Combiner des estimateurs en utilisant les mêmes données pour construire les experts et l'agrégé.

Frédéric Lavancier[1] & Paul Rochet[2]

[1] *Université de Nantes, Laboratoire de Mathématiques Jean Leray, 2 rue de la Houssinière, BP 92208, F-44322 Nantes Cedex 3. Frederic.lavancier@univ-nantes.fr*
[2] *Université de Nantes, Laboratoire de Mathématiques Jean Leray, 2 rue de la Houssinière, BP 92208, F-44322 Nantes Cedex 3. Paul.rochet@univ-nantes.fr*

Résumé. Etant donnés plusieurs estimateurs d'une même quantité, appelés experts, nous proposons une façon de les agréger afin de produire un meilleur estimateur. L'estimateur agrégé est une simple combinaison linéaire des experts, sous la contrainte minimale que la somme des poids vale un. Dans ce contexte, les poids optimaux minimisant le risque quadratique sont entièrement déterminés par la matrice des erreurs quadratiques des experts. L'estimateur agrégé est ainsi obtenu en utilisant une estimation de cette matrice, qui peut être calculée à partir du même jeu de données. Nous montrons que l'agrégé satisfait une inégalité oracle et qu'il est asymptotiquement optimal, pourvu que la matrice des erreurs quadratiques soit convenablement estimée. Cette méthode est illustrée sur des problèmes statistiques standards : l'estimation de la position d'une distribution symétrique, l'estimation dans un modèle paramétrique, l'estimation de densités. Dans la plupart des situations, l'agrégé surpassent tous les estimateurs initiaux.

Mots-clés. Agrégation, inégalité oracle, estimation paramétrique, modèle de Weibull, estimateur à noyaux.

Abstract. Given several estimators of the same quantity, called experts, we propose a way to aggregate them in order to produce a better estimate. The aggregated estimator is simply a linear combination of the experts, with the minimal requirement that the weights sum to one. In this framework, the optimal weights, minimizing the quadratic loss, are entirely determined by the mean square error matrix of the experts. The aggregation estimator is then obtained using an estimation of this matrix, which can be computed from the same dataset. We show that the aggregate satisfies a non-asymptotic oracle inequality and is asymptotically optimal, provided the mean square error matrix is suitably estimated. This method is illustrated on standard statistical problems: estimation of the position of a symmetric distribution, estimation in a parametric model, density estimation. In most situations, the aggregate outperforms the initial estimators.

Keywords. Averaging, Aggregation, Oracle inequality, Parametric estimation, Weibull model, Kernel density estimation.

1 Procédure d'agrégation

Soit $\mathbf{T} = (T_1, ..., T_k)$ une collection d'estimateurs (ou encore experts) d'un paramètre réel θ. L'objectif est de présenter une règle de décision permettant de combiner au mieux les T_i afin de construire un nouvel estimateur. Nous nous restreignons à ce cadre simple, mais on trouvera dans [1] la situation la plus générale où l'on souhaite estimer plusieurs paramètres en agrégeant simultanément plusieurs collections d'estimateurs, et où chaque paramètre peut être à valeur dans un espace de Hilbert. La procédure est sensiblement la même.

On remarque dans un premier temps que la meilleure fonction combinant les experts est la solution triviale mais inexploitable $f(\mathbf{T}) = \theta$. Une restriction naturelle est alors de ne s'intéresser qu'aux combinaisons linéaires des experts

$$\hat{\theta}_\lambda = \lambda^\top \mathbf{T},\ \lambda \in \Lambda,$$

où $\lambda^\top$ désigne la transposée de λ et Λ est un sous-ensemble de $\mathbb{R}^k$ dont le choix est discuté ci-dessous. La meilleure combinaison linéaire estimant θ, au sens du coût quadratique, est l'oracle $\hat{\theta}^* = \lambda^{*\top}\mathbf{T}$ où

$$\lambda^* = \arg\min_{\lambda \in \Lambda}\ \mathbb{E}(\lambda^\top \mathbf{T} - \theta)^2. \tag{1}$$

En pratique λ^* est inconnu et doit être estimé par un certain $\hat{\lambda}$ conduisant à l'agrégé $\hat{\theta} = \hat{\lambda}^\top \mathbf{T}$.

Clairement, plus Λ est grand, meilleur sera l'oracle. Cependant, pour que l'agrégé soit comparable à l'oracle, il faut que λ^* puisse être estimé au moins aussi bien que θ. Pour cette raison, on se rend compte rapidement que le choix $\Lambda = \mathbb{R}^k$ n'est pas pertinent. En effet la solution du problème d'optimisation (1) est alors

$$\lambda^*_{\text{lin}} = \arg\min_{\lambda \in \mathbb{R}^k}\ \mathbb{E}(\lambda^\top \mathbf{T} - \theta)^2 = \theta \left[\mathbb{E}(\mathbf{T}\mathbf{T}^\top)\right]^{-1} \mathbb{E}(\mathbf{T}),$$

et la présence de θ dans cette expression rend la procédure d'agrégation inefficace. Une solution naturelle consiste à considérer pour Λ des sous-ensembles de

$$\Lambda_{\max} = \{\lambda \in \mathbb{R}^k : \lambda^\top \mathbf{1} = 1\},$$

où $\mathbf{1}$ désigne le vecteur unité $\mathbf{1} = (1, ..., 1)^\top$, en d'autres termes d'imposer que les poids somment à un. Dans la suite, seul le choix $\Lambda = \Lambda_{\max}$ est considéré. D'autres choix de $\Lambda \subseteq \Lambda_{\max}$ sont discutés dans [1], permettant notamment l'agrégation convexe.

En notant Σ la matrice des erreurs quadratiques des experts, i.e.

$$\Sigma = \mathbb{E}\left[(\mathbf{T} - \theta\mathbf{1})(\mathbf{T} - \theta\mathbf{1})^\top\right],$$

on obtient comme solution de (1) lorsque $\Lambda = \Lambda_{\max}$:

$$\lambda^*_{\max} = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}^\top \Sigma^{-1}\mathbf{1}}. \tag{2}$$

En pratique Σ est inconnu et il convient de l'estimer, ce qui fournit l'agrégé

$$\hat{\theta}_{\max} = \frac{\mathbf{1}^\top \hat{\Sigma}^{-1}}{\mathbf{1}^\top \hat{\Sigma}^{-1}\mathbf{1}}\mathbf{T}. \tag{3}$$

où l'estimateur $\hat{\Sigma}$ peut être construit à partir des mêmes données que celles ayant servies à calculer $\mathbf{T}$.

Il y a plusieurs manières de construire l'estimateur $\hat{\Sigma}$, qui diffèrent essentiellement selon que le modèle sous-jacent est paramétrique ou non. Dans un modèle entièrement paramétrique dans lequel Σ est connu à θ près, une simple méthode plug-in permet d'obtenir $\hat{\Sigma}$, où un estimateur initial de θ est utilisé (par exemple la moyenne des experts). Une procédure par bootstrap paramétrique est aussi envisageable. Dans ces deux derniers cas, il est remarquable que l'agrégation ne nécessite que les valeurs des experts et pas les données sous jacentes. Dans un contexte non-paramétrique, Σ peut être estimé par bootstrap non-paramétrique. De façon alternative, une forme paramétrique asymptotique peut être disponible pour Σ et les techniques paramétriques précédentes s'appliquent. Ces différentes stratégies sont illustrées à travers plusieurs exemples dans [1].

2 Résultats théoriques

2.1 Inégalité Oracle

En remarquant que le problème d'optimisation (1) peut se récrire

$$\lambda^* = \arg\min_{\lambda\in\Lambda} \ \mathbb{E}(\lambda^\top \mathbf{T} - \theta)^2 = \arg\min_{\lambda\in\Lambda} \ \lambda^\top \Sigma \lambda,$$

il devient clair que la performance de l'agrégé dépend de la qualité de l'estimation de l'erreur $\lambda^\top\Sigma\lambda$, lorsque $\lambda \in \Lambda$. C'est pourquoi nous introduisons la distance suivante, pour A et B deux matrices symétriques définies positives et pour tout ensemble Λ ne contenant pas 0,

$$\delta_\Lambda(A|B) = \sup_{\lambda\in\Lambda} \left|1 - \frac{\lambda^\top A\lambda}{\lambda^\top B\lambda}\right|,$$

et $\delta_\Lambda(A,B) = \max\{\delta_\Lambda(A|B), \delta_\Lambda(B|A)\}$. Le théorème suivant fournit une inégalité oracle pour l'agrégé (3) où $\Lambda = \Lambda_{\max}$. Le même type de résultat est donné dans [1] pour d'autres choix de $\Lambda \subseteq \Lambda_{\max}$.

Théorème 2.1 *Soit $\hat{\Sigma}$ une matrice symétrique définie positive. L'estimateur agrégé $\hat{\theta}_{\max}$ donné en* (3) *vérifie*

$$(\hat{\theta}_{\max} - \hat{\theta}^*_{\max})^2 \leq \left[\inf_{\lambda\in\Lambda_{\max}} \ \mathbb{E}(\lambda^\top \mathbf{T} - \theta)^2\right] \ \left(2\delta_{\Lambda_{\max}}(\hat{\Sigma},\Sigma) + \delta_{\Lambda_{\max}}(\hat{\Sigma},\Sigma)^2\right) \ \ \|\Sigma^{-\frac{1}{2}}(\mathbf{T} - \mathbf{J}\theta)\|^2, \tag{4}$$

où $\hat{\theta}^* = \lambda^{*\top}_{\max}\mathbf{T}$ est l'oracle determiné par (2).

Le dernier terme dans (4) joue le rôle d'une constante dépendant du nombre d'experts, au vu de l'égalité $\mathbb{E}\|\Sigma^{-\frac{1}{2}}(\mathbf{T}-\mathbf{1}\theta)\|^2 = k$.
Il est à noter que le théorème précédent est valide sans aucune hypothèse de dépendance entre $\mathbf{T}$ et $\hat{\Sigma}$, en particulier elle a lieu même si $\mathbf{T}$ et $\hat{\Sigma}$ sont calculés à partir des mêmes données.

2.2 Comportement asymptotique

Le théorème 2.1 ne s'appuie sur aucune hypothèse concernant la construction des experts $\mathbf{T}$ et de l'estimateur $\hat{\Sigma}$. En pratique, pour peu qu'au moins un expert estime correctement θ, on s'attend à ce que l'oracle ait de bonnes propriétés asymptotiques telles que la consistance ou la normalité asymptotique. L'estimateur agrégé devrait alors hériter de ces propriétés, pourvu que $\hat{\Sigma}$ estime suffisamment bien Σ. La proposition suivante clarifie cette idée.

On suppose que $\mathbf{T}$ et $\hat{\Sigma}$ sont construits à partir d'un jeu d'observations $X_1, \ldots, X_n$ dont la taille n tend vers l'infini. Pour souligner la dépendance en n, les notations précédentes deviennent $\mathbf{T}_n$, $\hat{\Sigma}_n$, Σ_n, λ_n^*, $\hat{\lambda}_n$, $\hat{\theta}_n$ et $\hat{\theta}_n^*$, où l'ensemble Λ est encore une fois ici égal à $\Lambda_{\max}$ pour simplifier la présentation (voir [1] pour le cas général).

Soit

$$\alpha_n := \mathbb{E}(\hat{\theta}_n^* - \theta)^2 = \lambda_n^{*\top}\Sigma_n\lambda_n^*, \qquad \hat{\alpha}_n = \hat{\lambda}_n^\top\hat{\Sigma}_n\hat{\lambda}_n.$$

On note I la matrice identité de taille k et $\xrightarrow{p}$ (resp. $\xrightarrow{d}$) la convergence en probabilité (resp. en loi) lorsque $n \to \infty$.

Proposition 2.2 *Si*

$$\hat{\Sigma}_n\Sigma_n^{-1} \xrightarrow{p} \mathrm{I}, \tag{5}$$

alors

$$(\hat{\theta}_n - \theta)^2 = (\hat{\theta}_n^* - \theta)^2 + o_p(\alpha_n). \tag{6}$$

De plus, s'il existe une variable aléatoire $\mathcal{Z}$ telle que $\alpha_n^{-\frac{1}{2}}(\hat{\theta}_n^* - \theta) \xrightarrow{d} \mathcal{Z}$, alors

$$\hat{\alpha}_n^{-\frac{1}{2}}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{Z}. \tag{7}$$

Le dernier résultat permet de construire un intervalle de confiance asymptotique pour θ, dès que $\mathcal{Z}$ est connu. Par exemple, si le vecteur des experts $\mathbf{T}_n$ est asymptotiquement gaussien et asymptotique sans biais, alors $\mathcal{Z} \overset{d}{=} \mathcal{N}(0,1)$, ce qui est garanti par le choix de α_n. De plus, aucune estimation supplémentaire n'est requise pour obtenir cet intervalle de confiance, car l'estimation de la variance asymptotique $\hat{\alpha}_n$ ne repose que sur $\hat{\lambda}_n$ et $\hat{\Sigma}_n$, qui sont déjà utilisés pour obtenir l'agrégé.

3 Un exemple d'application

Plusieurs applications sont présentées dans [1]. Outre celle détaillée ci-dessous, la méthode est appliquée à l'estimation des paramètres d'une loi de Weibull, en combinant trois estimateurs standards. Elle est aussi utilisée pour l'estimation d'une densité, où plusieurs estimateurs à noyaux ayant des fenêtres différentes sont agrégés. Dans chaque situation, l'agrégé surpasse les experts.

On s'intéresse dans cette partie à l'estimation de la position θ d'une distribution symétrique f à partir de l'observation d'un échantillon i.i.d $x_1, \ldots, x_n$. Deux choix naturels s'offrent à nous : la moyenne empirique $\bar{x}_n$ ou la médiane $x_{(n/2)}$. Si $\sigma^2 := \int (x-\theta)^2 f(x)dx$ est fini, ces deux estimateurs sont consistants.

L'idée de combiner ces deux estimateurs pour en construire un meilleur remonte aux travaux de Pierre Simon de Laplace au début du 19ème siècle, voir [2] et [3]. P. S. de Laplace obtient les poids optimaux garantissant une variance asymptotique minimale pour l'agrégé. En particulier, il remarque que pour la distribution gaussienne, la meilleure combinaison est de ne choisir que $\bar{x}_n$, montrant ainsi pour la première fois l'efficacité de $\bar{x}_n$ dans le cas gaussien. Pour les autres distributions, il note que les poids ne sont pas accessibles car ils dépendent de la distribution inconnue. Notre agrégé (3) est simplement construit en estimant ces poids.

Avec les notations des sections précédentes, on a $T_1 = \overline{x}_n$, $T_2 = x_{(n/2)}$ et il convient d'estimer Σ_n. Pour cette estimation, on peut procéder de deux manières : soit on utilise la forme asymptotique de Σ_n, soit on utilise une procédure de bootstrap standard. La forme asymptotique de Σ_n (obtenue par P. S. de Laplace) est $n^{-1}W$ où

$$W = \begin{pmatrix} \sigma^2 & \frac{\mathbb{E}|X-\theta|}{2f(\theta)} \\ \frac{\mathbb{E}|X-\theta|}{2f(\theta)} & \frac{1}{4f(\theta)^2} \end{pmatrix}.$$

Chaque élément de W peut s'estimer de façon naturelle, en utilisant un estimateur initial $\hat{\theta}_0$ de θ, voir [1] pour les détails. Pour des questions de robustesse, on a choisi $\hat{\theta}_0 = x_{(n/2)}$. L'estimateur agrégé (3) s'en déduit et on le note $\hat{\theta}_{AG}$. On peut montrer que $\hat{\theta}_{AG}$ remplit toutes les hypothèses de la proposition 2.2, prouvant son optimalité asymptotique. Pour la procédure par bootstrap, on estime Σ_n à partir de 1000 réplications et on note l'estimateur agrégé $\hat{\theta}_{AGB}$.

La table 1 résume l'erreur quadratique moyenne (EQM) de $\bar{x}_n$, $x_{(n/2)}$, $\hat{\theta}_{AG}$ et $\hat{\theta}_{AGB}$, estimée à partir de 10^4 réplications, pour différentes distributions et lorsque $n = 30,\ 50,\ 100$. Les distributions considérés sont : la loi de Cauchy, la loi de Student à 5 et 7 degrés de liberté, la loi logistique, la loi normale standard, et le mélange équilibré d'une $\mathcal{N}(-2, 1)$ et d'une $\mathcal{N}(2, 1)$. Dans chaque situation $\theta = 0$.

L'agrégé $\hat{\theta}_{AG}$ ou $\hat{\theta}_{AGB}$ a une EQM inférieure à $\bar{x}_n$ et $x_{(n/2)}$ pour toutes les tailles d'échantillon et toutes les distributions considérées, sauf pour la loi gaussienne. Dans ce dernier cas, on sait que l'oracle est la moyenne empirique, donc notre agrégé n'a aucune

chance d'améliorer les performances de $\bar{x}_n$. Néanmoins l'EQM de $\hat{\theta}_{AG}$ et $\hat{\theta}_{AGB}$ est très proche de celle de $\bar{x}_n$, ce qui montre que les poids optimaux $(1, 0)$ sont bien estimés. Par ailleurs, les performances de l'agrégé sont remarquables pour la loi de Cauchy. Cette dernière n'a pas de moment d'ordre 2 et $\bar{x}_n$ ne devrait pas être utilisé. Mais il se trouve que l'agrégé est très robuste dans ce cas et est capable de sélectionner $x_{(n/2)}$.

	n=30				n=50				n=100			
	MEAN	MED	AG	AGB	MEAN	MED	AG	AGB	MEAN	MED	AG	AGB
Cauchy	2.10^6	9	8.95	8.99	4.10^7	5.07	4.92	4.9	2.10^7	2.56	2.49	2.49
	(1.10^6)	(0.14)	(0.15)	(0.15)	(4.10^7)	(0.08)	(0.08)	(0.08)	(2.10^7)	(0.04)	(0.04)	(0.04)
St(4)	6.68	5.71	5.4	5.43	4.12	3.53	3.33	3.34	1.99	1.74	1.61	1.62
	(0.1)	(0.08)	(0.08)	(0.08)	(0.06)	(0.05)	(0.05)	(0.05)	(0.03)	(0.02)	(0.02)	(0.02)
St(7)	4.8	5.51	4.6	4.64	2.82	3.32	2.74	2.8	1.42	1.67	1.37	1.38
	(0.07)	(0.08)	(0.07)	(0.07)	(0.04)	(0.05)	(0.04)	(0.04)	(0.02)	(0.02)	(0.02)	(0.02)
Logistic	10.89	12.7	10.76	10.87	6.64	7.93	6.52	6.6	3.3	4	3.2	3.26
	(0.16)	(0.18)	(0.16)	(0.16)	(0.09)	(0.11)	(0.09)	(0.09)	(0.05)	(0.06)	(0.05)	(0.05)
Gauss	3.39	5.11	3.53	3.61	2.04	3.1	2.1	2.15	1	1.51	1.02	1.06
	(0.05)	(0.07)	(0.05)	(0.05)	(0.03)	(0.04)	(0.03)	(0.03)	(0.01)	(0.02)	(0.01)	(0.01)
Mix	16.79	87	15.03	13.41	10.08	66.53	7.57	6.68	5.05	42.35	3.09	2.36
	(0.23)	(0.82)	(0.29)	(0.3)	(0.14)	(0.64)	(0.15)	(0.18)	(0.07)	(0.43)	(0.06)	(0.07)

Table 1: Estimation par Monte Carlo de l'EQM de $\overline{x}_n$ (MEAN), $x_{(n/2)}$ (MED), $\hat{\theta}_{AG}$ (AG) et $\hat{\theta}_{AGB}$ (AGB). Le nombre de réplications vaut 10^4 et l'écart-type des EQM est donné entre parenthèses. Chaque valeur a été multipliée par 100.

Bibliographie

[1] Lavancier, F. et Rochet, P. (2014), Combining estimators using the same dataset to produce both the experts and the aggregate, arxiv:1401.6371.

[2] Laplace, P.-S., *Théorie analytique des probabilités. Vol. II*, Éditions Jacques Gabay, Paris, 1995. Reprint of the 1820 third edition (Book II) and of the 1816, 1818, 1820 and 1825 originals (Supplements).

[3] Stigler, S. M. (1973), Laplace, Fisher, and the discovery of the concept of sufficiency. *Biometrika 60*, 3, 439–445.