Combining estimators using the same dataset to produce both the experts and the aggregate

Frédéric Lavancier, Paul Rochet

Combining estimators using the same dataset to produce both the experts and the aggregate

F. Lavancier and P. Rochet

University of Nantes, France

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 Introduction

We are interested in estimating a parameter θ in a statistical model, based on a collection of preliminary estimators $\mathbf{T} = (T_1, ..., T_k)$, referred to as *experts*. The issue of dealing with several possibly competing estimators of the same quantity arises in numerous situations in which aggregation procedures aim to produce a single final estimator that hopefully performs as well as possible, given the experts. Different approaches are possible. For instance, one can search for the best estimator among the T_i's, or allow combinations of the experts in order to pursue the best performance possible. Model selection aggregation as well as linear or convex combinations of the experts have been extensively studied in the literature, some of the main references are [12], [3], [4], [26]. All these frameworks can

be described in a similar setting where an *aggregate* is obtained as a linear combination of the T_i's,

$$\hat{\theta}_\lambda = \lambda^\top \mathbf{T} = \sum_{i=1}^{k} \lambda_i T_i,$$

for λ in a particular subset Λ of $\mathbb{R}^k$. For example, model selection aggregation corresponds to the set of vertices $\Lambda = \{(1,0,...,0)^\top, (0,1,0,...,0)^\top, ..., (0,...,0,1)^\top\}$, convex aggregation corresponds to $\Lambda = \{\lambda : \sum_{i=1}^k \lambda_i = 1, \lambda_i \geq 0\}$ while linear aggregation corresponds to the case $\Lambda = \mathbb{R}^k$.

Ever since the paper of Juditsky and Nemirovsky [12], aggregation literature has been developed essentially in the context of non parametric regression. Goldenshluger [8] and Bunea et al. [3] propose general methods adapted to different aggregation sets, assuming that the experts are non random or independent from the observations. In [7], the authors develop an aggregation procedure dealing with dependent experts that are affine functions of the observations. Taniguchi and Tresp [23, 25] propose to combine neural estimators in a non parametric regression framework, but do not investigate the efficiency of the aggregate. Aggregation methods have also been extensively studied for density estimation, as it provides for instance an alternative to classical bandwidth selection methods. In [17] and [5], the authors propose natural methods for dealing with several estimates of a density, based on a quadratic minimization, that result in aggregation estimators. Here again, the experts are assumed to be non random. A different approach was proposed almost simultaneously by Yang [27] and Catoni [6] to aggregate density estimators using a sequential process.

As seen from the above references, most of the literature assume that the experts are non random, which is generally achieved by dividing the sample in order to separate training and validation. Without this splitting procedure, the experts $T_1, \ldots, T_k$ are built from the same data as those used to aggregate them, so they are random and generally dependent. The question is then to find the best way to combine these estimators, when the nature of the dependency can be unknown. In this range of work, the issue of estimating the common mean of Gaussian variables has been given a particular interest, with references going back to the 1960's such as [9] and [15], for which an averaging of two independent Gaussian variables with unknown variances is considered. Keller and Olkin [13] study the same problem with more than two Gaussian variables for which the covariance matrix is unknown and estimated independently. However, the problems studied in these papers remain very specific, as they investigate the estimation of the mean in a Gaussian model. To our knowledge, the general question of estimating an unknown quantity given a collection of experts built from the same set of data, without further assumptions on the model, has not been addressed in these terms in the literature.

The aggregation method discussed in this paper aims at providing a solution to the latter general problem. The experts are thus random and possibly dependent. Our aggregated estimator approaches the best linear combination of the experts under the

minimal requirement that the weights sum to one. However more constraints on the weights, or equivalently on the definition of the set Λ, can be added, leading for instance to convex aggregation. We discuss the optimality of the aggregate with respect to the quadratic loss. This allows us to deal with many aggregation problems, since this choice of cost function is not specific to an underlying model. As a result, the aggregated estimator relies on the estimation of the mean square error matrix of the experts. In most cases, this estimation can be carried out by standard methods (e.g. plug-in or bootstrap methods), and does not require the tuning of any extra parameter.

The aggregation procedure is detailed in Section 2, both for the estimation of one parameter, belonging to a Hilbert space, and for the estimation of several parameters. In Section 3, we discuss some examples of natural aggregation frameworks, i.e. choices of the set of weights Λ. In Section 4 we derive a non asymptotic oracle inequality for the aggregate and we discuss its asymptotic optimality. Section 5 is devoted to some examples of aggregation problems, where we show that the aggregate performs almost always better than the best expert. These examples deal with the estimation of the position of a symmetric distribution, estimation in a parametric model, and kernel density estimators. Proofs of our results are postponed to the Appendix.

2 Construction of the aggregate

For ease of comprehension, we present separately the aggregation procedure for one parameter and for several parameters simultaneously, although the former is a particular case of the latter.

2.1 Aggregation for one parameter

Let $\mathbf{T} = (T_1, ..., T_k)$ be a collection of estimators, or experts, of a parameter θ lying in some Hilbert space $(\mathcal{H}, \langle ., . \rangle)$. We search for a decision rule that combines suitably the T_i's to provide a unique estimate of θ. Remark that considering all transformations $f(\mathbf{T})$ is pointless, since the best transformation in this class is the trivial $f(\mathbf{T}) = \theta$. Nevertheless, a reasonable alternative is to settle for linear transformations

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

where $\lambda^\top$ denotes the transpose of λ and Λ is a given subset of $\mathbb{R}^k$. In this linear setting, a convenient way to measure the performance of an aggregate is to compare it to the best non random combination $\hat{\theta}^*$ in the class $\{\hat{\theta}_\lambda, \lambda \in \Lambda\}$, called *oracle*. Specifically, we define the oracle as the linear combination $\hat{\theta}^* = {\lambda^*}^\top \mathbf{T}$ minimizing the mean square error (MSE), i.e.

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

where $\|.\|$ denotes the norm on $\mathcal{H}$, i.e. for any $x \in \mathcal{H}$, $\|x\|^2 = \langle x, x \rangle$. Of course in practice λ^* is unknown and needs to be approximated by an estimator, say $\hat{\lambda}$.

Clearly, the larger the set Λ, the better the oracle. However, choosing the whole space $\Lambda = \mathbb{R}^k$ (which corresponds to linear aggregation) is generally not exploitable. Indeed, assuming that the Gram matrix $\mathbb{E}\langle \mathbf{T}, \mathbf{T}^\top \rangle$ (with entries $\mathbb{E}\langle T_i, T_j \rangle$) exists and is non-singular, the oracle over $\Lambda = \mathbb{R}^k$ is given by

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

But for the aggregate $\hat{\theta} = \hat{\lambda}^\top \mathbf{T}$ to be comparable to the oracle, we need to be able to approach the optimal weights at least as well as we can estimate θ. The presence of θ in the above expression shows that λ^*_{lin} should be at least as difficult to estimate as θ, rending linear aggregation inefficient. In fact, the performance of the aggregate highly relies on the choice of the set Λ. Indeed, choosing a too large set Λ might increase the accuracy of the oracle but make it difficult to estimate λ^*. On the contrary, a too small set Λ might lead to a poorly efficient oracle but easy to approximate. Therefore, a good balance must be found for the oracle to be both accurate and reachable.

Writing the estimation error

$$\hat{\theta} - \theta = \hat{\theta}^* - \theta + (\hat{\lambda} - \lambda^*)^\top \mathbf{T},$$

the objective is to consider a set Λ for which the residual term $(\hat{\lambda} - \lambda^*)^\top \mathbf{T}$ can be made negligible (in a sense to be made precise) compared to the error of the oracle $\hat{\theta}^* - \theta$. A natural way to deal with this issue (see for instance [23]) is to consider for Λ a subset of

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

where $\mathbf{1}$ denotes the unit vector $\mathbf{1} = (1, ..., 1)^\top$. This choice enables a better control of the error. To see this, write the equality

$$(\hat{\lambda} - \lambda^*)^\top \mathbf{T} = (\hat{\lambda} - \lambda^*)^\top (\mathbf{T} - \theta \mathbf{1}),$$

which always holds if λ^* and $\hat{\lambda}$ are in $\Lambda_{\max}$. The residual term $(\hat{\lambda} - \lambda^*)^\top (\mathbf{T} - \theta \mathbf{1})$ appears to be more easily manageable compared to linear aggregation, as long as the initial estimators T_i are sufficiently accurate.

In the sequel, we assume that the experts have finite order-two moments and $1, T_1, ..., T_k$ are linearly independent so that the Gram matrix

$$\Sigma = \mathbb{E}\ \langle \mathbf{T} - \theta \mathbf{1}, (\mathbf{T} - \theta \mathbf{1})^\top \rangle = (\mathbb{E}\ \langle T_i - \theta, T_j - \theta \rangle)_{i,j=1,...,k}$$

is well defined and non-singular. Let Λ be a non-empty closed convex subset of $\Lambda_{\max}$, the oracle is defined as the linear combination $\hat{\theta}^* = {\lambda^*}^\top \mathbf{T}$ where

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

the last equality holding due to the identity $\lambda^\top \mathbf{1} = 1$. Remark that the assumptions made on Λ ensure both existence and unicity of the minimizer. In the particular important example where $\Lambda = \Lambda_{\max}$, we get as the explicit solution of the above optimization problem,

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

Of course, in practice, the MSE matrix Σ is unknown and has to be approximated by some estimator $\hat{\Sigma}$ to yield the aggregate $\hat{\theta} = \hat{\lambda}^\top \mathbf{T}$, where

$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda} \ \lambda^\top \hat{\Sigma} \lambda.$$

There are natural methods to construct $\hat{\Sigma}$ that essentially differ whether the model is parametric or not. In a fully specified parametric model in which Σ is known up to θ, the MSE can be estimated by plugging in an initial estimate of θ. Precisely, assuming that the MSE can be expressed as the image of θ through a known map $\Sigma(.) : \mathcal{H} \to \mathbb{R}^{k \times k}$, one can choose $\hat{\Sigma} = \Sigma(\hat{\theta}_0)$, where $\hat{\theta}_0$ is an initial estimate of θ. A natural choice for $\hat{\theta}_0$ is to take one of the expert or the average $\bar{\mathbf{T}} = \frac{1}{k}\sum_{i=1}^k T_i$. In this case, the aggregation procedure does not require any other information than the experts. Remark that even if the map $\Sigma(.)$ is not explicitly known, $\Sigma(\hat{\theta}_0)$ may be approximated by parametric bootstrap. On the other hand, in a non-parametric setting, an estimation of Σ may be achieved by standard (non-parametric) bootstrap. Alternatively, a parametric closed-form expression for Σ may be available asymptotically, i.e. when the sample size on which the experts are built tends to infinity, and the plugging method explained above then becomes possible. Some of these methods are illustrated in our examples in Section 5.

2.2 Aggregation for several parameters

We now investigate the simultaneous aggregation procedure for several parameters. Let $\theta = (\theta_1, \ldots, \theta_d)^\top \in \mathcal{H}^d$ and assume we have access to several collections of experts, $\mathbf{T}_1, \ldots, \mathbf{T}_d$, one for each component θ_j. For sake of generality we allow the collections $\mathbf{T}_j$'s to have different sizes denoted $k_1, \ldots, k_d$ respectively. So, let $\mathbf{T}_1 \in \mathcal{H}^{k_1}, \ldots, \mathbf{T}_d \in \mathcal{H}^{k_d}$ and denote $\mathbf{T} = (\mathbf{T}_1^\top, \ldots, \mathbf{T}_d^\top)^\top \in \mathcal{H}^k$, with $k = \sum_{j=1}^d k_j$. We consider aggregation estimators of θ of the form

$$\hat{\theta}_\lambda = \lambda^\top \mathbf{T} \in \mathcal{H}^d,$$

where here, λ is a $k \times d$ matrix. For similar reasons as previously, we choose to make some restrictions on the set of authorized values for λ. In this purpose, let

$$\mathbf{J} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \in \mathbb{R}^{k \times d},$$

where the j-th column of $\mathbf{J}$ contains exactly k_j ones, and define the maximal restriction set

$$\Lambda_{\max} = \{\lambda \in \mathbb{R}^{k \times d} : \lambda^\top \mathbf{J} = \mathrm{I}\}, \tag{1}$$

with I the identity matrix. Let $\Pi_j(\lambda)$ denote the j-th column of $\lambda \in \mathbb{R}^{k \times d}$. For each component θ_j, the aggregate is given by

$$\hat{\theta}_{\lambda,j} = \Pi_j(\lambda)^\top \mathbf{T} = \underline{\lambda}_{j,1}^\top \mathbf{T}_1 + \cdots + \underline{\lambda}_{j,d}^\top \mathbf{T}_d,$$

where $\Pi_j(\lambda) = (\underline{\lambda}_{j,1}^\top, \dots, \underline{\lambda}_{j,d}^\top)^\top$ with $\underline{\lambda}_{j,\ell} \in \mathbb{R}^{k_\ell}$, $\ell = 1, \dots, d$. Imposing that $\lambda \in \Lambda_{\max}$ means that for any $j = 1, \dots, d$

$$\underline{\lambda}_{j,\ell}^\top \mathbf{1} = \begin{cases} 0 & \text{if } \ell \neq j \\ 1 & \text{if } \ell = j. \end{cases} \tag{2}$$

In particular, this condition does not rule out using the entire collection $\mathbf{T}$ to estimate each component θ_j, although the weights $\underline{\lambda}_{j,\ell}$ do not satisfy the same constraints depending on the relevance of $\mathbf{T}_\ell$. While it may seem more natural to impose that only $\mathbf{T}_j$ is involved in the estimation of θ_j (and this can be made easily through an appropriate choice of $\Lambda \subset \Lambda_{\max}$, letting $\underline{\lambda}_{j,\ell} = 0$ for $\ell \neq j$), allowing one to use the whole set $\mathbf{T}$ to estimate each component enables to take into account possible dependencies between the experts, which may improve the aggregate. Moreover, the condition $\lambda^\top \mathbf{J} = \mathrm{I}$ appears as a minimal requirement to obtain an oracle inequality, as shown further in Theorem 4.1.

Since there is no ambiguity, we shall use abusively the same notation $\|.\|$ to refer to the norm in $\mathcal{H}^d$, i.e. for any $a = (a_1, \dots, a_d)^\top \in \mathcal{H}^d$

$$\|a\| = \|a\|_{\mathcal{H}^d} = \sqrt{\|a_1\|_{\mathcal{H}}^2 + \dots + \|a_d\|_{\mathcal{H}}^2}.$$

Similarly, for a and b in $\mathcal{H}^d$, $\langle a^\top, b\rangle$ stands for $\sum\langle a_i, b_i\rangle$, while $\langle a, b^\top\rangle$ denotes as before the Gram matrix with entries $\langle a_i, b_j\rangle$.

Notice that the condition $\lambda^\top \mathbf{J} = \mathrm{I}$ implies that

$$\lambda^\top \mathbf{T} - \theta = \lambda^\top(\mathbf{T} - \mathbf{J}\theta),$$

and the expression of the mean square error can be rewritten

$$\begin{aligned}
\mathbb{E}\|\lambda^\top \mathbf{T} - \theta\|^2 &= \mathbb{E}\left[\langle (\mathbf{T} - \mathbf{J}\theta)^\top \lambda, \lambda^\top(\mathbf{T} - \mathbf{J}\theta)\rangle\right] \\
&= \mathbb{E}\left[\mathrm{tr}\left(\langle (\mathbf{T} - \mathbf{J}\theta)^\top \lambda, \lambda^\top(\mathbf{T} - \mathbf{J}\theta)\rangle\right)\right] \\
&= \mathbb{E}\left[\mathrm{tr}\left(\langle \lambda^\top(\mathbf{T} - \mathbf{J}\theta), (\mathbf{T} - \mathbf{J}\theta)^\top \lambda\rangle\right)\right] \\
&= \mathrm{tr}(\lambda^\top \Sigma \lambda),
\end{aligned}$$

where $\Sigma = \mathbb{E}\langle \mathbf{T} - \mathbf{J}\theta, (\mathbf{T} - \mathbf{J}\theta)^\top\rangle \in \mathbb{R}^{k\times k}$ and $\mathrm{tr}(.)$ denotes the trace operator. Here again, we assume that Σ exists and is non-singular.

The simultaneous aggregation process for several parameters generalizes the procedure presented in Section 2.1. In fact, aggregation for one parameter just becomes the particular case with $d = 1$. Given a subset $\Lambda \subseteq \Lambda_{\max}$, we define the oracle as the linear transformation $\hat{\theta}^* = \lambda^{*\top}\mathbf{T}$ with

$$\lambda^* = \arg\min_{\lambda\in\Lambda} \ \mathbb{E}\|\lambda^\top \mathbf{T} - \theta\|^2 = \arg\min_{\lambda\in\Lambda} \ \mathrm{tr}(\lambda^\top \Sigma \lambda). \tag{3}$$

Finally, assuming we have access to an estimator $\hat{\Sigma}$ of Σ, we define the aggregation estimator as $\hat{\theta} = \hat{\lambda}^\top \mathbf{T}$ where

$$\hat{\lambda} = \arg\min_{\lambda\in\Lambda} \ \mathrm{tr}(\lambda^\top \hat{\Sigma} \lambda). \tag{4}$$

If $\lambda^\top\Sigma\lambda$ is well approximated by $\lambda^\top\hat{\Sigma}\lambda$ for $\lambda \in \Lambda$, we may reasonably think that the aggregate $\hat{\theta}$ will be close to the oracle $\hat{\theta}^*$, regardless of the possible dependency between $\hat{\Sigma}$ and $\mathbf{T}$.

3 Examples of aggregation frameworks

3.1 Maximal constraint set

When a good estimation of Σ can be provided, it is natural to consider the maximal constraint set $\Lambda = \Lambda_{\max}$ defined in (1), thus aiming for the best possible oracle. This set is actually an affine subspace of $\mathbb{R}^{k\times d}$ and in particular, it is convex. The oracle, obtained by minimizing the convex map $\lambda \mapsto \mathrm{tr}(\lambda^\top\Sigma\lambda)$ subject to the constraint $\lambda^\top\mathbf{J} = \mathrm{I}$ is given by $\hat{\theta}^*_{\max} = \lambda^{*\top}_{\max}\mathbf{T}$ where

$$\lambda^*_{\max} = \Sigma^{-1}\mathbf{J}(\mathbf{J}^\top\Sigma^{-1}\mathbf{J})^{-1}, \tag{5}$$

generalizing the formula given in Section 2.1, and its mean-square error can be calculated directly

$$\mathbb{E}\langle\hat{\theta}^*_{\max} - \theta, (\hat{\theta}^*_{\max} - \theta)^\top\rangle = (\mathbf{J}^\top\Sigma^{-1}\mathbf{J})^{-1}.$$

This solution is a direct consequence of the equality

$$\lambda^\top\Sigma\lambda - (\mathbf{J}^\top\Sigma^{-1}\mathbf{J})^{-1} = \lambda^\top\Sigma\lambda - \lambda^{*\top}_{\max}\Sigma\lambda^*_{\max} = (\lambda - \lambda^*_{\max})^\top\Sigma(\lambda - \lambda^*_{\max}) \tag{6}$$

that holds for all $\lambda \in \Lambda_{\max}$ due to the condition $\lambda^\top\mathbf{J} = \mathrm{I}$, and where the last matrix is positive definite.

Moreover, (6) shows that the oracle is not only the solution of our optimization problem over $\Lambda_{\max}$, but it is optimal to estimate any linear transformation of θ. In particular each component $\hat{\theta}^*_{\max,j}$ of the oracle is the best linear transformation $\lambda^\top\mathbf{T}$, $\lambda \in \Lambda_{\max}$, that one can get to estimate θ_j. Another desirable property of the choice $\Lambda = \Lambda_{\max}$ is that due to the closed expression (5), the aggregate $\hat{\theta}_{\max}$ obtained by replacing Σ by its estimation $\hat{\Sigma}$ has also a closed expression which makes it easily computable, namely

$$\hat{\theta}_{\max} = (\mathbf{J}^\top\hat{\Sigma}^{-1}\mathbf{J})^{-1}\mathbf{J}\hat{\Sigma}^{-1}\mathbf{T}. \tag{7}$$

3.2 Component-wise aggregation

A natural and simpler aggregation framework is to consider component-wise aggregation, for which only the collection $\mathbf{T}_j$ is involved in the estimation of θ_j. The associated aggregation set is the set of matrices λ whose support is included in the support of $\mathbf{J}$, that is

$$\Lambda = \{\lambda \in \Lambda_{\max} :\ \mathrm{supp}(\lambda) \subseteq \mathrm{supp}(\mathbf{J})\},$$

where for a matrix $A = (A_{i,j}) \in \mathbb{R}^{k\times d}$, $\mathrm{supp}(A) := \{(i,j),\ A_{i,j} \neq 0\}$. In this particular framework, the covariance of two experts in different collection $\mathbf{T}_i, \mathbf{T}_j$, $i \neq j$ is not involved in the computation of the oracle, so that the corresponding entries of Σ need not be estimated. Consequently, each component of θ is aggregated regardless of the others and as a result, the oracle is given by

$$\hat{\theta}^*_j = \frac{\mathbf{1}^\top\Sigma_j^{-1}\mathbf{T}_j}{\mathbf{1}^\top\Sigma_j^{-1}\mathbf{1}},\ j = 1, ..., d.$$

where

$$\Sigma_j = \mathbb{E}\langle\mathbf{T}_j - \theta_j\mathbf{1}, (\mathbf{T}_j - \theta_j\mathbf{1})^\top\rangle \in \mathbb{R}^{k_j\times k_j},\ j = 1, ..., d.$$

In order to build the aggregate, it is sufficient to plug an estimate of Σ_j, $j = 1, ..., d$, in the above expression, which makes it easily computable. See Section 5.2 for further discussion.

3.3 Convex aggregation

Convex aggregation corresponds to the choice

$$\Lambda = \{\lambda \in \Lambda_{\max} : \ \lambda_{i,j} \geq 0, \ i = 1, ..., k, \ j = 1, ..., d\}.$$

Observe that the positivity restriction combined with the condition $\lambda^\top \mathbf{J}$ results in λ having its support included in that of $\mathbf{J}$, making convex aggregation a particular case of component-wise aggregation. This means in particular that each component of θ can be dealt with separately. So, for sake of simplicity in this example, we only consider the case $d = 1$.

This aggregation framework has been widely studied in the literature. An advantage of convex aggregation lies in the increased stability of the solution, due to the restriction of λ to a compact set, though the oracle may of course be less efficient than in the case $\Lambda = \Lambda_{\max}$. The use of convex combinations is also particularly convenient to preserve some properties of the experts, such as positivity or boundedness. Moreover, imposing non-negativity of the weights enables to construct sparse aggregates.

In this convex constrained optimization problem, the minimizer $\hat{\lambda} = \arg\min_{\lambda \in \Lambda} \ \lambda^\top \hat{\Sigma} \lambda$ can either lie in the interior of the domain, in which case $\hat{\lambda} = \hat{\Sigma}^{-1}\mathbf{1}/\mathbf{1}^\top \hat{\Sigma}^{-1}\mathbf{1}$ corresponds to the global minimizer over $\Lambda_{\max}$, or on the edge, meaning that it has at least one zero coordinate. Letting $\hat{m} \subseteq \{1, ..., k\}$ denote the support of $\hat{\lambda}$, it follows that the aggregation procedure obtained with the experts $\mathbf{T}_{\hat{m}} := (T_i)_{i \in \hat{m}}$ leads to a solution $\hat{\lambda}_{\hat{m}}$ with full support. As a result, it can be expressed as the global minimizer for the collection $\mathbf{T}_{\hat{m}}$,

$$\hat{\lambda}_{\hat{m}} = \frac{\hat{\Sigma}_{\hat{m}}^{-1}\mathbf{1}}{\mathbf{1}^\top \hat{\Sigma}_{\hat{m}}^{-1}\mathbf{1}},$$

where $\hat{\Sigma}_{\hat{m}}$ is the submatrix composed of the entries $\hat{\Sigma}_{i,j}$ for $(i, j) \in \hat{m}^2$. Since we have by construction $\hat{\lambda}_{\hat{m}}^\top \mathbf{T}_{\hat{m}} = \hat{\lambda}^\top \mathbf{T} = \hat{\theta}$, we deduce the following characterization of the convex aggregate:

$$\hat{\theta} = \frac{\mathbf{1}^\top \hat{\Sigma}_{\hat{m}}^{-1} \mathbf{T}_{\hat{m}}}{\mathbf{1}^\top \hat{\Sigma}_{\hat{m}}^{-1}\mathbf{1}},$$

where $\hat{m}$ must be the support which is both admissible and provides the minimal mean-square error of the aggregate, i.e. $\hat{m} = \arg\max_{m \subseteq \{1,...,k\}} \ \mathbf{1}^\top \hat{\Sigma}_m^{-1}\mathbf{1}$ subject to the constraint that $\hat{\Sigma}_m^{-1}\mathbf{1}$ has all its coordinates positive. This provides an easy method to implement the convex aggregate in practice. Remark that this method is only efficient if k is not too large, otherwise we recommend to use a standard quadratic programming solver to get $\hat{\lambda}$, see for instance [16].

4 Theoretical results

4.1 Oracle inequality

The performance of the aggregate relies on the accuracy of the estimator $\hat{\Sigma}$, but more precisely, on the ability to evaluate the error $\mathrm{tr}(\lambda^\top \Sigma \lambda)$ as λ ranges over Λ. As a result, it is not crucial that $\hat{\Sigma}$ be a perfect estimate of Σ as long as $\mathrm{tr}(\lambda^\top \hat{\Sigma} \lambda)$ is close to its true value for all $\lambda \in \Lambda$. In order to measure the accuracy of $\hat{\Sigma}$ for this particular purpose, we introduce the following criterion. For two symmetric positive definite matrices A and B and for any non-empty set Λ that does not contain 0, let $\delta_\Lambda(A|B)$ denote the maximal divergence of the ratio $\mathrm{tr}(\lambda^\top A \lambda)/\mathrm{tr}(\lambda^\top B \lambda)$ over Λ,

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

and $\delta_\Lambda(A, B) = \max\{\delta_\Lambda(A|B), \delta_\Lambda(B|A)\}$. We are now in position to state our main result.

Theorem 4.1 *Let Λ be a non-empty closed convex subset of $\Lambda_{\max}$ and $\hat{\Sigma}$ a symmetric positive definite $k \times k$ matrix. The aggregation estimator $\hat{\theta} = \hat{\lambda}^\top \mathbf{T}$ defined through* (4) *satisfies*

$$\|\hat{\theta} - \hat{\theta}^*\|^2 \leq \left[\inf_{\lambda \in \Lambda} \ \mathbb{E}\|\lambda^\top \mathbf{T} - \theta\|^2 \right] \ \left(2\delta_\Lambda(\hat{\Sigma}, \Sigma) + \delta_\Lambda(\hat{\Sigma}, \Sigma)^2 \right) \ \|\Sigma^{-\frac{1}{2}}(\mathbf{T} - \mathbf{J}\theta)\|^2, \quad (8)$$

where $\hat{\theta}^*$ is the oracle given by (3).

In this theorem, we provide an upper bound on the distance of the aggregate to the oracle. We emphasize that this result holds without requiring any condition on the joint behavior of $\mathbf{T}$ and $\hat{\Sigma}$ (in particular, they may be strongly dependent). The influence of the constraint set Λ in the aggregation process becomes apparent through both the minimal error $\inf_{\lambda \in \Lambda} \ \mathbb{E}\|\lambda^\top \mathbf{T} - \theta\|^2$ and the maximal divergence $\delta_\Lambda(\hat{\Sigma}, \Sigma)$. This result conveys that while the efficiency of the oracle is increased for large sets Λ, one must settle for combinations λ for which $\mathrm{tr}(\lambda^\top \Sigma \lambda)$ can be well evaluated, thus yielding a small value of $\delta_\Lambda(\hat{\Sigma}, \Sigma)$.

Remark moreover that the last term $\|\Sigma^{-\frac{1}{2}}(\mathbf{T} - \mathbf{J}\theta)\|^2$ influences the efficiency of the aggregate essentially through the number of experts used in the aggregation process, in view of the equality

$$\mathbb{E}\|\Sigma^{-\frac{1}{2}}(\mathbf{T} - \mathbf{J}\theta)\|^2 = k.$$

4.2 Asymptotic study

So far, we established properties of the aggregate that do not rely on any assumption on the construction of $\mathbf{T}$ or $\hat{\Sigma}$. In practice, we expect the oracle $\hat{\theta}^*$ to have good properties

such as consistency and asymptotic normality. As the oracle inequality in Theorem 4.1 suggests, the aggregate $\hat{\theta}$ should inherit these properties, provided $\hat{\Sigma}$ is a good estimator of Σ. In this section, we clarify the asymptotic properties of the aggregate in a situation where both $\mathbf{T}$ and $\hat{\Sigma}$ are computed from a set of observations $X_1, \ldots, X_n$ of size n growing to infinity. We modify our notations to $\mathbf{T}_n$, $\hat{\Sigma}_n$, Σ_n, λ_n^*, $\hat{\lambda}_n$, $\hat{\theta}_n$ and $\hat{\theta}_n^*$ to emphasize the dependency on n.

Let us introduce some definitions and notation. For each component θ_j, $j = 1, ..., d$, define

$$\alpha_{n,j} := \mathbb{E}\|\hat{\theta}_{n,j}^* - \theta_j\|^2 = \Pi_j(\lambda_n^*)^\top \Sigma_n \Pi_j(\lambda_n^*),$$

where we recall that $\Pi_j(\lambda_n^*)$ is the j-th column of λ_n^*. Similarly, let $\hat{\alpha}_{n,j} = \Pi_j(\hat{\lambda}_n)^\top \hat{\Sigma}_n \Pi_j(\hat{\lambda}_n)$. We assume that the quadratic error of the oracle, given by

$$\alpha_n := \mathbb{E}\|\hat{\theta}_n^* - \theta\|^2 = \mathrm{tr}(\lambda_n^{*\top} \Sigma_n \lambda_n^*) = \sum_{j=1}^{d} \alpha_{n,j},$$

converges to zero as $n \to \infty$.

For a given aggregation set $\Lambda \subset \mathbb{R}^{k \times d}$, we define $\Lambda_j = \{\Pi_j(\lambda) : \lambda \in \Lambda\} \subset \mathbb{R}^k$. We say that Λ is a *cylinder* if $\Lambda = \{\lambda : \Pi_1(\lambda) \in \Lambda_1, ..., \Pi_d(\lambda) \in \Lambda_d\}$, i.e., if Λ is the Cartesian product of its marginal sets Λ_j. We point out that choosing an aggregation set Λ that satisfies this property is very natural, as it simply states that each vector of weights $\Pi_j(\lambda_n^*)$ used to produce $\hat{\theta}_{n,j}^*$ can be computed independently of the others. In particular, all the aggregation sets discussed in Section 3 are cylinders.

We denote by $\xrightarrow{p}$ (resp. $\xrightarrow{d}$) the convergence in probability (resp. in distribution) as $n \to \infty$.

Proposition 4.2 *If*

$$\hat{\Sigma}_n \Sigma_n^{-1} \xrightarrow{p} \mathrm{I}, \tag{9}$$

then

$$\|\hat{\theta}_n - \theta\|^2 = \|\hat{\theta}_n^* - \theta\|^2 + o_p(\alpha_n). \tag{10}$$

Moreover, if Λ is a cylinder and $\alpha_{n,j}^{-\frac{1}{2}}(\hat{\theta}_{n,j}^* - \theta_j) \xrightarrow{d} \mathcal{Z}$ for some $j = 1, ..., d$, then

$$\hat{\alpha}_{n,j}^{-\frac{1}{2}}(\hat{\theta}_{n,j} - \theta_j) \xrightarrow{d} \mathcal{Z}. \tag{11}$$

This proposition establishes that building an estimator $\hat{\Sigma}_n$ for which (9) holds ensures that the error of the aggregate is asymptotically comparable to that of the oracle, up to $o_p(\alpha_n)$. If in addition Λ is a cylinder, it is possible to provide asymptotic confidence regions for θ_j, if $\mathcal{Z}$ is known. If $\mathcal{H} = \mathbb{R}$, this situation occurs for instance when $\mathbf{T}_n$ is asymptotically unbiased and asymptotically Gaussian. In this case, the normalization

$\alpha_{n,j} = \mathbb{E}(\hat{\theta}^*_{n,j} - \theta_j)^2$ guarantees that $\mathcal{Z} \overset{d}{=} \mathcal{N}(0,1)$ and (11) enables to build an asymptotic confidence interval for θ_j. From (10), this confidence interval is of minimal length amongst all possible confidence intervals based on a linear combination of $\mathbf{T}_n$. Note finally that no extra estimation is needed to approach the asymptotic variance, as $\hat{\alpha}_{n,j}$ is entirely determined by $\hat{\lambda}_n$ and $\hat{\Sigma}_n$, which are already used to compute the aggregate.

These properties rely on the assumption (9), that might be difficult to check in practice. In the following lemma, we discuss a particular situation where this condition is verified.

Lemma 4.3 *Assume there exist an orthogonal matrix P (i.e. with $P^\top P = \mathrm{I}$) and a known deterministic sequence $(A_n)_{n\in\mathbb{N}}$ of diagonal invertible matrices such that*

$$\lim_{n\to\infty} A_n P \Sigma_n P^\top = D,$$

for some non-singular diagonal matrix D. If $\hat{P}_n$ and $\hat{D}_n$ are consistent estimators of P and D respectively, then $\hat{\Sigma}_n = \hat{P}_n^\top A_n^{-1} \hat{D}_n \hat{P}_n$ satisfies (9).

This lemma concerns the situation in which the normed eigenvectors of Σ_n converge as $n \to \infty$, their limits being the rows of P. In this case, a natural estimator $\hat{\Sigma}_n$ is provided by the asymptotic expansion of Σ_n. This result covers the particular case where all constant combinations $\lambda^\top \mathbf{T}_n$ converge to θ at the same rate, as illustrated by the following result.

Corollary 4.4 *Assume there exists a sequence $(a_n)_{n\in\mathbb{N}}$ tending to zero such that*

$$\Sigma_n = a_n W + o(a_n), \tag{12}$$

for some non-singular matrix W. Then, if $\hat{W}_n$ is a consistent estimate of W, the aggregate $\hat{\theta}_n$ obtained by minimizing $\lambda \mapsto \mathrm{tr}(\lambda^\top \hat{W}_n \lambda)$ satisfies (10). *Moreover, if Λ is a cylinder and $\alpha_{n,j}^{-\frac{1}{2}}(\hat{\theta}^*_{n,j} - \theta_j) \xrightarrow{d} \mathcal{Z}$, then* (11) *holds.*

The proof follows directly from Proposition 4.2 and Lemma 4.3 with $A_n = a_n^{-1}\mathrm{I}$, in which case the scaling a_n has no influence on the value of $\hat{\theta}_n$, as shown by (4), and needs not be known. If (12) holds, the situation becomes particularly convenient if the limit matrix W follows a known parametric expression $W = W(\eta, \theta)$, with η a nuisance parameter. If the map $W(.,.)$ is continuous, plugging consistent estimators $\hat{\eta}_0$, $\hat{\theta}_0$ yields an estimator $\hat{W}_n = W(\hat{\eta}_0, \hat{\theta}_0)$ that fulfills the sufficient conditions for the aggregate $\hat{\theta}_n$ to satisfy (10).

In Proposition 4.2, the asymptotic optimality of the aggregate is stated in probability. Remark however that asymptotic optimality in quadratic loss can be obtained easily

under additional assumptions. If, for instance, $\hat{\Sigma}_n$ and $\mathbf{T}$ are computed from independent samples (which may be achieved by sample splitting), the aggregate $\hat{\theta}_n$ is asymptotically optimal in quadratic loss as soon as $\mathbb{E}[\delta_\Lambda(\hat{\Sigma}_n, \Sigma_n)^2]$ tends to 0. Indeed, we have in this case

$$\mathbb{E}\|\hat{\theta}_n - \theta\|^2 = \mathbb{E}\|\hat{\theta}_n^* - \theta\|^2 + o(\alpha_n), \tag{13}$$

which follows directly from (21) in the proof of Proposition 4.2, taking the expectation on both sides. We emphasize however that the use of sample splitting may considerably deteriorate the oracle, as it would be computed from fewer data. This is a high price to pay to obtain asymptotic optimality in $\mathbb{L}^2$. For this reason, we do not recommend to separate training and validation with this aggregation procedure.

Asymptotic optimality in $\mathbb{L}^2$ can also be achieved if one can show there exists $p > 1$ such that

$$\sup_{n\in\mathbb{N}} \mathbb{E}\|\Sigma_n^{-\frac{1}{2}}(\mathbf{T}_n - \mathbf{J}\theta)\|^{\frac{2p}{p-1}} < \infty \quad \text{and} \quad \lim_{n\to\infty} \mathbb{E}[\delta_\Lambda(\hat{\Sigma}_n, \Sigma_n)^{2p}] = 0.$$

In this case, Equation (13) follows directly by applying Hölder's inequality in (21). These conditions ensure the asymptotic optimality in $\mathbb{L}^2$ of the aggregate without sample splitting, but they remain nonetheless extremely difficult to check in practice.

5 Applications

5.1 Estimating the position of a symmetric distribution

Let us consider a continuous real distribution with density f, symmetric around some parameter θ. To estimate θ from a sample of n realisations $x_1, \ldots, x_n$, a natural choice is to use the mean $\overline{x}_n$ or the median $x_{(n/2)}$. Both estimators are consistent whenever $\sigma^2 = \int (x-\theta)^2 f(x)dx$ is finite.

Surprisingly enough, the idea of combining the mean and the median to construct a better estimator goes back to Pierre Simon de Laplace [14] in early 19th century, see the discussion in [22]. P. S. de Laplace obtains the expression of the weights in $\Lambda_{\max}$ that ensure a minimal asymptotic variance for the aggregated estimator. In particular, he deduced that for a Gaussian distribution, the better combination is to take the mean only, showing for the first time the efficiency of the latter. For other distributions, he noticed that the best combination is not available in practice because it depends on the unknown distribution.

Similarly, we consider the aggregation of the mean and the median over $\Lambda_{\max}$. In the setting of the previous sections, we have two experts $T_1 = \overline{x}_n$, $T_2 = x_{(n/2)}$, the space $\mathcal{H}$ is simply $\mathbb{R}$, and the aggregated estimator is given by (7) where $\mathbf{J}$ is just in this case the vector $(1,1)^\top$. The MSE matrix between the two estimators is denoted by Σ_n. We assume that the n realizations are independent and we propose two ways to estimate Σ_n:

1. *Based on the asymptotic equivalent of* Σ_n. The latter, obtained in P. S. de Laplace's work and recalled in [22], is $n^{-1}W$ where

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

Each entry of W may be naturally estimated from an initial estimate $\hat{\theta}_0$ of θ as follows: σ^2 by the empirical variance s_n^2; $\mathbb{E}|X-\theta|$ by $\hat{m} = 1/n \sum_{i=1}^n |x_i - \hat{\theta}_0|$; and $f(\theta)$ by the kernel estimator $\hat{f}(\hat{\theta}_0) = 1/(nh) \sum_{i=1}^n \exp(-(x_i - \hat{\theta}_0)^2/(2h^2))$, where h is chosen, e.g., by the so-called Silverman's rule of thumb (see [21]). With this estimation of Σ_n, we get the following aggregated estimator:

$$\hat{\theta}_{AG} = \frac{p_1}{p_1 + p_2}\overline{x}_n + \frac{p_2}{p_1 + p_2}x_{(n/2)} \tag{14}$$

where $p_1 = 1/(4\hat{f}(\hat{\theta}_0)) - \hat{m}/2$ and $p_2 = s_n^2 \hat{f}(\hat{\theta}_0) - \hat{m}/2$. This aggregated estimator corresponds to an empirical version of the best combination obtained by P. S. de Laplace.

2. *Based on non-parametric bootstrap.* We draw with replacement B samples of size n from the original dataset. We compute the mean and the median of each sample, respectively denoted $\overline{x}_n^{(b)}$ and $x_{(n/2)}^{(b)}$ for $b = 1, \ldots, B$. The MSE matrix Σ_n is then estimated by

$$\frac{1}{B}\begin{pmatrix} \sum_{b=1}^B (\overline{x}_n^{(b)} - \overline{x}_n)^2 & \sum_{b=1}^B (\overline{x}_n^{(b)} - \overline{x}_n)(x_{(n/2)}^{(b)} - x_{(n/2)}) \\ \sum_{b=1}^B (\overline{x}_n^{(b)} - \overline{x}_n)(x_{(n/2)}^{(b)} - x_{(n/2)}) & \sum_{b=1}^B (x_{(n/2)}^{(b)} - x_{(n/2)})^2 \end{pmatrix}.$$

This estimation of Σ_n leads to another aggregated estimator, denoted by $\hat{\theta}_{AGB}$.

Let us note that the first procedure above fits the asymptotic justification presented in Section 4.2, as (12) holds. For this reason, $\hat{\theta}_{AG}$ is asymptotically as efficient as the oracle, provided $\hat{\theta}_0$ is consistent. Moreover, since the experts are asymptotically Gaussian and unbiased, an optimal asymptotic confidence interval for θ can be provided without further estimation, see Section 4.2. For the second procedure, theory is lacking to study the behaviour of δ in (8) when $\hat{\Sigma}$ is estimated by non-parametric bootstrap, so no consistency can be claimed at this point. However the latter is a very natural procedure, easy to implement in practice, so it is natural to assess its performances in our simulation study.

Table 1 summarizes the estimated MSE of $\overline{x}_n$, $x_{(n/2)}$, $\hat{\theta}_{AG}$ and $\hat{\theta}_{AGB}$, for $n = 30$, 50, 100, and for different distributions, namely: Cauchy, Student with 5 degrees of freedom, Student with 7 degrees of freedom, Logistic, standard Gaussian, and an equal mixture distribution of a $\mathcal{N}(-2, 1)$ and a $\mathcal{N}(2, 1)$. For all distributions, $\theta = 0$. For the initial estimate

$\hat{\theta}_0$ in (14), we take the median $x_{(n/2)}$, because it is well defined and consistent for any continuous distribution. The number of bootstrap samples taken for $\hat{\theta}_{AGB}$ is $B = 1000$.

While the best estimator between $\overline{x}_n$ and $x_{(n/2)}$ depends on the underlying distribution, the aggregated estimators $\hat{\theta}_{AG}$ and $\hat{\theta}_{AGB}$ perform better than both $\overline{x}_n$ and $x_{(n/2)}$, for all distributions considered in Table 1 except the Gaussian law. For the latter distribution, we know that the oracle is the mean, so our aggregate cannot improve on $\overline{x}_n$. However the MSE of $\hat{\theta}_{AG}$ and $\hat{\theta}_{AGB}$ are very close to that of $\overline{x}_n$ in this case, proving that the optimal weights $(1, 0)$ are fairly well estimated. Moreover, note that the Cauchy distribution does not belong to our theoretical setting because it has no finite moments and $\overline{x}_n$ should not be used. But it turns out that the aggregated estimators are very robust in this case, as they manage to select $x_{(n/2)}$. Choosing the median $x_{(n/2)}$ as the initial estimator $\hat{\theta}_0$ is of course crucial in this case.

Finally, while $\hat{\theta}_{AGB}$ suffers from a lack of theoretical justification, it behaves pretty much like $\hat{\theta}_{AG}$, except for the mixture distribution where it performs slightly better than $\hat{\theta}_{AG}$. This may be explained by the fact that $\hat{\theta}_{AG}$ is more sensitive than $\hat{\theta}_{AGB}$ to the initial estimate $\hat{\theta}_0$, the variance of which is large for the mixture distribution because $f(0)$ is close to 0. Nevertheless, $\hat{\theta}_{AG}$ demonstrates very good performance in this case, for the sample sizes considered in Table 1 .

	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: Monte Carlo estimation of the MSE of $\overline{x}_n$ (MEAN), $x_{(n/2)}$ (MED), $\hat{\theta}_{AG}$ (AG) and $\hat{\theta}_{AGB}$ (AGB) in the estimation of the position of a symmetric distribution, depending on the distribution and the sample size. The number of replications is 10^4 and the standard deviation of the MSE estimations is given in parenthesis. Each entry has been multiplied by 100 for ease of presentation.

5.2 Estimating the parameters of a Weibull distribution

We consider aggregation of estimators in a parametric setting, namely the Weibull distribution with shape parameter $\beta > 0$ and scale parameter $\eta > 0$, the density function of which is

$$f(x) = \frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1} \mathrm{e}^{-(x/\eta)^\beta}, \quad x > 0.$$

Based on a sample of n independent realizations, many estimators of β and η are available (see [10]). We consider the following three standard methods:

- the maximum likelihood estimator (ML) is the solution of the system

$$\frac{n}{\beta} + \sum_{i=1}^{n} \log(x_i) - n\frac{\sum_{i=1}^n x_i^\beta \log(x_i)}{\sum_{i=1}^n x_i^\beta} = 0, \quad \eta = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^\beta\right)^{1/\beta}.$$

- the method of moments (MM), based on the two first moments, reduces to solve:

$$\frac{s_n^2}{\overline{x}_n^2} = \frac{\Gamma(1+2/\beta)}{\Gamma(1+1/\beta)^2} - 1, \quad \eta = \frac{\overline{x}_n}{\Gamma(1+1/\beta)},$$

  where $\overline{x}_n$ and s_n denote the empirical sample mean and the unbiased sample variance.

- the ordinary least squares method (OLS) is based on the fact that for any $x > 0$, $\log(-\log(1-F(x)) = \beta\log(x) - \beta\log\eta$, where F denotes the cumulative distribution function of the Weibull distribution. More precisely, denoting $x_{(1)}, \ldots, x_{(n)}$ the ordered sample, an estimation of β and η is deduced from the simple linear regression of $(\log(-\log(1 - F(x_{(i)})))_{i=1\ldots n}$ on $(\log x_{(i)})_{i=1\ldots n}$, where according to the "mean rank" method $F(x_{(i)})$ may be estimated by $i/(n+1)$. This fitting method is very popular in the engineer community (see [1]): the estimation of β simply corresponds to the slope in a "Weibull plot".

The performances of these three estimators are variable, depending on the value of the parameters and the sample size. In particular, no one is uniformly better than the others, see Figure 1 for an illustration.

Let us now consider the aggregation of these estimators. In the setting of the previous sections, we have $d = 2$ parameters in $\mathcal{H} = \mathbb{R}$ to estimate and $k_1 = 3$, $k_2 = 3$ estimators (i.e. experts) of each are available. The aggregation over the maximal constraint set $\Lambda_{\max}$ demands to estimate the 6×6 MSE matrix Σ, that involves 21 unknown values. The Weibull distribution is often used to model lifetimes, and typically only a low number of observations are available to estimate the parameters. As a consequence aggregation

over $\Lambda_{\max}$ of the 6 experts above could be too demanding. Moreover, between the two parameters β and η, the shape parameter β is often the most important to identify, as it characterizes for instance the type of failure rate in reliability engineering. For these reasons, we choose to aggregate the three estimators of β presented above, $\hat{\beta}_{ML}$, $\hat{\beta}_{MM}$ and $\hat{\beta}_{OLS}$, and to consider only one estimator of η: $\hat{\eta}_{ML}$ (where $\hat{\beta}_{ML}$ is used for its computation). The aggregation over $\Lambda_{\max}$ of these 4 estimators has three consequences: First, the number of unknown values in the MSE matrix is reduced to 10. Second, the aggregated estimator of β depends only on $\hat{\beta}_{ML}$, $\hat{\beta}_{MM}$ and $\hat{\beta}_{OLS}$, because $\hat{\eta}_{ML}$ has a zero weight from (2). This means that we actually implement a component-wise aggregation for β. Third, the aggregated estimator of η equals $\hat{\eta}_{ML}$ plus some linear combination of $\hat{\beta}_{ML}$, $\hat{\beta}_{MM}$ and $\hat{\beta}_{OLS}$ where the weights sum to zero. This particular situation will allow us to see if $\hat{\eta}_{ML}$ can be improved by exploiting the correlation with the estimators of β, or if it is deteriorated.

So we have $d = 2$, $k_1 = 3$, $k_2 = 1$, $\mathbf{T}_1 = (\hat{\beta}_{ML}, \hat{\beta}_{MM}, \hat{\beta}_{OLS})^\top$, $\mathbf{T}_2 = \hat{\eta}_{ML}$ and the aggregated estimator over $\Lambda_{\max}$ is given by (7), denoted by $(\hat{\beta}_{AG}, \hat{\eta}_{AG})^\top$. The matrix Σ is estimated by parametric bootstrap: Starting from initial estimates $\hat{\beta}_0$, $\hat{\eta}_0$, we simulate B samples of size n of a Weibull distribution with parameters $\hat{\beta}_0$, $\hat{\eta}_0$. Then the four estimators are computed, which gives $\hat{\beta}_{ML}^{(b)}$, $\hat{\beta}_{MM}^{(b)}$, $\hat{\beta}_{OLS}^{(b)}$ and $\hat{\eta}_{ML}^{(b)}$, for $b = 1, \ldots, B$, and each entry of Σ is estimated by its bootstrap counterpart. For instance the estimation of $\mathbb{E}(\hat{\beta}_{ML} - \beta)(\hat{\beta}_{MM} - \beta)$ is $(1/B)\sum_{b=1}^{B}(\hat{\beta}_{ML}^{(b)} - \hat{\beta}_0)(\hat{\beta}_{MM}^{(b)} - \hat{\beta}_0)$. In our simulations, we chose $\hat{\beta}_0$ as the mean of $\mathbf{T}_1$, and $\hat{\eta}_0 = \hat{\eta}_{ML}$. Note that Σ having a parametric form ensures that $(\hat{\beta}_{AG}, \hat{\eta}_{AG})^\top$ is asymptotically as efficient as the oracle, see Corollary 4.4.

Table 2 gives the MSE, estimated from 10^4 replications, of each estimator of β, for $n = 10, 20, 50$, and for $\beta = 0.5, 1, 2, 3$, $\eta = 10$. The aggregated estimator has by far the lowest MSE, even for small samples. As an illustration, the repartition of each estimator, for $n = 20$ and $\beta = 0.5, 3$, is represented in Figure 1.

Table 3 shows the MSE for $\hat{\eta}_{ML}$ and $\hat{\eta}_{AG}$ where only estimators of β were used in attempt to improve $\hat{\eta}_{ML}$ by aggregation. The performances of both estimators are similar, showing that the information coming from $\mathbf{T}_1$ did not help significantly improving $\hat{\eta}_{ML}$. On the other hand, the estimation of these (almost zero) weights might have deteriorated $\hat{\eta}_{ML}$, especially for small sample sizes. This did not happen.

5.3 Aggregating kernel density estimators

Let $x_1, \ldots, x_n$ be a sample from a real random variable with density f. The kernel density estimator of f at $x \in \mathbb{R}$ is

$$\hat{f}_{n,h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

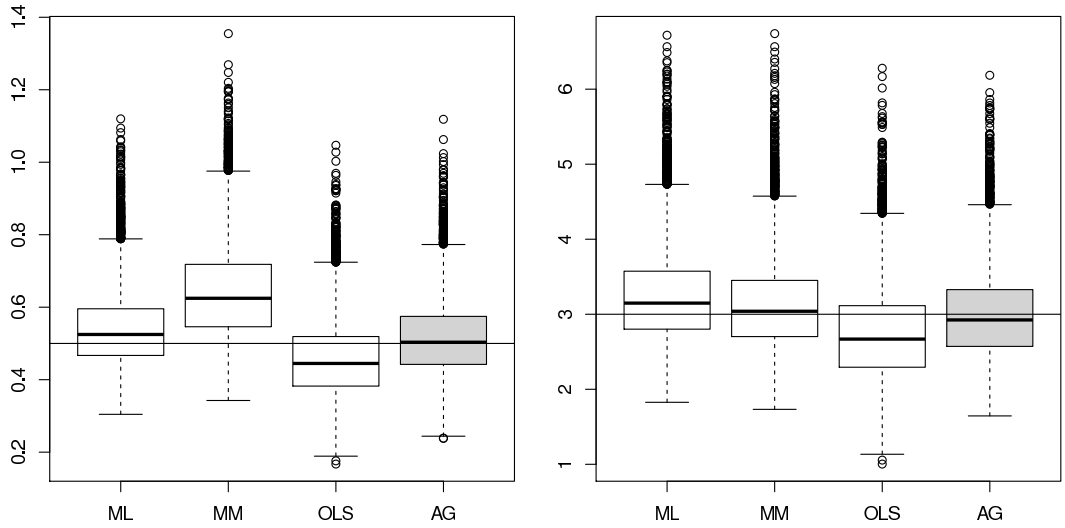

Figure 1: Repartition of $\hat{\beta}_{ML}$, $\hat{\beta}_{MM}$, $\hat{\beta}_{OLS}$ and $\hat{\beta}_{AG}$ (from left to right) based on 10^4 replications of a sample of size $n = 20$ from a Weibull distribution with $\beta = 0.5$ (left), $\beta = 3$ (right) and $\eta = 10$.

	n=10				n=20				n=50			
	ML	MM	OLS	AG	ML	MM	OLS	AG	ML	MM	OLS	AG
$\beta = 0.5$	35.53	76.95	24.41	25.27	12.06	35.57	13.74	10.5	3.7	14.19	6.04	3.52
	(0.91)	(1.27)	(0.40)	(0.64)	(0.26)	(0.52)	(0.19)	(0.19)	(0.07)	(0.20)	(0.08)	(0.06)
$\beta = 1$	152.4	131.6	98.1	85.5	49.2	53.6	54.2	36.9	14.4	19.3	23.9	12.8
	(3.8)	(3.1)	(1.5)	(1.7)	(1.1)	(1.1)	(0.7)	(0.7)	(0.2)	(0.3)	(0.3)	(0.2)
$\beta = 2$	596.4	444.6	399.4	355.5	194.5	164.5	218	163.3	57.9	53.9	94.8	54.3
	(14.4)	(11.9)	(6.3)	(6.7)	(3.8)	(3.3)	(2.8)	(2.7)	(1.0)	(0.9)	(1.3)	(0.9)
$\beta = 3$	1369	1080	905	770	452	394	486	343	128	122	211	120
	(34.6)	(29.7)	(14.6)	(18.1)	(9.8)	(8.9)	(6.7)	(6.2)	(2.2)	(2.0)	(2.7)	(1.9)

Table 2: Monte Carlo estimation of the MSE of $\hat{\beta}_{ML}$, $\hat{\beta}_{MM}$, $\hat{\beta}_{OLS}$ and $\hat{\beta}_{AG}$, based on 10^4 replications of a sample of size $n = 10$, 20, 50 from a Weibull distribution with parameters $\beta = 0.5$, 1, 2, 3 and $\eta = 10$. The standard deviation of the MSE estimations are given in parenthesis. Each entry has been multiplied by 100 for ease of presentation.

	n=10		n=20		n=50	
	ML	AG	ML	AG	ML	AG
$\beta = 0.5$	60.59	55.61	25.96	24.56	9.57	9.38
	(1.60)	(1.48)	(0.53)	(0.5)	(0.17)	(0.17)
$\beta = 1$	11.15	10.88	5.53	5.43	2.23	2.22
	(0.18)	(0.17)	(0.08)	(0.08)	(0.03)	(0.03)
$\beta = 2$	2.71	2.74	1.36	1.37	0.55	0.56
	(0.04)	(0.04)	(0.02)	(0.02)	(0.01)	(0.01)
$\beta = 3$	1.21	1.23	0.61	0.61	0.247	0.248
	(0.02)	(0.02)	(0.01)	(0.01)	(0.003)	(0.004)

Table 3: Monte Carlo estimation of the MSE of $\hat{\eta}_{ML}$ and $\hat{\eta}_{AG}$, based on 10^4 replications of a sample of size $n = 10$, 20, 50 from a Weibull distribution with parameters $\beta = 0.5$, 1, 2, 3 and $\eta = 10$. The standard deviation of the MSE estimations are given in parenthesis. Each entry has been multiplied by 100 for ease of presentation.

where the function K is the kernel and h is a smoothing parameter called bandwidth. It is well known that the choice of K has only a small impact on the performances of $\hat{f}_{n,h}$, while the choice of h is crucial. Many works are dedicated to propose some data-based bandwidth selection. We refer to [19] for a review.

For our simulations, we choose the standard Gaussian kernel $K(x) = \mathrm{e}^{-x^2/2}/\sqrt{2\pi}$ and we consider four choices of bandwidth, see [19] for more details and references:

- The Silverman's rule-of-thumb, namely $h_1 = 0.9 \min(s_n, IQR/1.34)\, n^{-1/5}$, where s_n denotes the standard deviation and IQR the interquartile range.
- The variation proposed in [18], where the constant 0.9 above is replaced by 1.06.
- The unbiased (or least squares) cross-validation method.
- The plug-in method of Sheather and Jones [20].

These four possible choices, called $h_1, \ldots, h_4$ in the following, correspond to the bandwidths proposed in the function `density` implemented in the freeware R: `nrd0`, `nrd`, `ucv` and `SJ` respectively.

Our goal is to aggregate the experts $\hat{f}_{n,h_i}$, $i = 1, \ldots, 4$ to obtain a better estimate of f. In most existing methods of density aggregation (see the references cited in introduction), the observations are assumed independent and the procedure mainly relies on a sample splitting, where a training sample is used to compute the experts $\hat{f}_{n,h_i}$ and the aggregation weights are estimated from the validation sample. In contrast, we propose to construct

an aggregate that minimizes the mean integrated square error (MISE), defined for an estimator $\hat{f}$ of f by $\int \mathbb{E}(\hat{f}(x) - f(x))^2 dx$. In the setting of the previous sections, $\mathcal{H} = L^2(\mathbb{R})$ and the oracle given by (3) involves the MISE matrix Σ with entries $\int \mathbb{E}(\hat{f}_{n,h_i}(x) - f(x))(\hat{f}_{n,h_j}(x) - f(x))dx$. The aggregated estimator over $\Lambda_{\max}$ is then given by (7), where $d = 1$ and $k_1 = 4$. In particular, our procedure does not require sample splitting, and we do not assume independence of the observations.

The main difficulty is the estimation of the MISE matrix Σ which is at the heart of most methods of bandwidth selection. A standard procedure consists in estimating the asymptotic form AMISE of the MISE. If the bandwidths h_i are deterministic, assuming that the observations are independent or weakly dependent (and under further mild assumptions, see for instance [2]), the asymptotic equivalent of each entry of Σ is given by:

$$AMISE(h_i, h_j) = \frac{1}{n\sqrt{h_i h_j}} \; I_K\left(\sqrt{\frac{h_j}{h_i}}\right) + \frac{h_i^2 h_j^2}{4}\mu_2^2(K)R(f'') \tag{15}$$

where for $\alpha > 0$, $I_K(\alpha) = \int K(\alpha u)K(\alpha^{-1}u)du$, $\mu_2(K) = \int x^2 K(x)dx$ and $R(f'') = \int (f''(x))^2 dx$. When K is the Gaussian kernel, we have $I_K(\alpha) = (1/\sqrt{2\pi})\,\alpha/\sqrt{\alpha^4 + 1}$ and $\mu_2(K) = 1$. If the bandwidths h_i depend on the observations, as for the four above choices, then (15) is the conditional AMISE given $h_1, \ldots, h_4$. We choose to estimate Σ by (15) where $R(f'')$ is estimated by the standard plug-in method proposed in [11], which turns out to be also used for the computation of h_4, see [20]. From a theoretical point of view, if all h_i's are deterministic and of the form $h_i = c_i n^{-1/5}$, $c_i > 0$, then (12) holds and our procedure provides a consistent aggregated estimator. In presence of random c_i's, some further investigations are necessary to prove consistency.

Our aggregated estimator, say $\hat{f}_{n,AG}$, is thus given by (7) with $\hat{\Sigma}$ obtained as explained above. Table 4 summarises the estimated MISE of $\hat{f}_{n,h_i}$, $i = 1, \ldots, 4$ and $\hat{f}_{n,AG}$, for some standard densities f and different sample sizes $n = 250, 500, 1000$, when the observations are independently drawn from f. Specifically, we consider the densities of the standard Gaussian law $\mathcal{N}(0, 1)$, of the equal mixture of a $\mathcal{N}(-1.5, 1)$ and a $\mathcal{N}(1.5, 1)$, of the Gamma distribution $\Gamma(2, 1)$ and of the Cauchy distribution. The MISE is estimated by averaging over 10^4 replications the integrated square error, obtained by the sum of the square error $(\hat{f}_{n,h_i}(x) - f(x))^2$ over 100 points x equally spaced on $[-2, 2]$, $[-3, 3]$, $[0.5, 4]$ and $[-4, 4]$, respectively. Note that Table 4 gives an estimation of the unconditional MISE (even if the estimation of the weights in $\hat{f}_{n,AG}$ is based on the conditional AMISE (15)). Moreover, Figure 2 shows the MSE functions $(\hat{f}_{n,h_i}(x) - f(x))^2$ for each distribution f when $n = 500$.

As a result, $\hat{f}_{n,AG}$ has a lower MISE than the experts, except for the mixture distribution when $n = 250$. In fact, our procedure works very well for large samples, but is less efficient for small samples. One obvious reason is that we do not actually estimate the MISE matrix Σ but its (conditional) asymptotic expression given by (15). Some simula-

	n=250					n=500					n=1000				
	h_1	h_2	h_3	h_4	AG	h_1	h_2	h_3	h_4	AG	h_1	h_2	h_3	h_4	AG
Gauss	29.9	27.2	26.8	29.9	24.9	17.7	16.2	16.2	17.3	14.4	10.5	9.7	9.8	10.1	8.4
Mix	24.0	27.5	27.1	25.2	26.7	14.8	17.6	15.3	14.9	14.2	9.1	11.1	8.9	8.8	7.4
Gamma	28.0	32.7	29.5	28.9	27.9	17.1	20.6	17.0	17.2	15.8	10.3	12.7	10.0	10.3	9.0
Cauchy	31.2	37.0	830	132	32.8	18.9	23.2	945	180	18.7	11.4	14.4	1068	226	10.6

Table 4: Monte Carlo estimation of the MISE of $\hat{f}_{n,h_i}$, for $i = 1, \ldots, 4$ and of the aggregate $\hat{f}_{n,AG}$, for different f and $n = 250, 500, 1000$. With the notation of R: $h_1 =$ `nrd0`, $h_2 =$ `nrd`, $h_3 =$ `ucv` and $h_4 =$ `SJ`. The MISE are estimated by the mean over 10^4 replications of the integrated square error, obtained by summing up the square error of 100 equally spaced points on the support of f. Each entry has been multiplied by $n.10^4$ for ease of presentation.

tions (not presented here) show that even the oracle estimator based on the asymptotic MISE matrix may have a larger MISE than the experts for small values of n. This is for instance the case for the mixture distribution when $n = 250$. Moreover, the performances of $\hat{f}_{n,AG}$ for the Cauchy distribution are remarkably good, in spite of the presence of two unadapted experts (namely $\hat{f}_{n,h_3}$ and $\hat{f}_{n,h_4}$).

An alternative procedure to estimate the MISE matrix is smooth bootstrap [24]. Recall that standard bootstrap fails to estimate the bias in a non-parametric setting. Smooth bootstrap amounts to resample according to a continuous density $\hat{f}$ that estimates f. At least two appealing features arise: First, smooth bootstrap estimates the MISE and not the asymptotic MISE, which could improve our procedure for small samples. Second, when K is the Gaussian kernel, closed-form formulas are available for the smooth bootstrap estimate of the conditional MISE given $h_1, \ldots, h_4$, and no Monte-Carlo step is needed. The unconditional MISE can also be estimated by smooth bootstrap, but some Monte-Carlo approximations are then needed. Unfortunately, smooth bootstrap depends on a pilot bandwidth for which we are unable to propose a satisfactory data-driven choice. Therefore, although smooth bootstrap seems a promising perspective for aggregating kernel density estimators based on small samples, its implementation deserves further analysis. Nevertheless, our simulation study shows that the aggregation based on the asymptotic approximation of the MISE produces satisfactory results for moderate sample sizes (e.g. $n \geq 250$).

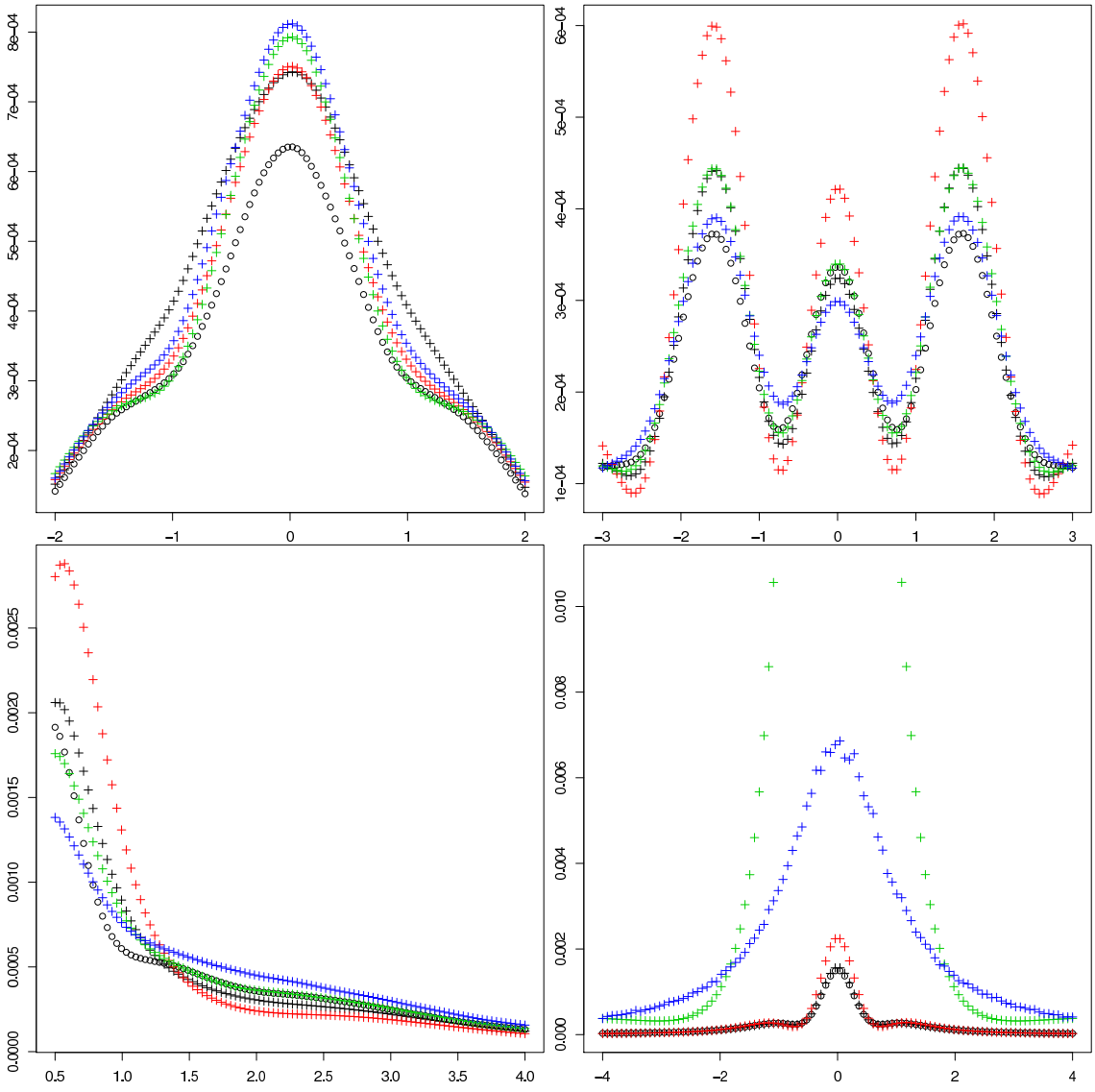

Figure 2: Monte Carlo estimation of the MSE of $\hat{f}_{n,h_1}$ (black crosses), $\hat{f}_{n,h_2}$ (red crosses), $\hat{f}_{n,h_3}$ (green crosses), $\hat{f}_{n,h_4}$ (blue crosses) and $\hat{f}_{n,AG}$ (black circles), based on 10^4 replications, for $n = 500$ and when f is the density of the Gaussian distribution (top left), the mixture distribution (top right), the Gamma distribution (bottom left) and the Cauchy distribution (bottom right). The estimated MISE in Table 4 (for $n = 500$) are the approximated integrals of these curves.

6 Appendix

Proof of Theorem 4.1

Since $\Lambda \subseteq \Lambda_{\max}$, we know that $\lambda^\top \mathbf{J} = \mathrm{I}$ for all $\lambda \in \Lambda$. Let $\mathbf{S} = \Sigma^{-\frac{1}{2}}(\mathbf{T} - \mathbf{J}\theta)$, we have

$$\|\hat{\theta} - \hat{\theta}^*\|^2 = \|(\hat{\lambda} - \lambda^*)^\top(\mathbf{T} - \mathbf{J}\theta)\|^2 = \|(\hat{\lambda} - \lambda^*)^\top \Sigma^{\frac{1}{2}}\mathbf{S}\|^2 \leq \|(\hat{\lambda} - \lambda^*)^\top \Sigma^{\frac{1}{2}}\|_F^2 \, \|\mathbf{S}\|^2, \quad (16)$$

where $\|A\|_F = \sqrt{\mathrm{tr}(A^\top A)}$ denotes the Frobenius norm of A. The map $\phi : \lambda \mapsto \mathrm{tr}(\lambda^\top \Sigma \lambda)$ is coercive, and strictly convex by assumption. So, since Λ is closed and convex, the minimum of ϕ on Λ is reached at a unique point $\lambda^* \in \Lambda$. Moreover, we know that for $\lambda \in \Lambda$, $\lambda^* + t(\lambda - \lambda^*)$ lies in Λ for all $t \in [0, 1]$, to which we deduce the optimality condition

$$\lim_{t \to 0^+} \frac{\phi(\lambda^* + t(\lambda - \lambda^*)) - \phi(\lambda^*)}{t} = \mathrm{tr}\left[\nabla\phi(\lambda^*)^\top(\lambda - \lambda^*)\right] = 2\mathrm{tr}\left[{\lambda^*}^\top \Sigma(\lambda - \lambda^*)\right] \geq 0,$$

for all $\lambda \in \Lambda$. It follows that

$$\begin{aligned} \|(\hat{\lambda} - \lambda^*)\Sigma^{\frac{1}{2}}\|_F^2 &= \mathrm{tr}(\hat{\lambda}^\top \Sigma \hat{\lambda}) - \mathrm{tr}({\lambda^*}^\top \Sigma \lambda^*) - 2\mathrm{tr}\left[{\lambda^*}^\top \Sigma(\hat{\lambda} - \lambda^*)\right] \\ &\leq \mathrm{tr}(\hat{\lambda}^\top \Sigma \hat{\lambda}) - \mathrm{tr}({\lambda^*}^\top \Sigma \lambda^*). \end{aligned} \quad (17)$$

By construction of $\hat{\lambda}$, we know that $\mathrm{tr}(\hat{\lambda}^\top \hat{\Sigma} \hat{\lambda}) \leq \mathrm{tr}({\lambda^*}^\top \hat{\Sigma} \lambda^*)$, yielding

$$\begin{aligned} \mathrm{tr}(\hat{\lambda}^\top \Sigma \hat{\lambda}) - \mathrm{tr}({\lambda^*}^\top \Sigma \lambda^*) &\leq \mathrm{tr}(\hat{\lambda}^\top \Sigma \hat{\lambda}) - \mathrm{tr}(\hat{\lambda}^\top \hat{\Sigma} \hat{\lambda}) + \mathrm{tr}({\lambda^*}^\top \hat{\Sigma} \lambda^*) - \mathrm{tr}({\lambda^*}^\top \Sigma \lambda^*) \\ &\leq \mathrm{tr}(\hat{\lambda}^\top \hat{\Sigma} \hat{\lambda}) \; \delta_\Lambda(\Sigma|\hat{\Sigma}) + \mathrm{tr}({\lambda^*}^\top \Sigma \lambda^*) \; \delta_\Lambda(\hat{\Sigma}|\Sigma) \\ &\leq \left[\mathrm{tr}(\hat{\lambda}^\top \hat{\Sigma} \hat{\lambda}) + \mathrm{tr}({\lambda^*}^\top \Sigma \lambda^*)\right] \delta_\Lambda(\hat{\Sigma}, \Sigma) \end{aligned}$$

where we recall $\delta_\Lambda(A|B) = \sup_{\lambda \in \Lambda} \left|1 - \frac{\mathrm{tr}(\lambda^\top A \lambda)}{\mathrm{tr}(\lambda^\top B \lambda)}\right|$ and $\delta_\Lambda(A, B) = \max\{\delta_\Lambda(A|B), \delta_\Lambda(B|A)\}$. Now using that $\mathrm{tr}(\hat{\lambda}^\top \hat{\Sigma} \hat{\lambda}) \leq \mathrm{tr}({\lambda^*}^\top \hat{\Sigma} \lambda^*)$ and

$$\mathrm{tr}({\lambda^*}^\top \hat{\Sigma} \lambda^*) = \mathrm{tr}({\lambda^*}^\top \Sigma \lambda^*) + \left[\mathrm{tr}({\lambda^*}^\top \hat{\Sigma} \lambda^*) - \mathrm{tr}({\lambda^*}^\top \Sigma \lambda^*)\right] \leq \mathrm{tr}({\lambda^*}^\top \Sigma \lambda^*) \left[1 + \delta_\Lambda(\hat{\Sigma}, \Sigma)\right],$$

we obtain

$$\mathrm{tr}(\hat{\lambda}^\top \Sigma \hat{\lambda}) - \mathrm{tr}({\lambda^*}^\top \Sigma \lambda^*) \leq \mathrm{tr}({\lambda^*}^\top \Sigma \lambda^*) \; \left[2\delta_\Lambda(\hat{\Sigma}, \Sigma) + \delta_\Lambda(\hat{\Sigma}, \Sigma)^2\right]. \quad (18)$$

Recall that $\mathrm{tr}({\lambda^*}^\top \Sigma \lambda^*) = \inf_{\lambda \in \Lambda} \mathbb{E}\|\lambda^\top \mathbf{T} - \theta\|^2$, the result follows from (16), (17) and (18).

Proof of Proposition 4.2

We use the following preliminary result.

Lemma 6.1 *Let A, B be two positive definite matrices of order k. For any non-empty set Λ that does not contain* 0*,*

$$\delta_\Lambda(A, B) \leq |||AB^{-1} - BA^{-1}|||,$$

where $|||A||| = \sup_{\|x\|_F=1} \|Ax\|_F$ stands for the operator norm.

Proof. By symmetry, it is sufficient to show that the result holds for $\delta_\Lambda(A|B)$. We have

$$\delta_\Lambda(A|B) = \sup_{\lambda\in\Lambda} \frac{|\mathrm{tr}[\lambda^\top(B-A)\lambda]|}{\mathrm{tr}(\lambda^\top B\lambda)} \leq \sup_{\lambda\neq 0} \frac{|\mathrm{tr}[\lambda^\top(B-A)\lambda]|}{\mathrm{tr}(\lambda^\top B\lambda)}.$$

By Cauchy-Schwarz inequality,

$$\begin{aligned}
|\mathrm{tr}[\lambda^\top(B-A)\lambda]| &= \left|\mathrm{tr}\left[\lambda^\top B^{\frac{1}{2}}\ (\mathrm{I} - B^{-\frac{1}{2}}AB^{-\frac{1}{2}})\ B^{\frac{1}{2}}\lambda\right]\right| \\
&\leq \|B^{\frac{1}{2}}\lambda\|_F \|(\mathrm{I} - B^{-\frac{1}{2}}AB^{-\frac{1}{2}})\ B^{\frac{1}{2}}\lambda\|_F \\
&\leq |||\mathrm{I} - B^{-\frac{1}{2}}AB^{-\frac{1}{2}}|||\|B^{\frac{1}{2}}\lambda\|_F^2. \qquad (19)
\end{aligned}$$

Recall that $\|B^{\frac{1}{2}}\lambda\|_F^2 = \mathrm{tr}(\lambda^\top B\lambda)$, it follows

$$\delta_\Lambda(A|B) \leq |||\mathrm{I} - B^{-\frac{1}{2}}AB^{-\frac{1}{2}}|||.$$

Since the matrix $C = I - B^{-\frac{1}{2}}AB^{-\frac{1}{2}}$ is symmetric, it is diagonalizable in an orthogonal basis. In particular, denoting sp(.) the spectrum, $|||C||| = \sup_{t\in\mathrm{sp}(C)} |t|$. Finally, observe that $\mathrm{sp}(C) = 1 - \mathrm{sp}(B^{-\frac{1}{2}}AB^{-\frac{1}{2}}) = 1 - \mathrm{sp}(AB^{-1})$, so that AB^{-1} has positive eigenvalues and

$$|||\mathrm{I} - B^{-\frac{1}{2}}AB^{-\frac{1}{2}}||| = \sup_{t\in\mathrm{sp}(AB^{-1})} |1-t| \leq \sup_{t\in\mathrm{sp}(AB^{-1})} |t - \frac{1}{t}| \leq |||AB^{-1} - BA^{-1}|||,$$

ending the proof. ■

By this lemma, we deduce that

$$\delta_\Lambda(\hat{\Sigma}_n, \Sigma_n) \leq |||\hat{\Sigma}_n\Sigma_n^{-1} - \Sigma_n\hat{\Sigma}_n^{-1}|||. \qquad (20)$$

In particular, $\delta_\Lambda(\hat{\Sigma}_n, \Sigma_n) = o_p(1)$ by the assumption $\hat{\Sigma}_n\Sigma_n^{-1} \xrightarrow{p} \mathrm{I}$. Write for $c > 0$,

$$\|\hat{\theta}_n - \theta\|^2 \leq (1+c)\|\hat{\theta}_n^* - \theta\|^2 + (1+c^{-1})\|\hat{\theta}_n - \hat{\theta}_n^*\|^2.$$

Applying Theorem 4.1, we get

$$\|\hat{\theta}_n - \theta\|^2 \leq (1+c)\|\hat{\theta}_n^* - \theta\|^2 + (1+c^{-1})\alpha_n \left(2\delta_\Lambda(\hat{\Sigma}_n, \Sigma_n) + \delta_\Lambda(\hat{\Sigma}_n, \Sigma_n)^2\right) \|\mathbf{S}_n\|^2, \quad (21)$$

where $\mathbf{S}_n = \Sigma_n^{-\frac{1}{2}}(\mathbf{T}_n - \mathbf{J}\theta)$. Since $\mathbb{E}\|\mathbf{S}_n\|^2 = k$, we know that $\|\mathbf{S}_n\|^2 = O_p(1)$. Equation (21) holds for all $c > 0$ so we can take $c = c_n$ such that $c_n \to 0$ and $\delta_\Lambda(\hat{\Sigma}_n, \Sigma_n)/c_n \xrightarrow{p} 0$ as $n \to \infty$, yielding

$$\|\hat{\theta}_n - \theta\|^2 \leq \|\hat{\theta}_n^* - \theta\|^2 + c_n\|\hat{\theta}_n^* - \theta\|^2 + o_p(\alpha_n) = \|\hat{\theta}_n^* - \theta\|^2 + o_p(\alpha_n).$$

We shall now prove the second part of the proposition. Write,

$$\hat{\alpha}_{n,j}^{-\frac{1}{2}}(\hat{\theta}_{n,j} - \theta_j) = \sqrt{\frac{\alpha_{n,j}}{\hat{\alpha}_{n,j}}}\ \alpha_{n,j}^{-\frac{1}{2}}\left[(\hat{\theta}_{n,j}^* - \theta_j) + (\hat{\theta}_{n,j} - \hat{\theta}_{n,j}^*)\right].$$

To prove the result, it suffices to show that $\alpha_{n,j}^{-\frac{1}{2}}\|\hat{\theta}_{n,j} - \hat{\theta}_{n,j}^*\| = o_p(1)$ and $\alpha_{n,j}/\hat{\alpha}_{n,j} \xrightarrow{p} 1$. When Λ is a cylinder, it is easy to see that the following holds

$$\Pi_j(\hat{\lambda}_n) = \arg\min_{\lambda \in \Lambda_j}\ \lambda^\top \hat{\Sigma}_n \lambda \ \text{ and } \ \Pi_j(\lambda_n^*) = \arg\min_{\lambda \in \Lambda_j}\ \lambda^\top \Sigma_n \lambda,$$

where we recall $\Lambda_j = \{\Pi_j(\lambda) : \lambda \in \Lambda\}$. Moreover, it is easy to adapt the proof of Theorem 4.1 to get

$$\|\hat{\theta}_{n,j} - \hat{\theta}_{n,j}^*\|^2 \leq \alpha_{n,j}\ \left(2\delta_{\Lambda_j}(\hat{\Sigma}, \Sigma) + \delta_{\Lambda_j}(\hat{\Sigma}, \Sigma)^2\right)\ \|\Sigma^{-\frac{1}{2}}(\mathbf{T} - \mathbf{J}\theta)\|^2.$$

We deduce that $\alpha_{n,j}^{-\frac{1}{2}}(\hat{\theta}_{n,j} - \hat{\theta}_{n,j}^*) = o_p(1)$ in view of (9) and Lemma 6.1. Now, remark that

$$\frac{\alpha_{n,j}}{\hat{\alpha}_{n,j}} = \frac{\Pi_j(\lambda_n^*)^\top \Sigma_n \Pi_j(\lambda_n^*)}{\Pi_j(\hat{\lambda}_n)^\top \hat{\Sigma}_n \Pi_j(\hat{\lambda}_n)} \leq \frac{\Pi_j(\hat{\lambda}_n)^\top \Sigma_n \Pi_j(\hat{\lambda}_n)}{\Pi_j(\hat{\lambda}_n)^\top \hat{\Sigma}_n \Pi_j(\hat{\lambda}_n)} - 1 + 1 \leq \delta_{\Lambda_j}(\hat{\Sigma}_n, \Sigma_n) + 1.$$

Similarly,

$$\frac{\hat{\alpha}_{n,j}}{\alpha_{n,j}} \leq \delta_{\Lambda_j}(\hat{\Sigma}_n, \Sigma_n) + 1.$$

So, we get

$$\frac{1}{1 + \delta_{\Lambda_j}(\hat{\Sigma}_n, \Sigma_n)} \leq \frac{\alpha_{n,j}}{\hat{\alpha}_{n,j}} \leq 1 + \delta_{\Lambda_j}(\hat{\Sigma}_n, \Sigma_n),$$

proving that $\alpha_{n,j}/\hat{\alpha}_{n,j} \xrightarrow{p} 1$.

Proof of Lemma 4.3

Using that by assumption $P\Sigma_n^{-1}P^\top A_n^{-1}$ tends to D^{-1}, and that $\hat{P}_n$ (resp. $\hat{D}_n$) converges in probability to P (resp. D), it follows that

$$\hat{\Sigma}_n\Sigma_n^{-1} = \hat{P}_n^\top A_n^{-1}\hat{D}_n\hat{P}_n\Sigma_n^{-1} = \hat{P}_n^\top A_n^{-1}\hat{D}_n\hat{P}_nP^\top P\Sigma_n^{-1}P^\top A_n^{-1}A_nP$$

converges in probability to I.

References

[1] Abernethy, R. *The New Weibull Handbook: Reliability and Statistical Analysis for Predicting Life, Safety, Supportability, Risk, Cost and Warranty Claims*, fifth ed. Barringer & Associates, 2006.

[2] Bosq, D. *Nonparametric statistics for stochastic processes*, vol. 110 of *Lecture Notes in Statistics*. Springer Verlag, 1998.

[3] Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. Aggregation and sparsity via l_1 penalized least squares. In *Learning theory*, vol. 4005 of *Lecture Notes in Comput. Sci.* Springer, Berlin, 2006, pp. 379–391.

[4] Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. Aggregation for Gaussian regression. *Ann. Statist. 35*, 4 (2007), 1674–1697.

[5] Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. Sparse density estimation with ℓ_1 penalties. In *Learning theory*, vol. 4539 of *Lecture Notes in Comput. Sci.* Springer, Berlin, 2007, pp. 530–543.

[6] Catoni, O. The mixture approach to universal model selection. Tech. rep., Ecole normale supérieure, 1997.

[7] Dalalyan, A. S., and Salmon, J. Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist. 40*, 4 (2012), 2327–2355.

[8] Goldenshluger, A. A universal procedure for aggregating estimators. *Ann. Statist. 37*, 1 (2009), 542–568.

[9] Graybill, F. A., and Deal, R. B. Combining unbiased estimators. *Biometrics 15* (1959), 543–550.

[10] Johnson, N. L., Kotz, S., and Balakrishnan, N. *Continuous univariate distributions. Vol. 1*, second ed. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1994.

[11] Jones, M. C., and Sheather, S. J. Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statistics & Probability Letters 11*, 6 (1991), 511–514.

[12] Juditsky, A., and Nemirovski, A. Functional aggregation for nonparametric regression. *Ann. Statist. 28*, 3 (2000), 681–712.

[13] Keller, T., and Olkin, I. Combining correlated unbiased estimators of the mean of a normal distribution. In *A festschrift for Herman Rubin*, vol. 45 of *IMS Lecture Notes Monogr. Ser.* Inst. Math. Statist., Beachwood, OH, 2004, pp. 218–227.

[14] Laplace, P.-S. d. *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).

[15] Mehta, J., and Gurland, J. On combining unbiased estimators of the mean. *Trabajos de estadística y de investigación operativa 20* (1969), 173–185.

[16] Nocedal, J., and Wright, S. Numerical optimization. *Springer, New York* (2006).

[17] Rigollet, P., and Tsybakov, A. Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics 16*, 3 (2007), 260–280.

[18] Scott, D. W. *Multivariate density estimation: theory, practice, and visualization.* Wiley, 1992.

[19] Sheather, S. J. Density estimation. *Statistical Science 19*, 4 (2004), 588–597.

[20] Sheather, S. J., and Jones, M. C. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological) 53*, 3 (1991), 683–690.

[21] Silverman, B. W. *Density estimation for statistics and data analysis.* Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986.

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

[23] Taniguchi, M., and Tresp, V. Averaging regularized estimators. *Neural Computation 9*, 5 (1997), 1163–1178.

[24] Taylor, C. C. Bootstrap choice of the smoothing parameter in kernel density estimation. *Biometrika 76*, 4 (1989), 705–712.

[25] Tresp, V., and Taniguchi, M. Combining estimators using non-constant weighting functions. In *Advances in Neural Information Processing Systems 7* (1995), MIT Press, pp. 419–426.

[26] Tsybakov, A. B. Optimal rates of aggregation. *Proceedings of 16th Annual Conference on Learning Theory (COLT) and 7th Annual Workshop on Kernel Machines. Lecture Notes in Artificial Intelligence 2777* (2003), 303–313.

[27] Yang, Y. Mixing strategies for density estimation. *Ann. Statist. 28*, 1 (2000), 75–87.