{f_0}\|_\infty^2\right)\int tr(P_t) d\rho(t)$$
$$+ 2\sigma^2\left(1 + \frac{2(1-\delta)\|\widetilde{f_0}\|_\infty^2}{\beta - 4\sigma^2}\right)\int tr(P_u) d\mu(u)$$
$$\left. + \beta\left(KL(\rho,\pi) + KL(\mu,\pi)\right)\right].$$

The use of this lemma is detailed in section 3.6.2. We focus now on its proof mixing control of exponential moments of a quadratic form of a Gaussian random variable with basic inequalities like Jensen, Fubini, and the important link between $\|f_0 - P_t f_0\|_2^2$ and $\|f_0 - P_t Y\|_2^2$. Note that this link is obvious in the case of orthogonal projections and need to be established differently in the general case, leading to technicalities (the introduction of γ).

3.6.1 Proof of Lemma 13

Proof. For the sake of clarity, for any $t, u \in \mathcal{T}$, let

$$\Delta_{t,u} = \|f_0 - \hat{f}_t\|_2^2 - r_t - \|f_0 - \hat{f}_u\|_2^2 + r_u.$$

A simple calculation yields

$$\Delta_{t,u} = 2\left(W^\top(P_t - P_u)W + W^\top(P_t - P_u)f_0 - \sigma^2 tr(P_t - P_u)\right).$$

Since $(P_t)_{t\in\mathcal{T}}$ are positive semi-definite matrices, $W^\top(P_t - P_u)W \leq W^\top P_t W$, and there exist an orthogonal matrix U and a diagonal matrix D such that $P_t = U^\top DU$.

For any $\beta > 0$,

$$\mathbb{E}\left[\exp\frac{\Delta_{t,u}}{\beta}\right] \leq \mathbb{E}\left[\exp\frac{2}{\beta}\left((UW)^\top D(UW) + (UW)^\top U(P_t - P_u)f - \sigma^2 tr(P_t - P_u)\right)\right].$$

Following lemma 2.4 of Hsu et al. [2012], if $\beta > 4\sigma^2$,

$$\mathbb{E}\left[\exp\frac{\Delta_{t,u}}{\beta}\right] \leq \exp\frac{2\sigma^2}{\beta}\left(tr(P_u) + \frac{2\sigma^2 tr(P_t) + \|(P_t - P_u)f_0\|_2^2}{\beta - 4\sigma^2}\right). \tag{3.3}$$

Note that

$$\|(P_t - P_u)f_0\|_2^2 \leq 2\left(\|f_0 - P_t f_0\|_2^2 + \|f_0 - P_u f_0\|_2^2\right) \leq 2\left(\|f_0 - P_t Y\|_2^2 + \|f_0 - P_u Y\|_2^2\right)$$

and

$$\|(P_t - P_u)f_0\|_2^2 \leq 2\left(\|P_t f_0\|_2^2 + \|P_u f_0\|_2^2\right) \leq 2\|\widetilde{f_0}\|_\infty^2\left(tr(P_t) + tr(P_u)\right).$$

Thus, for any $\beta > 4\sigma^2$, for any $\delta \in [0,1]$,

$$\mathbb{E}\exp\left[\frac{\Delta_{t,u}}{\beta} - \frac{2\sigma^2}{\beta}\left(tr(P_u) + \frac{2\sigma^2 tr(P_t)}{\beta - 4\sigma^2}\right) - \frac{4\sigma^2\delta}{\beta(\beta-4\sigma^2)}\left(\|f_0 - \hat{f}_t\|_2^2 + \|f_0 - \hat{f}_u\|_2^2\right) - \frac{4\sigma^2}{\beta(\beta-4\sigma^2)}(1-\delta)\|\widetilde{f_0}\|_\infty^2 (tr(P_t) + tr(P_u))\right] \leq 1.$$

Along the same lines as Alquier and Lounici [2011], we first integrate according to the prior π and use Fubini's theorem,

$$\mathbb{E}\int\int \exp\frac{1}{\beta}\left[\Delta_{t,u} - 2\sigma^2\left(tr(P_u) + \frac{2\sigma^2 tr(P_t)}{\beta - 4\sigma^2}\right) - \frac{4\sigma^2\delta}{\beta-4\sigma^2}\left(\|f_0 - \hat{f}_t\|_2^2 + \|f_0 - \hat{f}_u\|_2^2\right) - \frac{4\sigma^2}{\beta-4\sigma^2}(1-\delta)\|\widetilde{f_0}\|_\infty^2 (tr(P_t) + tr(P_u))\right] d\pi(t)d\pi(u) \leq 1,$$

then introduce the probability distributions ρ and μ, and $\eta > 0$

$$\mathbb{E}\int\int \exp\frac{1}{\beta}\left[\Delta_{t,u} - 2\sigma^2\left(tr(P_u) + \frac{2\sigma^2 tr(P_t)}{\beta - 4\sigma^2}\right) - \frac{4\sigma^2\delta}{\beta-4\sigma^2}\left(\|f_0 - \hat{f}_t\|_2^2 + \|f_0 - \hat{f}_u\|_2^2\right) - \frac{4\sigma^2(1-\delta)}{\beta-4\sigma^2}\|\widetilde{f_0}\|_\infty^2 (tr(P_t) + tr(P_u)) - \beta\left(\ln\frac{d\rho}{d\pi}(t) + \ln\frac{d\mu}{d\pi}(u) + \ln\frac{1}{\eta}\right)\right] d\rho(t)d\mu(u) \leq \eta,$$

before applying Jensen's inequality

$$\mathbb{E}\exp\frac{1}{\beta}\left[\int\int \Delta_{t,u}d\rho(t)d\mu(u) - \frac{4\delta\sigma^2}{\beta-4\sigma^2}\left(\int\|f_0 - \hat{f}_t\|_2^2 d\rho(t) + \int\|f_0 - \hat{f}_u\|_2^2 d\mu(u)\right) - \frac{4\sigma^2}{\beta-4\sigma^2}\left(\sigma^2 + (1-\delta)\|\widetilde{f_0}\|_\infty^2\right)\int tr(P_t)d\rho(t) - \beta\left(KL(\rho,\pi) + KL(\mu,\pi) + \ln\frac{1}{\eta}\right) - 2\sigma^2\left(1 + \frac{2(1-\delta)\|\widetilde{f_0}\|_\infty^2}{\beta-4\sigma^2}\right)\int tr(P_u)d\mu(u)\right] \leq \eta. \quad (3.4)$$

Finally, using the basic inequality $\exp(x) \geq \mathbb{1}_{\mathbb{R}_+}(x)$,

$$\mathbb{P}\left[\int\int \Delta_{t,u}d\rho(t)d\mu(u) \leq \frac{4\delta\sigma^2}{\beta-4\sigma^2}\left(\int\|f_0 - \hat{f}_t\|_2^2 d\rho(t) + \int\|f_0 - \hat{f}_u\|_2^2 d\mu(u)\right) + \frac{4\sigma^2}{\beta-4\sigma^2}\left(\sigma^2 + (1-\delta)\|\widetilde{f_0}\|_\infty^2\right)\int tr(P_t)d\rho(t) + \beta\left(KL(\rho,\pi) + KL(\mu,\pi) + \ln\frac{1}{\eta}\right) + \frac{2\sigma^2}{n}\left(1 + \frac{2(1-\delta)n\|\widetilde{f_0}\|_\infty^2}{\beta n - 4\sigma^2}\right)\int tr(P_u)d\mu(u)\right] \geq 1 - \eta.$$

The result in expectation is obtained by Equation (3.4) with $\eta = 1$:

$$\mathbb{E}\exp\frac{1}{\beta}\left[\int\int \Delta_{t,u}d\rho(t)d\mu(u) - \frac{4\delta\sigma^2}{\beta-4\sigma^2}\left(\int \|f_0-\hat{f}_t\|_2^2 d\rho(t) + \int \|f_0-\hat{f}_u\|_2^2 d\mu(u)\right)\right.$$
$$-\frac{4\sigma^2}{\beta-4\sigma^2}\left(\sigma^2+(1-\delta)\|\widetilde{f_0}\|_\infty^2\right)\int tr(P_t)d\rho(t) - \beta\left(KL(\rho,\pi)+KL(\mu,\pi)\right)$$
$$\left.-2\sigma^2\left(1+\frac{2(1-\delta)\|\widetilde{f_0}\|_\infty^2}{\beta-4\sigma^2}\right)\int tr(P_u)d\mu(u)\right] \leq 1,$$

combined with the inequality $t \leq \exp(t) - 1$. □

3.6.2 Proof of Theorem 4

We follow the scheme of proof given in section 3.5 and use Lemma 13, leading to the following result: for any $\eta > 0$, any prior probability distribution π, any $\delta \in [0,1]$ and any $\beta > 4\sigma^2(1+\delta)$, with probability at least $1-\eta$, for any probability distribution μ,

$$\int \|f_0-\hat{f}_t\|_2^2 d\rho(t) - \int \|f_0-\hat{f}_t\|_2^2 d\mu(t)$$
$$\leq \frac{4\delta\sigma^2}{\beta-4\sigma^2}\left(\int \|f_0-\hat{f}_t\|_2^2 d\rho(t) + \int \|f_0-\hat{f}_u\|_2^2 d\mu(u)\right)$$
$$+\frac{4\sigma^2}{\beta-4\sigma^2}\left(\sigma^2+(1-\delta)\|\widetilde{f_0}\|_\infty^2\right)\int tr(P_t)d\rho(t) - \int \mathrm{pen}(t)d\rho(t)$$
$$+2\sigma^2\left(1+\frac{2(1-\delta)\|\widetilde{f_0}\|_\infty^2}{\beta-4\sigma^2}\right)\int tr(P_t)d\mu(t) + \int \mathrm{pen}(t)d\mu(t)$$
$$+\beta\left(2KL(\mu,\pi)+\ln\frac{1}{\eta}\right).$$

With $\mathrm{pen}(t) \geq \frac{4\sigma^2}{\beta-4\sigma^2}\left(\sigma^2+(1-\delta)\|\widetilde{f_0}\|_\infty^2\right)tr(P_t)$, the previous inequality becomes

$$\left(1-\frac{4\delta\sigma^2}{\beta-4\sigma^2}\right)\int \|f_0-\hat{f}_t\|_2^2 d\rho(t) - \left(1+\frac{4\delta\sigma^2}{\beta-4\sigma^2}\right)\int \|f_0-\hat{f}_t\|_2^2 d\mu(t)$$
$$\leq 2\sigma^2\left(1+\frac{2(1-\delta)\|\widetilde{f_0}\|_\infty^2}{\beta-4\sigma^2}\right)\int tr(P_t)d\mu(t) + \int \mathrm{pen}(t)d\mu(t) + \beta\left(2KL(\mu,\pi)+\ln\frac{1}{\eta}\right).$$

Furthermore, using

$$\|f_0 - f_{EWA}\|_2^2 \leq \int \|f_0-\hat{f}_t\|_2^2 d\rho(t),$$

if $\beta > 4\sigma^2(\delta+1)$, we obtain

$$\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\mu\in\mathcal{M}_+^1(\mathcal{T})}\left(1+\frac{8\sigma^2\delta}{\beta-4\sigma^2(\delta+1)}\right)\int\|f_0-\hat{f}_t\|_2^2 d\mu(t)$$
$$+\left(1+\frac{4\sigma^2\delta}{\beta-4\sigma^2(\delta+1)}\right)2\sigma^2\left(1+\frac{2(1-\delta)\|\widetilde{f_0}\|_\infty^2}{\beta-4\sigma^2}\right)\int tr(P_t)d\mu(t)$$
$$+\left(1+\frac{4\sigma^2\delta}{\beta-4\sigma^2(\delta+1)}\right)\int \mathrm{pen}(t)d\mu(t)+\beta\left(1+\frac{4\sigma^2\delta}{\beta-4\sigma^2(\delta+1)}\right)\left(2KL(\mu,\pi)+\ln\frac{1}{\eta}\right).$$

In addition, taking $\epsilon = \frac{4\sigma^2\delta}{\beta-4\sigma^2(\delta+1)}$, gives

$$\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\mu\in\mathcal{M}_+^1(\mathcal{T})}(1+2\epsilon)\int\|f_0-\hat{f}_t\|_2^2 d\mu(t)$$
$$+2\sigma^2(1+\epsilon)\left(1+\frac{2(1-\delta)\|\widetilde{f_0}\|_\infty^2}{\beta-4\sigma^2}\right)\int tr(P_t)d\mu(t)$$
$$+(1+\epsilon)\left(\int \mathrm{pen}(t)d\mu(t)+2\beta KL(\mu,\pi)+\beta\ln\frac{1}{\eta}\right).$$

3.7 Appendix : Proofs in the sub-Gaussian case

3.7.1 Proof of Theorem 3

The proof follows from the scheme described in section 3.5. The main point is still to control

$$\int\left(\|f_0-\hat{f}_t\|_2^2 - r_t\right)d\rho(t) - \int\left(\|f_0-\hat{f}_t\|_2^2 - r_t\right)d\mu(t).$$

We recall that P_t is a symmetric positive semi-definite matrix, there exists $V>0$ such that $\sup_{t\in\mathcal{T}}\|P_t\|_2 \leq V$ and W is a centered sub-Gaussian noise. For any $t,u\in\mathcal{T}$, we still denote $\Delta_{t,u} = \|f_0-\hat{f}_t\|_2^2 - r_t - \|f_0-\hat{f}_u\|_2^2 + r_u$.

Lemma 14. *Let π be an arbitrary prior probability. For any $\delta\in[0,1]$, any $\beta > 4\sigma^2V$ and $\beta \geq 4\sigma^2V(1+4\delta)$, let*

$$\gamma = \frac{1}{16\sigma^2\delta V^2}\left(\beta - 4\sigma^2V(1+2\delta) - \sqrt{\beta-4\sigma^2V}\sqrt{\beta-4\sigma^2V(1+4\delta)}\right)\mathbb{1}_{\delta>0}.$$

Then, for any probability distributions ρ and μ, for any $\nu>0$,

— *for any $\eta\in(0,1]$, with probability at least $1-\eta$,*

$$\int\int\Delta_{t,u}d\rho(t)d\mu(u) \leq (1+\nu)\gamma\int\|P_tY-f_0\|_2^2 d\rho(t)$$
$$+\frac{4\sigma^2}{\beta-4\sigma^2V}\left(\sigma^2+(1-\delta)(1+2\gamma V)^2\|\widetilde{f_0}\|_\infty^2\right)\int tr(P_t^2)d\rho(t)$$
$$+2\sigma^2\left(\int tr(P_u)d\mu(u)+\frac{2(1-\delta)(1+2\gamma V)^2}{\beta-4\sigma^2V}\|\widetilde{f_0}\|_\infty^2\int tr(P_u^2)d\mu(u)\right)$$
$$+\left(1+\frac{1}{\nu}\right)\gamma\int\|P_uY-f_0\|_2^2 d\mu(u)+\beta\left(KL(\rho,\pi)+KL(\mu,\pi)+\ln\frac{1}{\eta}\right)$$

— *Moreover,*

$$\mathbb{E}\left[\int\int \Delta_{t,u} d\rho(t)d\mu(u)\right] \leq \mathbb{E}\Big[(1+\nu)\gamma \int \|P_t Y - f_0\|_2^2 d\rho(t)$$
$$+ \frac{4\sigma^2}{\beta - 4\sigma^2 V}\left(\sigma^2 + (1-\delta)(1+2\gamma V)^2 \|\widetilde{f_0}\|_\infty^2\right)\int tr(P_t^2)d\rho(t)$$
$$+ 2\sigma^2\left(\int tr(P_u)d\mu(u) + \frac{2(1-\delta)(1+2\gamma V)^2}{\beta - 4\sigma^2 V}\|\widetilde{f_0}\|_\infty^2 \int tr(P_u^2)d\mu(u)\right)$$
$$+ \left(1+\frac{1}{\nu}\right)\gamma\int \|P_u Y - f_0\|_2^2 d\mu(u) + \beta\left(KL(\rho,\pi) + KL(\mu,\pi)\right)\Big]$$

Under the assumptions of the previous lemma, with probability at least $1-\eta$,

$$\int \|f_0 - \hat{f}_t\|_2^2 d\rho(t) - \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t) \leq (1+\nu)\gamma\int \|\hat{f}_t - f_0\|_2^2 d\rho(t)$$
$$+ \frac{4\sigma^2}{\beta - 4\sigma^2 V}\left(\sigma^2 + (1-\delta)(1+2\gamma V)^2\|\widetilde{f_0}\|_\infty^2\right)\int tr(P_t^2)d\rho(t) - \int \mathrm{pen}(t)d\rho(t)$$
$$+ 2\sigma^2\left(\int tr(P_t)d\mu(t) + \frac{2(1-\delta)(1+2\gamma V)^2}{\beta - 4\sigma^2 V}\|\widetilde{f_0}\|_\infty^2\int tr(P_t^2)d\mu(t)\right) + \int \mathrm{pen}(t)d\mu(t)$$
$$+ \left(1+\frac{1}{\nu}\right)\gamma\int \|\hat{f}_t - f_0\|_2^2 d\mu(t) + \beta\left(2KL(\mu,\pi) + \ln\frac{1}{\eta}\right).$$

Taking $\mathrm{pen}(t) \geq \frac{4\sigma^2}{\beta - 4\sigma^2 V}\left(\sigma^2 + (1-\delta)(1+2\gamma V)^2\|\widetilde{f_0}\|_\infty^2\right) tr(P_t^2)$ and $\nu \in N = \{\nu > 0 | (1+\nu)\gamma < 1\}$, such that the inequality stays informative,

$$(1-(1+\nu)\gamma)\int \|f_0 - \hat{f}_t\|_2^2 d\rho(t) \leq \left(1+\left(1+\frac{1}{\nu}\right)\gamma\right)\int \|f_0 - \hat{f}_t\|_2^2 d\mu(t)$$
$$+ 2\sigma^2\left(\int tr(P_t)d\mu(t) + \frac{2(1-\delta)(1+2\gamma V)^2}{\beta - 4\sigma^2 V}\|\widetilde{f_0}\|_\infty^2\int tr(P_t^2)d\mu(t)\right)$$
$$+ \int \mathrm{pen}(t)d\mu(t) + \beta\left(2KL(\mu,\pi) + \ln\frac{1}{\eta}\right).$$

Finally, since $\|f_0 - f_{EWA}\|_2^2 \leq \int \|f_0 - \hat{f}_t\|_2^2 d\rho(t)$,

$$\|f_0 - f_{EWA}\|_2^2 \leq \left(1 + \frac{(1+\nu)^2\gamma}{\nu(1-(1+\nu)\gamma)}\right)\int \|f_0 - \hat{f}_t\|_2^2 d\mu(t)$$
$$+ \frac{2\sigma^2}{1-(1+\nu)\gamma}\left(\int tr(P_t)d\mu(t) + \frac{2(1-\delta)(1+2\gamma V)^2}{\beta - 4\sigma^2 V}\|\widetilde{f_0}\|_\infty^2\int tr(P_t^2)d\mu(t)\right)$$
$$+ \frac{1}{1-(1+\nu)\gamma}\left(\int \mathrm{pen}(t)d\mu(t) + \beta\left(2KL(\mu,\pi) + \ln\frac{1}{\eta}\right)\right).$$

The result in expectation is obtained in the same fashion.

3.7.2 Proof of Lemma 14

The exponential moment of $\Delta_{t,u}$ is easily controlled by a term involving $\|P_t f_0 - f_0\|_2^2$ (see Equation (3.3)). Since P_t are not projections, $\|P_t f_0 - f_0\|_2^2 \leq \|P_t Y - f_0\|_2^2$

does not hold any more. The presence of $\|P_tY - f_0\|_2^2$ allows us to obtain a weak oracle inequality. To overcome this difficulty, $\|(P_t - P_u)Y\|_2^2$ is introduced and for an arbitrary $\gamma \geq 0$, we try to control $\Delta_{t,u} - \gamma\|(P_t - P_u)Y\|_2^2$.

Proof. A simple calculation yields

$$\begin{aligned}\Delta_{t,u} - \gamma\|(P_t - P_u)Y\|_2^2 &= W^\top(2I - \gamma(P_t - P_u)^\top)(P_t - P_u)W \\ &\quad + 2W^\top(I - \gamma(P_t - P_u)^\top)(P_t - P_u)f_0 - 2\sigma^2 tr(P_t - P_u) - \gamma\|(P_t - P_u)f_0\|_2^2.\end{aligned}$$

Noting that $W^\top(2I - \gamma(P_t - P_u)^\top)(P_t - P_u)W \leq 2W^\top(P_t - P_u)W$ and since $(P_t)_{t\in\mathcal{T}}$ are positive semi-definite matrices, $2W^\top(P_t - P_u)W \leq 2W^\top P_t W$. Thus, for any $\beta > 0$, any $\gamma \geq 0$,

$$\begin{aligned}&\mathbb{E}\exp\left(\frac{\Delta_{t,u}}{\beta} - \frac{\gamma}{\beta}\|(P_t - P_u)Y\|_2^2\right) \\ &\qquad \leq \mathbb{E}\left[\exp\frac{2}{\beta}\left(W^\top P_t W + W^\top(I - \gamma(P_t - P_u))(P_t - P_u)f_0\right)\right] \\ &\qquad\qquad \times \exp\frac{-1}{\beta}\left(2\sigma^2 tr(P_t - P_u) + \gamma\|(P_t - P_u)f_0\|_2^2\right).\end{aligned}$$

The first step is to bring us back to the Gaussian case, using W's sub-Gaussianity and an idea of Hsu et al. [2012]. Let Z be a standard Gaussian random variable, independent of W. Then,

$$\begin{aligned}&\mathbb{E}\exp\left(\frac{2}{\sqrt{\beta}}W^\top\sqrt{P_t}Z + \frac{2}{\beta}W^\top(I - \gamma(P_t - P_u))(P_t - P_u)f_0\right) \\ &\quad = \mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{2}{\sqrt{\beta}}W^\top\sqrt{P_t}Z + \frac{2}{\beta}W^\top(I - \gamma(P_t - P_u))(P_t - P_u)f_0\right)\Big|W\right]\right] \\ &\quad = \mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{2}{\sqrt{\beta}}W^\top\sqrt{P_t}Z\right)\Big|W\right]\exp\left(\frac{2}{\beta}W^\top(I - \gamma(P_t - P_u))(P_t - P_u)f_0\right)\right] \\ &\quad = \mathbb{E}\exp\frac{2}{\beta}\left(W^\top P_t W + W^\top(I - \gamma(P_t - P_u))(P_t - P_u)f_0\right).\end{aligned}$$

On the other hand,

$$\begin{aligned}&\mathbb{E}\left[\exp\frac{2}{\beta}\left(W^\top P_t W + W^\top(I - \gamma(P_t - P_u))(P_t - P_u)f_0\right)\right] \\ &\qquad = \mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{2}{\sqrt{\beta}}W^\top\sqrt{P_t}Z + \frac{2}{\beta}W^\top(I - \gamma(P_t - P_u))(P_t - P_u)f_0\right)\Big|Z\right]\right].\end{aligned}$$

Since W is sub-Gaussian with parameter σ,

$$\begin{aligned}&\mathbb{E}\left[\exp\frac{2}{\beta}\left(W^\top P_t W + W^\top(I - \gamma(P_t - P_u))(P_t - P_u)f_0\right)\right] \\ &\qquad\qquad \leq \mathbb{E}\exp\left(\frac{\sigma^2}{2}\left\|\frac{2}{\sqrt{\beta}}\left(\sqrt{P_t}Z + \frac{1}{\sqrt{\beta}}(I - \gamma(P_t - P_u))(P_t - P_u)f_0\right)\right\|_2^2\right)\end{aligned}$$

Hence,

$$\mathbb{E}\exp\left(\frac{\Delta_{t,u}}{\beta}-\frac{\gamma}{\beta}\|(P_t-P_u)Y\|_2^2\right)$$
$$\leq \mathbb{E}\left[\exp\frac{2\sigma^2}{\beta}\left(Z^\top P_t Z+\frac{2}{\sqrt{\beta}}Z^\top\sqrt{P_t}(I-\gamma(P_t-P_u))(P_t-P_u)f_0\right)\right]$$
$$\times\exp\left(\frac{2\sigma^2}{\beta^2}\|(I-\gamma(P_t-P_u))(P_t-P_u)f_0\|_2^2-\frac{2\sigma^2}{\beta}tr(P_t-P_u)-\frac{\gamma}{\beta}\|(P_t-P_u)f_0\|_2^2\right).$$

The expectation is similar to the one obtained in the Gaussian case: the exponential of some quadratic form. The same recipe is applied. Since P_t is positive semi-definite, there exist an orthogonal matrix U and a diagonal matrix D such that $P_t = U^\top DU$. Note that UZ is a standard Gaussian variable. This diagonalization step and the non-negativity of the eigenvalues allow to apply Lemma 2.4 of Hsu et al. [2012]. Then, for any $\beta > 4\sigma^2 V$, any $\gamma \geq 0$,

$$\mathbb{E}\exp\left(\frac{\Delta_{t,u}}{\beta}-\frac{\gamma}{\beta}\|(P_t-P_u)Y\|_2^2\right)$$
$$\leq \exp\frac{2\sigma^2}{\beta}\left(tr(P_t)+\frac{2\sigma^2}{\beta(\beta-4\sigma^2V)}\left(\beta tr(P_t^2)+2\left\|\sqrt{P_t}(I-\gamma(P_t-P_u))(P_t-P_u)f_0\right\|_2^2\right)\right)$$
$$\times\exp\left(\frac{2\sigma^2}{\beta^2}\|(I-\gamma(P_t-P_u))(P_t-P_u)f_0\|_2^2-\frac{2\sigma^2}{\beta}tr(P_t-P_u)-\frac{\gamma}{\beta}\|(P_t-P_u)f_0\|_2^2\right).$$

Consequently,

$$\mathbb{E}\exp\left(\frac{\Delta_{t,u}}{\beta}+\frac{\gamma}{\beta}\left(\|(P_t-P_u)f_0\|_2^2-\|(P_t-P_u)Y\|_2^2\right)\right)$$
$$\leq \exp\frac{2\sigma^2}{\beta}\left(tr(P_u)+\frac{2\sigma^2}{\beta-4\sigma^2V}tr(P_t^2)\right)$$
$$\times\exp\left(\frac{2\sigma^2}{\beta^2}\left(\frac{4\sigma^2V}{\beta-4\sigma^2V}(1+2\gamma V)^2+(1+2\gamma V)^2\right)\|(P_t-P_u)f_0\|_2^2\right).$$
$$\leq \exp\frac{2\sigma^2}{\beta}\left(tr(P_u)+\frac{2\sigma^2}{\beta-4\sigma^2V}tr(P_t^2)+\frac{(1+2\gamma V)^2}{\beta-4\sigma^2V}\|(P_t-P_u)f_0\|_2^2\right).$$

If an exact oracle inequality is wished, $\|(P_t-P_u)f_0\|_2^2$ should be upper bounded by some constant and γ should be set to zero. Else, γ is used to *replace* the terms in $\|(P_t-P_u)f_0\|_2^2$ by $\|(P_t-P_u)Y\|_2^2$. Thus, the terms depending on f_0 will be upper bounded in two ways:

— on the one hand, using $\|\widetilde{f_0}\|_\infty^2$

$$\|(P_t-P_u)f_0\|_2^2 \leq 2\left(\|P_tf_0\|_2^2+\|P_uf_0\|_2^2\right) \leq 2\left(tr(P_t^2)+tr(P_u^2)\right)\|\widetilde{f_0}\|_\infty^2$$

For any $\delta \in [0,1]$,

$$\begin{aligned}
&\mathbb{E}\exp\left(\frac{\Delta_{t,u}}{\beta} + \frac{\gamma}{\beta}\left(\|(P_t - P_u)f_0\|_2^2 - \|(P_t - P_u)Y\|_2^2\right)\right) \\
&\leq \exp\frac{2\sigma^2}{\beta}\left(tr(P_u) + \frac{2\sigma^2}{\beta - 4\sigma^2 V}tr(P_t^2) + \frac{(1+2\gamma V)^2(1-\delta)}{\beta - 4\sigma^2 V}\|(P_t - P_u)f_0\|_2^2\right) \\
&\times \exp\left(\frac{2\sigma^2(1+2\gamma V)^2\delta}{\beta(\beta - 4\sigma^2 V)}\|(P_t - P_u)f_0\|_2^2\right) \\
&\leq \exp\frac{2\sigma^2}{\beta}\left(tr(P_u) + \frac{2\sigma^2}{\beta - 4\sigma^2 V}tr(P_t^2) + \frac{(1+2\gamma V)^2\delta}{\beta - 4\sigma^2 V}\|(P_t - P_u)f_0\|_2^2\right) \\
&\times \exp\left(\frac{4\sigma^2(1+2\gamma V)^2(1-\delta)}{\beta(\beta - 4\sigma^2 V)}\left(tr(P_t^2) + tr(P_u^2)\right)\|\widetilde{f_0}\|_\infty^2\right).
\end{aligned}$$

— on the other hand, introducing $\|P_tY - f_0\|_2^2$ to obtain a weak oracle inequality: conditions should be found on γ such that

$$\begin{aligned}
\frac{2\sigma^2(1+2\gamma V)^2\delta}{\beta - 4\sigma^2 V}\|(P_t - P_u)f_0\|_2^2 - \gamma\left(\|(P_t - P_u)f_0\|_2^2 - \|(P_t - P_u)Y\|_2^2\right) \\
\leq C_1\|P_tY - f_0\|_2^2 + C_2\|P_uY - f_0\|_2^2
\end{aligned}$$

for some non-negative constants C_1 and C_2 and with $\delta > 0$. Since for any $\nu > 0$, $\|(P_t - P_u)Y\|_2^2 \leq (1+\nu)\|P_tY - f_0\|_2^2 + \left(1 + \frac{1}{\nu}\right)\|P_uY - f_0\|_2^2$, it suffices that

$$\frac{2\sigma^2(1+2\gamma V)^2\delta}{\beta - 4\sigma^2 V}\|(P_t - P_u)f_0\|_2^2 - \gamma\|(P_t - P_u)f_0\|_2^2 \leq 0.$$

This condition may be fulfilled if $\beta \geq 4\sigma^2 V(1+4\delta)$. The smallest $\gamma \geq 0$ among all the possible ones is chosen :

$$\gamma = \frac{1}{16\sigma^2\delta V^2}\left(\beta - 4\sigma^2 V(1+2\delta) - \sqrt{\beta - 4\sigma^2 V}\sqrt{\beta - 4\sigma^2 V(1+4\delta)}\right)\mathbb{1}_{\delta>0}.$$

This leads to the following inequality : for any $\delta \in [0,1]$, for any $\beta > 4\sigma^2 V$ and $\beta \geq 4\sigma^2 V(1+4\delta)$, with γ previously defined, for any $\nu > 0$,

$$\begin{aligned}
&\mathbb{E}\exp\left(\frac{\Delta_{t,u}}{\beta} - \frac{\gamma}{\beta}\left((1+\nu)\|P_tY - f_0\|_2^2 + \left(1 + \frac{1}{\nu}\right)\|P_uY - f_0\|_2^2\right)\right) \\
&\leq \exp\frac{2\sigma^2}{\beta}\left(tr(P_u) + \frac{2\sigma^2}{\beta - 4\sigma^2 V}tr(P_t^2) + \frac{2(1+2\gamma V)^2(1-\delta)}{\beta - 4\sigma^2 V}\left(tr(P_t^2) + tr(P_u^2)\right)\|\widetilde{f_0}\|_\infty^2\right).
\end{aligned}$$

The rest of the proof follows the same steps as in the Gaussian case: we first integrate according to the prior π, use Fubini's theorem, introduce the probability

measures ρ and μ and apply Jensen's inequality to obtain that for any $\eta \in (0,1]$,

$$\mathbb{E}\exp\frac{1}{\beta}\Big[\int\int \Delta_{t,u}d\rho(t)d\mu(u) - (1+\nu)\gamma\int\|P_tY - f_0\|_2^2 d\rho(t) \\ - \frac{4\sigma^2}{\beta - 4\sigma^2 V}\left(\sigma^2 + (1-\delta)(1+2\gamma V)^2\|\widetilde{f_0}\|_\infty^2\right)\int tr(P_t^2)d\rho(t) \\ - 2\sigma^2\left(\int tr(P_u)d\mu(u) + \frac{2(1-\delta)(1+2\gamma V)^2}{\beta - 4\sigma^2 V}\|\widetilde{f_0}\|_\infty^2\int tr(P_u^2)d\mu(u)\right) \\ - \left(1+\frac{1}{\nu}\right)\gamma\int\|P_uY - f_0\|_2^2 d\mu(u) \\ - \beta\left(KL(\rho,\pi) + KL(\mu,\pi) + \ln\frac{1}{\eta}\right)\Big] \leq \eta. \quad (3.5)$$

Finally, using $\exp(x) \geq \mathbb{1}_{\mathbb{R}_+}(x)$, for any $\delta \in [0,1]$, any $\beta > 4\sigma^2 V$ and $\beta \geq 4\sigma^2 V(1+4\delta)$, with γ previously defined, for any $\eta \in (0,1]$, for any $\nu > 0$,

$$\mathbb{P}\Big[\int\int \Delta_{t,u}d\rho(t)d\mu(u) \leq (1+\nu)\gamma\int\|P_tY - f_0\|_2^2 d\rho(t) \\ + \frac{4\sigma^2}{\beta - 4\sigma^2 V}\left(\sigma^2 + (1-\delta)(1+2\gamma V)^2\|\widetilde{f_0}\|_\infty^2\right)\int tr(P_t^2)d\rho(t) \\ + 2\sigma^2\left(\int tr(P_u)d\mu(u) + \frac{2(1-\delta)(1+2\gamma V)^2}{\beta - 4\sigma^2 V}\|\widetilde{f_0}\|_\infty^2\int tr(P_u^2)d\mu(u)\right) \\ + \left(1+\frac{1}{\nu}\right)\gamma\int\|P_uY - f_0\|_2^2 d\mu(u) + \beta\left(KL(\rho,\pi) + KL(\mu,\pi) + \ln\frac{1}{\eta}\right)\Big] \geq 1-\eta.$$

The result in expectation comes from Equation (3.5) with $\eta = 1$, combined with the inequality $t \leq \exp(t) - 1$. □

Bibliography

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973. 15, 66

N. Akakpo and C. Lacour. Inhomogeneous and anisotropic conditional density estimation from dependent data. *Electron. J. Stat.*, 5:1618–1653, 2011. ISSN 1935-7524. doi: 10.1214/11-EJS653. URL http://dx.doi.org/10.1214/11-EJS653. 13

D. M. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127, 1974. ISSN 0040-1706. 15

P. Alquier and K. Lounici. PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electron. J. Stat.*, 5:127–145, 2011. ISSN 1935-7524. doi: 10.1214/11-EJS601. URL http://dx.doi.org/10.1214/11-EJS601. 78

Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Comput.*, 9(7):1545–1588, October 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.7.1545. URL http://dx.doi.org/10.1162/neco.1997.9.7.1545. 16, 66

A. Antoniadis, J. Bigot, and R. von Sachs. A multiscale approach for statistical characterization of functional images. *Journal of Computational and Graphical Statistics*, 18, 2009. 19, 25

S. Arlot. V-fold cross-validation improved: V-fold penalization. 40 pages, plus a separate technical appendix., 2008. URL http://hal.archives-ouvertes.fr/hal-00239182. 15

S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Stat. Surv.*, 4:40–79, 2010. ISSN 1935-7516. doi: 10.1214/09-SS054. URL http://dx.doi.org.revues.math.u-psud.fr:2048/10.1214/09-SS054. 15

J.-Y. Audibert. Aggregated estimators and empirical complexity for least square regression. *Ann. Inst. H. Poincaré Probab. Statist.*, 40(6):685–736, 2004. ISSN 0246-0203. doi: 10.1016/j.anihpb.2003.11.006. URL http://dx.doi.org/10.1016/j.anihpb.2003.11.006. 18

J.-Y. Audibert. Progressive mixture rules are deviation suboptimal. In *Advances in Neural Information Processing Systems*, 2008. 21

A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999. ISSN 0178-8051. doi: 10.1007/s004400050210. URL http://dx.doi.org.revues.math.u-psud.fr:2048/10.1007/s004400050210. 15

A. Barron, C. Huang, J. Li, and X. Luo. The mdl principle, penalized likelihoods, and statistical risk. *Festschrift for Jorma Rissanen. Tampere University Press, Tampere, Finland*, 2008. 31

D. M. Bashtannyk and R. J. Hyndman. Bandwidth selection for kernel conditional density estimation. *Comput. Statist. Data Anal.*, 36(3):279–298, 2001. ISSN 0167-9473. doi: 10.1016/S0167-9473(00)00046-3. URL http://dx.doi.org.revues.math.u-psud.fr:2048/10.1016/S0167-9473(00)00046-3. 12

J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: Overview and implementation. *Statistics and Computing*, 22, 2011. 34

P. Bellec. Concentration of quadratic forms and aggregation of affine estimators. *ArXiv e-prints*, October 2014. 67

A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011. ISSN 0006-3444. doi: 10.1093/biomet/asr043. URL http://dx.doi.org/10.1093/biomet/asr043. 21, 67

J. O. Berger and L. R. Pericchi. The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.*, 91(433):109–122, 1996. ISSN 0162-1459. doi: 10.2307/2291387. URL http://dx.doi.org/10.2307/2291387. 15

G. Biau. Analysis of a random forests model. *J. Mach. Learn. Res.*, 13:1063–1095, 2012. ISSN 1532-4435. 16, 66

G. Biau and L. Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *J. Multivariate Anal.*, 101(10):2499–2518, 2010. ISSN 0047-259X. doi: 10.1016/j.jmva.2010.06.019. URL http://dx.doi.org/10.1016/j.jmva.2010.06.019. 16, 66

G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.*, 9:2015–2033, 2008. ISSN 1532-4435. 16, 66

C. Biernacki and G. Castellan. A data-driven bound on variances for avoiding degeneracy in univariate gaussian mixtures. *Pub IRMA Lille*, 71, 2011. 34

L. Birgé and P. Massart. Minimal penalties for gaussian model selection. *Probability Theory and Related Fields*, 138(1-2):33–73, 2007. ISSN 0178-8051. doi: 10.1007/s00440-006-0011-8. 34

G. Blanchard, C. Schäfer, Y. Rozenholc, and K. R. Müller. Optimal dyadic decision trees. *Machine Learning*, 66(2):209–241, 2007. 13

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. ISSN 0885-6125. doi: 10.1023/A:1018054314350. 16, 66

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. 16, 66

N. D Brinkman. Ethanol fuel-a single-cylinder engine study of efficiency and exhaust emissions. *SAE Technical Paper*, 810345, 1981. 20, 26, 37

E. Brunel, F. Comte, and C. Lacour. Adaptive estimation of the conditional density in presence of censoring. *Sankhya*, 69(Part 4.):pages 734–763, December 2007. URL http://hal.archives-ouvertes.fr/hal-00152794. 13

F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007. ISSN 0090-5364. doi: 10.1214/009053606000001587. URL http://dx.doi.org/10.1214/009053606000001587. 17

K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference. A practical information-theoretic approach*. Springer-Verlag, New-York, 2nd edition, 2002. 18, 25

O. Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004. ISBN 3-540-22572-2. doi: 10.1007/b99352. URL http://dx.doi.org/10.1007/b99352. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001. 17, 18, 66

O. Catoni. *Pac-Bayesian supervised classification: the thermodynamics of statistical learning*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 56. Institute of Mathematical Statistics, Beachwood, OH, 2007. ISBN 978-0-940600-72-0; 0-940600-72-2. 17, 18, 66, 74

G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5), 1995. 30, 33, 34

A. Celisse. Optimal cross-validation in density estimation. 2008. URL http://hal.archives-ouvertes.fr/hal-00337058. 15

N. Cesa-Bianchi and G. Lugosi. On prediction of individual sequences. *Ann. Statist.*, 27(6):1865–1895, 1999. ISSN 0090-5364. doi: 10.1214/aos/1017939242. URL http://dx.doi.org/10.1214/aos/1017939242. 16, 66

N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *J. ACM*, 44(3):427–485, May 1997. ISSN 0004-5411. doi: 10.1145/258128.258179. URL http://doi.acm.org/10.1145/258128.258179. 16, 66

F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Time series modelling by a regression approach based on a latent process. *Neural Networks*, 22:593–602, 2009. 20

F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. A hidden process regression model for functional data description. application to curve discrimination. *Neurocomputing*, 73:1210–1221, March 2010. 20, 26, 27, 33

T. Choi. Convergence of posterior distribution in the mixture of regressions. *Journal of Nonparametric Statistics*, 20(4):337–351, may 2008. 19, 25

S. X. Cohen and E. Le Pennec. Conditional density estimation by penalized likelihood model selection and applications. Technical report, INRIA, 2011. 20, 27, 28, 42, 43, 48, 51, 59, 60

S. X. Cohen and E. Le Pennec. Partition-based conditional density estimation. *ESAIM: Probability and Statistics*, 17:672–697, 1 2013. ISSN 1262-3318. doi: 10.1051/ps/2012017. URL http://www.esaim-ps.org/action/article_S1292810012000171. 13, 19, 25, 41

D. Dai, P. Rigollet, and T. Zhang. Deviation optimal learning using greedy Q-aggregation. *Ann. Statist.*, 40(3):1878–1905, 2012. ISSN 0090-5364. doi: 10.1214/12-AOS1025. URL http://dx.doi.org/10.1214/12-AOS1025. 21, 67

D. Dai, P. Rigollet, L. Xia, and T. Zhang. Aggregation of affine estimators. *Electron. J. Stat.*, 8:302–327, 2014. ISSN 1935-7524. doi: 10.1214/14-EJS886. URL http://dx.doi.org/10.1214/14-EJS886. 21, 67, 68, 69, 72

A. S. Dalalyan. SOCP based variance free Dantzig selector with application to robust estimation. *C. R. Math. Acad. Sci. Paris*, 350(15-16):785–788, 2012. ISSN 1631-073X. doi: 10.1016/j.crma.2012.09.016. URL http://dx.doi.org/10.1016/j.crma.2012.09.016. 21, 67

A. S. Dalalyan and J. Salmon. Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.*, 40(4):2327–2355, 2012. ISSN 0090-5364. doi: 10.1214/12-AOS1038. URL http://dx.doi.org/10.1214/12-AOS1038. 18, 21, 67, 68, 72

A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In NaderH. Bshouty and Claudio Gentile, editors, *Learning Theory*, volume 4539 of *Lecture Notes in Computer Science*, pages 97–111. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-72925-9. doi: 10.1007/978-3-540-72927-3_9. URL http://dx.doi.org/10.1007/978-3-540-72927-3_9. 17, 66

A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Machine Learning*, 72(1-2):39–61, 2008. doi: 10.1007/s10994-008-5051-0. URL http://certis.enpc.fr/~dalalyan/Download/Dal_Tsyb2008.pdf. 17, 66

A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.*, 78(5):1423–1443, 2012. ISSN 0022-0000. doi: 10.1016/j.jcss.2011.12.023. URL http://dx.doi.org/10.1016/j.jcss.2011.12.023. 17, 66

A. S. Dalalyan, M. Hebiri, K. Meziani, and J. Salmon. Learning heteroscedastic models by convex programming under group sparsity. In *ICML*, 2013. URL papers/ICML13_DHMS.pdf. 21, 66

J. G. De Gooijer and D. Zerom. On conditional density estimation. *Statist. Neerlandica*, 57(2):159–176, 2003. ISSN 0039-0402. doi: 10.1111/1467-9574.00226. URL http://dx.doi.org.revues.math.u-psud.fr:2048/10.1111/1467-9574.00226. 12

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B.*, 39 (1), 1977. 33

D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994. ISSN 0006-3444. doi: 10.1093/biomet/81.3.425. URL http://dx.doi.org/10.1093/biomet/81.3.425. 16

D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 26(3):879–921, 1998. ISSN 0090-5364. doi: 10.1214/aos/1024691081. URL http://dx.doi.org/10.1214/aos/1024691081. 14

D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B*, 57(2):301–369, 1995. ISSN 0035-9246. URL http://links.jstor.org/sici?sici=0035-9246(1995)57:2<301:WSA>2.0.CO;2-S&origin=MSN. With discussion and a reply by the authors. 12, 16, 66

S. Efromovich. Oracle inequality for conditional density estimation and an actuarial example. *Ann. Inst. Statist. Math.*, 62(2):249–275, 2010. ISSN 0020-3157. doi: 10.1007/s10463-008-0185-1. URL http://dx.doi.org/10.1007/s10463-008-0185-1. 13

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1996. ISBN 0-412-98321-4. 12

J. Fan and T. H. Yim. A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4):819–834, 2004. ISSN 0006-3444. doi: 10.1093/biomet/91.4.819. URL http://dx.doi.org/10.1093/biomet/91.4.819. 12

J. Fan, Q. Yao, and H. Tong. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1):189–206, 1996. ISSN 0006-3444. doi: 10.1093/biomet/83.1.189. URL http://dx.doi.org/10.1093/biomet/83.1.189. 12

Y. Freund. Boosting a weak learning algorithm by majority. *Inform. and Comput.*, 121(2):256–285, 1995. ISSN 0890-5401. doi: 10.1006/inco.1995.1136. URL http://dx.doi.org/10.1006/inco.1995.1136. 16, 66

S. Gaïffas and G. Lecué. Hyper-sparse optimal aggregation. *J. Mach. Learn. Res.*, 12:1813–1833, 2011. ISSN 1532-4435. 21

Theo Gasser and Hans-Georg Müller. Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.*, 11(3):171–185, 1984. ISSN 0303-6898. 66

E. Gassiat and R. van Handel. The local geometry of finite mixtures. *Trans. Amer. Math. Soc.*, 366(2):1047–1072, 2014. 32

Y. Ge and W. Jiang. On consistency of bayesian inference with mixtures of logistic regression. *Neural Computation*, 18(1):224–243, January 2006. 19, 25

S. Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):pp. 320–328, 1975. ISSN 01621459. URL http://www.jstor.org/stable/2285815. 15

C. Genovese and L. Wasserman. Rates of convergence for the gaussian mixture sieve. *The Annals of Statistics*, 28(4):1105–1127, august 2000. 19, 25

R. Genuer. *Forêts aléatoires : aspects théoriques, sélection de variables et applications*. PhD thesis, Université Paris-Sud, 2011. 16, 66

C. Giraud. Mixing least-squares estimators when the variance is unknown. *Bernoulli*, 14(4):1089–1107, 2008. ISSN 1350-7265. doi: 10.3150/08-BEJ135. URL http://dx.doi.org/10.3150/08-BEJ135. 21, 66

C. Giraud, S. Huet, and N. Verzelen. High-dimensional regression with unknown variance. *Statist. Sci.*, 27(4):500–518, 2012. ISSN 0883-4237. doi: 10.1214/12-STS398. URL http://dx.doi.org/10.1214/12-STS398. 21, 67

A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3):1608–1632, 2011. ISSN 0090-5364. doi: 10.1214/11-AOS883. URL http://dx.doi.org.revues.math.u-psud.fr:2048/10.1214/11-AOS883. 15

B. Guedj and P. Alquier. PAC-Bayesian estimation and prediction in sparse additive models. *Electron. J. Stat.*, 7:264–291, 2013. ISSN 1935-7524. doi: 10.1214/13-EJS771. URL http://dx.doi.org/10.1214/13-EJS771. 72

L. Györfi and M. Kohler. Nonparametric estimation of conditional distributions. *IEEE Trans. Inform. Theory*, 53(5):1872–1879, 2007. ISSN 0018-9448. doi: 10.1109/TIT.2007.894631. URL http://dx.doi.org/10.1109/TIT.2007.894631. 13

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer series in statistics. Springer, New York, Berlin, Paris, 2002. ISBN 0-387-95441-4. URL http://opac.inria.fr/record=b1123996. Autre(s) tirage(s) : 2010. 12

P. Hall, R. C. L. Wolff, and Q. Yao. Methods for estimating a conditional distribution function. *J. Amer. Statist. Assoc.*, 94(445):154–163, 1999. ISSN 0162-1459. doi: 10.2307/2669691. URL http://dx.doi.org.revues.math.u-psud.fr:2048/10.2307/2669691. 12

P. Hall, J. Racine, and Q. Li. Cross-validation and the estimation of conditional probability densities. *J. Amer. Statist. Assoc.*, 99(468):1015–1026, 2004. ISSN 0162-1459. doi: 10.1198/016214504000000548. URL http://dx.doi.org/10.1198/016214504000000548. 12

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. ISBN 978-0-387-84857-0. doi: 10.1007/978-0-387-84858-7. URL http://dx.doi.org.revues.math.u-psud.fr:2048/10.1007/978-0-387-84858-7. Data mining, inference, and prediction. 13, 16, 66

D. Hsu, S. M. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.*, 17:no. 52, 6, 2012. ISSN 1083-589X. doi: 10.1214/ECP.v17-2079. URL http://dx.doi.org/10.1214/ECP.v17-2079. 77, 82, 83

M. Huang and W. Yao. Mixture of regression models with varying mixing proportions: a semiparametric approach. *J. Amer. Statist. Assoc.*, 107(498):711–724, 2012. ISSN 0162-1459. doi: 10.1080/01621459.2012.682541. 19, 25

M. Huang, R. Li, and S. Wang. Nonparametric mixtures of regressions models. *Journal of the American Statistical Association*, 108(503):929–941, 2013. 19, 25

D. R. Hunter and D. S. Young. Semiparametric mixtures of regressions. *J. Nonparametr. Stat.*, 24(1):19–38, 2012. ISSN 1048-5252. doi: 10.1080/10485252.2011.608430. 19, 25

R. J. Hyndman and Q. Yao. Nonparametric estimation and symmetry tests for conditional density functions. *J. Nonparametr. Stat.*, 14(3):259–278, 2002. ISSN 1048-5252. doi: 10.1080/10485250212374. URL http://dx.doi.org.revues.math.u-psud.fr:2048/10.1080/10485250212374. 12

R. J. Hyndman, D. M. Bashtannyk, and G. K. Grunwald. Estimating and visualizing conditional densities. *J. Comput. Graph. Statist.*, 5(4):315–336, 1996. ISSN 1061-8600. doi: 10.2307/1390887. URL http://dx.doi.org/10.2307/1390887. 12

M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the em algorithm. In Maria Marinaro and PietroG. Morasso, editors, *ICANN 94*, pages 479–486. Springer London, 1994. ISBN 978-3-540-19887-1. 19, 25

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):pp. 773–795, 1995. ISSN 01621459. URL http://www.jstor.org/stable/2291091. 15

E. D. Kolaczyk, J. Ju, and S. Gopal. Multiscale, multigranular statistical image segmentation. *Journal of the American Statistical Association*, 100:1358–1369, 2005. 19, 25

G. Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13(4):1000–1022, 2007. ISSN 1350-7265. doi: 10.3150/07-BEJ6044. URL http://dx.doi.org/10.3150/07-BEJ6044. 17, 66

G. Lecué and S. Mendelson. Aggregation via empirical risk minimization. *Probab. Theory Related Fields*, 145(3-4):591–613, 2009. ISSN 0178-8051. doi: 10.1007/s00440-008-0180-8. URL http://dx.doi.org/10.1007/s00440-008-0180-8. 21

H. K. H. Lee. Consistency of posterior distributions for neural networks. *Neural Networks*, 13:629–642, july 2000. 19, 25

A. M. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. Firmin Didot, Paris, 1805. "Sur la Méthode des moindres quarrés" appears as an appendix. 13

G. Leung. *Improving regression through model mixing*. ProQuest LLC, Ann Arbor, MI, 2004. ISBN 978-0496-72534-2. URL http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqdiss&rft_dat=xri:pqdiss:3125244. Thesis (Ph.D.)–Yale University. 18, 21

G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006. ISSN 0018-9448. doi: 10.1109/TIT.2006.878172. URL http://dx.doi.org/10.1109/TIT.2006.878172. 17, 66, 67

Q. Li and J. S. Racine. *Nonparametric Econometrics: Theory and Practice*. Princeton University Press, 2007. ISBN 9780691121611. URL http://books.google.fr/books?id=BI_PiWazY0YC. 12, 19

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212 – 261, 1994. ISSN 0890-5401. doi: http://dx.doi.org/10.1006/inco.1994.1009. URL http://www.sciencedirect.com/science/article/pii/S0890540184710091. 16, 66

K. Lounici. Generalized mirror averaging and D-convex aggregation. *Math. Methods Statist.*, 16(3):246–259, 2007. ISSN 1066-5307. doi: 10.3103/S1066530707030040. URL http://dx.doi.org/10.3103/S1066530707030040. 17, 66

C. L. Mallows. Some comments on cp. *Technometrics*, 15(4):pp. 661–675, 1973. ISSN 00401706. URL http://www.jstor.org/stable/1267380. 15

M. L. Martin-Magniette, T. Mary-Huard, C. Bérard, and S. Robin. Chipmix: mixture model of regressions for two-color chip-chip analysis. *Bioinformatics*, 24(16):i181–i186, 2008. doi: 10.1093/bioinformatics/btn280. 20, 40

P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. ISBN 978-3-540-48497-4; 3-540-48497-3. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. 15, 16, 18

P. Massart and C. Meynet. The Lasso as an ℓ_1-ball model selection procedure. *Electron. J. Stat.*, 5:669–687, 2011. ISSN 1935-7524. doi: 10.1214/11-EJS623. URL http://dx.doi.org/10.1214/11-EJS623. 16, 17

P. Massart and É. Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34 (5):2326–2366, 2006. ISSN 0090-5364. doi: 10.1214/009053606000000786. URL http://dx.doi.org/10.1214/009053606000000786. 13

C. Maugis and B. Michel. A non asymptotic penalized criterion for gaussian mixture model selection. *ESAIM Probability and Statistics*, 2011. 18, 25, 54

C. Maugis and B. Michel. Adaptive density estimation using finite gaussian mixtures. *ESAIM Probability and Statistics*, 2012. Accepted for publication. 37

D. A. McAllester. Some pac-bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, pages 230–234, New York, NY, USA, 1998. ACM. ISBN 1-58113-057-0. doi: 10.1145/279943.279989. URL http://doi.acm.org/10.1145/279943.279989. 18

P. McCullagh and J. A. Nelder. *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1983. ISBN 0-412-23850-0. doi: 10.1007/978-1-4899-3244-0. URL http://dx.doi.org.revues.math.u-psud.fr:2048/10.1007/978-1-4899-3244-0. 13

G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, 2000. 19, 25

L. Montuelle and E. Le Pennec. Mixture of gaussian regressions model with logistic weights, a penalized maximum likelihood approach. *Electron. J. Statist.*, 8(1): 1661–1695, 2014. ISSN 1935-7524. doi: 10.1214/14-EJS939. 23

É. Nadaraya. On non-parametric estimates of density functions and regression curves. *Theory of Probability & Its Applications*, 10(1):186–190, 1965. doi: 10.1137/1110024. URL http://dx.doi.org/10.1137/1110024. 12, 65

A. Nemirovski. Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998)*, volume 1738 of *Lecture Notes in Math.*, pages 85–277. Springer, Berlin, 2000. 17, 66

A. E. Raftery, D. Madigan, and J. A. Hoeting. Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc.*, 92(437):179–191, 1997. ISSN 0162-1459. doi: 10.2307/2291462. URL http://dx.doi.org/10.2307/2291462. 15

C. R. Rao and Y. Wu. *On model selection*, volume 38 of *Lecture Notes–Monograph Series*, pages 1–57. Institute of Mathematical Statistics, Beachwood, OH, 2001. doi: 10.1214/lnms/1215540960. URL http://dx.doi.org/10.1214/lnms/1215540960. 15

P. Rigollet. *Inégalités d'oracle, agrégration et adaptation*. PhD thesis, Université Pierre et Marie Curie- Paris VI, 2006. 17, 66

P. Rigollet. Kullback-Leibler aggregation and misspecified generalized linear models. *Ann. Statist.*, 40(2):639–665, 2012. ISSN 0090-5364. doi: 10.1214/11-AOS961. URL http://dx.doi.org/10.1214/11-AOS961. 21, 29

P. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Math. Methods Statist.*, 16(3):260–280, 2007. ISSN 1066-5307. doi: 10.3103/S1066530707030052. URL http://dx.doi.org/10.3103/S1066530707030052. 17, 66

P. Rigollet and A. B. Tsybakov. Sparse estimation by exponential weighting. *Statist. Sci.*, 27(4):558–575, 2012. ISSN 0883-4237. doi: 10.1214/12-STS393. URL http://dx.doi.org/10.1214/12-STS393. 17, 67

M. Rosenblatt. Conditional probability density and regression estimators. In *Multivariate Analysis, II (Proc. Second Internat. Sympos., Dayton, Ohio, 1968)*, pages 25–31. Academic Press, New York, 1969. 12

R. E. Schapire. The strength of weak learnability. *Mach. Learn.*, 5(2):197–227, July 1990. ISSN 0885-6125. doi: 10.1023/A:1022648800760. URL http://dx.doi.org/10.1023/A:1022648800760. 16, 66

G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978. ISSN 0090-5364. 16, 66

M. Seeger. PAC-Bayesian generalisation error bounds for Gaussian process classification. *J. Mach. Learn. Res.*, 3(2):233–269, 2003. ISSN 1532-4435. doi: 10.1162/153244303765208377. URL http://dx.doi.org/10.1162/153244303765208377. 18

J. Shawe-Taylor and R. C. Williamson. A pac analysis of a bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, COLT '97, pages 2–9, New York, NY, USA, 1997. ACM. ISBN 0-89791-891-6. doi: 10.1145/267460.267466. URL http://doi.acm.org/10.1145/267460.267466. 18

C. M. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 1981. ISSN 0090-5364. URL http://links.jstor.org/sici?sici=0090-5364(198111)9:6<1135:EOTMOA>2.0.CO;2-5&origin=MSN. 18

C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 12 1982. doi: 10.1214/aos/1176345969. URL http://dx.doi.org/10.1214/aos/1176345969. 12

C. J. Stone. The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.*, 22(1):118–184, 1994. ISSN 0090-5364. doi: 10.1214/aos/1176325361. URL http://dx.doi.org/10.1214/aos/1176325361. 13

M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974. ISSN 0035-9246. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors. 15

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012. ISSN 0006-3444. doi: 10.1093/biomet/ass043. URL http://dx.doi.org/10.1093/biomet/ass043. 21, 67

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994. 13, 16, 66

A. B. Tsybakov. Optimal rates of aggregation. In B. Schölkopf and M. K. Warmuth, editors, *Learning Theory and Kernel Machines*, volume 2777 of *Lecture Notes in Computer Science*, pages 303–313. Springer Berlin Heidelberg, 2003. ISBN 978-3-540-40720-1. doi: 10.1007/978-3-540-45167-9_23. URL http://dx.doi.org/10.1007/978-3-540-45167-9_23. 17, 66

A. B. Tsybakov. Agrégation d'estimateurs et optimisation stochastique. *J. Soc. Fr. Stat. & Rev. Stat. Appl.*, 149(1):3–26, 2008. ISSN 1962-5197. 17, 66

A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. ISBN 978-0-387-79051-0. doi: 10.1007/b13794. URL http://dx.doi.org.revues.math.u-psud.fr:2048/10.1007/b13794. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. 12, 13

A. B. Tsybakov. Aggregation and minimax optimality in high-dimensional estimation. In *Proceedings of the International Congress of Mathematicians*, August 2014. to appear. 21

A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer, 1996. 19, 25, 42, 43

I. van Keilegom and N. Veraverbeke. Density and hazard estimation in censored regression models. *Bernoulli*, 8(5):607–625, 2002. ISSN 1350-7265. 12

K. Viele and B. Tong. Modeling with mixtures of linear regressions. *Stat. Comput.*, 12(4):315–330, 2002. ISSN 0960-3174. doi: 10.1023/A:1020779827503. 19, 25

V. G. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, COLT '90, pages 371–386, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc. ISBN 1-55860-146-5. URL http://dl.acm.org/citation.cfm?id=92571.92672. 16, 66

G. Wahba. *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990. ISBN 0-89871-244-0. doi: 10.1137/1.9781611970128. URL http://dx.doi.org/10.1137/1.9781611970128. 12, 66

L. Wasserman. *All of statistics*. Springer Texts in Statistics. Springer-Verlag, New York, 2004. ISBN 0-387-40272-1. doi: 10.1007/978-0-387-21736-9. URL http://dx.doi.org/10.1007/978-0-387-21736-9. A concise course in statistical inference. 12

G. S. Watson. Smooth regression analysis. *Sankhya Ser. A*, 26:359–372, 1964. ISSN 0581-572X. 12, 65

Y. Yang. Mixing strategies for density estimation. *Ann. Statist.*, 28(1):75–87, 2000a. ISSN 0090-5364. doi: 10.1214/aos/1016120365. URL http://dx.doi.org/10.1214/aos/1016120365. 17, 66

Y. Yang. Combining different procedures for adaptive regression. *J. Multivariate Anal.*, 74(1):135–161, 2000b. ISSN 0047-259X. doi: 10.1006/jmva.1999.1884. URL http://dx.doi.org/10.1006/jmva.1999.1884. 17, 66

Y. Yang. Adaptive estimation in pattern recognition by combining different procedures. *Statist. Sinica*, 10(4):1069–1089, 2000c. ISSN 1017-0405. 17, 66

Y. Yang. Adaptive regression by mixing. *J. Amer. Statist. Assoc.*, 96(454):574–588, 2001. ISSN 0162-1459. doi: 10.1198/016214501753168262. URL http://dx.doi.org/10.1198/016214501753168262. 15, 17, 66

Y. Yang. Regression with multiple candidate models: selecting or mixing? *Statist. Sinica*, 13(3):783–809, 2003. ISSN 1017-0405. 17, 66

Y. Yang. Combining forecasting procedures: some theoretical results. *Econometric Theory*, 20(1):176–222, 2004a. ISSN 0266-4666. doi: 10.1017/S0266466604201086. URL http://dx.doi.org/10.1017/S0266466604201086. 17, 66

Y. Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10(1):25–47, 2004b. ISSN 1350-7265. doi: 10.3150/bj/1077544602. URL http://dx.doi.org/10.3150/bj/1077544602. 17, 66

D. S. Young. Mixtures of regressions with changepoints. *Statistics and Computing*, 24(2):265–281, 2014. ISSN 0960-3174. doi: 10.1007/s11222-012-9369-x. 19, 25, 26, 37, 39

D. S. Young and D. R. Hunter. Mixtures of regressions with predictor-dependent mixing proportions. *Computational Statistics & Data Analysis*, 54(10):2253 – 2266, 2010. ISSN 0167-9473. 19, 25, 63

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00503.x. URL http://dx.doi.org.revues.math.u-psud.fr:2048/10.1111/j.1467-9868.2005.00503.x. 16, 66