A general procedure to combine estimators

F. Lavancier[a,b,*], P. Rochet[a]

[a]*University of Nantes, Laboratoire de Mathématiques Jean Leray, 2 rue de la Houssinière, 44322 Nantes, France*
[b]*Inria, Centre Rennes Bretagne Atlantique, Campus universitaire de Beaulieu 35042 Rennes, France*

Abstract

A general method to combine several estimators of the same quantity is investigated. In the spirit of model and forecast averaging, the final estimator is computed as a weighted average of the initial ones, where the weights are constrained to sum to one. In this framework, the optimal weights, minimizing the quadratic loss, are entirely determined by the mean squared error matrix of the vector of initial estimators. The averaging estimator is built using an estimation of this matrix, which can be computed from the same dataset. A non-asymptotic error bound on the averaging estimator is derived, leading to asymptotic optimality under mild conditions on the estimated mean squared error matrix. This method is illustrated on standard statistical problems in parametric and semi-parametric models where the averaging estimator outperforms the initial estimators in most cases.

Keywords: Averaging, Parametric estimation, Weibull model, Boolean model

1. Introduction

We are interested in estimating a parameter θ in a statistical model, based on a collection of preliminary estimators $T_1, ..., T_k$. In general, the relative performance of each estimator depends on the true value of the parameter, the sample size, or other unknown factors, in which case deciding in advance

---

[*]Corresponding author
Email addresses: `lavancier@univ-nantes.fr` (F. Lavancier), `rochet@univ-nantes.fr` (P. Rochet)

what method to favor can be difficult. This situation occurs in numerous problems of modern statistics like forecasting or non-parametric regression, but it remains a major concern even in simple parametric problems. In this paper, we study a general methodology to combine linearly several estimators in order to produce a final single better estimator.

The issue of dealing with several possibly competing estimators of the same quantity has been extensively studied in the literature these past decades. One of the main solution retained is to consider a weighted average of the T_i's. The idea of estimator averaging actually goes back to the early 19th century with Pierre Simon de Laplace [19], who was interested in finding the best combination between the mean and the median to estimate the location parameter of a symmetric distribution, see the discussion in [29]. More generally, the solution can be expressed as a linear combination of the initial estimators

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

for λ a vector of weights lying in a subset Λ of $\mathbb{R}^k$ and $\mathbf{T} = (T_1, \ldots, T_k)^\top$. A large number of statistical frameworks fit with this description. For example, model selection can be viewed as a particular case for Λ the set of vertices. Similarly, convex combinations corresponds to the simplex $\Lambda = \{\lambda : \sum_{i=1}^{k} \lambda_i = 1, \lambda_i \geq 0\}$ while linear combinations to $\Lambda = \mathbb{R}^k$. Another well-used framework consists in relaxing the positivity condition of convex combination, corresponding to the set $\Lambda = \{\lambda : \sum_{i=1}^{k} \lambda_i = 1\}$.

Estimator averaging has received a particular attention for prediction purposes. Ever since the paper of Bates and Granger [2], dealing with forecast averaging for time series, the literature on this subject has greatly developed, see for instance [10, 30] in econometrics and [7] in machine learning. In this framework, the parameter θ represents the future observation of a series to be predicted and $\mathbf{T} = (T_1, ..., T_k)$ a collection of predictors. Averaging methods have also been widely used for prediction in a regression framework. In this case, θ is the response variable to predict given some regressors, and $\mathbf{T}$ is a collection of models output. These so-called model averaging procedures are shown to provide good alternatives to model selection for parametric regression, see [23] for a survey. Model averaging has been studied in both Bayesian [26, 32] and frequentist contexts [3, 13, 15]. In close relation, func-

tional aggregation deals with the same problem in non-parametric regression [4, 9, 24, 31, 35]. Aggregation methods have also been extensively studied for density estimation, as an alternative to classical bandwidth selection methods [5, 6, 27, 34].

In [13], Hansen introduced a least squares model average estimator, in the same spirit as the forecast average estimator proposed in [2]. Loosely speaking, this estimator aims to mimic the oracle, defined as the linear combination $\hat{\theta}_\lambda$ that minimizes the quadratic loss $\mathbb{E}(\hat{\theta}_\lambda - \theta)^2$, under the constraint on the weights $\sum_{i=1}^k \lambda_i = 1$. Under this constraint, the oracle expresses in terms of the mean squared error matrix Σ of $\mathbf{T}$. The averaging estimator is then defined by replacing Σ by an estimator $\hat{\Sigma}$.

The main objective of this paper is to apply the latter idea to classical estimation problems, not restricted to prediction. Although it can be applied to non-parametric models, our procedure is essentially designed for parametric or semi-parametric models, where the number k of available estimators is small compared to the sample size n and does not vary with n. The procedure works well in these situations because the estimation of Σ can be carried out efficiently by standard methods (e.g. plug-in or Monte-Carlo), and does not require the tuning of extra parameters. While it recovers some results of [11, 12, 20] and more recently [18] on estimator averaging for the mean in a Gaussian model, the method applies to a wide range of statistical models. It is implemented in Section 4 on four other examples. In the first one, θ represents the position of an unknown distribution, which can be estimated by both the empirical mean and median, as initially addressed by P. S. de Laplace in [19]. In the second example, θ is the two-dimensional parameter of a Weibull distribution, for which several competing estimators exist. In the third one, we consider a stochastic process, namely the Boolean model, that also depends on a two-dimensional parameter and we apply averaging to get a better estimate. The fourth example deals with estimation of a quantile from the combination of a non-parametric estimator and possibly misspecified parametric estimators.

An important contribution of our approach is to include the case where several parameters $\theta_1, \ldots, \theta_d$ have to be estimated, and a collection of estimators is available for each of them. In order to fully exploit the available information to estimate say θ_1, it may be profitable to average all estimators, including those designed for θ_j, $j \neq 1$. We show that a minimal requirement

3

is that the weights associated to the latter estimators sum to 0, while the weights associated to the estimators of θ_1 sum to one (additional constraints on the weights can also be added as discussed in Section 2.3). To our knowledge, estimator averaging including estimators of other parameters is a new idea. Our simulation study shows that it can produce conclusive results in some specific situations such as the Boolean model treated in the third example of Section 4.

From a theoretical point of view, we provide an upper bound on the deviation of the averaging estimator to the oracle. Our result is non-asymptotic and involves the error to the oracle for the actual event, in contrast with usual criteria based on expected loss functions. In particular, our result strongly differs from classical oracle inequalities derived in the literature on aggregation for non-parametric regression [4, 9, 17, 35] or density estimation [5, 6, 27, 34]. Moreover, we deduce that under mild assumptions, our averaging estimator behaves asymptotically as the oracle, generalizing the asymptotic optimality result proved by Hansen and Racine [14] in the frame of model averaging where Σ is estimated by jackknife. Our result applies in particular if $\sqrt{n}(\mathbf{T} - \theta)$ converges in quadratic mean to a Gaussian law and a consistent estimator of the asymptotic covariance matrix is available, though these conditions are far from being necessary. This situation makes it possible to construct an asymptotic confidence interval based on the averaging estimator, the length of which is necessarily smaller than all confidence intervals based on the initial estimators.

The remainder of the paper is organized as follows. The averaging procedure is detailed in Section 2, where we give some examples for the choice of the set of weights Λ and we detail some methods for the estimation of Σ. In Section 3 we prove a non-asymptotic bound on the error to the oracle and discuss the asymptotic optimality of the averaging estimator. Section 4 is devoted to some numerical applications where we show that the method performs almost always better than the best estimator in the initial collection $\mathbf{T}$ when the model is well-specified, and is quite robust to misspecification problems. Proofs of our results are postponed to the Appendix.

2. The averaging procedure

The method is different whether it is applied to one parameter or several. For ease of comprehension, we first present the averaging procedure for one

parameter, which follows the idea introduced in [2] for forecast averaging, though our choice of the set of weights Λ may be different. We then introduce a generalization of the procedure for averaging several parameters simultaneously. Finally, we discuss in Sections 2.3 and 2.4 the choice of Λ and the construction of $\hat{\Sigma}$.

2.1. Averaging for one parameter

Let $\mathbf{T} = (T_1, ..., T_k)^\top$ be a collection of estimators of a real parameter θ. We search for a decision rule that combines suitably the T_i's to provide a unique estimate of θ. A widely spread idea is to consider 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 $\hat{\theta}_\lambda$ is to compare it to the *oracle* $\hat{\theta}^*$, defined as the best linear combination $\hat{\theta}_\lambda$ obtained for a non-random vector $\lambda \in \Lambda$. Specifically, the oracle is the linear combination $\hat{\theta}^* = \lambda^{*\top} \mathbf{T}$ minimizing the mean squared error (MSE), i.e.

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

Of course, λ^* is unknown in practice and needs to be approximated by an estimator, say $\hat{\lambda}$.

The performance of the averaging procedure 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. In this purpose, a choice proposed in [2] and widely used in the averaging literature is to impose the condition $\lambda^\top \mathbf{1} = 1$ on the weights, where $\mathbf{1}$ denotes the unit vector $\mathbf{1} = (1, ..., 1)^\top$. We explain in Section 2.3 why this condition is minimal for the efficiency of the averaging procedure when $\theta \in \mathbb{R}$. Moreover, if one wants to impose additional constraints on the weights, such as positivity for instance, the method proposed in this paper allows one to consider as the constraint set Λ, any non-empty closed subset of

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

Some examples of constraint sets Λ are discussed in Section 2.3.

We assume that the initial estimators have finite order-two moments and $1, T_1, ..., T_k$ are linearly independent so that the Gram matrix

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

is well defined and non-singular. From the identity $\lambda^\top\mathbf{1} = 1$, we see that the optimal weight λ^* defining the oracle $\hat{\theta}^* = \lambda^{*\top}\mathbf{T}$ writes

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

Remark that the assumptions made on Λ ensure the existence of a minimizer. If Λ is convex, the solution is unique, otherwise we agree that λ^* refers to one of the minimizers. In the particular important example where $\Lambda = \Lambda_{\max}$, we get the explicit solution

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

considered for instance in [2], [10] or [14]. In practice, the MSE matrix Σ is unknown and has to be approximated by some estimator $\hat{\Sigma}$ to yield the averaging estimator $\hat{\theta} = \hat{\lambda}^\top\mathbf{T}$, where

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

Possible methods to construct $\hat{\Sigma}$ are discussed in Section 2.4. While it may seem paradoxical to shift our attention from θ to the less accessible Σ, the effectiveness of the averaging process can be explained by a lesser sensibility to the errors on $\hat{\Sigma}$. As a result, the averaging estimator improves on the original collection as soon as we are able to build $\hat{\Sigma}$ sufficiently close from the true value, without stronger requirement such as consistency. On the contrary, the chances of considerably deteriorating the estimation of θ are expected to be small due to the smoothing effect of averaging.

2.2. Averaging for several parameters

We now discuss a generalization of the method that deals with several parameters simultaneously. Let $\theta = (\theta_1, \ldots, \theta_d)^\top \in \mathbb{R}^d$ and assume we have access to a collection of estimators $\mathbf{T}_j$ for each component θ_j. For sake of generality we allow the collections $\mathbf{T}_1, \ldots, \mathbf{T}_d$ to have different sizes $k_1, \ldots, k_d$

with $k_j \geq 1$. So, let $\mathbf{T}_1 \in \mathbb{R}^{k_1}, \ldots, \mathbf{T}_d \in \mathbb{R}^{k_d}$ and set $\mathbf{T} = (\mathbf{T}_1^\top, \ldots, \mathbf{T}_d^\top)^\top \in \mathbb{R}^k$, with $k = \sum_{j=1}^d k_j \geq d$. We consider averaging estimators of θ of the form

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

where here, λ is a $k \times d$ matrix. In order to make the oracle more accessible, we impose some restrictions on the set of authorized values for λ. In this purpose, define the matrix

$$J = \begin{pmatrix} \mathbf{1}_{k_1} & 0 & \ldots & 0 \\ 0 & \mathbf{1}_{k_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \mathbf{1}_{k_d} \end{pmatrix} \in \mathbb{R}^{k \times d},$$

where $\mathbf{1}_{k_j}$ is the vector composed of k_j ones (we simply denote it $\mathbf{1}$ in the sequel to ease notation). We consider the maximal constraint set

$$\Lambda_{\max} = \{\lambda \in \mathbb{R}^{k \times d} : \lambda^\top J = I\}, \tag{2}$$

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

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

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

$$\underline{\lambda}_{j,\ell}^\top \mathbf{1} = \begin{cases} 0 & \text{if } \ell \neq j \\ 1 & \text{if } \ell = j. \end{cases} \tag{3}$$

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, which may improve the results. Finally, remark that if the collections $\mathbf{T}_j$ are uncorrelated, the two frameworks are identical.

From a technical point of view, the condition $\lambda^\top J = I$ is imposed to have the equality $\lambda^\top \mathbf{T} - \theta = \lambda^\top (\mathbf{T} - J\theta)$, which is used to derive the optimality result of Theorem 3.1 in Section 3.1. Letting $\|.\|$ denote the usual Euclidean norm on $\mathbb{R}^d$, the mean squared error then becomes, using the classical trick of switching trace and expectation,

$$\mathbb{E}\|\lambda^\top \mathbf{T} - \theta\|^2 = \mathbb{E}\big[\, \mathrm{tr}\,\big[(\mathbf{T} - J\theta)^\top \lambda \lambda^\top (\mathbf{T} - J\theta)\big]\big] = \mathrm{tr}(\lambda^\top \Sigma \lambda), \qquad (4)$$

where $\Sigma = \mathbb{E}\big[(\mathbf{T} - J\theta)(\mathbf{T} - J\theta)^\top\big] \in \mathbb{R}^{k \times k}$. Here again, we assume that Σ exists and is non-singular.

Ideally, one would want to minimize the matrix mean squared error $\mathbb{E}\big[(\lambda^\top \mathbf{T} - \theta)(\lambda^\top \mathbf{T} - \theta)^\top\big] = \lambda^\top \Sigma \lambda$, and not only its trace, for λ^* to satisfy the stronger property

$$\forall \lambda \in \Lambda, \quad \lambda^\top \Sigma \lambda - \lambda^{*\top} \Sigma \lambda^* \text{ is non-negative definite.} \qquad (5)$$

Notice however that comparing λ and λ^* according to this criterion is not always possible since it involves a partial order relation over the matrices and a solution to (5) might not exist. By considering its trace, we are guaranteed to reach an admissible solution, that is, a solution for which no other value is objectively better in this sense. In particular, minimizing $\lambda \mapsto \mathrm{tr}(\lambda^\top \Sigma \lambda)$ reaches the unique solution λ^* of (5) whenever one exists. This occurs for instance for $\Lambda = \Lambda_{\max}$, as we point out in Section 2.3.

The simultaneous averaging process for several parameters generalizes the procedure presented in Section 2.1. In fact, averaging 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

$$\lambda^* = \arg\min_{\lambda \in \Lambda}\ \mathbb{E}\|\lambda^\top \mathbf{T} - \theta\|^2 = \arg\min_{\lambda \in \Lambda}\ \mathrm{tr}(\lambda^\top \Sigma \lambda). \qquad (6)$$

Finally, assuming we have access to an estimator $\hat{\Sigma}$ of Σ, see Section 2.4, we define the averaging estimator as $\hat{\theta} = \hat{\lambda}^\top \mathbf{T}$ where

$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda}\ \mathrm{tr}(\lambda^\top \hat{\Sigma} \lambda). \qquad (7)$$

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

2.3. Choice of the constraint set

The constraint set plays a crucial part in the averaging procedure. As discussed in the previous sections, an appropriate choice of Λ must take into account both the accuracy of the oracle and the ability to estimate λ^*. Writing the estimation error as

$$\hat{\theta} - \theta = \hat{\theta}^* - \theta + (\hat{\lambda} - \lambda^*)^\top \mathbf{T}, \tag{8}$$

a good rule of thumb is to choose a set Λ as large as possible, but for which the residual term $(\hat{\lambda} - \lambda^*)^\top \mathbf{T}$ can be made negligible compared to the error of the oracle $\hat{\theta}^* - \theta$. Without restrictions on λ, Equation (8) suggests that we would have to estimate the optimal combination λ^* more efficiently than we can estimate θ. This condition can be reasonably expected for high-dimensional parameters θ, e.g. for non-parametric regression or density estimation, and linear aggregation has indeed been shown to be particularly well adapted to these frameworks, see for instance [4, 5, 9, 24, 27, 31, 34]. Nevertheless, hoping for $\hat{\lambda} - \lambda^*$ to be negligible compared to $\hat{\theta}^* - \theta$ seems rather unrealistic when θ is vector-valued, especially given that the optimal combination needs to be estimated for every component θ_j. Even in the simple case $d = 1$, the oracle obtained over $\Lambda = \mathbb{R}^k$ can be written as

$$\hat{\theta}^* = \theta\, \mathbb{E}(\mathbf{T}^\top)\big[\mathbb{E}(\mathbf{T}\mathbf{T}^\top)\big]^{-1}\mathbf{T},$$

where the term $\mathbb{E}(\mathbf{T}^\top)\big[\mathbb{E}(\mathbf{T}\mathbf{T}^\top)\big]^{-1}\mathbf{T}$ appears as an inadequate estimate of 1. One can argue that aiming for the oracle in this case may divert from the primary objective to estimate θ.

Besides the reasons to rule out linear averaging in this framework, one can provide some additional arguments in favor of the affine constraint $\lambda^\top \mathrm{J} = \mathrm{I}$ as a minimal requirement on λ. For instance, remark that if $\hat{\lambda}$ and λ^* satisfy this condition, the error term can be written as

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

where both $(\hat{\lambda} - \lambda^*)$ and $(\mathbf{T} - \mathrm{J}\,\theta)$ can contribute to make the error term negligible. Moreover, this restriction leads to an expression of the mean squared error matrix that only involves Σ, as pointed out in (4). Thus, building an approximation $\hat{\lambda}$ of the optimal combination only requires to estimate Σ (this step is discussed in details in Section 2.4). Finally, the error bound proved

in Theorem 3.1 only holds if Λ is a subset of $\Lambda_{\max}$ which tends to confirm the necessity of this constraint in this framework.

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

$$\lambda^*_{\max} = \Sigma^{-1} J (J^\top \Sigma^{-1} J)^{-1}, \tag{9}$$

generalizing the formula given in Section 2.1. Its mean squared error can be calculated directly

$$\mathbb{E}\big[(\hat{\theta}^*_{\max} - \theta)(\hat{\theta}^*_{\max} - \theta)^\top\big] = (J^\top \Sigma^{-1} J)^{-1}.$$

One verifies that $\lambda^*_{\max}$ is a minimizer by

$$\lambda^\top \Sigma \lambda - (J^\top \Sigma^{-1} J)^{-1} = \lambda^\top \Sigma \lambda - \lambda^{*\top}_{\max} \Sigma \lambda^*_{\max} = (\lambda - \lambda^*_{\max})^\top \Sigma (\lambda - \lambda^*_{\max}) \tag{10}$$

which holds for all $\lambda \in \Lambda_{\max}$ due to the condition $\lambda^\top J = I$, and where the last matrix is non-negative definite. Moreover, (10) shows that the oracle is not only the solution of our optimization problem (6), but it also fulfills the stronger requirement (5). 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 (9), the averaging estimator $\hat{\theta}_{\max}$ obtained by replacing Σ by its estimation $\hat{\Sigma}$ has also a closed expression which makes it easily computable, namely

$$\hat{\theta}_{\max} = (J^\top \hat{\Sigma}^{-1} J)^{-1} J^\top \hat{\Sigma}^{-1} \mathbf{T}. \tag{11}$$

As mentioned earlier, the maximal constraint set allows one to use the information contained in external collections to estimate each parameter. This requires to estimate the whole MSE matrix, including the cross correlations between different collections $\mathbf{T}_i$. While this can produce surprisingly good results in some cases (see Section 4), it may deteriorate the estimator if the external collections do not contain significant additional information on the parameter.

10

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

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

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 initial estimators 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 combined 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

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

In order to build the averaging estimator, it is sufficient to plug an estimate of Σ_j for $j = 1, ..., d$ in the above expression, which makes it easily computable.

Another well-used framework is convex averaging, which corresponds to the choice

$$\Lambda = \{\lambda \in \Lambda_{\max} : \ \lambda_{i,j} \geq 0, \ i = 1, ..., k, \ j = 1, ..., d\}. \tag{12}$$

Observe that the positivity restriction combined with the condition $\lambda^\top \mathrm{J} = \mathrm{I}$ results in λ having its support included in that of J, making convex averaging a particular case of component-wise averaging. This means that each component of θ can be dealt with separately, thus reducing to the case $d = 1$.

Convex combination of estimators is a natural choice that has been widely studied in the literature. An advantage 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 maximal case $\Lambda = \Lambda_{\max}$. The use of convex combinations is also particularly convenient to preserve some properties of the initial estimators, such as positivity or boundedness. Moreover, imposing non-negativity often leads to sparse solutions.

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 averaging procedure obtained with the estimators $\mathbf{T}_{\hat{m}} := (T_i)_{i\in\hat{m}}$ 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 averaging solution:

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

where $\hat{m}$ is the admissible support with minimal mean squared error, 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 convex averaging 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 [25].

The issue of selecting the best estimator in the collection can also be viewed as a particular case of averaging. Performing estimator selection based on an estimation of the MSE is an approach used in numerous practical situations. In the univariate case, estimator selection corresponds to the constraint set

$$\Lambda = \left\{ (1, 0, ..., 0)^\top, (0, 1, 0, ..., 0)^\top, ..., (0, ..., 0, 1)^\top \right\}.$$

The main advantage of this framework is that it only requires to estimate the mean squared error of each estimator T_j, i.e., the diagonal entries of Σ. Applying the procedure in this case simply consists in selecting the T_j with minimal estimated mean squared error. While the oracle is easier to approach in this framework, estimator selection may suffer from the poor efficiency of the oracle, compared to less constrained situations. For this reason, we do not recommend to settle for estimator selection if one can provide a reasonable estimation of Σ, as the oracle under larger constraint

sets can be much more efficient while remaining reachable. This observation is confirmed by the numerical study of Section 4 where the average estimator appears to be better than the best estimator in the initial collection in most cases. Finally, remark that the theoretical performance of estimator selection is not covered by Theorem 3.1, due to the non-convexity of the constraint set.

2.4. Estimation of the MSE matrix

The accuracy of $\hat{\Sigma}$ is clearly a main factor to the performance of the averaging method. There exist several methods to construct $\hat{\Sigma}$, whether the model is parametric or not. In all cases, the estimation of Σ can be carried out from the same data as those used to produce the initial estimators T_i and no sample splitting is needed.

In a fully specified parametric model, the MSE matrix Σ can be estimated by plugging an initial estimate of θ. Precisely, assuming that the MSE matrix can be expressed as the image of θ through a known continuous map $\Sigma(.) : \mathbb{R}^d \to \mathbb{R}^{k \times k}$, one can choose $\hat{\Sigma} = \Sigma(\hat{\theta}_0)$, where $\hat{\theta}_0$ is a consistent estimate of θ. A suitable choice for $\hat{\theta}_0$ is to take one of the initial estimators if it is known to be consistent, or the average $\frac{1}{k} \sum_{i=1}^{k} T_i$ provided all initial estimators are consistent. If the map $\Sigma(.)$ is not explicitly known, $\Sigma(\hat{\theta}_0)$ may be approximated by Monte-Carlo simulations of the model using the estimated parameter $\hat{\theta}_0$, a procedure sometimes called parametric bootstrap. This method is illustrated in our examples in Sections 4.2 and 4.3 and reveals to be efficient whenever the model is well specified. Remark that in this parametric situation, the averaging procedure does not require any information other than the initial collection $\mathbf{T}$.

In some cases, Σ may also depend on a nuisance parameter η. In this situation, $\hat{\Sigma}$ can be built similarly by plugging or Monte-Carlo, provided η can be estimated from the observations. This situation requires the sample $X_1, ..., X_n$ used to built the initial estimators T_i to be available to the user.

In a semi and non-parametric setting, a parametric closed-form expression for Σ may be available asymptotically, i.e. when the sample size on which $\mathbf{T}$ is built tends to infinity, and the above plugging method then becomes possible, see also the discussion in Section 3.2. Alternatively, Σ can be estimated by standard bootstrap if no extra information is available. These two methods are implemented in the first example of Section 4.

3. Theoretical results

3.1. Non-asymptotic error bound

The performance of the averaging estimator relies on the accuracy of $\hat{\Sigma}$, but more specifically, on the ability to evaluate $\operatorname{tr}(\lambda^\top \Sigma \lambda)$ as λ ranges over Λ. As a result, it is not crucial that $\hat{\Sigma}$ be a perfect estimate of Σ as long as the error $|\operatorname{tr}(\lambda^\top \hat{\Sigma} \lambda) - \operatorname{tr}(\lambda^\top \Sigma \lambda)|$ is small for $\lambda \in \Lambda$, which can hopefully be achieved by a suitable choice of the constraint set. 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 $\operatorname{tr}(\lambda^\top A \lambda)/\operatorname{tr}(\lambda^\top B \lambda)$ over Λ,

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

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 3.1. *Let Λ be a non-empty closed convex subset of $\Lambda_{\max}$ with associated oracle $\hat{\theta}^*$ defined through (6), and $\hat{\Sigma}$ a symmetric positive definite $k \times k$ matrix. The averaging estimator $\hat{\theta} = \hat{\lambda}^\top \mathbf{T}$ defined through (7) satisfies*

$$\|\hat{\theta} - \hat{\theta}^*\|^2 \le \tilde{\delta}_\Lambda(\hat{\Sigma}, \Sigma) \, \|\mathbf{S}\|^2 \, \mathbb{E}\|\hat{\theta}^* - \theta\|^2 \tag{13}$$

where $\tilde{\delta}_\Lambda(\hat{\Sigma}, \Sigma) = 2\delta_\Lambda(\hat{\Sigma}, \Sigma) + \delta_\Lambda(\hat{\Sigma}, \Sigma)^2$ and $\mathbf{S} = \Sigma^{-\frac{1}{2}}(\mathbf{T} - \mathrm{J}\,\theta)$.

In this theorem, we provide an upper bound on the distance of the averaging estimator 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). Moreover, we point out that the upper bound applies to the actual error to the oracle (for the current event ω), contrary to classical oracle inequalities which generally involve an expected loss of some kind. Nonetheless, the following corollary compares the mean squared errors of the averaging estimator and the oracle.

Corollary 3.2. *Under the assumptions of Theorem 3.1, for all $\epsilon > 0$,*

$$\mathbb{E}\|\hat{\theta} - \theta\|^2 \le (1 + \epsilon) \, \mathbb{E}\|\hat{\theta}^* - \theta\|^2 \left[1 + \epsilon^{-1}\mathbb{E}\big(\tilde{\delta}_\Lambda(\hat{\Sigma}, \Sigma) \, \|\mathbf{S}\|^2\big)\right]. \tag{14}$$

14

Some comments on Theorem 3.1 and its corollary are in order. Recall that a main concern when selecting the constraint set Λ is to be able to make the distance to the oracle $\|\hat{\theta} - \hat{\theta}^*\|$ negligible compared to the error of the oracle $\|\hat{\theta}^* - \theta\|$. This objective is achieved whenever the term $\tilde{\delta}_\Lambda(\hat{\Sigma}, \Sigma)\|\mathbf{S}\|^2$ in (13) can be shown to be negligible. Corollary 3.2 makes this remark more specific in the $\mathbb{L}^2$ sense: by choosing ϵ tending to 0 not too fast, the mean squared errors of the averaging estimator and the oracle are seen to be asymptotically equivalent whenever $\mathbb{E}\big(\tilde{\delta}_\Lambda(\hat{\Sigma}, \Sigma)\|\mathbf{S}\|^2\big)$ tends to 0. Section 3.2 details more consequences of Theorem 3.1 from an asymptotic point of view.

The factor $\tilde{\delta}_\Lambda(\hat{\Sigma}, \Sigma)$, which only involves the divergence $\delta_\Lambda(\hat{\Sigma}, \Sigma)$, emphasizes the influence of the constraint set Λ and the accuracy of $\hat{\Sigma}$ to estimate Σ. It appears 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, in order to get a small value of $\delta_\Lambda(\hat{\Sigma}, \Sigma)$. Actually, as stated in Lemma 5.1 of the Appendix, we have

$$\delta_\Lambda(\hat{\Sigma}, \Sigma) \leq \||\hat{\Sigma}\Sigma^{-1} - \Sigma\hat{\Sigma}^{-1}\||, \tag{15}$$

where $\||.\||$ denotes the operator norm. This shows that the efficiency of the averaging procedure can be measured by how well $\hat{\Sigma}\Sigma^{-1}$ approximates the identity matrix. It is nevertheless difficult to study the behavior of $\delta_\Lambda(\hat{\Sigma}, \Sigma)$ or even $\hat{\Sigma}\Sigma^{-1}$ without more information on the statistical model from which $\mathbf{T}$ is computed. In particular, we are not able to derive oracle-like inequalities involving minimax rates of convergence without further specification. For instance, $\delta_\Lambda(\hat{\Sigma}, \Sigma)$ can be expected to converge in probability to zero at a typical rate of $\sqrt{n}$ in a well specified parametric sampling model, while similar properties should be seldom verified in semi or non-parametric models.

The term $\|\mathbf{S}\|^2$ in (13) shows the price to pay for averaging too many estimators, in view of the equality

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

This suggests that the averaging procedure might be improved by performing a preliminary selection to keep only the relevant estimators. Moreover, including too many estimators to the initial collection increases the possibilities of strong correlations, which may lead to a near singular matrix Σ and result in amplified errors when computing $\hat{\Sigma}^{-1}$ and a larger value of $\delta_\Lambda(\hat{\Sigma}, \Sigma)$.

15

3.2. Asymptotic study

The properties of the averaging estimator established in Theorem 3.1 do not rely on any assumption on the construction of $\mathbf{T}$ or $\hat{\Sigma}$. In this section, we investigate the asymptotic properties of the averaging estimator 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. From now on, 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.

In practice, we expect the oracle $\hat{\theta}_n^*$ to satisfy good properties such as consistency and asymptotic normality. Theorem 3.1 suggests that $\hat{\theta}_n$ should inherit these asymptotic properties if Σ_n can be sufficiently well estimated. Remark that if the initial estimators T_i are consistent in quadratic mean, Σ_n converges to the null matrix as $n \to \infty$. In this case, providing an estimator $\hat{\Sigma}_n$ such that $\hat{\Sigma}_n - \Sigma_n \xrightarrow{p} 0$ is clearly not sufficient for $\hat{\theta}_n$ to achieve the asymptotic performance of the oracle (here p stands for the convergence in probability while d is used for distribution). On the contrary, requiring that $\hat{\Sigma}_n^{-1} - \Sigma_n^{-1} \xrightarrow{p} 0$ is unnecessarily too strong and would be nearly impossible to achieve. In fact, we show in Proposition 3.3 below that the condition

$$\hat{\Sigma}_n \Sigma_n^{-1} \xrightarrow{p} \mathrm{I} \tag{16}$$

appears as a simple compromise, both sufficient for asymptotic optimality and reasonable enough to be verified in numerous situations with regular estimators of Σ_n. Typically, if $\mathbf{T}_n$ is asymptotically Gaussian with

$$\sqrt{n}(\mathbf{T}_n - \mathrm{J}\,\theta) \xrightarrow{\mathbb{L}^2} \mathcal{N}(0, W) \tag{17}$$

for a non-singular matrix W, providing a consistent estimator $\hat{W}_n$ of W is sufficient to achieve (16), taking $\hat{\Sigma}_n = n^{-1}\hat{W}_n$. The situation becomes particularly convenient if the limit matrix W follows a known parametric expression $W = W(\eta, \theta)$, with η a nuisance parameter (see the first example in Section 4). 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 (16). If the map $W(.,.)$ is unknown, parametric bootstrap based on the initial estimates $\hat{\eta}_0$ and $\hat{\theta}_0$ may be used and the same theoretical justifications apply. Observe that knowing the rate $\sqrt{n}$ in this case is not necessary as it simplifies in the expression of $\hat{\theta}_n$. In fact, a different rate of convergence, even unknown, would lead to the exact same result. In this case, the asymptotic normality

16

can make it possible to construct asymptotic confidence intervals of minimal length for the parameter, as shown in Proposition 3.3 below.

More generally, if one can provide an asymptotic expression for Σ_n of the form

$$\Sigma_n = a_n W + o(a_n), \tag{18}$$

for some vanishing sequence a_n, building a consistent estimator of W is sufficient to achieve (16). Here again, the rate of convergence needs not be known.

Finally, if there are different rates of convergence within the collection $\mathbf{T}_n$, the condition (16) can still be verified if the normed eigenvectors of Σ_n converge as $n \to \infty$. Precisely, if 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

$$\lim_{n \to \infty} A_n P \Sigma_n P^\top = D, \tag{19}$$

for some non-singular diagonal matrix D, producing consistent estimators $\hat{P}_n$ and $\hat{D}_n$ of P and D respectively enables to verify (16) by $\hat{\Sigma}_n = \hat{P}_n^\top A_n^{-1} \hat{D}_n \hat{P}_n$. Here, the limit of the normed eigenvectors of Σ_n are given by the rows of P and the estimator $\hat{\Sigma}_n$ is constructed from the asymptotic expansion of Σ_n. This situations allows one to have different rates of convergence within the collection $\mathbf{T}_n$ but also covers the case where all constant combinations $\lambda^\top \mathbf{T}_n$ converge to θ at the same rate.

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

$$\alpha_{n,j} := \mathbb{E}\|\hat{\theta}_{n,j}^* - \theta_j\|^2 = \underline{\lambda}_{n,j}^{*\top} \Sigma_n \underline{\lambda}_{n,j}^*,$$

where we recall that $\underline{\lambda}_{n,j}^*$ is the j-th column of λ_n^*. Similarly, let $\hat{\alpha}_{n,j} = \underline{\hat{\lambda}}_{n,j}^\top \hat{\Sigma}_n \underline{\hat{\lambda}}_{n,j}$. 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 constraint set $\Lambda \subset \mathbb{R}^{k \times d}$, we denote by $\Lambda_j = \{\underline{\lambda}_j : \lambda \in \Lambda\} \subset \mathbb{R}^k$ its marginal set. We say that Λ is a

17

cylinder if $\Lambda = \{\lambda : \underline{\lambda}_1 \in \Lambda_1, ..., \underline{\lambda}_d \in \Lambda_d\}$, i.e., if Λ is the Cartesian product of its marginal sets Λ_j. We point out that choosing a constraint set Λ that satisfies this property is not restrictive in general, as it simply states that each vector of weights $\underline{\lambda}_j$ used to estimate θ_j can be computed independently of the others. In particular, all the constraint sets discussed in Section 2.3 are cylinders.

Proposition 3.3. *If* (16) *holds, then*

$$\|\hat{\theta}_n - \theta\|^2 = \|\hat{\theta}_n^* - \theta\|^2 + o_p(\alpha_n). \tag{20}$$

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 $j = 1, ..., d$, where $\mathcal{Z}$ is a random variable, then

$$\hat{\alpha}_{n,j}^{-\frac{1}{2}}(\hat{\theta}_{n,j} - \theta_j) \xrightarrow{d} \mathcal{Z}. \tag{21}$$

This proposition establishes that building an estimate $\hat{\Sigma}_n$ for which (16) holds ensures that the error of the average $\hat{\theta}_n$ is asymptotically comparable to that of the oracle, up to $o_p(\alpha_n)$. In addition, under the mild assumption that Λ is a cylinder, it is possible to provide an asymptotic confidence interval for each θ_j when the limit distribution $\mathcal{Z}$ is known. In most applications, this distribution is a standard Gaussian but other more complicated situations may be handled as discussed in the following remark. From (7), we know these confidence intervals are of minimal length amongst all possible confidence intervals based on a linear combination of $\mathbf{T}_n$. Note that no extra estimation is needed to compute $\hat{\alpha}_{n,j}$, as it is entirely determined by $\hat{\lambda}_n$ and $\hat{\Sigma}_n$.

Remark 3.4. *As noticed above, $\mathcal{Z}$ can be expected to follow a standard Gaussian distribution in numerous situations. For instance, this happens if* (17) *holds, with a possibly different normalization than $\sqrt{n}$, or under the condition* (19) *provided $\mathbf{T}_n$ is asymptotically normal with vanishing bias. We refer to Section 4 for actual examples. However, in presence of misspecified initial estimators, the distribution of $\mathcal{Z}$ can be different. In [15], the authors study a specific local misspecification framework where the asymptotic law of the oracle is not Gaussian, neither centered (see their Theorem 4.1 and Section 5.4 for an averaging procedure based on the mean squared error). In these cases, the quantiles of $\mathcal{Z}$ may depend on some unknown extra parameters that have to be estimated to provide asymptotic confidence intervals.*

In such a misspecified framework though, the challenge is to estimate accurately Σ. Conditions implying (16) in a misspecified framework are difficult to establish and are beyond the scope of the present paper.

In Proposition 3.3, the asymptotic optimality of $\hat{\theta}_n$ is stated in probability. In view of Corollary 3.2, it is not difficult to strengthen this result to get the asymptotic optimality in quadratic loss, i.e.

$$\mathbb{E}\|\hat{\theta}_n - \theta\|^2 = \mathbb{E}\|\hat{\theta}_n^* - \theta\|^2(1 + o(1)), \tag{22}$$

provided additional assumptions hold. If for instance $\hat{\Sigma}_n$ and $\mathbf{T}_n$ are computed from independent samples (e.g. by sample splitting), then (22) holds as soon as $\mathbb{E}[\tilde{\delta}_\Lambda(\hat{\Sigma}_n, \Sigma_n)]$ tends to 0, which occurs whenever both $\hat{\Sigma}_n\Sigma_n^{-1}$ and $\Sigma_n\hat{\Sigma}_n^{-1}$ converge to the identity matrix in $\mathbb{L}^2$. We emphasize however that the use of sample splitting may reduce the performance of the oracle, as it would be computed from fewer data. One can argue that this is a high price to pay to obtain asymptotic optimality in $\mathbb{L}^2$ and is not to be recommended in this framework. Asymptotic optimality in $\mathbb{L}^2$ can also be achieved if one can show there exists $p > 1$ such that

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

which is a direct consequence of Corollary 3.2 by applying Hölder's inequality. These conditions ensure the asymptotic optimality in $\mathbb{L}^2$ of the averaging estimator without sample splitting, but they remain nonetheless extremely difficult to check in practice.

4. Applications

This section gathers four examples of models where we apply our averaging procedure. Depending on the situation, we combine parametric, semiparametric or non-parametric estimators. The examples in Section 4.2 and 4.3 involve a bivariate parameter which allows us to assess the multivariate procedure introduced in Section 2.2. For the estimation of the MSE matrix Σ_n, we use either parametric bootstrap, (non-parametric) bootstrap or the asymptotic expression Σ_∞ when it has a parametric form. From a theoretical point of view, all our examples satisfy the main condition (16) implying the asymptotic optimality of the average estimator in the sense of Proposition 3.3, since they verify either (17) or (19), except when non-parametric

bootstrap or misspecified models are used. The last two settings are common situations, so we chose to include them in our simulation study, however they are out of the scope of the theoretical study of this paper and will be the subject of future investigations.

4.1. 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 $x_1, \ldots, x_n$, simple solutions are 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.

As noticed in Section 1, the idea of combining the mean and the median to construct a better estimator goes back to Pierre Simon de Laplace [19]. P. S. de Laplace obtains the expression of the weights in $\Lambda_{\max}$ that ensure a minimal asymptotic variance for the averaging 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 averaging of the mean and the median over $\Lambda_{\max}$. We have two initial estimators $T_1 = \overline{x}_n$, $T_2 = x_{(n/2)}$ and the averaging estimator is given by (11) where J is just in this case the vector $(1, 1)^\top$. We assume that the n realisations are independent and we propose two ways to estimate the MSE matrix Σ_n. The first is based on the asymptotic equivalent $\Sigma_n = n^{-1}W + o(n^{-1})$, obtained in P. S. de Laplace's work and recalled in [29], 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 estimated from an initial consistent 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 [28]). With this estimation of Σ_n, we get the following averaging estimator:

$$\hat{\theta}_{AV} = \frac{p_1}{p_1 + p_2} \overline{x}_n + \frac{p_2}{p_1 + p_2} x_{(n/2)} \tag{23}$$

20

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 estimator corresponds to an empirical version of the best combination obtained by P. S. de Laplace. The second way to estimate Σ_n is by 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 leads to another averaging estimator, denoted by $\hat{\theta}_{AVB}$.

Let us note that the first procedure verifies (17). For this reason, $\hat{\theta}_{AV}$ is asymptotically as efficient as the oracle, provided $\hat{\theta}_0$ is consistent. Moreover, the oracle is asymptotically Gaussian and an asymptotic confidence interval for θ can be provided without further estimation, see Section 3.2 and particularly Remark 3.4. For the second procedure, theory is lacking to study the behaviour of $\tilde{\delta}_\Lambda$ in (13) when $\hat{\Sigma}_n$ is estimated by bootstrap, so no consistency can be claimed at this point.

Table 1 summarizes the estimated MSE of $\overline{x}_n$, $x_{(n/2)}$, $\hat{\theta}_{AV}$ and $\hat{\theta}_{AVB}$, 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 (23), 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 $\hat{\theta}_{AVB}$ is $B = 1000$.

While the best estimator between $\overline{x}_n$ and $x_{(n/2)}$ depends on the underlying distribution, the averaging estimators $\hat{\theta}_{AV}$ and $\hat{\theta}_{AVB}$ 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 the averaging estimator cannot improve on $\overline{x}_n$. However the MSE of $\hat{\theta}_{AV}$ and $\hat{\theta}_{AVB}$ are 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 averaging estimators are robust in this case, as they manage to highly favor $x_{(n/2)}$.

21

	$n=30$				$n=50$				$n=100$			
	MEAN	MED	AV	AVB	MEAN	MED	AV	AVB	MEAN	MED	AV	AVB
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), $\hat{\theta}_{AV}$ (AV) and $\hat{\theta}_{AVB}$ (AVB) 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.

Choosing the median $x_{(n/2)}$ as the initial estimator $\hat{\theta}_0$ is of course crucial in this case.

Since the condition (17) holds, the oracle is asymptotically normal and Proposition 3.3 yields an asymptotic confidence interval without any further estimation. By construction, the length of these intervals is smaller than the length of any similar confidence interval based on $\overline{x}_n$ or $x_{(n/2)}$. Further, the empirical rate of coverage of these intervals is reported in Table 2 for the previous simulations, and turns out to be close to the nominal level 95%.

	$n=30$		$n=50$		$n=100$	
	AV	AVB	AV	AVB	AV	AVB
Cauchy	98.08	96.18	98.21	95.55	97.75	95.22
St(4)	93.59	91.45	94.38	92.71	94.71	92.55
St(7)	93.34	91.25	93.93	91.77	94.27	92.73
Logistic	92.48	90.33	93.96	92.05	93.91	92.21
Gauss	92.97	91.13	93.54	91.94	94.09	92.59
Mix	93.19	93.83	94.77	95.97	94.94	97.91

Table 2: Empirical rate of coverage (in %) of the asymptotic 95% confidence intervals based on $\hat{\theta}_{AV}$ and $\hat{\theta}_{AVB}$ in the estimation of the position of a symmetric distribution, deduced from the same simulations as in Table 1.

Finally, while $\hat{\theta}_{AVB}$ suffers from a lack of theoretical justification, it be-

haves pretty much like $\hat{\theta}_{AV}$, except for the mixture distribution where it performs slightly better than $\hat{\theta}_{AV}$. This may be explained by the fact that $\hat{\theta}_{AV}$ is more sensitive than $\hat{\theta}_{AVB}$ 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, $\hat{\theta}_{AV}$ demonstrates very good performance in this case, for the sample sizes considered in Tables 1 and 2.

4.2. Estimating the parameters of a Weibull distribution

The Weibull distribution with shape parameter $\beta > 0$ and scale parameter $\eta > 0$ has the density

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

Based on a sample of n independent realisations, many estimators of β and η are available (see [16]). We consider three standard methods. The maximum likelihood estimator (ML) is 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. Finally, 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 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, none is uniformly better than the others, see Figure 1 for an illustration.

Let us now consider the averaging of these estimators. In the setting of the previous sections, we have $d = 2$ parameters to estimate and $k_1 = 3$, $k_2 = 3$ initial estimators of each are available. The averaging 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 averaging over $\Lambda_{\max}$ of the 6 initial estimators 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 average 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 averaging 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 averaging 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 (3). This means that we actually implement a component-wise averaging for β. Third, the averaging 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 averaging estimator over $\Lambda_{\max}$ is given by (11), denoted by $(\hat{\beta}_{AV}, \hat{\eta}_{AV})^\top$. The matrix Σ is estimated by Monte Carlo simulations: 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 empirical 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 $(\hat{\beta}_{AV}, \hat{\eta}_{AV})^\top$ is asymptotically as efficient as the oracle, as explained in Section 3.2.

Table 3 gives the MSE, estimated from 10^4 replications, of each estimator of β, for $n = 10, 20, 50$, and for $\beta = 0.5, 1, 2, 3$, $\eta = 10$, where for each replication $B = 1000$. The averaging estimator has by far the lowest MSE,

24

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 4 shows the MSE for $\hat{\eta}_{ML}$ and $\hat{\eta}_{AV}$ where only estimators of β were used in attempt to improve $\hat{\eta}_{ML}$ by averaging. 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.

Finally, the empirical rate of coverage of the asymptotic confidence intervals based on $\hat{\beta}_{AV}$ and $\hat{\eta}_{AV}$ is given in Table 5, showing that it is not far from the nominal level 95%, even for the small sample sizes considered in this simulation. On the other hand, the length of these intervals are by construction smaller than the length of similar confidence intervals based on the initial estimators.

	$n = 10$				$n = 20$				$n = 50$			
	ML	MM	OLS	AV	ML	MM	OLS	AV	ML	MM	OLS	AV
$\beta = 0.5$	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)	(0.2)	(0.2)	(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 3: Monte Carlo estimation of the MSE of $\hat{\beta}_{ML}$, $\hat{\beta}_{MM}$, $\hat{\beta}_{OLS}$ and $\hat{\beta}_{AV}$, 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	AV	ML	AV	ML	AV
$\beta = 0.5$	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 4: Monte Carlo estimation of the MSE of $\hat{\eta}_{ML}$ and $\hat{\eta}_{AV}$, 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.
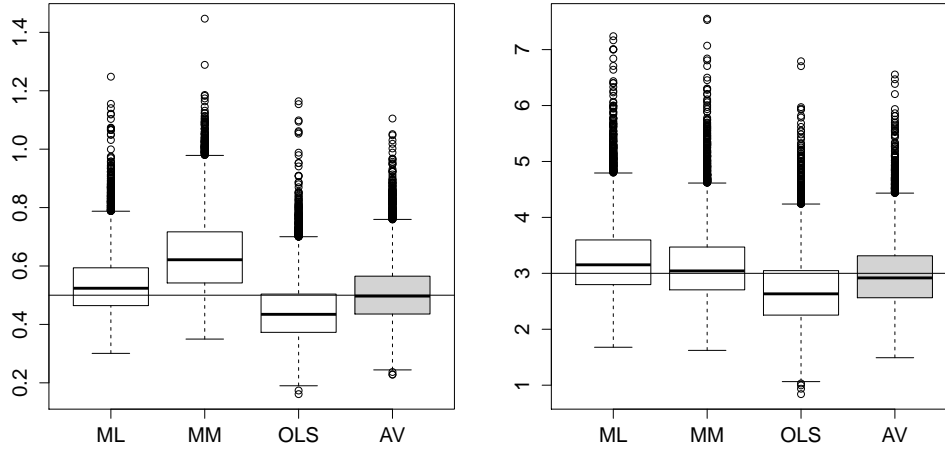

Figure 1: Repartition of $\hat{\beta}_{ML}$, $\hat{\beta}_{MM}$, $\hat{\beta}_{OLS}$ and $\hat{\beta}_{AV}$ (from left to right) based on 10^4 replications of a sample of size $n = 20$ from a Weibull distribution with $\beta = 0.5$ (left), $\beta = 3$ (right) and $\eta = 10$.

	$n = 10$		$n = 20$		$n = 50$	
	$\hat{\beta}_{AV}$	$\hat{\eta}_{AV}$	$\hat{\beta}_{AV}$	$\hat{\eta}_{AV}$	$\hat{\beta}_{AV}$	$\hat{\eta}_{AV}$
$\beta = 0.5$	89.84	87.48	93.43	90.01	95.41	93.07
$\beta = 1$	87.25	89.24	90.98	91.61	93.81	93.62
$\beta = 2$	89.96	91.36	91.77	93.39	93.09	94.20
$\beta = 3$	92.19	92.38	92.86	93.83	94.25	94.77

Table 5: Empirical rate of coverage (in %) of the asymptotic 95% confidence intervals based on $\hat{\beta}_{AV}$ and $\hat{\eta}_{AV}$ for the parameters of a Weibull distribution, deduced from the same simulations as in Tables 3 and 4.

4.3. Estimation in a Boolean model

The Boolean model is the main model of random sets used in spatial statistics and stochastic geometry, see [8]. It is a germ-grain model where, in the planar and stationary case, the germs come from a homogeneous Poisson point process on $\mathbb{R}^2$ with intensity ρ and the grains are independent random discs, the radii of which are distributed according to a probability law μ. Figure 2 contains four realisations of a Boolean model on $[0,1]^2$ where $\rho = 25, 50, 100, 150$ respectively and the law of the radii μ is the uniform distribution over $[0, 0.1]$. We assume in the following that μ is the beta distribution over $[0, 0.1]$ with parameter $(1, \alpha)$, $\alpha > 0$, denoted by $B(1, \alpha)$, i.e. μ has density $10\alpha (1 - 10x)^{\alpha-1}$ on $[0, 0.1]$. The simulations of Figure 2 correspond to $\alpha = 1$.
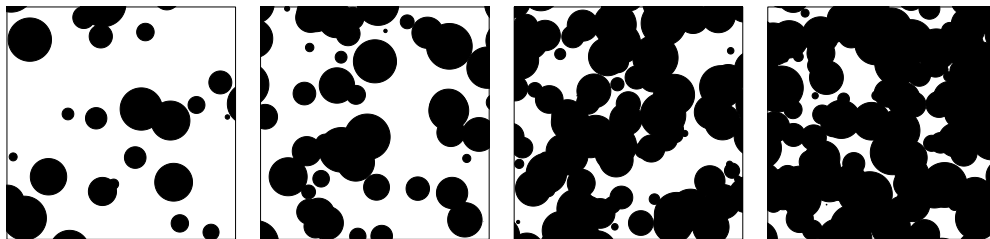

Figure 2: Samples from a Boolean model on $[0,1]^2$ with intensity, from left to right, $\rho = 25, 50, 100, 150$ and law of radii $B(1, \alpha)$ where $\alpha = 1$.

The estimation of parameters ρ and α from the observation of random sets as in Figure 2 is challenging, since the individual grains cannot be identified and likelihood-based inference is impossible. The standard method of

27

inference, see [22], is based on the following equations proved in [33]. They relate the expected area per unit area $\mathcal{A}$ and the expected perimeter per unit area $\mathcal{P}$ of the random set to the intensity ρ and the two first moments of μ, namely

$$\mathcal{A} = 1 - \exp(-\pi\rho E_\mu(R^2)), \qquad \mathcal{P} = 2\pi\rho E_\mu(R)\exp(-\pi\rho E_\mu(R^2)),$$

where R denotes a random variable with distribution μ. Developing $E_\mu(R)$ and $E_\mu(R^2)$ in terms of α, we obtain the following estimates of α and ρ :

$$\hat{\alpha}_1 = \frac{\mathcal{P}_{obs}}{10(\mathcal{A}_{obs}-1)\log(1-\mathcal{A}_{obs})} - 2, \qquad \hat{\rho}_1 = \frac{5\,(\hat{\alpha}_1+1)\mathcal{P}_{obs}}{\pi(1-\mathcal{A}_{obs})},$$

where $\mathcal{A}_{obs}$ and $\mathcal{P}_{obs}$ denote the observed area and perimeter per unit area of the set.

An alternative procedure to estimate the intensity ρ is based on the number of tangent points to the random set in a given direction. Let u be a vector in $\mathbb{R}^2$. We denote by $N(u)$ the number of tangent points to the random set such that the associated tangent line is orthogonal to u and the boundary of the set is convex in direction u. Considering k distinct vectors $u_1, \ldots, u_k$, an estimator of ρ, studied in [21], is

$$\hat{\rho}_2 = \frac{\frac{1}{k}\sum_{i=1}^{k} N(u_i)}{|W|(1-\mathcal{A}_{obs})},$$

where $|W|$ denotes the area of the observation window. Although this estimator is consistent and asymptotically normal for $k = 1$, it becomes more efficient as k increases, see [21]. In the following, we consider $k = 100$ and the directions of $u_1, \ldots, u_k$ are randomly drawn from an uniform distribution over $[0, 2\pi]$.

Let us now consider the combination of the above estimators. In connection with the previous sections, we have $d = 2$, $k_1 = 2$, $k_2 = 1$, $\mathbf{T}_1 = (\hat{\rho}_1, \hat{\rho}_2)$ and $\mathbf{T}_2 = \hat{\alpha}_1$. The averaging estimator over $\Lambda_{\max}$ is denoted by $(\hat{\rho}_{AV}, \hat{\alpha}_{AV})$. In this setting, we recall that $\hat{\rho}_{AV}$ is a linear combination of $\hat{\rho}_1$ and $\hat{\rho}_2$ where the weights sum to one, whereas $\hat{\alpha}_{AV}$ equals $\hat{\alpha}_1$ plus a linear combination of $\hat{\rho}_1$ and $\hat{\rho}_2$ where the weights sum to zero. The weights are estimated according to (9), where Σ is obtained from Monte-Carlo simulations of the model with parameters $0.5(\hat{\rho}_1+\hat{\rho}_2)$ and $\hat{\alpha}_1$ (see the previous section for more details).

28

Table 6 reports the MSE of each estimator, estimated from 10^4 replications from a Boolean model with parameters $\rho = 25, 50, 100, 150$ and $\alpha = 1$. For each replication, 100 Monte-Carlo samples were used. The averaging estimators have better performances than the initial estimators. It is worth noticing the improvement of $\hat{\alpha}_1$ when it is corrected by $\hat{\rho}_1$ and $\hat{\rho}_2$ through $\hat{\alpha}_{AV}$. Though this procedure might seem unnatural, the result is conclusive for this model. More simulations with other values of α (not reported in this paper) gave similar results.

	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_{AV}$	$\hat{\alpha}_1$	$\hat{\alpha}_{AV}$
$\rho = 25$	34.15	14.63	14.60	8.09	6.70
	(0.55)	(0.22)	(0.22)	(0.15)	(0.13)
$\rho = 50$	131.63	47.41	45.65	4.69	3.24
	(2.26)	(0.72)	(0.67)	(0.067)	(0.048)
$\rho = 100$	949	272	223	5.70	2.29
	(21.8)	(4.9)	(3.6)	(0.086)	(0.034)
$\rho = 150$	7606	1656	1005	14.7	4.1
	(341)	(46.5)	(24.4)	(0.34)	(0.11)

Table 6: Monte Carlo estimation of the MSE of $\hat{\rho}_1$, $\hat{\rho}_2$, $\hat{\rho}_{AV}$ and $\hat{\alpha}_1$, $\hat{\alpha}_{AV}$ based on 10^4 replications of a Boolean model with parameters $\rho = 25, 50, 100, 200$ and $\mu \sim B(1, \alpha)$ with $\alpha = 1$. The standard deviation of the MSE estimations are given in parenthesis. The two last columns have been multiplied by 100 for ease of presentation.

As explained in Section 3.2, we can deduce from $\hat{\rho}_{AV}$ and $\hat{\alpha}_{AV}$ an asymptotic confidence interval without any further estimation. The length of this interval is smaller than the length of any similar confidence interval based on the initial estimators and Table 7 reports the empirical rate of coverage of these intervals, showing that it is close to the nominal level 95%.

	$\rho = 25$	$\rho = 50$	$\rho = 100$	$\rho = 150$
$\hat{\rho}_{AV}$	98.3 %	97.6 %	96.5 %	93.4 %
$\hat{\alpha}_{AV}$	95.9 %	94.3 %	93.9 %	94.9 %

Table 7: Empirical rate of coverage (in %) of the asymptotic 95% confidence intervals based on $\hat{\rho}_{AV}$ and $\hat{\alpha}_{AV}$ for the parameters of the Boolean model, deduced from the same simulations as in Table 6.

4.4. Estimation of a quantile under misspecification

In this section we consider a situation where we combine a non-parametric estimator with several parametric estimators coming from possibly misspecified models. In this setting, our main condition (16) implying the asymptotic optimality of the average estimator is unlikely to hold and further investigations would be necessary to well understand the implications of a model misspecification. The following simulations give nonetheless an idea of the robustness of the averaging procedure.

Specifically, given an iid sample $x_1, \ldots, x_n$, we build four estimators of the p-th quantile of the unknown underlying distribution: the non-parametric estimator $\hat{q}_{NP} = x_{(\lfloor np \rfloor)}$; the parametric estimator in the Weibull model (i.e. $\hat{q}_W = F^{-1}_{\hat{\alpha}, \hat{\beta}}(p)$ where $F_{\alpha, \beta}$ denotes the Weibull cdf with parameter (α, β) and $(\hat{\alpha}, \hat{\beta})$ is the MLE estimator); the parametric estimator $\hat{q}_G$ in the Gamma model and the parametric estimator $\hat{q}_B$ in the Burr model.

The three parametric models above have a different right-tail behavior: The Weibull distribution is not heavy-tailed when $\beta > 1$, while the Gamma distribution is heavy-tailed but not fat-tailed and the Burr distribution is fat-tailed.

In this misspecified framework, we choose to use convex averaging, i.e. the set of weights is given by (12), and we denote by $\hat{q}_{AV}$ the average estimator. The MSE matrix is estimated by bootstrap where for the initial estimator we take $\hat{q}_{NP}$.

Table 8 reports the mean squared error of each estimator when the ground truth distribution is either a Weibull distribution with parameter $(3, 2)$, or the Gamma distribution with parameter $(3, 2)$, or the Burr distribution with parameter $(2, 1)$, or the standard lognormal distribution. Note that in the three first cases, one of the parametric estimators is well specified while the other parametric estimators are misspecified, and in the last situation all parametric estimators are misspecified. The table is concerned with the estimation of the p-th quantile with $p = 0.99$, based on $n = 100$ and $n = 1000$ observations. The mean squared errors are estimated from 10^4 replications and the standard deviation of these estimations are given in parenthesis. Further simulations have been conducted for other values of p, leading to the same conclusions as below, so we do not report them in this article.

The average estimator outperforms the non-parametric estimator as soon

30

as there is one well-specified parametric estimator in the initial collection, that is for the three first rows of Table 8. When no parametric model is well-specified, as this is the case for the last row of Table 8, then the average estimator has a similar mean squared error as $\hat{q}_{NP}$.

| | $n = 100$ | | | | | $n = 1000$ | | | | |
	$\hat{q}_W$	$\hat{q}_G$	$\hat{q}_B$	$\hat{q}_{NP}$	$\hat{q}_{AV}$	$\hat{q}_W$	$\hat{q}_G$	$\hat{q}_B$	$\hat{q}_{NP}$	$\hat{q}_{AV}$
Weibull	22	255	1.10^5	52	42	2.1	234	1.10^5	5.7	4.7
	(0.30)	(2.0)	(405)	(0.7)	(0.6)	(0.03)	(0.61)	(1.10^2)	(0.08)	(0.07)
Gamma	3.6	1.8	3.10^6	5.3	4.2	1.7	0.18	3.10^6	0.62	0.60
	(0.04)	(0.02)	(2.10^4)	(0.07)	(0.06)	(0.01)	(0.003)	(5.10^4)	(0.008)	(0.008)
Burr	15	18	7	24	16	8.9	12.7	0.6	2.4	1.8
	(0.14)	(0.35)	(0.17)	(1.46)	(0.82)	(0.04)	(0.06)	(0.01)	(0.04)	(0.03)
Lognormal	9.9	11.6	30.2	10.9	9.2	7.29	9.87	13.39	1.38	1.38
	(0.08)	(0.07)	(0.51)	(0.24)	(0.13)	(0.03)	(0.03)	(0.09)	(0.02)	(0.02)

Table 8: Monte Carlo estimation of the MSE of $\hat{q}_W$, $\hat{q}_G$, $\hat{q}_B$, $\hat{q}_{NP}$ and $\hat{q}_{AV}$ when $p = 0.99$, $n = 100$ (left) and $n = 1000$ (right), based on 10^4 replications of a Weibull distribution (first row), a Gamma distribution (second row), a Burr distribution (third row) and a lognormal distribution (last row). The standard deviation of the MSE estimations are given in parenthesis. The first row has been multiplied by 1000 for ease of presentation.

5. Appendix

Proof of Theorem 3.1.
Since $\Lambda \subseteq \Lambda_{\max}$, we know that $\lambda^\top J = I$ for all $\lambda \in \Lambda$. Let $\mathbf{S} = \Sigma^{-\frac{1}{2}}(\mathbf{T} - J\theta)$, we have

$$\|\hat{\theta} - \hat{\theta}^*\|^2 = \|(\hat{\lambda} - \lambda^*)^\top (\mathbf{T} - J\theta)\|^2 = \|(\hat{\lambda} - \lambda^*)^\top \Sigma^{\frac{1}{2}} \mathbf{S}\|^2 \leq \|(\hat{\lambda} - \lambda^*)^\top \Sigma^{\frac{1}{2}}\|_F^2 \|\mathbf{S}\|^2, \tag{24}$$

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

$$\|(\hat{\lambda} - \lambda^*)\Sigma^{\frac{1}{2}}\|_F^2 = \mathrm{tr}(\hat{\lambda}^\top \Sigma \hat{\lambda}) - \mathrm{tr}(\lambda^{*\top}\Sigma\lambda^*) - 2\,\mathrm{tr}\left[\lambda^{*\top}\Sigma(\hat{\lambda} - \lambda^*)\right]$$
$$\leq \mathrm{tr}(\hat{\lambda}^\top \Sigma \hat{\lambda}) - \mathrm{tr}(\lambda^{*\top}\Sigma\lambda^*). \tag{25}$$

31

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

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

where $\delta_\Lambda(A|B)$ and $\delta_\Lambda(A, B)$ are defined in Section 3.1. Now using that $\mathrm{tr}(\hat{\lambda}^\top \hat{\Sigma} \hat{\lambda}) \le \mathrm{tr}(\lambda^{*\top} \hat{\Sigma} \lambda^*)$ and

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

we obtain

$$
\mathrm{tr}(\hat{\lambda}^\top \Sigma \hat{\lambda}) - \mathrm{tr}(\lambda^{*\top} \Sigma \lambda^*) \le \mathrm{tr}(\lambda^{*\top} \Sigma \lambda^*)\,\left[2\delta_\Lambda(\hat{\Sigma}, \Sigma) + \delta_\Lambda(\hat{\Sigma}, \Sigma)^2\right]. \qquad (26)
$$

Recall that $\mathrm{tr}(\lambda^{*\top} \Sigma \lambda^*) = \inf_{\lambda \in \Lambda} \mathbb{E}\|\lambda^\top \mathbf{T} - \theta\|^2 = \mathbb{E}\|\hat{\theta}^* - \theta\|^2$, the result follows from (24), (25) and (26). $\qquad \square$

Proof of Corollary 3.2.
Write for $\epsilon > 0$,

$$
\|\hat{\theta} - \theta\|^2 \le (1 + \epsilon)\|\hat{\theta}^* - \theta\|^2 + (1 + \epsilon^{-1})\|\hat{\theta} - \hat{\theta}^*\|^2.
$$

Applying Theorem 3.1, we get

$$
\|\hat{\theta} - \theta\|^2 \le (1 + \epsilon)\|\hat{\theta}^* - \theta\|^2 + (1 + \epsilon^{-1})\mathbb{E}\|\hat{\theta}^* - \theta\|^2\,\left(\tilde{\delta}_\Lambda(\hat{\Sigma}, \Sigma)\,\|\mathbf{S}\|^2\right), \quad (27)
$$

and the result follows by taking the expectation. $\qquad \square$

Lemma 5.1. *Let A, B be two positive definite matrices of order k. For any non-empty set Λ that does not contain 0,*

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

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

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

By Cauchy-Schwarz inequality,

$$
\begin{aligned}
|\operatorname{tr}[\lambda^\top (B - A)\lambda]| &= \left| \operatorname{tr}\left[\lambda^\top B^{\frac{1}{2}}\,(\mathrm{I} - B^{-\frac{1}{2}}AB^{-\frac{1}{2}})\,B^{\frac{1}{2}}\lambda\right]\right| \\
&\leq \|B^{\frac{1}{2}}\lambda\|_F\|(\mathrm{I} - B^{-\frac{1}{2}}AB^{-\frac{1}{2}})\,B^{\frac{1}{2}}\lambda\|_F \\
&\leq |\!|\!|\mathrm{I} - B^{-\frac{1}{2}}AB^{-\frac{1}{2}}|\!|\!|\,\|B^{\frac{1}{2}}\lambda\|_F^2.
\end{aligned}
\tag{28}
$$

Recall that $\|B^{\frac{1}{2}}\lambda\|_F^2 = \operatorname{tr}(\lambda^\top B\lambda)$, 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 $\mathrm{sp}(.)$ 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. $\qquad\square$

Proof of Proposition 3.3.
By Lemma 5.1, we know that $\delta_\Lambda(\hat\Sigma_n, \Sigma_n) = o_p(1)$ whenever $\hat\Sigma_n\Sigma_n^{-1} \xrightarrow{p} \mathrm{I}$. Letting $\mathbf{S}_n = \Sigma_n^{-\frac{1}{2}}(\mathbf{T}_n - \mathrm{J}\,\theta)$, the fact that $\mathbb{E}\|\mathbf{S}_n\|^2 = k$ implies $\|\mathbf{S}_n\|^2 = O_p(1)$. Equation (27) holds for all $\epsilon > 0$ so we can take $\epsilon = \epsilon_n$ such that $\epsilon_n \to 0$ and $\delta_\Lambda(\hat\Sigma_n, \Sigma_n)/\epsilon_n \xrightarrow{p} 0$ as $n \to \infty$, yielding

$$
\|\hat\theta_n - \theta\|^2 \leq \|\hat\theta_n^* - \theta\|^2 + \epsilon_n\|\hat\theta_n^* - \theta\|^2 + o_p(\alpha_n) = \|\hat\theta_n^* - \theta\|^2 + o_p(\alpha_n).
$$

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

$$
\underline{\hat\lambda}_{n,j} = \arg\min_{\lambda\in\Lambda_j}\ \lambda^\top\hat\Sigma_n\lambda \quad\text{and}\quad \underline{\lambda}_{n,j}^* = \arg\min_{\lambda\in\Lambda_j}\ \lambda^\top\Sigma_n\lambda,
$$

33

where we recall $\Lambda_j = \{\underline{\lambda}_j : \lambda \in \Lambda\}$. From the proof of Theorem 3.1, we get

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

We deduce that $\alpha_{n,j}^{-\frac{1}{2}}(\hat{\theta}_{n,j} - \hat{\theta}_{n,j}^*) = o_p(1)$ in view of (16) and Lemma 5.1. Now, remark that

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

Similarly,

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

proving that $\alpha_{n,j}/\hat{\alpha}_{n,j} \xrightarrow{p} 1$ by the squeeze theorem. $\qquad\square$

Acknowledgments

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] BATES, J. M., AND GRANGER, C. W. The combination of forecasts. *Operations Research 20*, 4 (1969), 451–468.

[3] BUCKLAND, S. T., BURNHAM, K. P., AND AUGUSTIN, N. H. Model selection: an integral part of inference. *Biometrics* (1997), 603–618.

[4] 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.

[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. Statistical learning theory and stochastic optimization, Ecole d'été de Probabilités de Saint-Flour XXXI–2001. *Lecture Notes in Mathematics 1851* (2004), 1–269.

[7] CESA-BIANCHI, N., AND LUGOSI, G. *Prediction, learning, and games.* Cambridge University Press, New York, 2006.

[8] CHIU, S. N., STOYAN, D., KENDALL, W. S., AND MECKE, J. *Stochastic geometry and its applications*, 3 ed. John Wiley & Sons, 2013.

[9] DALALYAN, A. S., AND SALMON, J. Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist. 40*, 4 (2012), 2327–2355.

[10] ELLIOTT, G. Averaging and the optimal combination of forecasts. Tech. rep., UCSD Working Paper, 2011.

[11] GRAYBILL, F. A., AND DEAL, R. B. Combining unbiased estimators. *Biometrics 15* (1959), 543–550.

[12] HALPERIN, M. Almost linearly-optimum combination of unbiased estimates. *Journal of the American Statistical Association 56*, 293 (1961), 36–43.

[13] HANSEN, B. E. Least squares model averaging. *Econometrica 75*, 4 (2007), 1175–1189.

[14] HANSEN, B. E., AND RACINE, J. S. Jackknife model averaging. *Journal of Econometrics 167*, 1 (2012), 38–46.

[15] HJORT, N. L., AND CLAESKENS, G. Frequentist model average estimators. *Journal of the American Statistical Association 98*, 464 (2003), 879–899.

[16] 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.

[17] JUDITSKY, A., AND NEMIROVSKI, A. Functional aggregation for nonparametric regression. *Ann. Statist. 28*, 3 (2000), 681–712.

[18] 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.

[19] LAPLACE, P.-S. DE. *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).

[20] MEHTA, J., AND GURLAND, J. On combining unbiased estimators of the mean. *Trabajos de estadística y de investigación operativa 20* (1969), 173–185.

[21] MOLCHANOV, I. S. Statistics of the boolean model: from the estimation of means to the estimation of distributions. *Advances in applied probability* (1995), 63–86.

[22] MOLCHANOV, I. S. *Statistics of the Boolean Model for Practitioners and Mathematicians.* Wiley, Chichester, 1997.

[23] MORAL-BENITO, E. Model averaging in economics: An overview. *Journal of Economic Surveys* (2013), 1–30.

[24] NEMIROVSKI, A. Topics in Non-Parametric Statistics, Ecole d'été de Probabilités de Saint-Flour XXVIII–1998. *Lecture Note in Mathematics 1738* (2000).

[25] NOCEDAL, J., AND WRIGHT, S. Numerical optimization. *Springer, New York* (2006).

[26] RAFTERY, A. E., MADIGAN, D., AND HOETING, J. A. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association 92*, 437 (1997), 179–191.

[27] RIGOLLET, P., AND TSYBAKOV, A. Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics 16*, 3 (2007), 260–280.

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

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

[30] TIMMERMANN, A. Forecast combinations. In *Handbook of Economic Forecasting*, G. Elliott, C. Granger, and A. Timmermann, Eds. North Holland, Amsterdam, 2006, pp. 135–196.

[31] WANG, Z., PATERLINI, S., GAO, F., AND YANG, Y. Adaptive minimax regression estimation over sparse ℓ_q-hulls. *J. Mach. Learn. Res. 15* (2014), 1675–1711.

[32] WASSERMAN, L. Bayesian model selection and model averaging. *Journal of mathematical psychology 44*, 1 (2000), 92–107.

[33] WEIL, W., AND WIEACKER, J. A. Densities for stationary random sets and point processes. *Advances in applied probability* (1984), 324–346.

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

[35] YANG, Y. Aggregating regression procedures to improve performance. *Bernoulli 10*, 1 (2004), 25–47.