The Cost-free Nature of Optimally Tuning Tikhonov Regularizers and Other Ordered Smoothers

Pierre C. Bellec [1] **Dana Yang** [2]

Abstract

We consider the problem of selecting the best estimator among a family of Tikhonov regularized estimators, or, alternatively, to select a linear combination of these regularizers that is as good as the best regularizer in the family. Our theory reveals that if the Tikhonov regularizers share the same penalty matrix with different tuning parameters, a convex procedure based on Q-aggregation achieves the mean square error of the best estimator, up to a small error term no larger than $C\sigma^2$, where σ^2 is the noise level and $C > 0$ is an absolute constant. Remarkably, the error term does not depend on the penalty matrix or the number of estimators as long as they share the same penalty matrix, i.e., it applies to any grid of tuning parameters, no matter how large the cardinality of the grid is. This reveals the surprising "cost-free" nature of optimally tuning Tikhonov regularizers, in striking contrast with the existing literature on aggregation of estimators where one typically has to pay a cost of $\sigma^2 \log(M)$ where M is the number of estimators in the family. The result holds, more generally, for any family of ordered linear smoothers; this encompasses Ridge regression as well as Principal Component Regression. The result is extended to the problem of tuning Tikhonov regularizers with different penalty matrices.

1. Introduction

Consider a learning problem where one is given a response vector $y \in \mathbb{R}^n$ and a design matrix $X \in \mathbb{R}^{n\times p}$. Given a positive definite matrix $K \in \mathbb{R}^{p\times p}$ and a regularization parameter $\lambda > 0$, the Tikhonov regularized estimator $\hat{w}(K, \lambda)$ is defined as the solution of the quadratic program

$$\hat{w}(K,\lambda) = \operatorname*{arg\,min}_{w\in\mathbb{R}^p} \left(\|Xw - y\|^2 + \lambda w^T K w \right), \quad (1.1)$$

where $\|\cdot\|$ is the Euclidean norm. Since we assume that the penalty matrix K is positive definite, the above optimization problem is strongly convex and the solution is unique. In the special case $K = I_{p\times p}$, the above estimator reduces to Ridge regression. It is well known that the above optimization problem can be explicitly solved and that

$$\begin{aligned}&\hat{w}(K,\lambda) = (X^TX + \lambda K)^{-1}X^Ty \\ =&K^{-1/2}(K^{-1/2}X^TXK^{-1/2} + \lambda I_{p\times p})^{-1}K^{-1/2}X^Ty.\end{aligned}$$

Problem statement. Consider the Gaussian mean model

$$y = \mu + \varepsilon \qquad \text{with} \qquad \varepsilon \sim N(0, \sigma^2 I_{n\times n}) \quad (1.2)$$

where $\mu \in \mathbb{R}^n$ is an unknown mean, and consider a deterministic design matrix $X \in \mathbb{R}^{n\times p}$. We are given a grid of tuning parameters $\lambda_1, ..., \lambda_M \geq 0$ and a penalty matrix K as above. Our goal is to construct an estimator $\tilde{w}$ such that the *regret* or *excess risk*

$$\mathbb{E}[\|X\tilde{w} - \mu\|^2] - \min_{j=1,...,M} \mathbb{E}[\|X\hat{w}(K,\lambda_j) - \mu\|^2] \quad (1.3)$$

is small. Beyond the construction of an estimator $\tilde{w}$ that has small regret, we aim to answer the following questions:

- How does the regret scale with M, the number of tuning parameters on the grid?
- How does the regret scale with $R^* = \min_{j=1,...,M} \mathbb{E}[\|X\hat{w}(K,\lambda_j) - \mu\|^2]$, the minimal mean squared error among the tuning parameters $\lambda_1, ..., \lambda_M$?

Ordered linear smoothers. If $A_j = X(X^TX + \lambda_j K)^{-1}X^T$ is the matrix such that $A_j y = X\hat{w}(K, \lambda_j)$, the family of estimators $\{A_j y, j = 1, ..., M\}$ is an example of ordered linear smoothers, introduced (Kneip, 1994).

Definition 1. The set of $n \times n$ matrices $F \subset \mathbb{R}^{n\times n}$ is referred to as a family of ordered linear smoothers if (i) for

[1]Department of Statistics, Busch Campus, Rutgers University, Piscataway, New Jersey, USA [2]The Fuqua School of Business, Duke University, Durham, North Carolina, USA. Correspondence to: Pierre C. Bellec <pierre.bellec@rutgers.edu>, Dana Yang <xiaoqian.yang@duke.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

all $A \in F$, A is symmetric and $0 \leq w^T A w \leq \|w\|^2$ for all $w \in \mathbb{R}^p$, (ii) the matrices commute: $AB = BA$ for all $A, B \in F$, and (iii) either $A \preceq B$ or $B \preceq A$ holds for all $A, B \in F$, where $\preceq$ denotes the partial order of positive symmetric matrices, i.e., $A \preceq B$ if and only if $B - A$ is positive semi-definite.

Condition (i) is mild: if the matrix A is not symmetric then it is not admissible and there exists a symmetric matrix A' such that $\mathbb{E}[\|A'y - \mu\|^2] \leq \mathbb{E}[\|Ay - \mu\|^2]$ with a strict inequality for at least one $\mu \in \mathbb{R}^n$ (Cohen, 1966), so we may as well replace A with the symmetric matrix A'. Similarly, if A is symmetric with some eigenvalues outside of $[0, 1]$, then A is not admissible and there exists another symmetric matrix A' with eigenvalues in $[0, 1]$ and smaller prediction error for all $\mu \in \mathbb{R}^n$, and strictly smaller prediction error for at least one $\mu \in \mathbb{R}^n$ if $n \geq 3$ (Cohen, 1966).

Conditions (ii) and (iii) are more stringent: they require that the matrices can be diagonalized in the same orthogonal basis $(u_1, ..., u_n)$ of $\mathbb{R}^n$, and that the matrices are ordered in the sense that there exists n functions $\alpha_1, ..., \alpha_n : \mathbb{R} \to [0, 1]$, either all non-increasing or all non-decreasing, such that

$$F \subset \{\alpha_1(\lambda) u_1 u_1^T + ... + \alpha_n(\lambda) u_n u_n^T, \lambda \in \mathbb{R}\}. \quad (1.4)$$

See (Kneip, 1994) for a proof of the equivalence between (ii)-(iii) and the existence of $u_1, ..., u_n$ and $\alpha_1, ..., \alpha_n$ as above such that (1.4) holds. This easily follows from the fact that symmetric matrices that commute can be diagonalized in the same orthonormal basis. A special case of particular interest is the above Tikhonov regularized estimators, which satisfies conditions (i)-(ii)-(iii). In this case, $F = \{A_1, ..., A_M\}$ for the matrices $A_j = X(X^T X + \lambda_j K)^{-1} X^T$ that satisfy $A_j y = X\hat{w}(K, \lambda_j)$. To see that for any grid of tuning parameters $\lambda_1, ..., \lambda_M$, the Tikhonov regularizers $F = \{A_1, ..., A_M\}$ form a family of ordered linear smoothers, the matrix A_j can be rewritten as $A_j = B(B^T B + \lambda_j I_{p \times p})^{-1} B^T$ where B is the matrix $XK^{-1/2}$. From this expression of A_j, it is clear that A_j is symmetric, that A_j can be diagonalized in the orthogonal basis made of the left singular vectors of B, and that the eigenvalues of A_j are decreasing functions of the tuning parameter. Namely, the i-th eigenvalue of A_j is equal to $\alpha_i(\lambda_j) = \mu_i(B)^2/(\mu_i(B)^2 + \lambda_j)$ where $\mu_i(B)$ is the i-th singular value of B.

Overview of the literature. There is a substantial amount of literature related to this problem, starting with (Kneip, 1994) where ordered linear smoothers are introduced and their properties were first studied. Kneip (1994) proves that if $A_1, ..., A_M$ are ordered linear smoothers, then selecting the estimate with the smallest C_p criterion (Mallows, 1973), i.e.,

$$\hat{k} = \underset{j=1,...,M}{\arg\min}\, C_p(A_j), \quad (1.5)$$
$$\text{where} \quad C_p(A) = \|Ay - y\|^2 + 2\sigma^2 \operatorname{trace}(A_j),$$

leads to the regret bound (sometimes referred to as *oracle inequality*)

$$\mathbb{E}[\|A_{\hat{k}} y - \mu\|^2] - R^* \leq C\sigma\sqrt{R^*} + C\sigma^2, \quad (1.6)$$
$$\text{where} \quad R^* = \min_{j=1,...,M} \mathbb{E}[\|A_j y - \mu\|^2]$$

for some absolute constant $C > 0$. See Künzel et al. (2014) for a discussion on the analysis in Kneip (1994). The regret bound (1.6) was later improved in (Golubev et al., 2010, Theorem 3) (Chernousova et al., 2013) using an estimate based on exponential weighting, showing that the regret is bounded from above by $\sigma^2 \log(2 + R^*/\sigma^2)$.

Another line of research has obtained regret bounds that scales with the cardinality M of the given family of linear estimators. Using an exponential weight estimate with a well chosen temperature parameter, (Leung & Barron, 2006; Dalalyan & Salmon, 2012) showed that if $A_1, ..., A_M$ are square matrices of size n that are either orthogonal projections, or that satisfies some commutative property, then a data-driven convex combination $\hat{A}_{EW}$ of the matrices $A_1, ..., A_M$ satisfies

$$\mathbb{E}[\|\hat{A}_{EW} y - \mu\|^2] - R^* \leq C\sigma^2 \log M. \quad (1.7)$$

where $C > 0$ is an absolute constant. This was later improved in (Bellec, 2018) using an estimate from the Q-aggregation procedure of (Dai et al., 2012; 2014). Namely, Theorem 2.1 in (Bellec, 2018) states that if $A_1, ..., A_M$ are square matrices with operator norm at most 1, then

$$\mathbb{P}\Big(\|\hat{A}_Q y - \mu\|^2] - \min_{j=1,...,M} \|A_j y - \mu\|^2 \leq C\sigma^2 \log(M/\delta)\Big)$$
$$\geq 1 - \delta \quad (1.8)$$

for any $\delta \in (0, 1)$, where $\hat{A}_Q$ is a data-driven convex combination of the matrices $A_1, ..., A_M$. A result similar to (1.7) can then be deduced from the above high probability bound by integration. It should be noted that the linear estimators in (1.7) and (1.8) need not be ordered smoothers (the only assumption in (1.8) is that the operator norm of A_j is at most one), unlike (1.6) where the ordered smoothers assumption is key.

Another popular approach to select a good estimate among a family of linear estimators is the Generalized Cross-Validation (GCV) criterion of (Craven & Wahba, 1978; Golub et al., 1979). If we are given M linear estimators defined by square matrices $A_1, ..., A_M$, Generalized Cross-Validation selects the estimator

$$\hat{k} = \underset{j=1,...,M}{\arg\min} \frac{\|A_j y - y\|^2}{(\operatorname{trace}[I_{n \times n} - A_j])^2}.$$

We could not pinpoint in the literature an oracle inequality satisfied by GCV comparable to (1.6)-(1.7)-(1.8), though we mention that (Li, 1986) exhibits asymptotic frameworks where GCV is suboptimal while, in the same asymptotic frameworks, Mallows C_p is optimal.

The problem of optimally tuning Tikhonov regularizers, Ridge regressors or smoothning splines has received considerable attention in the last four decades (for instance, the GCV paper (Golub et al., 1979) is cited more than four thousand times) and the authors of the present paper are guilty of numerous omissions of important related works. We refer the reader to the recent surveys (Arlot & Celisse, 2010; Arlot & Bach, 2009) and the references therein for the problem of tuning linear estimators, and to (Tsybakov, 2014) for a survey of aggregation results.

Coming back to our initial problem of optimally tuning a family of Tikhonov regularizers $\hat{w}(K, \lambda_1), ..., \hat{w}(K, \lambda_M)$, the results (1.6), (1.7) and (1.8) above suggest that one must pay a price that depends either on the cardinality M of the grid of tuning parameters, or on $R^* = \min_{j=1,...,M} \mathbb{E}[\|X\hat{w}(K, \lambda_j) - \mu\|^2]$, the minimal mean squared error on this grid.

Optimally tuning ordered linear smoothers incurs no statistical cost. Surprisingly, our theoretical results of the next sections reveal that if $A_1, ..., A_M$ are ordered linear smoothers, for example Tikhonov regularizers sharing the same penalty matrix K, then it is possible to construct a data-driven convex combination $\hat{A}$ of $A_1, ..., A_M$ such that the regret satisfies

$$\mathbb{E}[\|\hat{A}y - \mu\|^2] - \min_{j=1,...,M} \mathbb{E}[\|A_j y - \mu\|^2] \leq C_1 \sigma^2$$

for some absolute constant $C_1 > 0$. Hence the regret in (1.3) is bounded by $C_1\sigma^2$, an upper bound that is (a) independent of the cardinality M of the grid of tuning parameters and (b) independent of the minimal risk $R^* = \min_{j=1,...,M} \mathbb{E}[\|A_j y - \mu\|^2]$. No matter how coarse the grid of tuning parameter is, no matter the number of tuning parameters to choose from, no matter how large the minimal risk R^* is, the regret of the procedure constructed in the next section is always bounded by $C_1\sigma^2$.

Notation. Throughout the paper, $C_1, C_2, C_3...$ denote absolute positive constants. The norm $\|\cdot\|$ is the Euclidean norm of vectors. Let $\|\cdot\|_{op}$ and $\|\cdot\|_F$ be the operator and Frobenius norm of matrices.

2. Construction of the estimator

Assume that we are given M matrices $A_1, ..., A_M$, each matrix corresponding to the linear estimator $A_j y$. Mallows (1973) C_p criterion is given by (1.5).

Following several works on aggregation of estimators (Nemirovski, 2000; Tsybakov, 2003; Leung & Barron, 2006; Rigollet & Tsybakov, 2007; Dalalyan & Salmon, 2012; Dai et al., 2012; Bellec, 2018) we parametrize the convex hull of the matrices $A_1, ..., A_M$ as follows:

$$A_\theta \triangleq \sum_{j=1}^M \theta_j A_j, \quad \text{for each } \theta \in \Lambda_M \tag{2.1}$$

$$\text{where } \Lambda_M = \Big\{\theta \in \mathbb{R}^M : \theta_j \geq 0, \sum_{j=1}^M \theta_j = 1\Big\}.$$

Above, Λ_M is the simplex in $\mathbb{R}^M$ and the convex hull of the matrices $A_1, ..., A_M$ is exactly the set $\{A_\theta, \theta \in \Lambda_M\}$. Finally, define the weights $\hat{\theta} \in \Lambda_M$ by

$$\hat{\theta} = \operatorname*{arg\,min}_{\theta \in \Lambda_M} \Big(C_p(A_\theta) + \frac{1}{2} \sum_{j=1}^M \theta_j \|(A_\theta - A_j)y\|^2 \Big). \tag{2.2}$$

The first term of the objective function is Mallows C_p from (1.5), while the second term is a penalty derived from the Q-aggregation procedure from (Rigollet, 2012; Dai et al., 2012). The estimate (2.2) is equivalent to the procedure (Dai et al., 2012) with the least-squares term replaced by C_p; it is also close to the estimate from of (Dai et al., 2014) although (Dai et al., 2014) presents an unncessary penalty term that leads to worse guarantees than (2.2), cf. the discussion in (Bellec, 2018).

The penalty is minimized at the vertices of the simplex and thus penalizes the interior of Λ_M. Although convexity of the above optimization problem is unclear at first sight because the penalty is non-convex, the objective function can be rewritten, thanks to a bias-variance decomposition, as

$$\tfrac{1}{2}\|A_\theta y - y\|^2 + 2\sigma^2 \operatorname{trace}(A_\theta) + \tfrac{1}{2}\sum_{j=1}^M \theta_j \|A_j y - y\|^2. \tag{2.3}$$

The first term is a convex quadratic function in θ, while both the second term ($2\sigma^2 \operatorname{trace}[A_\theta]$) and the last term are linear in θ. It is now clear that the objective function is convex and (2.2) is a convex quadratic program (QP) over the canonical simplex, with M variables and $M + 1$ linear constraints. The computational complexity of such convex QP is polynomial and well studied, e.g., (Vavasis, 2001, page 304), (Frank et al., 1956), (Nocedal & Wright, 2006, Section 16.5).

The final estimator is

$$\hat{\mu} \triangleq A_{\hat{\theta}} y = \textstyle\sum_{j=1}^M \hat{\theta}_j A_j y, \tag{2.4}$$

a weighted sum of the values predicted by the linear estimators $A_1, ..., A_j$. The performance of this procedure is studied in (Dai et al., 2014; Bellec, 2018); (Bellec, 2018) derived the oracle inequality (1.8) which is optimal for certain

collections $\{A_1, ..., A_M\}$. However, we are not aware of previous analysis of this procedure in the context of ordered linear smoothers.

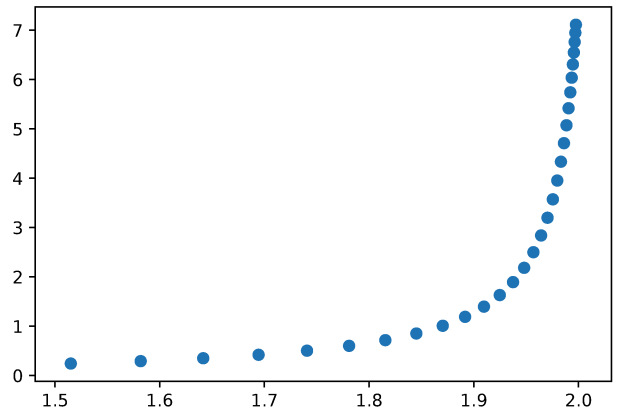

Figure 1. $A_j y = X\hat{w}(I_p, \lambda_j)y$ for $X = \text{diag}(10, 1), y = (2, 8)$ and different values of λ_j

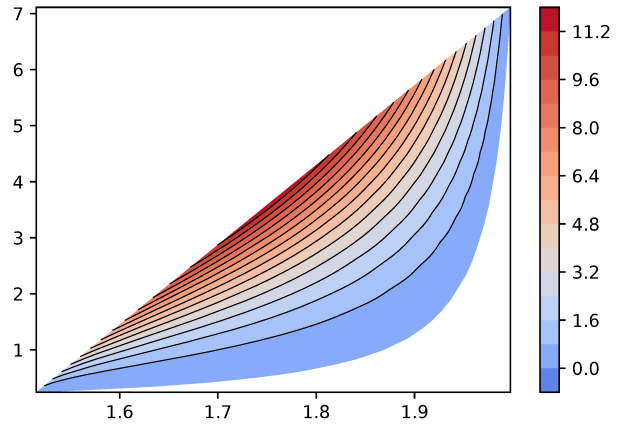

Figure 2. Heatmap of penalty values in the convex hull for the estimators in Figure 1

For $n = 2$, a visualization of the penalty $\text{pen}(\theta) = \frac{1}{2}\sum_{j=1}^{M} \theta_j \|(A_\theta - A_j)y\|^2$ in (2.2) is displayed in the figures above. The discrete points in Figure 1 correspond to predicted values $X\hat{w}(K, \lambda_j)$ for the same matrix K and many different tuning parameters $\{\lambda_j, j = 1, ..., M\}$. Figure 2 is a heatmap of the penalty defined in the convex hull of $\{X\hat{w}(K, \lambda_j), j = 1, ..., M\}$. For each point $\hat{\mu}$ in the convex hull of $\{Xw(K, \lambda_j), j = 1, ..., M\}$, the value of the penalty of $\hat{\mu}$ in the heatmap is defined as $\min_{\theta \in \Lambda_M : \hat{\mu} = A_\theta y} \text{pen}(\theta)$. This minimum is computed by linear programming for each $\hat{\mu}$ in a dense grid in the convex hull. We used $n = 2$, $y = (1, -1)^T$, $X = \text{diag}(2, 1)$ and $K = Q\,\text{diag}(1, 3)\,Q^T$ where $Q \in \mathbb{R}^{n \times n}$ is a randomly generated orthonormal matrix. The code is provided in the supplementary material.

3. Constant regret for ordered linear smoothers

Theorem 3.1. *The following holds for absolute constants $C_1, C_2, C_3 > 0$. Consider the Gaussian mean model (1.2). Let $\{A_1, ..., A_M\}$ be a family of ordered linear smoothers as in Definition 1. Let $\hat{\theta}$ be the solution to the optimization problem (2.2). Then $\hat{\mu} = A_{\hat{\theta}} y$ enjoys the regret bound*

$$\mathbb{E}[\|A_{\hat{\theta}} y - \mu\|^2] - \min_{j=1,...,M} \mathbb{E}[\|A_j y - \mu\|^2] \le C_1 \sigma^2. \quad (3.1)$$

Furthermore, if $j_* = \arg\min_{j=1,...,M} \mathbb{E}[\|A_j y - \mu\|^2]$ has minimal risk then for all $x \ge 1$,

$$\mathbb{P}\left\{\|A_{\hat{\theta}} y - \mu\|^2 - \|A_{j_*} y - \mu\|^2 \le C_2 \sigma^2 x\right\} \ge 1 - C_3 e^{-x}. \quad (3.2)$$

Let us explain the "cost-free" nature of the above result. In the simplest, one-dimensional regression problem where the design matrix X has only one column and $\mu = X\beta^*$ for some unknown scalar β^*, the prediction error of the Ordinary Least Squares estimator is $\mathbb{E}[\|X(\hat{\beta}^{ols} - \beta^*)\|^2] = \sigma^2$ because the random variable $\|X(\hat{\beta}^{ols} - \beta^*)\|^2/\sigma^2$ has chi-square distribution with one degree-of-freedom. As indicated by our result, the regret of $\hat{\mu}$ never exceedes a constant multiple of the prediction error in a one-dimensional linear model. The right hand side of (3.1) is independent of the minimal risk R^*, independent of the cardinality M of the family of estimators, and if the estimators were constructed from a linear model with p covariates, the right hand side of (3.1) is also independent of the dimension p.

The following result shows that selecting an estimator based on C_p as in $\hat{k}$ in (1.5) cannot enjoy the same property: (1.5) will suffer an error term of order $\sigma(R^*)^{1/2}$ in some cases. The intuition is that because $\hat{k}$ in (1.5) is discontinuous, $\hat{k}$ can be mistaken with small but positive probability around the discontinuity boundary and this mistake induces an error of order $\sigma(R^*)^{1/2}$ and (1.6) is not improvable for $\hat{k}$. Convex minimization over the simplex as in (2.2), together with the penalty of the previous result, has a smoother behavior and does not suffer this instability.

Theorem 3.2. *There exists $\mu \in \mathbb{R}^n$ and a family of ordered linear smoothers $F = \{A_1, A_2\}$ with $M = 2$ such that if $y \sim N(\mu, I_n)$ with $\sigma^2 = 1$,*

$$\mathbb{P}(\|A_{\hat{k}} y - \mu\|^2 - R^* \ge C_4 (R^*)^{1/2}) \ge C_5 > 0$$

where $R^* = \min_{j=1,2} \mathbb{E}[\|A_j y - \mu\|^2] = \Theta(n)$.

Since the most commonly used ordered linear smoothers are Tikhonov regularizers (which encompass Ridge regression

and smoothing splines), we provide the following corollary of Theorem 3.1 for convenience.

Corollary 3.3 (Application to Tikhonov regularizers)**.** *Let K be a positive definite matrix of size $p \times p$ and let $\lambda_1, ..., \lambda_M \geq 0$ be distinct tuning parameters. Define $\hat{\theta}$ as the minimizer of*

$$\hat{\theta} = \underset{\theta \in \Lambda_M}{\arg\min} \Big(\frac{1}{2} \| \sum_{j=1}^{M} \theta_j X \hat{w}(K, \lambda_j) - y \|^2 + 2\sigma^2 \sum_{j=1}^{M} \theta_j \mathrm{df}_j + \frac{1}{2} \sum_{j=1}^{M} \theta_j \| X \hat{w}(K, \lambda_j) - y \|^2 \Big), \tag{3.3}$$

where $\mathrm{df}_j = \mathrm{trace}[X^T (X^T X + \lambda_j K)^{-1} X^T]$. Then the weight vector $\tilde{w} = \sum_{j=1}^{M} \hat{\theta}_j \hat{w}(K, \lambda_j)$ in $\mathbb{R}^p$ is such that the regret (1.3) is bounded from above by $C_1 \sigma^2$ for some absolute constant $C_1 > 0$.

This corollary is a direct consequence of Theorem 3.1 with $A_j = X^T (X^T X + \lambda_j K)^{-1} X^T$. The fact that this forms a family of ordered linear smoothers is explained after (1.4). The objective function (3.3) corresponds to the formulation (2.3) of the objective function in (2.2); we have chosen this formulation so that (3.3) can be easily implemented as a convex quadratic program with linear constraints, the first term of the objective function being quadratic in θ while the second and third terms are linear in θ.

The procedure above requires knowledge of σ^2, which needs to be estimated beforehand in practice. Estimators of σ^2 are available depending on the underlying context, e.g., difference based estimates for observations on a grid (Dette et al., 1998; Hall et al., 1990; Munk et al., 2005; Brown et al., 2007), or pivotal estimators of σ in sparse linear regression, e.g., (Belloni et al., 2014; Sun & Zhang, 2012; Owen, 2007). On a low-bias model, (Hastie et al., 2001, Section 7.5) recommends estimating σ^2 by the squared residuals. It was later suggested in (Arlot, 2019) that the procedure in (Hastie et al., 2001) could overfit, and that using slope heuristic to estimate σ^2 is likely to provide a better estimator. We also note that procedure (2.2) is robust to misspecified σ if each A_j is an orthogonal projection (Bellec, 2018, Section 6.2).

4. Multiple families of ordered smoothers or Tikhonov penalty matrices

Theorem 4.1. *The following holds for absolute constants $C_1, C_2, C_3 > 0$. Consider the Gaussian mean model (1.2). Let $\{A_1 y, ..., A_M y\}$ be a set of linear estimators such that*

$$\{A_1, ..., A_M\} \subset F_1 \cup ... \cup F_q,$$

where F_k is a family of ordered linear smoothers as in Definition 1 for each $k = 1, ..., q$. Let $\hat{\theta}$ be the solution to the optimization problem (2.2). Then $\hat{\mu} = A_{\hat{\theta}} y$ enjoys the regret bound

$$\begin{aligned} &\mathbb{E}[\|A_{\hat{\theta}} y - \mu\|^2] - \min_{j=1,...,M} \mathbb{E}[\|A_j y - \mu\|^2] \\ \leq & C_1 \sigma^2 + C_3 \sigma^2 \log q. \end{aligned} \tag{4.1}$$

Furthermore, if $j_* = \arg\min_{j=1,...,M} \mathbb{E}[\|A_j y - \mu\|^2]$ has minimal risk then for all $x \geq 1$,

$$\begin{aligned} &\mathbb{P}\left\{ \|A_{\hat{\theta}} y - \mu\|^2 - \|A_{j_*} y - \mu\|^2 \leq C_2 \sigma^2 (x + \log q) \right\} \\ \geq & 1 - C_3 e^{-x}. \end{aligned} \tag{4.2}$$

We now allow not only one family of ordered linear smoothers, but several. Above, q denotes the number of families. This setting was considered in (Kneip, 1994), although with a regret bound of the form $\sqrt{R^*} \sigma \log(q)^2 + \sigma^2 \log(q)^4$ where $R^* = \min_{j=1,...,M} \mathbb{E}[\|A_j y - \mu\|^2]$; Theorem 4.1 improves both the dependence in R^* and in q. Let us also note that the dependence in q in the above bound (4.2) is optimal (Bellec, 2018, Proposition 2.1).

The above result is typically useful in situations where several Tikhonov penalty matrices $K_1, ..., K_q$ are candidate. For each $m = 1, ..., q$, the penalty matrix is K_m, the practitioner chooses a grid of $b_m \geq 1$ tuning parameters, say, $\{\lambda_a^{(m)}, a = 1, ..., b_m\}$. If the matrices $A_1, ..., A_M$ are such that

$$\begin{aligned} &\{A_1, ..., A_M\} \\ = &\cup_{m=1}^{q} \{X(X^T X + \lambda_a^{(m)} K_m)^{-1} X^T, a = 1, ..., b_m\}, \end{aligned}$$

so that $M = \sum_{m=1}^{q} b_m$, the procedure (2.2) enjoys the regret bound

$$\begin{aligned} &\mathbb{E}[\|A_{\hat{\theta}} y - \mu\|^2] - \min_{m \leq q} \min_{a=1,...,b_m} \mathbb{E}[\|X \hat{w}(K_m, \lambda_a) - \mu\|^2] \\ &\leq C_6 \sigma^2 (1 + \log q) \end{aligned}$$

and a similar bound in probability. That is, the procedure of Section 2 automatically adapts to both the best penalty matrix and the best tuning parameter. The error term $\sigma^2(1 + \log q)$ only depends on the number of regularization matrices used, not on the cardinality of the grids of tuning parameters.

5. Proofs

We start the proof with the following deterministic result.

Lemma 5.1 (Deterministic inequality)**.** *Let $A_1, ..., A_M$ be square matrices of size $n \times n$ and consider the procedure (2.2) in the unknown mean model (1.2). Then for any $\bar{A} \in$*

$\{A_1, ..., A_M\}$,

$$\begin{aligned}&\|A_{\hat\theta}y-\mu\|^2-\|\bar A y-\mu\|^2\\ \le&\max_{j=1,...,M}\Big(2\varepsilon^T(A_j-\bar A)y-2\sigma^2\operatorname{trace}(A_j-\bar A)\\ &-\tfrac12\|(A_j-\bar A)y\|^2\Big).\end{aligned}$$

Proof. The above is proved in (Bellec, 2018, Proposition 3.2). We reproduce the short proof here for completeness: If $H:\Lambda_M\to\mathbb{R}$ is the convex objective of (2.2) and $\bar A = A_k$ for some $k=1,...,M$, the optimality condition of (2.2) states that $\nabla H(\hat\theta)(e_k-\hat\theta)\ge 0$ holds (cf. (Boyd & Vandenberghe, 2009, (4.21))). Then $\nabla H(\hat\theta)(e_k-\hat\theta)\ge 0$ can be equivalently rewritten as

$$\begin{aligned}&\|A_{\hat\theta}y-\mu\|^2-\|\bar A y-\mu\|^2\\ \le&\sum_{j=1}^M\hat\theta_j\Big(2\varepsilon^T(A_j-\bar A)y-2\sigma^2\operatorname{trace}(A_j-\bar A)\\ &-\tfrac12\|(A_j-\bar A)y\|^2\Big).\end{aligned}$$

The proof is completed by noting that the average $\sum_{j=1}^M\hat\theta_j a_j$ with weights $\hat\theta=(\hat\theta_1,...,\hat\theta_M)\in\Lambda_M$ is smaller than the maximum $\max_{j=1,...,M}a_j$ for all $a_1,...,a_M$. □

Throughout the proof, $\bar A$ is a fixed deterministic matrix with $\|\bar A\|_{op}\le 1$. Our goal is to bound from above the right hand side of Lemma 5.1 with high probability. To this end, define the process $(Z_B)_B$ indexed by symmetric matrices B of size $n\times n$, by

$$\begin{aligned}Z_B=&2\varepsilon^T(B-\bar A)y-2\sigma^2\operatorname{trace}(B-\bar A)\\ &-\tfrac12(\|(B-\bar A)y\|^2-d(B,\bar A)^2)\end{aligned}$$

where d is the metric

$$d(B,A)^2\triangleq\mathbb{E}[\|(B-A)y\|^2]=\sigma^2\|B-A\|_F^2+\|(B-A)\mu\|^2, \tag{5.1}$$

where $A,B\in\mathbb{R}^{n\times n}$.

With this definition, the quantity inside the parenthesis in the right hand side of Lemma 5.1 is exactly $Z_{A_j}-\frac12 d(A_j,\bar A)^2$. The appearance of the term involving the metric d is thanks to the penalty term in the definition of $\hat\theta$ (cf. (2.2)). To show that the regret of the estimator $A_{\hat\theta}y$ does not scale with the model dimension, we need to establish that the growth of the centered process Z_B is surpassed by the growth of the term $\frac12 d(B,\bar A)^2$.

The main technical challenge is in controlling the process Z_B. We rely on the fact that the smoothers are ordered, which allows us to use generic chaining to bound the supremum of Z_B in a very efficient manner.

Split the process Z_B into a Gaussian part and a quadratic part. Define the processes $(G_B)_B$ and $(W_B)_B$ by

$$\begin{aligned}G_B&=\varepsilon^T[2I_{n\times n}-(B-\bar A)](B-\bar A)\mu,\\ W_B&=2\varepsilon^T(B-\bar A)\varepsilon-2\sigma^2\operatorname{trace}(B-\bar A)\\ &\quad-\tfrac12\varepsilon^T(B-\bar A)^2\varepsilon+\tfrac{\sigma^2}{2}\|B-\bar A\|_F^2.\end{aligned}$$

Before bounding supremum of the above processes, we need to derive the following metric property of ordered linear smoothers. If T is a subset of the space of symmetric matrices of size $n\times n$ and if d is a metric on T, the diameter $\Delta(T,d)$ of T and the Talagrand generic chaining functionals for each $\alpha=1,2$ are defined by

$$\begin{aligned}\Delta(T,d)&=\sup_{A,B\in T}d(A,B),\\ \gamma_\alpha(T,d)&=\inf_{(T_k)_{k\ge0}}\sup_{t\in T}\sum_{k=1}^{+\infty}2^{k/\alpha}d(t,T_k)\end{aligned}\tag{5.2}$$

where the infimum is over all sequences $(T_k)_{k\ge0}$ of subsets of T such that $|T_0|=1$ and $|T_k|\le 2^{2^k}$.

Lemma 5.2. *Let $a\ge0$ and let $\mu\in\mathbb{R}^n$. Let $F\subset\mathbb{R}^{n\times n}$ be a family of ordered linear smoothers (cf. Definition 1) and let d be any semi-metric of the form $d(A,B)^2=a\|A-B\|_F^2+\|(A-B)\mu\|^2$. Then $\gamma_2(F,d)+\gamma_1(F,d)\le C_7\Delta(F,d)$ where C_7 is an absolute constant.*

Proof. We have to specify a sequence $(T_k)_{k\ge0}$ of subsets of F with $|T_k|\le 2^{2^k}$. Since F satisfies Definition 1, there exists a basis of eigenvectors $u_1,...,u_n$, non-decreasing functions $\alpha_1,...,\alpha_n:\mathbb{R}\to[0,1]$ and a set $\Lambda\subset\mathbb{R}$ such that $F=\{B_\lambda,\lambda\in\Lambda\}$ where $B_\lambda=\sum_{i=1}^n\alpha_i(\lambda)u_iu_i^T$, cf. (1.4). Hence for any $\lambda_0,\lambda,\nu\in\Lambda$,

$$\begin{aligned}d(B_\lambda,B_\nu)^2&=\sum_{i=1}^n w_i(\alpha_i(\lambda)-\alpha_i(\nu))^2\\ \text{for weights}\quad w_i&=(a+(u_i^T\mu)^2)\ge0,\\ d(B_{\lambda_0},B_\lambda)^2+d(B_\lambda,B_\nu)^2&=d(B_{\lambda_0},B_\nu)^2\\ +2\sum_{i=1}^n w_i(\alpha_i(\lambda)-\alpha_i(\lambda_0))&(\alpha_i(\lambda)-\alpha_i(\nu)).\end{aligned}$$

If $\lambda_0\le\lambda\le\nu$, since each $\alpha_i(\cdot)$ is non-decreasing, the sum in the right hand side of the previous display is non-positive and $d(B_\lambda,B_\nu)^2\le d(B_\nu,B_{\lambda_0})^2-d(B_\lambda,B_{\lambda_0})^2$ holds. Let $N=2^{2^k}$ and $\delta=\Delta(F,d)/N$. We construct a δ-covering of F by considering the bins $\text{BIN}_j=\{B\in F:\delta^2 j\le d(B,B_{\lambda_0})^2<\delta^2(j+1)\}$ for $j=0,...,N-1$ where $\lambda_0=\inf\Lambda$. If BIN_j is non-empty, then any of its elements is a δ-covering of BIN_j thanks to

$$\begin{aligned}d(B_\lambda,B_\nu)^2\le&d(B_\nu,B_{\lambda_0})^2-d(B_\lambda,B_{\lambda_0})^2\\ \le&(j+1)\delta^2-j\delta^2=\delta^2.\end{aligned}$$

for $B_\nu, B_\lambda \in \text{BIN}_j$ with $\lambda \le \nu$. This constructs a δ-covering of F with $N = 2^{2^k}$ elements. Hence $\gamma_2(F,d) \le \Delta(F,d)\sum_{k=1}^{\infty} 2^{k/2}/2^{2^k} = \Delta(F,d)C_8$ and the same holds for $\gamma_1(F,d)$ for a different absolute constant. □

The two following lemmas are proved in the supplement by leveraging the bound in Lemma 5.2 on the complexity of ordered smoothers.

Lemma 5.3 (The Gaussian process G_B)**.** *Let T^* be a family of ordered linear smoothers (cf. Definition 1) such that $\sup_{B\in T^*} d(\bar{A}, B) \le \delta^*$ for the metric (5.1). Then for all $x > 0$,*

$$\mathbb{P}(\sup_{B\in T^*} G_B \le \sigma(C_9 + 3\sqrt{2x})\delta^*) \ge 1 - e^{-x}.$$

Lemma 5.4 (The Quadratic process W_B)**.** *Let T^* be a family of ordered smoothers (cf. Definition 1) such that $\sigma\|B - \bar{A}\|_F \le \delta^*$ for all $B \in T^*$. Then for all $x > 0$,*

$$\mathbb{P}\Big(\sup_{B\in T^*} W_B \le C_{10}\sigma\delta^* + C_{11}\sigma\sqrt{x}\delta^* + C_{12}\sigma^2 x\Big) \ge 1 - 2e^{-x}.$$

The goal is to now combine the two previous lemmas to obtain tail bounds on $\sup_{B\in F}\left(Z_B - \frac{1}{2}d(B,\bar{A})^2\right)$. We first proceed on a *slice* $\{B \in F :\ \delta_* \le d(B,\bar{A}) < \delta^*\}$ for two reals $\delta_* < \delta^*$.

Lemma 5.5. *Suppose F is a family of $n \times n$ ordered linear smoothers (cf. Definition 1), and $\bar{A}$ is a fixed matrix with $\|\bar{A}\|_{op} \le 1$ which may not belong to F. Let d be the metric (5.1). Then for any reals $u \ge 1$, and $\delta^* > \delta_* \ge 0$, we have with probability at least $1 - 3e^{-u}$,*

$$\begin{aligned}&\sup_{B\in F:\ \delta_*\le d(B,\bar{A})<\delta^*}\left(Z_B - \tfrac{1}{2}d(B,\bar{A})^2\right)\\ &\le C_{13}\left[\sigma^2 u + \delta^*\sigma\sqrt{u}\right] - \tfrac{1}{2}\delta_*^2\\ &\le C_{14}\sigma^2 u + \tfrac{1}{16}(\delta^*)^2 - \tfrac{1}{2}\delta_*^2.\end{aligned}$$

Proof. First note that $-d(B,\bar{A})^2 \le -\delta_*^2$ for any B as in the supremum.

Now $Z_B = G_B + W_B$ where G_B and W_B are the processes studied in Lemmas 5.3 and 5.4. These lemmas applied to $T^* = \{B \in F : d(B,\bar{A}) \le \delta^*\}$ yields that on an event of probability at least $1 - 3e^{-u}$ we have

$$\begin{aligned}\sup_{B\in T^*} Z_B &\le \sup_{B\in T^*}(G_B + W_B)\\ &\le C_{15}(\sigma\delta^*(1+\sqrt{u}) + \sigma^2 u).\end{aligned}$$

Since $u \ge 1$, we have established the first inequality by adjusting the absolute constant. For the second inequality, we use that $C_{13}\delta_*\sigma\sqrt{u} \le 4C_{13}^2\sigma^2 u + \frac{1}{16}(\delta^*)^2$ and set $C_{14} = C_{13} + 4C_{13}^2$. □

We use here a method known as *slicing*, we refer the reader to Section 5.4 in (van Handel, 2014) for an introduction.

Lemma 5.6 (Slicing)**.** *Suppose F is a family of $n \times n$ ordered linear smoothers (cf. Definition 1), and $\bar{A}$ is a fixed matrix with $\|\bar{A}\|_{op} \le 1$ which may not belong to F. Let d be the metric (5.1). Then for any $x \ge 1$, we have with probability at least $1 - C_3 e^{-x}$*

$$\sup_{B\in F}\left(Z_B - \tfrac{1}{2}d(B,\bar{A})^2\right) \le C_2\sigma^2 x.$$

Proof. Write F as the union $F = \cup_{k=1}^{\infty} T_k$ where T_k is the slice

$$T_k = \{B \in F : \delta_{k-1} \le \tilde{d}(B,\bar{A}) \le \delta_k\},$$

with $\delta_0 = 0$ and $\delta_k = 2^k\sigma$ for $k \ge 1$. By definition of the geometric sequence $(\delta_k)_{k\ge 0}$, inequality $\frac{1}{16}\delta_k^2 - \frac{1}{2}\delta_{k-1}^2 \le \frac{1}{2}\sigma^2 - \frac{1}{16}\delta_k^2 \le \frac{1}{2}\sigma^2 x - \frac{1}{16}\delta_k^2$ holds for all $k \ge 1$. With $\delta_* = \delta_{k-1}, \delta^* = \delta_k$, Lemma 5.5 yields that for all $k \ge 1$,

$$\begin{aligned}&\mathbb{P}\Big(\sup_{B\in T_k}\left(Z_B - \tfrac{1}{2}d(B,\bar{A})^2\right) \le C_{14}\sigma^2 u_k - \tfrac{1}{16}\delta_k^2 + \tfrac{\sigma^2 x}{2}\Big)\\ &\ge 1 - 3e^{-u_k},\end{aligned}$$

for all $u_k \ge 1$. The above holds simultaneously over all slices $(T_k)_{k\ge 1}$ with probability at least $1 - 3\sum_{k=1}^{\infty} e^{-u_k}$ by the union bound. It remains to specify a sequence $(u_k)_{k\ge 1}$ of reals greater than 1. We choose $u_k = x + \delta_k^2/(16\sigma^2 C_{14})$ which is greater than 1 since $x \ge 1$. Then by construction, $C_{14}\sigma^2 u_k - \frac{1}{16}\delta_k^2 + \frac{\sigma^2 x}{2} = (C_{14} + 1/2)\sigma^2 x$ and we set $C_2 = C_{14} + 1/2$. Furthermore, $\sum_{k=1}^{\infty} e^{-u_k} = e^{-x}\sum_{k=1}^{\infty} e^{-2^{2k}/(16C_{14})}$. The sum $3\sum_{k=1}^{\infty} e^{-2^{2k}/(16C_{14})}$ is equal to a finite absolute constant named C_3 in the statement of the Lemma. □

Proof of Theorem 3.1. Let $F = \{A_1, ..., A_M\}$ and $\bar{A} = A_{j_*}$ where j_* is defined in the statement of Theorem 3.1. The conclusion of Lemma 5.1 can be rewritten as

$$\|A_{\hat{\theta}}y - \mu\|^2 - \|\bar{A}y - \mu\|^2 \le \sup_{B\in F}(Z_B - \tfrac{1}{2}d(B,\bar{A})^2)$$

where $F = \{A_1, ..., A_M\}$ is a family of ordered linear smoothers. Lemma 5.6 completes the proof of (3.2). Then (3.1) is obtained by integration of (3.2) using $\mathbb{E}[Z] \le \int_0^{\infty}\mathbb{P}(Z > t)dt$ for any $Z \ge 0$. □

Proof of Theorem 4.1. As in the proof of Theorem 3.1, we use Lemma 5.1 to deduce that a.s.,

$$\begin{aligned}&\|A_{\hat{\theta}}y - \mu\|^2 - \|\bar{A}y - \mu\|^2\\ \le\ &\max_{j=1,...,M}(Z_{A_j} - \tfrac{1}{2}d(A_j,\bar{A})^2)\\ =\ &\max_{k=1,...,q}\max_{B\in F_k}(Z_B - \tfrac{1}{2}d(B,\bar{A})^2).\end{aligned}$$

Since each F_k is a family of ordered linear smoothers, by Lemma 5.6 we have

$$\mathbb{P}\big(\max_{B\in F_k}(Z_B - \tfrac{1}{2}d(B,\bar{A})^2) > C_2\sigma^2 x\big) \le C_3 e^{-x}$$

for each $k = 1,\dots,q$. The union bound yields (4.2) and we use $\mathbb{E}[Z] \le \int_0^\infty \mathbb{P}(Z > t)dt$ for $Z \ge 0$ to deduce (4.1). □

6. Numerical experiments

6.1. Experiment 1: example from Theorem 3.2

In this section we use the example discussed in Theorem 3.2 to illustrate the advantage of the Q-aggregation estimator $A_{\hat{\theta}}y$ over selection estimators. Following the example constructed for the proof of Theorem 3.2, we let μ be any vector in $\mathbb{R}^n$ with $\|\mu\|^2 = n(1 - c/\sqrt{n})$. Let $A_1 = 0$, $A_2 = I_n$ and $\sigma^2 = 1$. From Theorem 3.2, the Mallows C_p selection estimator $A_{\hat{j}}y$ has a regret of order $\Omega(\sqrt{n})$, while the regret of $A_{\hat{\theta}}y$ is of order $O(1)$ from Theorem 3.1. Thus our theory guarantees that the Q-aggregation estimator (2.2) outperforms Mallows C_p in this example.

Take $n = 500$, $c = 1$. We generate $N = 5000$ *i.i.d.* copies of $y = \mu + \epsilon$. Each time we compute the two estimators $A_{\hat{\theta}}y$ and $A_{\hat{j}}y$, and their squared errors $\|A_{\hat{\theta}}y - \mu\|^2$ and $\|A_{\hat{j}}y - \mu\|^2$. The histograms of the squared errors are shown in figure 3.

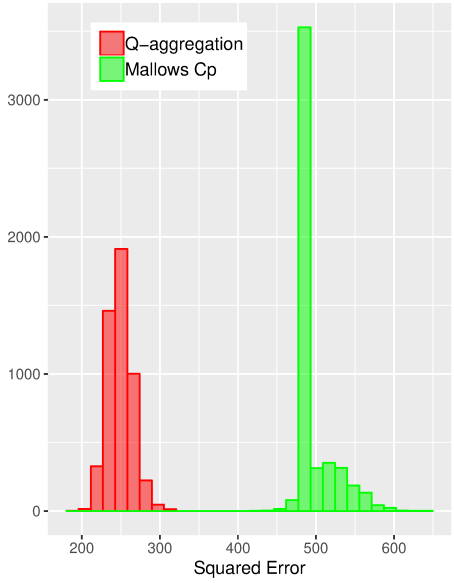

Figure 3. histograms of squared errors from experiment 1

The estimator A_2y has risk $\mathbb{E}\|A_2y - \mu\|^2 = n$, which is larger than $\mathbb{E}\|A_1y - \mu\|^2 = \|\mu\|^2$. However we argued in the proof of Theorem 3.2 that even for large n, Mallows C_p selects A_2 with some constant probability, as is reflected in figure 3.

It is also shown in figure 3 that the Q-aggregation estimator achieves much more accurate estimation of μ compared to Mallows C_p. That is due to its freedom to explore any convex combination of A_1y and A_2y. In fact our experiment shows that even if the preferred $\hat{j} = 2$ is always selected, $A_{\hat{j}}y$ still incurs a higher loss than $A_{\hat{\theta}}y$. That is, for this particular example, no selection procedure can match the performance of the Q-aggregation estimator.

6.2. Experiment 2: Tikhonov regularizers

In this section we test the performance of the Q-aggregation estimator (2.2) on a family of Tikhonov regularized estimators, in the special case of Ridge regression, *i.e.* $K = I_{p\times p}$. Specify the Tikhonov regularizers as follows.

1. Set $n = 500$, $p = 1000$, $M = 3$.
2. Generate $X \in \mathbb{R}^{n\times p}$ with all entries of X distributed *i.i.d.* standard normal.
3. Set $\lambda_1 = 200$, $\lambda_2 = 2000$, $\lambda_3 = 20000$.
4. Define $A_j = X(X^TX + \lambda_j I_{p\times p})^{-1}X^T$.

Set the ground truth $\mu \in \mathbb{R}^n$ as the all ones vector. We generate $N = 5000$ *i.i.d.* copies of $y = \mu + \epsilon$, and compare the performances of the following three estimators of μ:

1. $\hat{\mu}$: the Q-aggregation estimator (2.2);
2. $\hat{\mu}_{cp}$: selection based on Mallows C_p;
3. $\hat{\mu}_{gcv}$: selection based the Generalized Cross-Validation criterion.

Figure 4 shows the boxplots for the squared errors of $\hat{\mu}$, $\hat{\mu}_{cp}$ and $\hat{\mu}_{gcv}$ across all $N = 5000$ randomizations. From the figure it is clear that the estimator $\hat{\mu}$ which takes convex combinations of the Tikhonov regularizers outperforms both selection procedures by a significant margin.

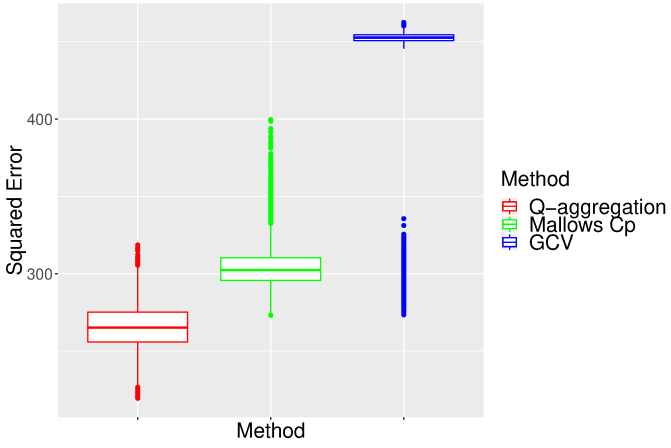

Figure 4. boxplots of squared errors from experiment 2

Acknowledgements

P. C. B.'s research was partially supported by NSF Grants DMS-1811976 and DMS-1945428. D. Y. is supported by NSF Grant CCF-1850743.

References

Arlot, S. Minimal penalties and the slope heuristics: a survey. *arXiv preprint arXiv:1901.07277*, 2019.

Arlot, S. and Bach, F. R. Data-driven calibration of linear estimators with minimal penalties. In *Advances in Neural Information Processing Systems*, pp. 46–54, 2009.

Arlot, S. and Celisse, A. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

Bellec, P. C. Optimal bounds for aggregation of affine estimators. *Ann. Statist.*, 46(1):30–59, 02 2018. doi: 10.1214/17-AOS1540. URL https://arxiv.org/pdf/1410.0346.pdf.

Belloni, A., Chernozhukov, V., and Wang, L. Pivotal estimation via square-root lasso in nonparametric regression. *Ann. Statist.*, 42(2):757–788, 04 2014. URL http://dx.doi.org/10.1214/14-AOS1204.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2009.

Brown, L. D., Levine, M., et al. Variance estimation in nonparametric regression via the difference sequence method. *The Annals of Statistics*, 35(5):2219–2232, 2007.

Chernousova, E., Golubev, Y., Krymova, E., et al. Ordered smoothers with exponential weighting. *Electronic journal of statistics*, 7:2395–2419, 2013.

Cohen, A. All admissible linear estimates of the mean vector. *The Annals of Mathematical Statistics*, pp. 458–463, 1966.

Craven, P. and Wahba, G. Smoothing noisy data with spline functions. *Numerische mathematik*, 31(4):377–403, 1978.

Dai, D., Rigollet, P., and Zhang, T. Deviation optimal learning using greedy Q-aggregation. *The Annals of Statistics*, 40(3):1878–1905, 2012.

Dai, D., Rigollet, P., L., X., and T., Z. Aggregation of affine estimators. *Electon. J. Stat.*, 8:302–327, 2014.

Dalalyan, A. S. and Salmon, J. Sharp oracle inequalities for aggregation of affine estimators. *The Annals of Statistics*, 40(4):2327–2355, 2012.

Dette, H., Munk, A., and Wagner, T. Estimating the variance in nonparametric regression–what is a reasonable choice? *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(4):751–764, 1998.

Frank, M., Wolfe, P., et al. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2): 95–110, 1956.

Golub, G. H., Heath, M., and Wahba, G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Golubev, Y. et al. On universal oracle inequalities related to high-dimensional linear models. *The Annals of Statistics*, 38(5):2751–2780, 2010.

Hall, P., Kay, J., and Titterinton, D. Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77(3):521–528, 1990.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, 2001.

Kneip, A. Ordered linear smoothers. *The Annals of Statistics*, 22(2):835–866, 1994.

Künzel, S. R., Pollard, D., and Yang, D. Remarks on kneip's linear smoothers. *arXiv preprint arXiv:1405.1744*, 2014.

Leung, G. and Barron, A. R. Information theory and mixing least-squares regressions. *Information Theory, IEEE Transactions on*, 52(8):3396–3410, 2006.

Li, K.-C. Asymptotic optimality of c_l and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics*, 14(3):1101–1112, 1986.

Mallows, C. L. Some comments on c p. *Technometrics*, 15 (4):661–675, 1973.

Munk, A., Bissantz, N., Wagner, T., and Freitag, G. On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):19–41, 2005.

Nemirovski, A. Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998)*, volume 1738 of *Lecture Notes in Mathematics*. Springer, Berlin, 2000.

Nocedal, J. and Wright, S. *Numerical optimization*. Springer Science & Business Media, 2006.

Owen, A. B. A robust hybrid of lasso and ridge regression. Technical report, Stanford University, 2007.

Rigollet, P. Kullback–Leibler aggregation and misspecified generalized linear models. *Ann. Statist.*, 40(2):639–665, 2012. doi: 10.1214/11-AOS961.

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

Sun, T. and Zhang, C.-H. Scaled sparse linear regression. *Biometrika*, 2012.

Tsybakov, A. Aggregation and minimax optimality in high dimensional estimation. *Proceedings of International Congress of Mathematicians (Seoul, 2014)*, 3:225–246, 2014.

Tsybakov, A. B. Optimal rates of aggregation. In *Learning Theory and Kernel Machines*, pp. 303–313. Springer, 2003.

van Handel, R. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

Vavasis, S. A. *Complexity Theory: Quadratic Programming*, pp. 304–307. Springer US, Boston, MA, 2001. ISBN 978-0-306-48332-5. doi: 10.1007/0-306-48332-7_65. URL https://doi.org/10.1007/0-306-48332-7_65.