Multi-task Regression using Minimal Penalties

Matthieu Solnon[*] matthieu.solnon@ens.fr
ENS; Sierra Project-team
Laboratoire d'Informatique de l'École Normale Supérieure
(CNRS/ENS/INRIA UMR 8548)
23, avenue d'Italie, CS 81321
75214 Paris Cedex 13, France

Sylvain Arlot[†] sylvain.arlot@ens.fr
CNRS; Sierra Project-team
Laboratoire d'Informatique de l'École Normale Supérieure
(CNRS/ENS/INRIA UMR 8548)
23, avenue d'Italie, CS 81321
75214 Paris Cedex 13, France

Francis Bach[‡] francis.bach@ens.fr
INRIA; Sierra Project-team
Laboratoire d'Informatique de l'École Normale Supérieure
(CNRS/ENS/INRIA UMR 8548)
23, avenue d'Italie, CS 81321
75214 Paris Cedex 13, France

Editor:

Abstract

In this paper we study the kernel multiple ridge regression framework, which we refer to as multi-task regression, using penalization techniques. The theoretical analysis of this problem shows that the key element appearing for an optimal calibration is the covariance matrix of the noise between the different tasks. We present a new algorithm to estimate this covariance matrix, based on the concept of minimal penalty, which was previously used in the single-task regression framework to estimate the variance of the noise. We show, in a non-asymptotic setting and under mild assumptions on the target function, that this estimator converges towards the covariance matrix. Then plugging this estimator into the corresponding ideal penalty leads to an oracle inequality. We illustrate the behavior of our algorithm on synthetic examples.

Keywords: multi-task, oracle inequality, learning theory.

1. Introduction

A classical paradigm in statistics is that increasing the sample size (that is, the number of observations) improves the performance of the estimators. However, in some cases it may be impossible to increase this sample size, for instance because of experimental limitations.

∗. http://www.di.ens.fr/∼solnon/
†. http://www.di.ens.fr/∼arlot/
‡. http://www.di.ens.fr/∼fbach/

Hopefully, in many situations practictioners can find many related and similar problems, and might want to use those other problems as if it gave more observations for his initial problem. The techniques using this heuristic are called "multi-task" techniques. In this paper we study the kernel ridge regression procedure in a multi-task framework.

One-dimensional kernel ridge regression, which we refer to as "single-task" regression has been widely studied. As we briefly review in Section 3 one has, given n data points $(X_i, Y_i)_{i=1}^n$, to estimate a function f, often the conditional expectation $f(X_i) = \mathbb{E}[X_i|Y_i]$, by minimizing the quadratic risk of the estimator regularized by a certain norm. A practically important task is to calibrate a regularization parameter, i.e., to estimate the regularization parameter directly from data. For kernel ridge regression (a.k.a. smoothing splines), many methods have been proposed based on different principles, e.g., Bayesian criteria through a Gaussian process interpretation (see, e.g., Rasmussen and Williams, 2006) or generalized cross-validation (see, e.g., Wahba, 1990). In this paper, we focus on the concept of minimal penalty, which was first introduced by Birgé and Massart (2007) and Arlot and Massart (2009) for model selection, then extended to linear estimators such as kernel ridge regression by Arlot and Bach (2011).

In this article we consider $p \geq 2$ different (but related) regression tasks, a framework we refer to as "multi-task" regression. This setting has already been studied in different papers. Some of those (Thrun and O'Sullivan, 1996; Caruana, 1997; Bakker and Heskes, 2003) empirically show that it can lead to performance improvement. Liang et al. (2010) also obtained a theoretical criterion (unfortunately non observable) which tells when this phenomenon asymptotically occurs. Several different paths have been followed to deal with this setting. Some (see for instance Obozinski et al., 2011; Lounici et al., 2010), consider a setting where $p \gg n$, and formulate a sparsity assumption which enables them to use the group Lasso, assuming that all the different functions have a small set of common active covariates. We exclude this setting from our analysis, because of the kernel nature of our problem, and thus will not consider the similarity between the tasks in terms of sparsity, but rather in terms of an Euclidiean similarity. An other theoretical approach has been also taken (see for example, Brown and Zidek (1980), Evgeniou et al. (2005) or Ando and Zhang (2005) on semi-supervised learning), the authors often defining a theoretical framework where the multi-task problem can easily be expressed, and where sometimes solutions can be computed. The main remaining theoretical problem is the calibration of a matricial parameter (typically of size p), which characterizes the relationship bewteen the tasks and extends the regularization parameter from the single-task regression. Because of the high dimensional nature of the problem (i.e., the small number of training observations) usual techniques, like cross-validation, are not likely to succeed. Argyriou et al. (2008) have a similar approach to ours, but solve this problem by adding a convex constraint to the matrix, which will be discussed at the end of Section 5. Through a penalization technique we show in Section 2 that the only element we have to estimate is the correlation matrix Σ of the noise between the tasks. We give here a new algorithm to estimate Σ, and show that the estimation is sharp enough to derive an oracle inequality, both with high probability and in expectation. Finally we give some simulation experiment results and show that our technique correctly deals with the multi-tasks settings with a low sample-size.

Notations. We now introduce some notations, which will be used throughout the article.

- The integer n is the sample size, the integer p is the number of tasks.
- For any $n \times p$ matrix Y, we define

$$y = \mathrm{vec}(Y) := (Y_{1,1}, \ldots, Y_{n,1}, Y_{1,2}, \ldots, Y_{n,2}, \ldots, Y_{1,p}, \ldots, Y_{n,p}) \in \mathbb{R}^{np},$$

  that is, the columns $Y^j := (Y_{i,j})_{1 \leq i \leq n}$ are stacked.
- $\mathcal{M}_n(\mathbb{R})$ is the set of all matrices of size n.
- $\mathcal{S}_p(\mathbb{R})$ is the set of symmetric matrices of size p.
- $\mathcal{S}_p^+(\mathbb{R})$ is the set of symmetric positive-semidefinite matrices of size p.
- $\mathcal{S}_p^{++}(\mathbb{R})$ is the set of symmetric positive-definite matrices of size p.
- $\preceq$ denotes the partial ordering on $\mathcal{S}_p(\mathbb{R})$ defined by: $A \preceq B$ if and only if $B - A \in \mathcal{S}_p^+(\mathbb{R})$.
- $\mathbf{1}$ is the vector of size p whose components are all equal to 1.
- $\|\cdot\|_2$ is the usual Euclidean norm on $\mathbb{R}^k$ for any $k \in \mathbb{N}$: $\forall u \in \mathbb{R}^k$, $\|u\|_2^2 := \sum_{i=1}^k u_i^2$.

2. Multi-task regression: problem set-up

We consider p kernel ridge regression tasks. Treating them simultaneously and sharing their common structure (e.g., being close in some metric space) will help in reducing the overall prediction error.

Let $\mathcal{X}$ be some set and $\mathcal{F}$ a set of real-valued functions over $\mathcal{X}$. We suppose $\mathcal{F}$ has a reproducing kernel Hilbert space (RKHS) structure (Aronszajn, 1950), with kernel k and feature map $\Phi : \mathcal{X} \to \mathcal{F}$. We observe $\mathcal{D}_n = (X_i, Y_i^1, \ldots, Y_i^p)_{i=1}^n \in (\mathcal{X} \times \mathbb{R}^p)^n$, which give us the positive semidefinite kernel matrix $K = (k(X_i, X_j))_{1 \leq i,j \leq n} \in \mathcal{S}_n^+(\mathbb{R})$. For each task $j \in \{1, \ldots, p\}$, $\mathcal{D}_n^j = (X_i, y_i^j)_{i=1}^n$ is a sample with distribution $\mathcal{P}_j$, for which a simple regression problem has to be solved. In this paper we consider for simplicity that the different tasks have the same design $(X_i)_{i=1}^n$. When the designs of the different tasks are different the analysis is similar, but the notations would be more complicated.

We now define the model. We assume $(f^1, \ldots, f^p) \in \mathcal{F}^p$, Σ is a symmetric positive-definite matrix of size p such that the vectors $(\varepsilon_i^j)_{j=1}^p$ are i.i.d. with normal distribution $\mathcal{N}(0, \Sigma)$, with mean zero and covariance matrix Σ, and

$$\forall i \in \{1, \ldots, n\}, \forall j \in \{1, \ldots, p\},\ y_i^j = f^j(X_i) + \varepsilon_i^j \ .$$

This means that, while the observations are independent, the different tasks are correlated, with correlation matrix Σ between the tasks. We now place ourselves in the fixed-design setting, that is, $(X_i)_{i=1}^n$ is deterministic and the goal is to estimate $\big(f^1(X_i), \ldots, f^p(X_i)\big)_{i=1}^p$. Let us introduce some notation:

- $\mu_{\min} = \mu_{\min}(\Sigma)$ (resp. $\mu_{\max}$) denotes the smallest (resp. largest) eigenvalue of Σ.
- $c(\Sigma) := \mu_{\max}/\mu_{\min}$ is the condition number of Σ.

To obtain compact equations, we will use the following definition:

Definition 1. *We denote by F the $n \times p$ matrix $(f^j(X_i))_{1 \leq i \leq n\,,\, 1 \leq j \leq p}$ and introduce the vector $f := \mathrm{vec}(F) = (f^1(X_1), \ldots, f^1(X_n), \ldots, f^p(X_n)) \in \mathbb{R}^{np}$, obtained by stacking the columns of F. Similarly we define $Y := (y_i^j) \in \mathcal{M}_{n \times p}(\mathbb{R})$, $y := \mathrm{vec}(Y)$, $E := (\varepsilon_i^j) \in \mathcal{M}_{n \times p}(\mathbb{R})$ and $\varepsilon := \mathrm{vec}(E)$.*

In order to estimate f, we use a regularization procedure, which extends the classical ridge regression of the single-task setting. Let M be a $p \times p$ matrix, symmetric and positive-definite. Generalising the work of Evgeniou et al. (2005), we estimate $f = (f^1, \ldots, f^p) \in \mathcal{F}^p$ by

$$\widehat{f}_M \in \operatorname*{argmin}_{g \in \mathcal{F}^p} \left\{ \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} (y_i^j - g^j(X_i))^2 + \sum_{j=1}^{p} \sum_{\ell=1}^{p} M_{j,l} \langle g^j, g^\ell \rangle_{\mathcal{F}} \right\} \,. \tag{2.1}$$

Remark 1. *Requiring that $M \succeq 0$ implies that Eq. (2.1) is a convex optimization problem, which here, because we consider the square loss, can be solved through the resolution of a linear system, as explained later. Moreover it allows an RKHS interpretation, which will also be explained later.*

Example 1. *The case where the p tasks are treated independently can be considered in this setting: taking $M = M_{\mathrm{ind}}(\lambda) := \mathrm{Diag}(\lambda_1, \ldots, \lambda_p)$ for any $\lambda \in \mathbb{R}^p$, which leads to the criterion*

$$\frac{1}{p} \sum_{j=1}^{p} \left[\frac{1}{n} \sum_{i=1}^{n} (y_i^j - g^j(X_i))^2 + \lambda_j \|g^j\|_{\mathcal{F}}^2 \right] \,, \tag{2.2}$$

that is, the sum of the single-task criteria described in Section 3. Hence, minimizing Eq. (2.2) over $\lambda \in \mathbb{R}^p$ amounts to solve independently *p single task problems.*

Example 2. *As done by Evgeniou et al. (2005), for every $\lambda, \mu \in (0, +\infty)^2$, define*

$$M_{\mathrm{similar}}(\lambda, \mu) := (\lambda + p\mu) I_p - \mu \mathbf{1}\mathbf{1}^\top = \begin{pmatrix} \lambda + (p-1)\mu & & -\mu \\ & \ddots & \\ -\mu & & \lambda + (p-1)\mu \end{pmatrix} \,. \tag{2.3}$$

Taking $M = M_{\mathrm{similar}}(\lambda, \mu)$ in Eq. (2.1) leads to the criterion

$$\frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} (y_i^j - g^j(X_i))^2 + \lambda \sum_{j=1}^{p} \|g^j\|_{\mathcal{F}}^2 + \mu \sum_{j=1}^{p} \sum_{k=1}^{p} \left\| g^j - g^k \right\|_{\mathcal{F}}^2 \,. \tag{2.4}$$

Minimizing Eq. (2.4) enforces a regularization on both the norms of the functions g^j and the norms of the differences $g^j - g^k$. Thus, matrices of the form $M_{\mathrm{similar}}(\lambda, \mu)$ are useful when the functions g^j are assumed to be similar in $\mathcal{F}$. One of the main contribution of the paper is to go beyond this case and learn from data a a similarity matrix M between tasks.

Example 3. *We extend Example 2 to the case where the p tasks consist of two groups of close tasks. Let I be a subset of $\{1, \ldots, p\}$, of cardinality $1 \leq k \leq p-1$. Let us denote*

by I^c the complementary of I in $\{1,\dots,p\}$, $\mathbf{1}_I$ the vector v with component $v_i = \mathbf{1}_{i\in I}$, and $\mathrm{Diag}(I)$ *the diagonal matrix d with component $d_{i,i} = \mathbf{1}_{i\in I}$. We then define*

$$M_I(\lambda,\mu,\nu) := \lambda I_p + \mu\,\mathrm{Diag}(I) + \nu\,\mathrm{Diag}(I^c) - \frac{\mu}{k}\mathbf{1}_I\mathbf{1}_I^\top - \frac{\nu}{p-k}\mathbf{1}_{I^c}\mathbf{1}_{I^c}^\top \quad . \tag{2.5}$$

This matrix leads to the following criterion, which enforces a regularization on both the norms of the functions g^j and the norms of the differences $g^j - g^k$ inside the groups I and I^c :

$$\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(y_i^j - g^j(X_i))^2 + \lambda\sum_{j=1}^{p}\left\|g^j\right\|_{\mathcal{F}}^2 + \frac{\mu}{k}\sum_{j\in I}\sum_{k\in I}\left\|g^j - g^k\right\|_{\mathcal{F}}^2 + \frac{\nu}{p-k}\sum_{j\in I^c}\sum_{k\in I^c}\left\|g^j - g^k\right\|_{\mathcal{F}}^2 \quad . \tag{2.6}$$

As shown in Section 6, we can estimate the set I from data (see Jacob et al. (2008) for a more general formulation).

Remark 2. *Since I_p and $\mathbf{1}\mathbf{1}^\top$ can be diagonalized simultaneously, minimizing Eq.* (2.4) *and Eq.* (2.6) *is quite easy: it only demands optimization over two independent parameters, which can be done with the procedure of Arlot and Bach (2011).*

Remark 3. *As stated below (Proposition 2), M acts as a scalar product between the tasks. Selecting a general matrix M is thus a way to express a similarity between tasks.*

Following Evgeniou et al. (2005), we define the vector-space $\mathcal{G}$ of real-valued functions over $\mathcal{X}\times\{1,\dots,p\}$ by

$$\mathcal{G} := \{g : \mathcal{X}\times\{1,\dots,p\} \to \mathbb{R} \,/\, \forall j \in \{1,\dots,p\}\ ,\ g(\cdot,j)\in\mathcal{F}\} \quad .$$

We now define a bilinear symmetric form over $\mathcal{G}$,

$$\forall g,h\in\mathcal{G}\ \ ,\quad \langle g,h\rangle_{\mathcal{G}} := \sum_{j=1}^{p}\sum_{l=1}^{p} M_{j,l}\langle g(\cdot,j), h(\cdot,l)\rangle_{\mathcal{F}},$$

which is a scalar product (see proof in Appendix A) and leads to a RKHS (see proof in Appendix B):

Proposition 2. *With the preceding notations $\langle\cdot,\cdot\rangle_{\mathcal{G}}$ is a scalar product on $\mathcal{G}$.*

Corollary 1. *$(\mathcal{G},\langle\cdot,\cdot\rangle_{\mathcal{G}})$ is a RKHS.*

In order to write down the kernel matrix in compact form, we introduce the following notations.

Definition 3 (Kronecker Product)**.** *Let $A\in\mathcal{M}_{m,n}(\mathbb{R})$, $B\in\mathcal{M}_{p,q}(\mathbb{R})$. We define the Kronecker product $A\otimes B$ as being the $(mp)\times(nq)$ matrix built with $p\times q$ blocs, the block of index (i,j) being $A_{i,j}\cdot B$:*

$$A\otimes B = \begin{pmatrix} A_{1,1}B & \dots & A_{1,n}B \\ \vdots & \ddots & \vdots \\ A_{m,1}B & \dots & A_{m,n}B \end{pmatrix} \quad .$$

The Kronecker product is a widely used tool to deal with matrices and tensor products. Some of its classical properties are given in Section D; see also Horn and Johnson (1991).

Proposition 4. *The kernel matrix associated with the design* $\widetilde{X} := (X_i, j)_{i,j} \in \mathcal{X} \times \{1, \dots, p\}$ *and the RKHS* $(\mathcal{G}, \langle \cdot, \cdot \rangle_{\mathcal{G}})$ *is* $\widetilde{K}_M := M^{-1} \otimes K$.

We can then apply the representer's theorem (Schölkopf and Smola, 2002) to the minimization problem (2.1) and deduce that $\widehat{f}_M = A_M y$ with

$$A_M = A_{M,K} := \widetilde{K}_M (\widetilde{K}_M + npI_{np})^{-1} = (M^{-1} \otimes K) \left((M^{-1} \otimes K) + nI_{np} \right)^{-1} \ .$$

Now when working in multi-task regression, a set $\mathcal{M} \subset \mathcal{S}_p^{++}(\mathbb{R})$ of matrices M is given, and the goal is to select the "best" one, that is, minimizing over M the quadratic risk $n^{-1} \| \widehat{f}_M - f \|_2^2$. For instance, the single-task framework corresponds to $p = 1$ and $\mathcal{M} = (0, +\infty)$. The multi-task case is far richer. The ideal choice, called the oracle, is

$$M^\star \in \operatorname*{argmin}_{M \in \mathcal{M}} \left\{ \left\| \widehat{f}_M - f \right\|_2^2 \right\} \ .$$

However $M^\star$ is not an estimator, since it depends on f. As explained by Arlot and Bach (2011), we choose $\widehat{M}$ as a minimizer over $\mathcal{M}$ of

$$\operatorname{crit}(M) = \frac{1}{np} \left\| y - \widehat{f}_M \right\|_2^2 + \operatorname{pen}(M) \ ,$$

where the penalty term $\operatorname{pen}(M)$ has to be chosen appropriately. The unbiased risk estimation principle (introduced by Akaike, 1970) requires

$$\mathbb{E}\left[\operatorname{crit}(M) \right] \approx \mathbb{E}\left[\frac{1}{np} \left\| \widehat{f}_M - f \right\|_2^2 \right] \ ,$$

which leads to the (deterministic) *ideal penalty*

$$\operatorname{pen}_{\mathrm{id}}(M) := \mathbb{E}\left[\frac{1}{np} \| \widehat{f}_M - f \|_2^2 \right] - \mathbb{E}\left[\frac{1}{np} \left\| y - \widehat{f}_M \right\|_2^2 \right] \ .$$

Since $\widehat{f}_M = A_M y$ and $y = f + \varepsilon$, we can write

$$\left\| \widehat{f}_M - y \right\|_2^2 = \left\| \widehat{f}_M - f \right\|_2^2 + \|\varepsilon\|_2^2 - 2\langle \varepsilon, A_M \varepsilon \rangle + 2\langle \varepsilon, (I_{np} - A_M) f \rangle \ .$$

Since ε is centered and M is deterministic, we get, up to an additive factor independent of M,

$$\operatorname{pen}_{\mathrm{id}}(M) = \frac{2\mathbb{E}\left[\langle \varepsilon, A_M \varepsilon \rangle \right]}{np} \ ,$$

that is, as the covariance matrix of ε is $\Sigma \otimes I_n$,

$$\operatorname{pen}_{\mathrm{id}}(M) = \frac{2 \operatorname{tr}\left(A_M \cdot (\Sigma \otimes I_n) \right)}{np} \ . \tag{2.7}$$

In order to approach this penalty as precisely as possible, we have to sharply estimate Σ. In the single-task case, such a problem reduces to estimating the variance σ^2 of the noise and was tackled by Arlot and Bach (2011). Since our approach for estimating Σ heavily relies on these results, they are summarized in the next section.

3. Single task framework: estimating a single variance

This section recalls some of the main results from Arlot and Bach (2011) and can be considered as a special case of Section 2, with $p = 1$, $\Sigma = \sigma^2 > 0$ and $\mathcal{M} = [0, +\infty]$. Writing $M = \lambda$ with $\lambda \in [0, +\infty]$, the regularization matrix is

$$\forall \lambda \in (0, +\infty), \quad A_\lambda = A_{\lambda,K} = K(K + n\lambda I_n)^{-1} \ , \tag{3.1}$$

$A_0 = I_n$ and $A_{+\infty} = 0$; the ideal penalty becomes

$$\mathrm{pen}_{\mathrm{id}}(\lambda) = \frac{2\sigma^2 \operatorname{tr}(A_\lambda)}{n} \ .$$

By analogy with the case where A_λ is an orthogonal projection matrix, $\mathrm{df}(\lambda) := \operatorname{tr}(A_\lambda)$ is called the effective degree of freedom, first introduced by Hastie and Tibshirani (1990) and generalized by Zhang (2005). The ideal penalty however depends on σ^2; in order to have a fully data-driven penalty we have to replace σ^2 by an estimator $\widehat{\sigma}^2$ inside $\mathrm{pen}_{\mathrm{id}}(\lambda)$. For every $\lambda \in [0, +\infty]$, define

$$\mathrm{pen}_{\min}(\lambda) = \mathrm{pen}_{\min}(\lambda, K) := \frac{(2\operatorname{tr}(A_{\lambda,K}) - \operatorname{tr}(A_{\lambda,K}^\top A_{\lambda,K}))}{n} \ .$$

Theoretical arguments show that when a penalty proportionnal to $\mathrm{pen}_{\min}(\lambda)$ is chosen, then if the proportionnality coefficient is smaller than σ^2/n the procedure overfits, while when this coefficient is greater than σ^2/n the procedure leads to good estimation properties and a low effective degree of freedom. The following algorithm was introduced in Arlot and Bach (2011) and uses this fact to estimate σ^2.

Algorithm 1. **Input:** $Y \in \mathbb{R}^n$, $K \in \mathcal{S}_n^{++}(\mathbb{R})$

1. *For every $C > 0$, compute*

$$\widehat{\lambda}_0(C) \in \operatorname*{argmin}_{\lambda \in [0,+\infty]} \left\{ \frac{1}{n} \left\| A_{\lambda,K} Y - Y \right\|_2^2 + C \, \mathrm{pen}_{\min}(\lambda, K) \right\} \ .$$

2. **Output:** *$\widehat{C}$ such that $\mathrm{df}(\widehat{\lambda}_0(\widehat{C})) \in [n/10, n/3]$.*

An efficient algorithm for the first step of Algorithm 1 is detailed in Arlot and Massart (2009), and we discuss the way we implemented Algorithm 1 in Section 6. The output $\widehat{C}$ of Algorithm 1 is a provably consistent estimator of σ^2, as stated in the following theorem.

Theorem 5 (Corollary of Theorem 1 of Arlot and Bach (2011))**.** *Let $\beta = 150$ and $\alpha = 2$. Suppose $\varepsilon \in \mathcal{N}(0, \sigma^2 I_n)$ with $\sigma^2 > 0$, and that $\lambda_0 \in (0, +\infty)$ and $d_n \geq 1$ exist such that*

$$\mathrm{df}(\lambda_0) \leq \sqrt{n} \text{ and } \frac{1}{n} \left\| (A_{\lambda_0} - I_n) F \right\|_2^2 \leq d_n \sigma^2 \sqrt{\frac{\ln n}{n}} \ . \tag{3.2}$$

Then for every $\delta \geq 2$, some constants $n_0(\delta), \kappa > 0$ and an event Ω exist such that $\mathbb{P}(\Omega) \geq 1 - \kappa n^{\delta}$ and for $n \geq n_0(\delta)$, on Ω,

$$\left(1 - \beta(\alpha + \delta) \sqrt{\frac{\ln n}{n}} \right) \sigma^2 \leq \widehat{C} \leq \left(1 + \beta(\alpha + \delta) d_n \sqrt{\frac{\ln(n)}{n}} \right) \sigma^2 \ . \tag{3.3}$$

Remark 4. *The values $n/10$ and $n/3$ have no particular meaning and can be replaced by n/k, n/k', with $k > k' > 2$. Only β depends on k and k'.*

4. Estimation of the noise covariance matrix Σ

Thanks to the results developped by Arlot and Bach (2011) (recapitulated in Section 3), we know how to estimate a variance for any one-dimensional problem. In order to estimate Σ, which has $p(p+1)/2$ parameters, we can use several one-dimensional problems. Projecting Y onto some direction $z \in \mathbb{R}^p$ yields

$$Y_z := Y \cdot z = F \cdot z + E \cdot z = F_z + \varepsilon_z \ , \qquad \textbf{(Pz)}$$

with $\varepsilon_z \sim \mathcal{N}(0, \sigma_z^2 I_n)$ and $\sigma_z^2 := \mathrm{Var}[\varepsilon \cdot z] = z^\top \Sigma z$. Therefore, we will estimate σ_z^2 for $z \in \mathcal{Z}$ a well chosen set, and use these estimators to build back an estimation of Σ.

We now explain how to estimate Σ using those one-dimensional projections.

Definition 6. *Let $a(z)$ be the output $\widehat{C}$ of Algorithm 1 applied to problem* (**Pz**)*, that is, with input $Y_z \in \mathbb{R}^n$ and $K \in \mathcal{S}_n^{++}(\mathbb{R})$.*

The idea is to apply Algorithm 1 to the elements z of a carefully chosen set $\mathcal{Z}$. Noting e_i the i-th vector of the canonical basis of $\mathbb{R}^p$, we introduce $\mathcal{Z} = \{e_i,\ i \in \{1, \dots, p\}\} \cup \{e_i + e_j,\ 1 \leq i < j \leq p\}$. We can see that $a(e_i)$ estimates $\Sigma_{i,i}$, while $a(e_i + e_j)$ estimates $\Sigma_{i,i} + \Sigma_{j,j} + 2\Sigma_{i,j}$. Henceforth, $\Sigma_{i,j}$ can be estimated by $(a(e_i + e_j) - a(e_i) - a(e_j))/2$. This leads to the definition of the following map J, which builds a symmetric matrix using the latter construction.

Definition 7. *Let $J : \mathbb{R}^{\frac{p(p+1)}{2}} \to \mathcal{S}_p(\mathbb{R})$ be defined by*

$$J(a_1, \dots, a_p, a_{1,2}, \dots, a_{1,p}, \dots, a_{p-1,p})_{i,i} = a_i \ \text{if}\ 1 \leq i \leq p \ ,$$
$$J(a_1, \dots, a_p, a_{1,2}, \dots, a_{1,p}, \dots, a_{p-1,p})_{i,j} = \frac{a_{i,j} - a_i - a_j}{2} \ \text{if}\ \leq i < j \leq p \ .$$

This map is bijective, and for all $B \in \mathcal{S}_p(\mathbb{R})$

$$J^{-1}(B) = (B_{1,1}, \dots, B_{p,p}, B_{1,1} + B_{2,2} + 2B_{1,2}, \dots, B_{p-1,p-1} + B_{p,p} + 2B_{p-1,p}) \ .$$

This leads us to defining the following estimator of Σ :

$$\widehat{\Sigma} := J\left(a(e_1), \dots, a(e_p), a(e_1 + e_2), \dots, a(e_1 + e_p), \dots, a(e_{p-1} + e_p)\right) \ . \qquad (4.1)$$

Let us recall that $\forall \lambda \in (0, +\infty)$, $A_\lambda = A_{\lambda,K} = K(K + n\lambda I_n)^{-1}$. Following Arlot and Bach (2011) we make the following assumption from now on:

$$\left.\begin{array}{l} \forall j \in \{1, \dots, p\}\,,\ \exists \lambda_{0,j} \in (0, +\infty)\,, \\ \qquad \mathrm{df}(\lambda_{0,j}) \leq \sqrt{n} \quad \text{and} \quad \dfrac{1}{n}\left\|(A_{\lambda_{0,j}} - I_n) F_{e_j}\right\|_2^2 \leq \Sigma_{j,j}\sqrt{\dfrac{\ln n}{n}} \end{array}\right\} \qquad \textbf{(Hdf)}$$

We can now state the main result of the paper.

Theorem 8. *Let $\widehat{\Sigma}$ be defined by Eq.* (4.1)*, $\alpha = 2$, $\kappa > 0$ be the numerical constant defined in Theorem 5 and assume* (**Hdf**) *holds. For every $\delta \geq 2$, a constant $n_0(\delta)$, an absolute*

constant $L_1 > 0$ and an event Ω exist such that $\mathbb{P}(\Omega) \geq 1 - \kappa p^2 n^{-\delta}$ and for every $n \geq n_0(\delta)$, on Ω,

$$(1-\eta)\Sigma \preceq \widehat{\Sigma} \preceq (1+\eta)\Sigma \ , \tag{4.2}$$
$$\text{where} \qquad \eta := L_1(\alpha+\delta)p\sqrt{\frac{\ln(n)}{n}}c(\Sigma)^2 \ .$$

Theorem 8 is proved in Section D. It shows $\widehat{\Sigma}$ estimates Σ with a "multiplicative" error controlled with large probability, in a non-asymptotic setting. The multiplicative nature of the error is crucial for deriving the oracle inequality stated in Section 5, since it allows to show the ideal penalty defined in Eq. (2.7) is precisely estimated when Σ is replaced by $\widehat{\Sigma}$.

An important feature of Theorem 8 is that it holds under very mild assumptions on the mean f of the data (see Remark 5). Therefore, it shows $\widehat{\Sigma}$ is able to estimate a covariance matrix *without prior knowledge on the regression function*, which, to the best of our knowledge, has never been obtained in multi-task regression/

Remark 5 (On assumption (**Hdf**))**.** *Assumption* (**Hdf**) *is a single-task assumption (made independently for each task). The upper bound $\sqrt{\ln(n)/n}$ can be multiplied by any factor $1 \leq d_n \ll \sqrt{n/\ln(n)}$ (as in Theorem 5), at the price of multiplying η_1 by d_n in the upper bound of Eq.* (4.2)*.*

Assumption (**Hdf**) *is rather classical in model selection, see Arlot and Bach (2011) for instance. In particular, (a weakened version of)* (**Hdf**) *holds if the bias $n^{-1}\|(A_\lambda - I_n)F_{e_i}\|_2^2$ is bounded by $C_1 \operatorname{tr}(A_\lambda)^{-C_2}$, for some $C_1, C_2 > 0$.*

Remark 6 (Scaling of (n,p) for consistency)**.** *A sufficient condition for ensuring $\widehat{\Sigma}$ is a consistent estimator of Σ is*

$$pc(\Sigma)^2\sqrt{\frac{\ln(n)}{n}} \longrightarrow 0 \ ,$$

which enforces a scaling between n, p and $c(\Sigma)$. Nevertheless, this condition is probably not necessary since the simulation experiments of Section 6 show that Σ can be well estimated (at least for estimator selection purposes) in a setting where $\eta \gg 1$.

Remark 7 (Choice of the set $\mathcal{Z}$)**.** *Other choices could have been made for $\mathcal{Z}$, however ours seems easier in terms of computation, since $|\mathcal{Z}| = p(p+1)/2$. Choosing a larger set $\mathcal{Z}$ leads to theoretical difficulties in the reconstruction of $\widehat{\Sigma}$, while taking other basis vectors leads to more complex computations. We can also note that increasing $|\mathcal{Z}|$ decreases the probability in Theorem 8, since it comes from an union bound over the one-dimensional estimations.*

5. Oracle inequality

We now show that the estimator introduced in Eq. (4.1) is precise enough to derive an oracle inequality when plugged in the penalty defined in Eq. (2.7).

Definition 9. *Let $\widehat{\Sigma}$ be the estimator of Σ defined by Eq.* (4.1) *. We define*

$$\widehat{M} \in \operatorname*{argmin}_{M \in \mathcal{M}} \left\{ \left\| \widehat{f}_M - y \right\|_2^2 + 2\operatorname{tr}\left(A_M \cdot (\widehat{\Sigma} \otimes I_n) \right) \right\} \ . \tag{5.1}$$

We assume the following assumption, which means that the matrices of $\mathcal{M}$ are jointly diagonalisable, holds true:

$$\exists P \in O_p(\mathbb{R})\,, \quad \mathcal{M} \subseteq \left\{ P^\top \operatorname{Diag}(d_1, \dots, d_p) P\,,\, (d_i)_{i=1}^p \in (0, +\infty)^p \right\}\,. \tag{\textbf{HM}}$$

Theorem 10. *Let $\alpha = 2$, $\delta \geq 2$ and assume* (**Hdf**) *and* (**HM**) *hold true.Absolute constants $L_2 > 0$ and κ', a constant $n_1(\delta)$ and an event $\widetilde{\Omega}$ exist such that $\mathbb{P}(\widetilde{\Omega}) \geq 1 - \kappa' p^2 n^{-\delta}$ and the following holds as soon as $n \geq n_1(\delta)$.First, on $\widetilde{\Omega}$,*

$$\frac{1}{np}\left\|\widehat{f}_{\widehat{M}} - f\right\|_2^2 \leq \left(1 + \frac{1}{\ln(n)}\right)^2 \inf_{M \in \mathcal{M}} \left\{ \frac{1}{np}\left\|\widehat{f}_M - f\right\|_2^2 \right\} + L_2 c(\Sigma)^4 \operatorname{tr}(\Sigma)(\alpha+\delta)^2 \frac{p^3 \ln(n)^3}{np}\,. \tag{5.2}$$

Second,

$$\begin{aligned} \mathbb{E}\left[\frac{1}{np}\left\|\widehat{f}_{\widehat{M}} - f\right\|_2^2\right] \leq \left(1 + \frac{1}{\ln(n)}\right)^2 \mathbb{E}\left[\inf_{M \in \mathcal{M}} \left\{ \frac{1}{np}\left\|\widehat{f}_M - f\right\|_2^2 \right\}\right] \\ + L_2 c(\Sigma)^4 \operatorname{tr}(\Sigma)(\alpha+\delta)^2 \frac{p^3 \ln(n)^3}{np} + \frac{p}{n^{\delta/2}} \frac{\|f\|_2^2}{np}\,. \end{aligned} \tag{5.3}$$

Theorem 10 is proved in Section E. Taking $p = 1$ (hence $c(\Sigma) = 1$ and $\operatorname{tr}(\Sigma) = \sigma^2$), we recover Theorem 3 of Arlot and Bach (2011) as a corollary.

Remark 8. *Our result is a non asymptotic oracle inequality, with a mutliplicative term of the form $1 + o(1)$. This allows us to claim that our selection procedure is nearly optimal, since our estimator is close (with regard to the empirical quadratic norm) to the oracle one. Furthermore the term $1 + (\ln(n))^{-1}$ in front of the infima in Eq.* (5.2) *and* (5.3) *can be further diminished, but this yields a greater rest as a consequence.*

Remark 9 (On assumption (**HM**))**.** *Assumption* (**HM**) *actually means all matrices in $\mathcal{M}$ can be diagonalized in a unique orthogonal basis, and thus can be parametrized by their eigenvalues. In that case the optimization problem is quite easy to solve. If not, solving* (5.1) *may turn out to be a hard problem, and our theoretical results do not cover this setting. However, it is always possible to discretize the set $\mathcal{M}$ as in Arlot and Bach (2011) or, in paractise, to use gradient descent; we conjecture Theorem 10 still holds without* (**HM**) *as long as $\mathcal{M}$ is not "too large", which could be proved similarly up to some uniform concentration inequalities.*

Note also that if $\mathcal{M}_1, \dots, \mathcal{M}_K$ all satisfy (**HM**) *(with different matrices P), then Theorem 10 still holds for $\mathcal{M} = \bigcup_{k=1}^K \mathcal{M}_k$ with $\mathbb{P}(\widetilde{\Omega}) \geq 1 - 9Kp^2 n^{-\delta}$, by applying the union bound in the proof.*

Remark 10 (Scaling of (n, p))**.** *Eq.* (5.2) *implies the asymptotic optimality of the estimator $\widehat{f}_{\widehat{M}}$ when*

$$c(\Sigma)^4 \frac{\operatorname{tr}\Sigma}{p} \times \frac{p^3 \left(\ln(n)\right)^3}{n} \ll \inf_{M \in \mathcal{M}} \left\{ \frac{1}{np}\left\|\widehat{f}_M - f\right\|_2^2 \right\}\,.$$

In particular, only (n, p) such that $p^3 \ll n/(\ln(n))^3$ are admissible.

Remark 11 (Relationship with the trace norm)**.** *Our framework relies on the minimization of Eq. (2.1) with respect to f. Argyriou et al. (2008) has shown that if we also minimize with respect to the matrix M subject to the constraint $\operatorname{tr} M^{-1} = 1$, then we obtain an equivalent regularization by the nuclear norm (a.k.a. trace norm), which implies the prior knowledge that our p prediction functions may be obtained as the linear combination of $r \ll p$ basis functions. This situation corresponds to cases where the matrix M^{-1} is singular, and we allow this explictly in our experiments.*

Note that the link between our framework and trace norm (i.e., nuclear norm) regularization is the same than between multiple kernel learning and the single task framework of Arlot and Bach (2011). In the multi-task case, the trace-norm regularization, though efficient computationally, does not lead to oracle inequality, while our criterion is an unbiased estimate of the generalization error, which turns out to be non-convex in the matrix M. While DC programming techniques (see, e.g. Gasso et al., 2009, and references therein) could be brought to bear to find local optima, the goal of the present work is to study the theoretical properties of our estimators, assuming we can minimize the cost function (e.g., in special cases, where we consider spectral variants, or by brute force enumeration).

6. Simulation experiments

In all the experiments presented in this section, we consider the framework of Section 2 with $\mathcal{X} = \mathbb{R}^d$, $d = 4$, and the kernel defined by $\forall x, y \in \mathcal{X}$, $k(x,y) = \prod_{j=1}^{d} e^{-|x_j - y_j|}$. The design points $X_1, \dots, X_n \in \mathbb{R}^d$ are drawn (repeatedly and independently for each sample) independently from the multivariate standard Gaussian distribution. For every $j \in \{1, \dots, p\}$, $f^j(\cdot) = \sum_{i=1}^{m} \alpha_i^j k(\cdot, z_i)$ where $m = 4$ and $z_1, \dots, z_m \in \mathbb{R}^d$ are drawn (once for all) independently from the multivariate standard Gaussian distribution, independent from the design $(X_i)_{1 \le i \le n}$. Thus, the expectations that will be considered are taken conditionally to the z_i. The coefficients $(\alpha_i^j)_{1 \le i \le m\,,\, 1 \le j \le p}$ differ according to the setting.

Settings. Four experimental settings are considered:

A⌋ Various numbers of tasks: $n = 100$ and $\forall i, j$, $\alpha_i^j = 1$, that is, $\forall j$, $f^j = f_A := \sum_{i=1}^{m} k(\cdot, z_i)$. The number of tasks is varying: $p \in \{2k \,/\, k = 1, \dots, 25\}$. The covariance matrix is $\Sigma = \Sigma_{A,p}$ defined as the first $p \times p$ block of $\Sigma_{A,50}$, where $\Sigma_{A,50}$ has been drawn (once for all) from the Wishart $W(I_{50}, 50, 100)$ distribution. The condition number of $\Sigma_{A,p}$ increases from $c(\Sigma_{A,2}) \approx 1.17$ to $c(\Sigma_{A,50}) \approx 22.50$ as p increases.

B⌋ Various sample sizes: $p = 5$, $\forall j$, $f^j = f_A$ and $\Sigma = \Sigma_B$ has been drawn (once for all) from the Whishart $W(I_5, 10, 5)$ distribution; the condition number of Σ_B is $c(\Sigma_B) \approx 22.05$. The only varying parameter is $n \in \{50k \,/\, k = 1, \dots, 20\}$.

C⌋ Various noise levels: $n = 100$, $p = 5$ and $\forall j$, $f^j = f_A$. The varying parameter is $\Sigma = \Sigma_{C,t} := t\Sigma_B$ with $t \in \{0.5k \,/\, k = 1, \dots, 20\}$.

D⌋ Clustering of two groups of functions $p = 10$, $n = 100$, $\Sigma = \Sigma_E$ has been drawn (once for all) from the Whishart $W(I_{10}, 20, 10)$ distribution; the condition number of Σ_E is $c(\Sigma_E) \approx 24.95$. We pick the function $f := \sum_{i=1}^{m} \alpha_i k(\cdot, z_i)$ by drawing

$(\alpha_1, \ldots, \alpha_m)$ from standard multivariate normal distributions and finally $f^1 = \cdots = f^5 = f$, $f^6 = \cdots = f^{10} = -f$.

Collections of matrices. Two different sets of matrices $\mathcal{M}$ are considered in the Experiments A-C, following Examples 1 and 2:

$$\mathcal{M}_{\text{similar}} := \left\{ M_{\text{similar}}(\lambda, \mu) = (\lambda + p\mu) I_p - \frac{\mu}{p} \mathbf{1}\mathbf{1}^\top \,/\, (\lambda, \mu) \in (0, +\infty)^2 \right\}$$
$$\text{and} \quad \mathcal{M}_{\text{ind}} := \{ M_{\text{ind}}(\lambda) = \text{Diag}(\lambda_1, \ldots, \lambda_p) \,/\, \lambda \in (0, +\infty)^p \} \quad .$$

In Experiment D, we also use two different sets of matrices, following Examples 3 :

$$\mathcal{M}_{\text{clus}} := \bigcup_{I \subset \{1,\ldots,p\}, I \notin \{\{1,\ldots,p\}, \emptyset\}} \left\{ M_I(\lambda, \mu, \mu) \,/\, (\lambda, \mu) \in (0, +\infty)^2 \right\} \bigcup \mathcal{M}_{\text{similar}}$$
$$\text{and} \quad \mathcal{M}_{\text{interval}} := \bigcup_{1 \le k \le p-1} \left\{ M_I(\lambda, \mu, \mu) \,/\, (\lambda, \mu) \in (0, +\infty)^2, I = \{1, \ldots, k\} \right\} \bigcup \mathcal{M}_{\text{similar}} \quad .$$

Remark 12. *The set $\mathcal{M}_{\text{clus}}$ contains $2^p - 1$ models, a case we will denote by "clustering". The other set, $\mathcal{M}_{\text{interval}}$, only has p models, and should take advantage of the knowledge of the structure of the Setting D. We call this setting "segmentation into intervals".*

Estimators. Concerning Experiments A-C combining the two possible sets of matrices with two penalization procedures (that is, with the penalty defined in Eq. (2.7) and either Σ known or estimated by $\widehat{\Sigma}$) leads to four estimators defined by

$$\forall \alpha \in \{\text{similar}, \text{ind}\} \ , \forall S \in \left\{ \Sigma, \widehat{\Sigma} \right\} , \quad \widehat{f}_{\alpha,S} := \widehat{f}_{\widehat{M}_{\alpha,S}} = A_{\widehat{M}_{\alpha,S}} y$$
$$\text{where} \quad \widehat{M}_{\alpha,S} \in \operatorname*{argmin}_{M \in \mathcal{M}_\alpha} \left\{ \frac{1}{np} \left\| y - \widehat{f}_M \right\|_2^2 + \frac{2}{np} \operatorname{tr}\left(A_M \cdot (S \otimes I_n) \right) \right\}$$

and Σ is defined by Eq. (4.1). As detailed in Examples 1–2, $\widehat{f}_{\text{ind},\widehat{\Sigma}}$ and $\widehat{f}_{\text{ind},\Sigma}$ are concatenations of single-task estimators, whereas $\widehat{f}_{\text{similar},\widehat{\Sigma}}$ and $\widehat{f}_{\text{similar},\Sigma}$ should take advantage of a setting where the functions f^j are close in $\mathcal{F}$ thanks to the regularization term $\sum_{j,k} \|f^j - f^k\|_{\mathcal{F}}^2$.

Concerning Experiment D, given the two possible sets of matrices, plus the single-task matrices, we consider the following three estimators :

$$\forall \beta \in \{\text{clus}, \text{interval}, \text{ind}\} \ , \quad \widehat{f}_\beta := \widehat{f}_{\widehat{M}_\beta} = A_{\widehat{M}_\beta} y$$
$$\text{where} \quad \widehat{M}_\beta \in \operatorname*{argmin}_{M \in \mathcal{M}_\beta} \left\{ \frac{1}{np} \left\| y - \widehat{f}_M \right\|_2^2 + \frac{2}{np} \operatorname{tr}\left(A_M \cdot (S \otimes I_n) \right) \right\}$$

Remark 13 (Finding the jump in Algorithm 1)**.** *The Pseudo-Algorithm 1 raises the question of how to detect the jump of astimated dimensionality of* $\text{df}(\lambda)$*, which happens around the variance we want to estimate. We chose to select an estimator of $\widehat{C}$ of σ^2 such that it was the smallest index such that $\text{df}(\widehat{\lambda}_0(\widehat{C})) < n/2$. An other approach was attempted, namely by choosing the index corresponding to the largest instantaneous jump of* $\text{df}(\widehat{\lambda}_0(C))$

(which is piece-wise constant and non-increasing). This approach had a major drawback, because it sometimes selected a jump far away from the "real" jump, which consisted of several small jumps. Both these approaches gave similar results in terms of prediction error, and we chose the first one because of its direct link to our theoretical criterion given in Algorithm 1.

Results. The results of the Experiments A-C are reported in Figures 1–8. In each experiment, $N = 1000$ independent samples $y \in \mathbb{R}^{np}$ have been generated. Due to computation time, only $N = 100$ samples have been generated in simulations B. Expectations are estimated thanks to empirical means over the N samples, and error bars correspond to the classical Gaussian 5% level difference test (that is, empirical variance over the N samples multiplied by $1.96/\sqrt{N}$). The results of Experiment D are reported in Table 1 . Here also the expectations are estimated thanks to empirical means over the 1000 samples. The p-value corresponds to the classical Gaussian difference test, where the hypotheses tested are of the shape $\mathbb{H}_0 = \{q < 1\}$ against the hypotheses $\mathbb{H}_1 = \{q \geq 1\}$, where the different quantities q are detailed later. We compute the p-value of the test that is, the minimal risk that leads the tests to reject the tested hypothesis.

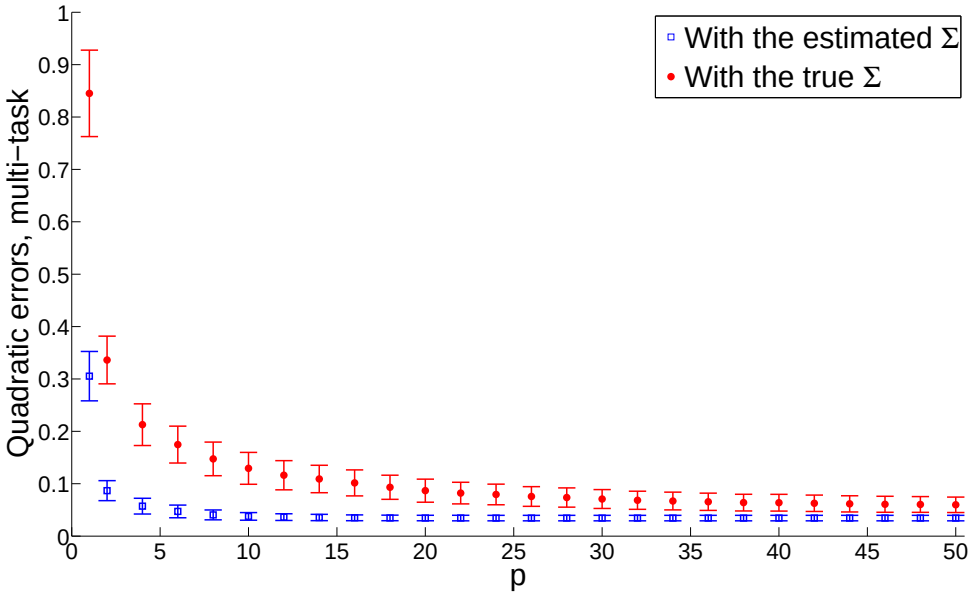

Figure 1: Increasing the number of tasks p (Setting A), quadratic errors of multi-task estimators $(np)^{-1}\mathbb{E}[\|\widehat{f}_{\text{similar},S} - f\|^2]$. Blue: $S = \widehat{\Sigma}$. Red: $S = \Sigma$.

Comments. As expected, multi-task learning significantly helps when all f^j are equal, as soon as p is large enough (Figure 3), especially for small n (Figure 6) and large noise-levels (Figure 8). Increasing the number of tasks rapidly reduces the quadratic error with multi-task estimators (Figure 1) contrary to what happens with single-task estimators (Figure 2). A noticeable phenomenon also occurs in Figure 1 and even more in Figure 2: the estimator

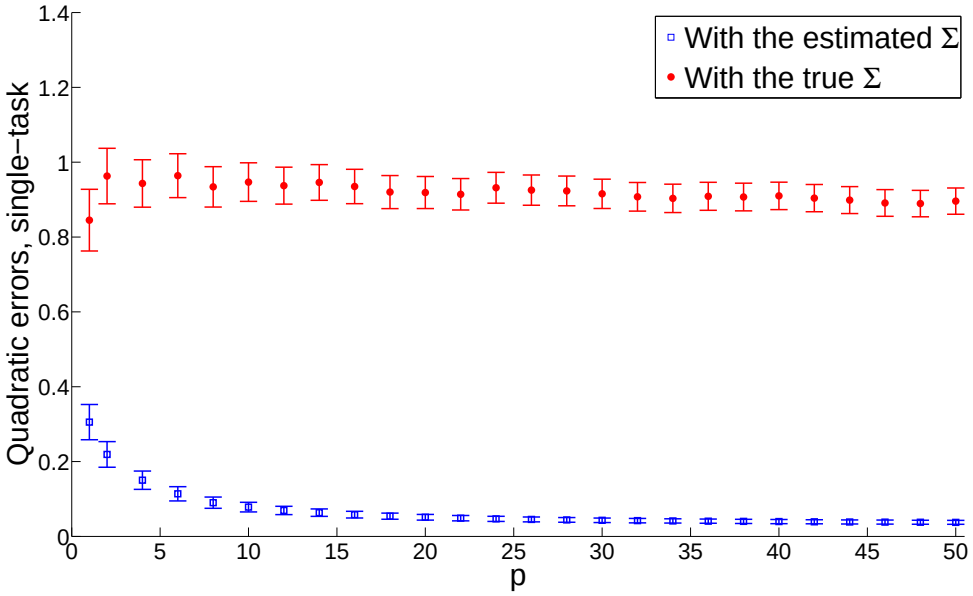

Figure 2: Increasing the number of tasks p (Setting A), quadratic errors of single-task estimators $(np)^{-1}\mathbb{E}[\|\widehat{f}_{\mathrm{ind},S} - f\|^2]$. Blue: $S = \widehat{\Sigma}$. Red: $S = \Sigma$.

Quantity estimated : q	Mean	Empirical variance	$\mathbb{H}_0$	p-value
$\mathbb{E}[\|\widehat{f}_{\mathrm{clus}} - f\|^2 / \|\widehat{f}_{\mathrm{ind}} - f\|^2]$	0.6683	0.2935	$q > 1$	$< 10^{-15}$
$\mathbb{E}[\|\widehat{f}_{\mathrm{interval}} - f\|^2 / \|\widehat{f}_{\mathrm{ind}} - f\|^2]$	0.6596	0.2704	$q > 1$	$< 10^{-15}$
$\mathbb{E}[\|\widehat{f}_{\mathrm{interval}} - f\|^2 / \|\widehat{f}_{\mathrm{clus}} - f\|^2]$	0.99998	0.1645	$q > 1$	0.5006

Table 1: Clustering and segmentation (Setting D).

$\widehat{f}_{\mathrm{ind},\Sigma}$ (that is, obtained knowing the true covariance matrix Σ) is less efficient than $\widehat{f}_{\mathrm{ind},\widehat{\Sigma}}$ where the covariance matrix is estimated. It corresponds to the combination of two facts: (i) multiplying the ideal penalty by a small factor $1 < C_n < 1 + o(1)$ is known to often improve performances in practice when the sample size is small (see Section 6.3.2 of Arlot (2009)), and (ii) minimal penalty algorithms like Algorithm 1 are conjectured to overpenalize slightly when n is small or the noise-level is large (Lerasle, 2011) (as confirmed by Figure 7). Interestingly, this phenomenon is stronger for single-task estimators (differences are smaller in Figure 1) and disappears when n is large enough (Figure 5), which is consistent with the heuristic motivating multi-task learning: "increasing the number of tasks p amounts to increase the sample size". However the advantage of the multi-task procedure (compared to the single task one) seems to decrease when p becomes very large, as seen in Figure 3. This seems reasonable since the multi-task procedure requires the estimation of the matrix Σ that is, $p(p+1)/2$ parameters, and thus induces a large variance when n is small (here, $p = 50$ and $n = 100$).

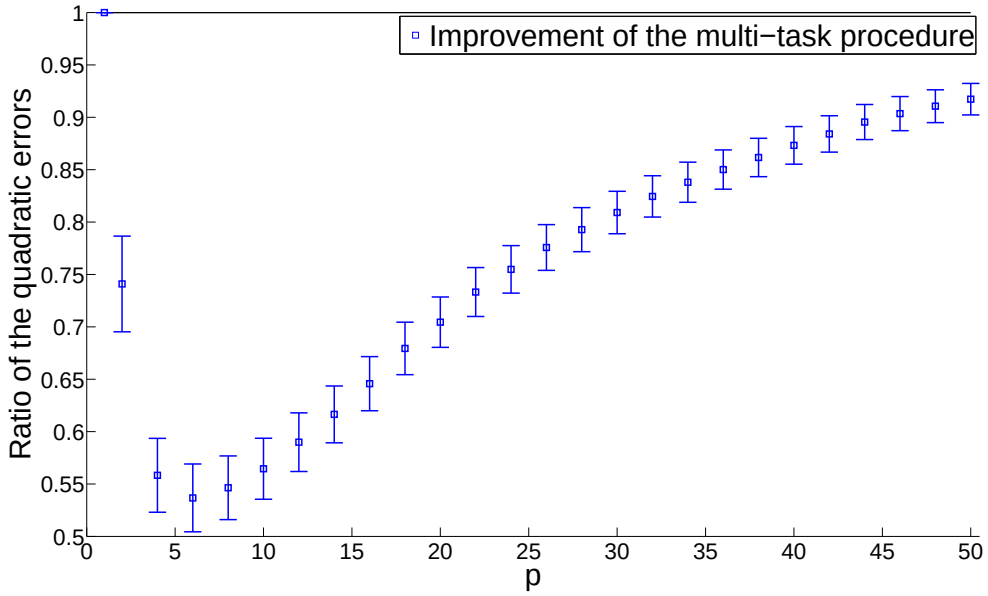

Figure 3: Increasing the number of tasks p (Setting A), improvement of multi-task compared to single-task: $\mathbb{E}[\|\widehat{f}_{\text{similar},\widehat{\Sigma}} - f\|^2 / \|\widehat{f}_{\text{ind},\widehat{\Sigma}} - f\|^2]$.

Figures 4 and 5 show us that our procedure works well with small n, and that increasing n does not seem to significantly improve the performance of our estimators, except in the single-task setting with Σ known, where the under-penalization phenomenon discussed above disappears.

Table 1 shows us that using the multitask procedure benefits the estimation accuracy, both in the clustering setting and in the segmentation setting. The last line of Table 1 does not show that the clustering setting improves over the "segmentation into intervals" one, which was awaited if both select a model close to the oracles, which are the same on both cases.

7. Conclusion and future work

This paper shows that taking into account the unknown similarity between p regression tasks can be done optimally (Theorem 10). The crucial point is to estimate the $p \times p$ covariance matrix Σ of the noise (covariance between tasks). An estimator of Σ is defined in Section 4, where non-asymptotic bounds on its error are provided under very mild assumptions on the mean of the sample (Theorem 8), which is probably the most important theoretical result of the paper.

Simulation experiments show that our algorithm works with reasonable sample sizes, and that our multi-task estimator often perform much better than its single-task counterpart. Up to the best of our knowledge, a theoretical proof of this point remains an open problem that we intend to investigate in a future work.

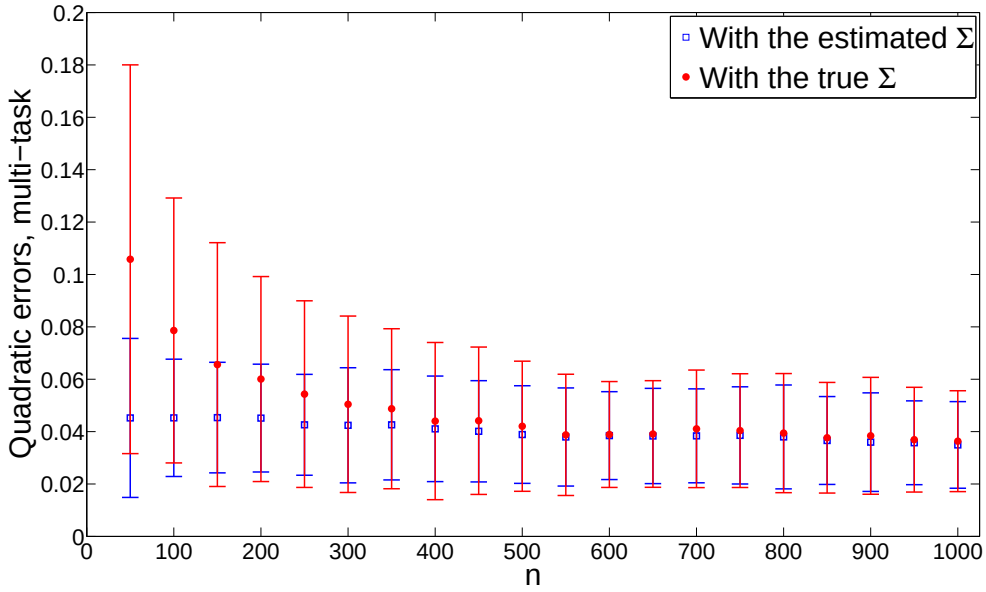

Figure 4: Increasing the sample size n (Setting B), quadratic errors of multi-task estimators $(np)^{-1}\mathbb{E}[\|\widehat{f}_{\text{similar},S} - f\|^2]$. Blue: $S = \widehat{\Sigma}$. Red: $S = \Sigma$.

Theorem 10 only holds when matrices $\mathcal{M}$ can be diagonalized simultaneously (assumption (**HM**)), which often corresponds to cases where we have a prior knowledge of what the relations between the tasks would be, and which is the only known case where the optimization is quite easy. We do plan to expand our results to larger sets $\mathcal{M}$, which may require new concentration inequalities and new optimization algorithms.

Acknowledgments. This paper was supported by grants from the Agence Nationale de la Recherche (Detect project, reference ANR-09-JCJC-0027-01) and from the European Research Council (SIERRA Project ERC-239993).

We give in Appendix the proofs of the different results stated in Sections 2, 4 and 5. The proofs of our main results are contained in Sections D and E.

Appendix A. Proof of Proposition 2

Proof It is sufficient to show that $\langle \cdot, \cdot \rangle_{\mathcal{G}}$ is positive-definite on $\mathcal{G}$. Take $g \in \mathcal{G}$ and $S = (S_{i,j})_{1 \leq i \leq j \leq p}$ the symmetric postive-definite matrix of size p verifying $S^2 = M$, and denote $T = S^{-1} = (T_{i,j})_{1 \leq i,j \leq p}$. Let f be the element of $\mathcal{G}$ defined by $\forall i \in \{1 \dots p\}$, $g(\cdot, i) =$

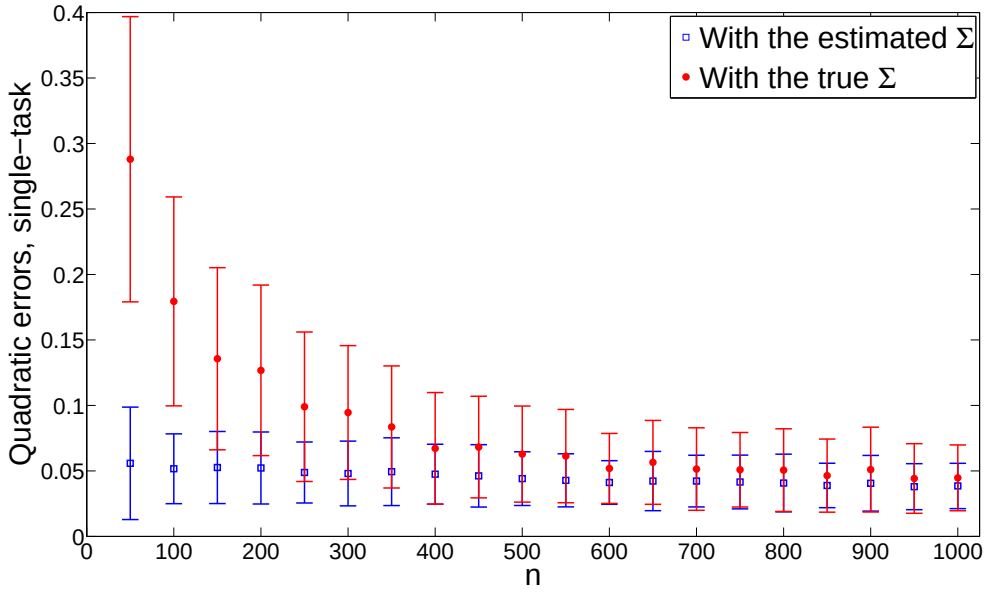

Figure 5: Increasing the sample size n (Setting B), quadratic errors of single-task estimators $(np)^{-1}\mathbb{E}[\|\widehat{f}_{\mathrm{ind},S} - f\|^2]$. Blue: $S = \widehat{\Sigma}$. Red: $S = \Sigma$.

$\sum_{k=1}^{n} T_{i,k} f(\cdot, k)$. We then have:

$$
\begin{aligned}
\langle g, g\rangle_{\mathcal{G}} &= \sum_{i=1}^{p}\sum_{j=1}^{p} M_{i,j}\langle g(\cdot,i), g(\cdot,j)\rangle_{\mathcal{F}} \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p} M_{i,j}T_{i,k}T_{j,l}\langle f(\cdot,k), f(\cdot,l)\rangle_{\mathcal{F}} \\
&= \sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p} T_{l,j}\langle f(\cdot,k), f(\cdot,l)\rangle_{\mathcal{F}} \sum_{i=1}^{p} M_{j,i}T_{i,k} \\
&= \sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p} T_{l,j}\langle f(\cdot,k), f(\cdot,l)\rangle_{\mathcal{F}} (M\cdot T)_{j,k} \\
&= \sum_{k=1}^{p}\sum_{l=1}^{p} T_{l,j}\langle f(\cdot,k), f(\cdot,l)\rangle_{\mathcal{F}} \sum_{j=1}^{p} T_{l,j}(M\cdot T)_{j,k} \\
&= \sum_{k=1}^{p}\sum_{l=1}^{p} \langle f(\cdot,k), f(\cdot,l)\rangle_{\mathcal{F}} (T\cdot M\cdot T)_{k,l} \\
&= \sum_{k=1}^{p} \|f(\cdot,k)\|_{\mathcal{F}}^2.
\end{aligned}
$$

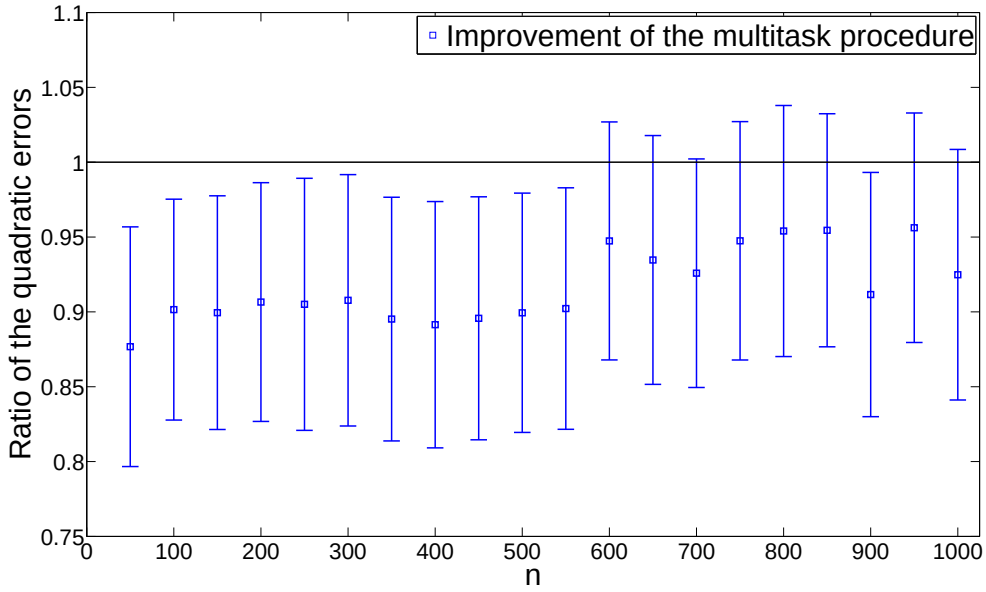

Figure 6: Increasing the sample size n (Setting B), improvement of multi-task compared to single-task: $\mathbb{E}[\|\widehat{f}_{\text{similar},\widehat{\Sigma}} - f\|^2 / \|\widehat{f}_{\text{ind},\widehat{\Sigma}} - f\|^2]$.

This shows that $\langle g, g\rangle_{\mathcal{G}} \geq 0$ and that $\langle g, g\rangle_{\mathcal{G}} = 0 \Rightarrow f = 0 \Rightarrow g = 0$. ■

Appendix B. Proof of Corollary 1

Proof If $(x, j) \in \mathcal{X} \times \{1, \ldots, p\}$, the application $(f^1, \ldots, f^p) \mapsto f^j(x)$ is clearly continuous. We now show that $(\mathcal{G}, \langle \cdot, \cdot \rangle_{\mathcal{G}})$ is complete. If $(g_n)_{n\in\mathbb{N}}$ is a Cauchy sequence of $\mathcal{G}$ and if we define, as in Section A , the functions f_n by $\forall n \in \mathbb{N},\ \forall i \in \{1 \ldots p\},\ g_n(\cdot, i) = \sum_{k=1}^{p} T_{i,k} f_n(\cdot, k)$. The same computations show that $(f_n(\cdot, i))_{n\in\mathbb{N}}$ are Cauchy sequences of $\mathcal{F}$, and thus converge. So the sequence $(f_n)_{n\in\mathbb{N}}$ converges in $\mathcal{G}$, and $(g_n)_{n\in\mathbb{N}}$ does likewise. ■

Appendix C. Proof of Proposition 4

Proof We define

$$\widetilde{\Phi}(x, j) = M^{-1} \cdot \begin{pmatrix} \delta_{1,j}\Phi(x) \\ \vdots \\ \delta_{p,j}\Phi(x) \end{pmatrix} ,$$

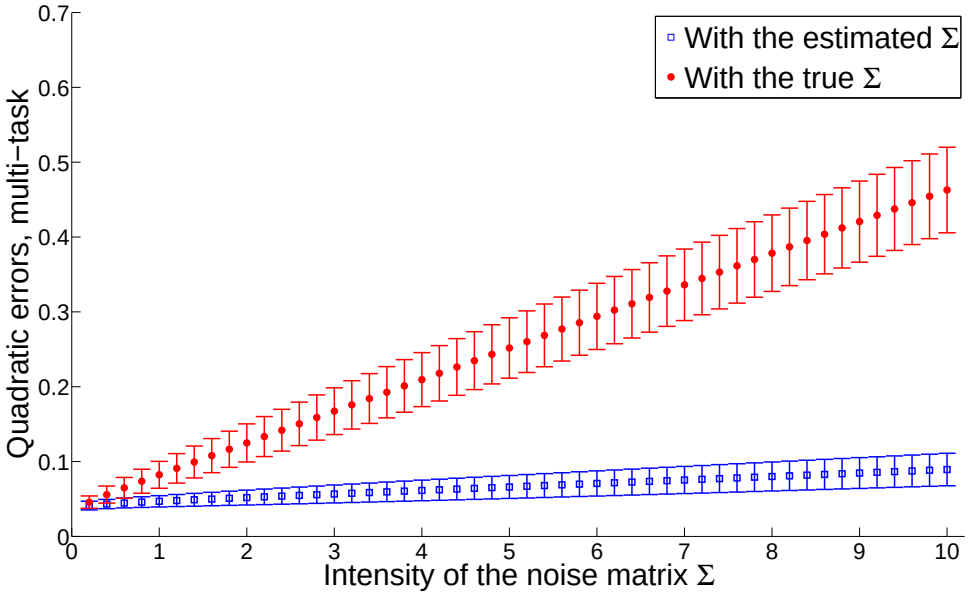

Figure 7: Increasing the signal-to-noise ratio (Setting C), quadratic errors of multi-task estimators $(np)^{-1}\mathbb{E}[\|\widehat{f}_{\text{similar},S} - f\|^2]$. Blue: $S = \widehat{\Sigma}$. Red: $S = \Sigma$.

with $\delta_{i,j} = \mathbf{1}_{i=j}$ being the Kronecker symbol, that is, $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise. We now show that $\widetilde{\Phi}$ is the feature function of the RKHS. For $g \in \mathcal{G}$ and $(x, l) \in \mathcal{X} \times \{1, \dots, p\}$, we have:

$$\begin{aligned}
\langle g, \widetilde{\Phi}(x,l)\rangle_{\mathcal{G}} &= \sum_{j=1}^{p}\sum_{i=1}^{p} M_{j,i}\langle g(\cdot,j), \widetilde{\Phi}(x,l)^i\rangle_{\mathcal{F}} \\
&= \sum_{j=1}^{p}\sum_{i=1}^{p}\sum_{m=1}^{p} M_{j,i}M^{-1}_{i,m}\delta_{m,l}\langle g(\cdot,j), \Phi(x)\rangle_{\mathcal{F}} \\
&= \sum_{j=1}^{p}\sum_{m=1}^{p} (M\cdot M^{-1})_{j,m}\delta_{m,l}g(x,j) \\
&= \sum_{j=1}^{p} \delta_{j,l}g(x,j) = g(x,l) \ .
\end{aligned}$$

Thus we can write:

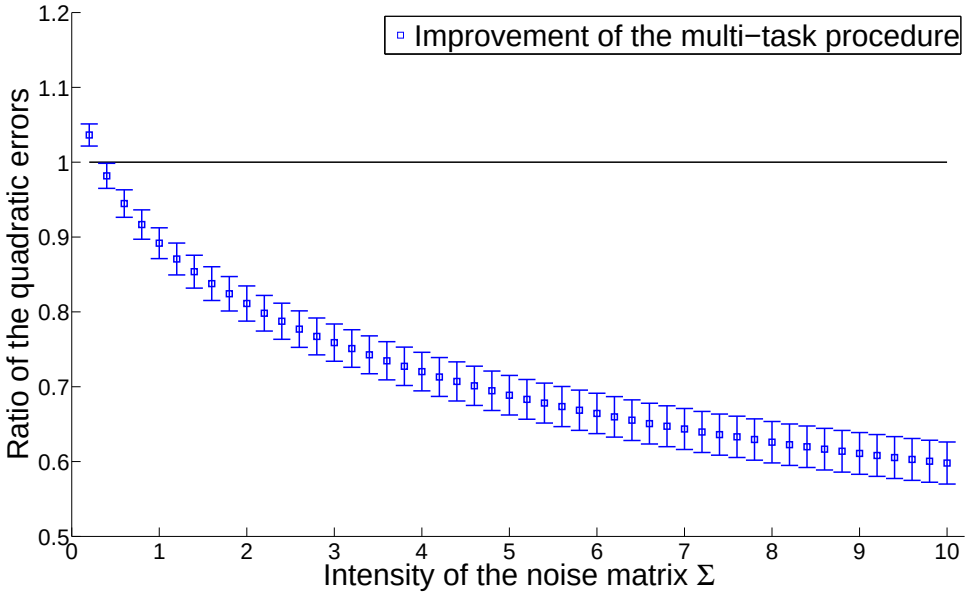

Figure 8: Increasing the signal-to-noise ratio (Setting C), improvement of multi-task compared to single-task: $\mathbb{E}[\|\widehat{f}_{\text{similar},\widehat{\Sigma}} - f\|^2 / \|\widehat{f}_{\text{ind},\widehat{\Sigma}} - f\|^2]$.

$$
\begin{aligned}
\widetilde{k}((x,i),(y,j)) &= \langle \widetilde{\Phi}(x,i), \widetilde{\Phi}(y,j) \rangle_{\mathcal{G}} \\
&= \sum_{h=1}^{p} \sum_{h'=1}^{p} M_{h,h'} \langle M_{h,i}^{-1} \Phi(x), M_{h',j}^{-1} \Phi(y) \rangle_{\mathcal{F}} \\
&= \sum_{h=1}^{p} \sum_{h'=1}^{p} M_{h,h'} M_{h,i}^{-1} M_{h',j}^{-1} K(x,y) \\
&= \sum_{h=1}^{p} M_{h,i}^{-1} (M \cdot M^{-1})_{h,j} K(x,y) \\
&= \sum_{h=1}^{p} M_{h,i}^{-1} \delta_{h,j} K(x,y) = M_{i,j}^{-1} K(x,y) \ .
\end{aligned}
$$

■

Appendix D. Proof of Theorem 8

D.1 Some useful tools

We now give two properties of the Kronecker product, and then introduce a useful norm on $\mathcal{S}_p(\mathbb{R})$, upon which we give several properties. Those are the tools needed to prove Theorem 8.

Property 1. *The Kronecker product is bilinear, associative and for every matrices A, B, C, D such that the dimensions fit, $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$.*

Property 2. *Let $A \in \mathcal{M}_n(\mathbb{R})$, $(A \otimes I_n)^\top = (A^\top \otimes I_n)$.*

Definition 11. *We now introduce the norm $|\!|\!| \cdot |\!|\!|$ on $\mathcal{S}_p(\mathbb{R})$, which is the modulus of the eigenvalue of largest magnitude, and can be defined by*

$$|\!|\!| S |\!|\!| := \sup_{z \in \mathbb{R}^p, \|z\|_2 = 1} \left| z^\top S z \right| \ .$$

This norm has several interesting properties, some of which we will use and which are stated below.

Property 3. *The norm $|\!|\!| \cdot |\!|\!|$ is a matricial norm: $\forall (A, B) \in \mathcal{S}_p(\mathbb{R})^2,\ |\!|\!| AB |\!|\!| \leq |\!|\!| A |\!|\!| |\!|\!| B |\!|\!|$.*

We will use the following result, which is a consequence of the preceding Property.

$$\forall S \in \mathcal{S}_p(\mathbb{R}),\ \forall T \in \mathcal{S}_p^{++}(\mathbb{R}),\ |\!|\!| T^{-\frac{1}{2}} S T^{-\frac{1}{2}} |\!|\!| \leq |\!|\!| S |\!|\!| |\!|\!| T^{-1} |\!|\!| \ .$$

We also have:

Proposition 12.

$$\forall \Sigma \in \mathcal{S}_p(\mathbb{R}),\ |\!|\!| \Sigma \otimes I_n |\!|\!| = |\!|\!| \Sigma |\!|\!| \ .$$

Proof We can diagonalize Σ in an orthonormal basis: $\exists U \in \mathcal{O}_n(\mathbb{R}),\ \exists D = \mathrm{Diag}(\mu_1, \ldots, \mu_p),\ \Sigma = U^\top D U$. We then have, using the properties of the Kronecker product:

$$\begin{aligned} \Sigma \otimes I_n &= (U^\top \otimes I_n)(D \otimes I_n)((U \otimes I_n) \\ &= (U \otimes I_n)^\top (D \otimes I_n)((U \otimes I_n) \ . \end{aligned}$$

We just have to notice that $U \otimes I_n \in \mathcal{O}_{np}(\mathbb{R})$ and that:

$$D \otimes I_n = \mathrm{Diag}(\underbrace{\mu_1, \ldots, \mu_1}_{n \text{ times}}, \ldots, \underbrace{\mu_p, \ldots, \mu_p}_{n \text{ times}}) \ .$$

■

This norm can also be written in other forms:

Property 4. *If $M \in \mathcal{M}_n(\mathbb{R})$, the operator norm $\|M\|_2 := \sup_{t \in \mathbb{R}^n \setminus \{0\}} \left\{ \frac{\|Mt\|_2}{\|t\|_2} \right\}$ is equal to the greatest singular value of M: $\sqrt{\rho(M^\top M)}$. Henceforth, if S is symmetric, we have $|\!|\!| S |\!|\!| = \|S\|_2$*

D.2 The proof

We now give a proof of Theorem 8, using Lemmas 13, 14 and 15, which are stated and proved in Section D.3. The outline of the prove is the following:

1. Apply Theorem 5 to problem (**Pz**) for every $z \in \mathcal{Z}$ in order to
2. control $\|s - \zeta\|_\infty$ with a large probability, where $s, \zeta \in \mathbb{R}^{p(p+1)/2}$ are defined by

$$s := (\Sigma_{1,1}, \ldots, \Sigma_{p,p}, \Sigma_{1,1} + \Sigma_{2,2} + 2\Sigma_{1,2}, \ldots, \Sigma_{i,i} + \Sigma_{j,j} + 2\Sigma_{i,j}, \ldots)$$
$$\text{and} \quad \zeta := (a(e_1), \ldots, a(e_p), a(e_1 + e_2), \ldots, a(e_1 + e_p), a(e_2 + e_3), \ldots, a(e_{p-1} + e_p)) \ .$$

3. Deduce that $\widehat{\Sigma} = J(\zeta)$ is close to $\Sigma = J(s)$ by controlling the Lipschitz norm of J.

Proof 1. Apply Theorem 5: We start by noticing that Assumption (**Hdf**) actually holds true with all $\lambda_{0,j}$ equal. Indeed, let $(\lambda_{0,j})_{1 \leq j \leq p}$ be given by Assumption (**Hdf**) and define $\lambda_0 := \min_{j=1,\ldots,p} \lambda_{0,j}$. Then, $\lambda_0 \in (0, +\infty)$ and $\mathrm{df}(\lambda_0)$ since all $\lambda_{0,j}$ satisfy these two conditions. For the last condition, remark that for every $j \in \{1, \ldots, p\}$, $\lambda_0 \leq \lambda_{0,j}$ and $\lambda \mapsto \|(A_\lambda - I)F_{e_j}\|_2^2$ is a nonincreasing function (as noticed in Arlot and Bach (2011) for instance), so that

$$\frac{1}{n} \left\|(A_{\lambda_0} - I_n)F_{e_j}\right\|_2^2 \leq \frac{1}{n} \left\|(A_{\lambda_{0,j}} - I_n)F_{e_j}\right\|_2^2 \leq \Sigma_{j,j} \sqrt{\frac{\ln(n)}{n}} \ . \tag{D.1}$$

In particular, Eq. (3.2) holds with $d_n = 1$ for problem (**Pz**) whatever $z \in \{e_1, \ldots, e_p\}$.

Let us now consider the case $z = e_i + e_j$ with $i \neq j \in \{1, \ldots, p\}$. Using Eq. (D.1) and that $F_{e_i + e_j} = F_{e_i} + F_{e_j}$, we have

$$\left\|(B_{\lambda_0} - I_n)F_{e_i+e_j}\right\|_2^2 \leq \|(B_{\lambda_0} - I_n)F_{e_i}\|_2^2 + \left\|(B_{\lambda_0} - I_n)F_{e_j}\right\|_2^2 + 2\langle (B_{\lambda_0} - I_n)F_{e_i}, (B_{\lambda_0} - I_n)F_{e_j} \rangle \ .$$

The last term is bounded as follows:

$$\begin{aligned}
2\langle (B_{\lambda_0} - I_n)F_{e_i}, (B_{\lambda_0} - I_n)F_{e_j} \rangle &\leq 2\|(B_{\lambda_0} - I_n)F_{e_i}\| \cdot \|(B_{\lambda_0} - I_n)F_{e_j}\| \\
&\leq 2\sqrt{n \ln(n)} \sqrt{\Sigma_{i,i}\Sigma_{j,j}} \\
&\leq \sqrt{n \ln(n)} (\Sigma_{i,i} + \Sigma_{j,j}) \\
&\leq (1 + c(\Sigma)) \sqrt{n \ln(n)} (\Sigma_{i,i} + \Sigma_{j,j} + 2\Sigma_{i,j}) \\
&= (1 + c(\Sigma)) \sqrt{n \ln(n)} \sigma_{e_i+e_j}^2 \ ,
\end{aligned}$$

because Lemma 13 shows

$$2(\Sigma_{i,i} + \Sigma_{j,j}) \leq (1 + c(\Sigma))(\Sigma_{i,i} + \Sigma_{j,j} + 2\Sigma_{i,j}) \ .$$

Therefore, Eq. (3.2) holds with $d_n = 1 + c(\Sigma)$ for problem (**Pz**) whatever $z \in \mathcal{Z}$.

2. Control $\|s - \zeta\|_\infty$: Let us define

$$\eta_1 := \beta(\alpha + \delta)(1 + c(\Sigma)) \sqrt{\frac{\ln(n)}{n}} \ .$$

By Theorem 5, for every $z \in \mathcal{Z}$, an event Ω_z of probability greater than $1 - \kappa n^{-\delta}$ exists on which, if $n \geq n_0(\delta)$,

$$(1 - \eta_1)\sigma_z^2 \leq a(z) \leq (1 + \eta_1)\sigma_z^2 \ .$$

So, on $\Omega := \bigcap_{z \in Z} \Omega_z$,

$$\|\zeta - s\|_\infty \leq \eta_1 \|s\|_\infty \ , \tag{D.2}$$

and $\mathbb{P}(\Omega) \geq 1 - \kappa p(p+1)/2n^{-\delta}$ by the union bound. Let

$$\|\Sigma\|_\infty := \sup_{i,j} |\Sigma_{i,j}| \quad \text{and} \quad C_1(p) := \sup_{\Sigma \in \mathcal{S}_p(\mathbb{R})} \left\{ \frac{\|\Sigma\|_\infty}{|||\Sigma|||} \right\} \ .$$

Since $\|s\|_\infty \leq 4 \|\Sigma\|_\infty$ and $C_1(p) = 1$ by Lemma 14, Eq. (D.2) implies that on Ω,

$$\|\zeta - s\|_\infty \leq 4\eta_1 \|\Sigma\|_\infty \leq 4\eta_1 |||\Sigma||| \ . \tag{D.3}$$

3. Conclusion of the proof: Let

$$C_2(p) := \sup_{\zeta \in \mathbb{R}^{p(p+1)/2}} \left\{ \frac{|||J(\zeta)|||}{\|\zeta\|_\infty} \right\} \ .$$

By Lemma 15, $C_2(p) \leq \frac{3}{2}p$. By Eq. (D.3), on Ω,

$$|||\widehat{\Sigma} - \Sigma||| = |||J(\zeta) - J(s)||| \leq C_2(p) \|\zeta - s\|_\infty \leq 4\eta_1 C_2(p) |||\Sigma||| \ . \tag{D.4}$$

Since

$$|||\Sigma^{-\frac{1}{2}} \widehat{\Sigma} \Sigma^{-\frac{1}{2}} - I_p||| = |||\Sigma^{-\frac{1}{2}} (\Sigma - \widehat{\Sigma}) \Sigma^{-\frac{1}{2}}||| \leq |||\Sigma^{-1}||| |||\Sigma - \widehat{\Sigma}||| \ ,$$

and $|||\Sigma||| |||\Sigma^{-1}||| = c(\Sigma)$, Eq. (D.4) implies that on Ω,

$$|||\Sigma^{-\frac{1}{2}} \widehat{\Sigma} \Sigma^{-\frac{1}{2}} - I_p||| \leq 4\eta_1 C_2(p) |||\Sigma||| |||\Sigma^{-1}||| = 4\eta_1 C_2(p) c(\Sigma) \leq 6\eta_1 p c(\Sigma) \ .$$

To conclude, Eq. (4.2) holds on Ω with

$$\eta = 6pc(\Sigma)\beta(\alpha + \delta)(1 + c(\Sigma))\sqrt{\frac{\ln(n)}{n}} \leq L_1(\alpha + \delta)p\sqrt{\frac{\ln(n)}{n}} c(\Sigma)^2 \tag{D.5}$$

for some numerical constant L_1. ■

Remark 14. *As stated in Arlot and Bach (2011), we need* $\sqrt{n_0(\delta)/\ln(n_0(\delta))} \geq 504$ *and* $\sqrt{n_0(\delta)}/\ln(n_0(\delta)) \geq 24(290 + \delta)$.

Remark 15. *To ensure that the estimated matrix* $\widehat{\Sigma}$ *is positive-definite we need that* $\eta < 1$, *that is,*

$$\sqrt{\frac{n}{\ln(n)}} > 6\beta(\alpha + \delta)pc(\Sigma)\left(1 + c(\Sigma)\right) \ .$$

D.3 Useful Lemmas

Lemma 13. *Let $p \geq 1$, $\Sigma \in \mathcal{S}_p^{++}(\mathbb{R})$ and $c(\Sigma)$ its condition number. Then,*

$$\forall 1 \leq i < j \leq p\,, \quad \Sigma_{i,j} \geq -\frac{c(\Sigma)-1}{c(\Sigma)+1}\frac{\Sigma_{i,i}+\Sigma_{j,j}}{2} \ , \tag{D.6}$$

Remark 16. *The proof of Lemma 13 shows the constant $\frac{c(\Sigma)-1}{c(\Sigma)+1}$ cannot be improved without additional assumptions on Σ.*

Proof It suffices to show the result when $p = 2$. Indeed, (D.6) only involves 2×2 submatrices $\widetilde{\Sigma}(i,j) \in \mathcal{S}_2^{++}(\mathbb{R})$ for which

$$1 \leq c(\widetilde{\Sigma}) \leq c\,(\Sigma) \quad \text{hence} \quad 0 \leq \frac{c(\widetilde{\Sigma})-1}{c(\widetilde{\Sigma})+1} \leq \frac{c(\Sigma)-1}{c(\Sigma)+1} \ .$$

So, some $\theta \in \mathbb{R}$ exists such that $\Sigma = |\!|\!|\Sigma|\!|\!| R_\theta^\top D R_\theta$ where

$$R_\theta := \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \qquad D = \begin{pmatrix} 1 & 0 \\ 0 & \lambda \end{pmatrix} \quad \text{and} \quad \lambda := \frac{1}{c(\Sigma)} \ .$$

Therefore,

$$\Sigma = |\!|\!|\Sigma|\!|\!| \begin{pmatrix} \cos^2(\theta) + \lambda \sin^2(\theta) & \frac{1-\lambda}{2}\sin(2\theta) \\ \frac{1-\lambda}{2}\sin(2\theta) & \lambda\cos^2(\theta) + \sin^2(\theta) \end{pmatrix} \ .$$

So, Eq. (D.6) is equivalent to

$$\frac{(1-\lambda)\sin(2\theta)}{2} \geq -\frac{1-\lambda}{1+\lambda}\frac{1+\lambda}{2} \ ,$$

which holds true for every $\theta \in \mathbb{R}$, with equality for $\theta \equiv \pi/2 \ (\text{mod. } \pi)$. ■

Lemma 14. *For every $p \geq 1$, $C_1(p) := \sup_{\Sigma \in \mathcal{S}_p(\mathbb{R})} \frac{\|\Sigma\|_\infty}{|\!|\!|\Sigma|\!|\!|} = 1$.*

Proof With $\Sigma = I_p$ we have $\|\Sigma\|_\infty = |\!|\!|\Sigma|\!|\!| = 1$, so $C_1(p) \geq 1$.
Let us introduce (i,j) such that $|\Sigma_{i,j}| = \|\Sigma\|_\infty$. We then have, with e_k being the k^{th} vector of the canonical basis of $\mathbb{R}^p$,

$$|\Sigma_{i,j}| = |e_i^\top \Sigma e_j| \leq |e_i^\top \Sigma e_i|^{1/2} |e_j^\top \Sigma e_j|^{1/2} \leq (\|\Sigma\|_2^{1/2})^2 \ .$$

■

Lemma 15. *For every $p \geq 1$, let $C_2(p) := \sup_{\zeta \in \mathbb{R}^{p(p+1)/2}} \frac{|\!|\!|J(\zeta)|\!|\!|}{\|\zeta\|_\infty}$. Then,*

$$\frac{p}{4} \leq C_2(p) \leq \frac{3}{2}p \ .$$

Proof For the lower bound, we consider

$$\zeta_1 = (\underbrace{1,\dots,1}_{p \text{ times}}, \underbrace{4,\dots,4}_{\frac{p(p-1)}{2} \text{ times}}), \quad \text{then} \quad J(\zeta_1) = \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix}$$

so that $\|\!|J(\zeta)|\!\| = p$ and $\|\zeta\|_\infty = 4$.

For the upper bound, we have for every $\zeta \in \mathbb{R}^{p(p+1)/2}$ and $z \in \mathbb{R}^p$ such that $\|z\|_2 = 1$

$$z^\top J(\zeta) z = \left| \sum_{1 \le i,j \le p} z_i z_j J(\zeta)_{i,j} \right| \le \sum_{1 \le i,j \le p} |z_i|\,|z_j|\,|J(\zeta)| \le \|J(\zeta)\|_\infty \|z\|_1^2 \ .$$

By definition of J, $\|J(\zeta)\|_\infty \le 3/2\,\|\zeta\|_\infty$. Remarking that $\|z\|_1^2 \le p\,\|z\|_2^2$ yields the result. ■

Appendix E. Proof of Theorem 10

The proof of Theorem 10 is similar to the proof of Theorem 3 in Arlot and Bach (2011). We give it here out of completeness.

E.1 Key quantities and their concentration around their means

Definition 16. *We introduce, for $S \in \mathcal{S}_p^{++}(\mathbb{R})$,*

$$\widehat{M}_o(S) \in \operatorname*{argmin}_{M \in \mathcal{M}} \left\{ \left\| \widehat{F}_M - Y \right\|_2 + 2\,\mathrm{tr}\left(A_M \cdot (S \otimes I_n)\right) \right\} \tag{E.1}$$

Definition 17. *Let $S \in \mathcal{S}_p(\mathbb{R})$, we note S_+ the symmetric matrix where the eigenvalues of S have been thresholded at 0. That is, if $S = U^\top D U$, with $U \in \mathcal{O}_p(\mathbb{R})$ and $D = \mathrm{Diag}(d_1,\dots,d_p)$, then*

$$S_+ := U^\top \mathrm{Diag}\left(\max\{d_1,0\},\dots,\max\{d_n,0\}\right) U \ .$$

Definition 18. *For every $M \in \mathcal{M}$, we define*

$$\begin{aligned}
b(M) &= \|(A_M - I_{np}) f\|_2^2 \ , \\
v_1(M) &= \mathbb{E}\left[\langle \varepsilon, A_M \varepsilon \rangle\right] = \mathrm{tr}(A_M \cdot (\Sigma \otimes I_n)) \ , \\
\delta_1(M) &= \langle \varepsilon, A_M \varepsilon \rangle - \mathbb{E}\left[\langle \varepsilon, A_M \varepsilon \rangle\right] = \langle \varepsilon, A_M \varepsilon \rangle - \mathrm{tr}(A_M \cdot (\Sigma \otimes I_n)) \ , \\
v_2(M) &= \mathbb{E}\left[\|A_M \varepsilon\|_2^2\right] = \mathrm{tr}(A_M^\top A_M \cdot (\Sigma \otimes I_n)) \ , \\
\delta_2(M) &= \|A_M \varepsilon\|_2^2 - \mathbb{E}\left[\|A_M \varepsilon\|_2^2\right] = \|A_M \varepsilon\|_2^2 - \mathrm{tr}(A_M^\top A_M \cdot (\Sigma \otimes I_n)) \ , \\
\delta_3(M) &= 2\langle A_M \varepsilon, (A_M - I_{np}) f \rangle \ , \\
\delta_4(M) &= 2\langle \varepsilon, (I_{np} - A_M) f \rangle \ , \\
\widehat{\Delta}(M) &= -2\delta_1(M) + \delta_4(M) \ .
\end{aligned}$$

Definition 19. *Let $C_A, C_B, C_C, C_D, C_E, C_F$ be fixed nonnegative constants. For every $x \geq 0$ we define the event*

$$\Omega_x = \Omega_x(\mathcal{M}, C_A, C_B, C_C, C_D, C_E, C_F)$$

on which, for every $M \in \mathcal{M}$ and $\theta_1, \theta_2, \theta_3, \theta_4 \in (0,1]$:

$$|\delta_1(M)| \leq \theta_1 \operatorname{tr}\left(A_M^\top A_M \cdot (\Sigma \otimes I_n)\right) + (C_A + C_B\theta_1^{-1})x\|\Sigma\| \tag{E.2}$$

$$|\delta_2(M)| \leq \theta_2 \operatorname{tr}\left(A_M^\top A_M \cdot (\Sigma \otimes I_n)\right) + (C_C + C_D\theta_2^{-1})x\|\Sigma\| \tag{E.3}$$

$$|\delta_3(M)| \leq \theta_3 \|(I_{np} - A_M)f\|_2^2 + C_E\theta_3^{-1}x\|\Sigma\| \tag{E.4}$$

$$|\delta_4(M)| \leq \theta_4 \|(I_{np} - A_M)f\|_2^2 + C_F\theta_4^{-1}x\|\Sigma\| \tag{E.5}$$

Of key interest is the concentration of the empirical processes δ_i, uniformly over $M \in \mathcal{M}$. The following Lemma introduces such a result, when $\mathcal{M}$ contains symmetric matrices parametrized with their eigenvalues (with fixed eigenvectors).

Lemma 20. *Let $P \in \mathcal{O}_p(\mathbb{R})$, and suppose that* (**HM**) *holds. Then $\mathbb{P}(\Omega_x(\mathcal{M}, C_A, C_B, C_C, C_D, C_E, C_F)) \geq 1 - pe^{1027+\ln(n)}e^{-x}$ if*

$$C_A = 2,\ C_B = 1,\ C_C = 2,\ C_D = 1,\ C_E = 306.25,\ C_F = 306.25\ .$$

Proof We can write

$$\begin{aligned} A_M = A_{d_1,\ldots,d_p} &= (P \otimes I_n)^\top \left[(D^{-1} \otimes K)\left(D^{-1} \otimes K + npI_{np}\right)^{-1}\right](P \otimes I_n) \\ &= Q^\top \widetilde{A}_{d_1,\ldots,d_p} Q\ , \end{aligned}$$

with $Q = P \otimes I_n$ and $\widetilde{A}_{d_1,\ldots,d_p} = (D^{-1} \otimes K)(D^{-1} \otimes K + npI_{np})^{-1}$. Remark that $\widetilde{A}_{d_1,\ldots,d_p}$ is block-diagonal, with diagonal blocks being $B_{d_1}, \ldots, B_{d_p}$ using the notations of Section 3. With $\widetilde{\varepsilon} = Q\varepsilon = (\widetilde{\varepsilon_1}^\top, \ldots, \widetilde{\varepsilon_p}^\top)^\top$ and $\widetilde{f} = Qf = (\widetilde{f_1}^\top, \ldots, \widetilde{f_p}^\top)^\top$ we can write

$$\begin{aligned} |\delta_1(M)| &= \langle \widetilde{\varepsilon}, \widetilde{A}_{d_1,\ldots,d_p}\widetilde{\varepsilon}\rangle - \mathbb{E}\left[\langle \widetilde{\varepsilon}, \widetilde{A}_{d_1,\ldots,d_p}\widetilde{\varepsilon}\rangle\right]\ , \\ |\delta_2(M)| &= \left\|\widetilde{A}_{d_1,\ldots,d_p}\widetilde{\varepsilon}\right\|_2^2 - \mathbb{E}\left[\left\|\widetilde{A}_{d_1,\ldots,d_p}\widetilde{\varepsilon}\right\|_2^2\right]\ , \\ |\delta_3(M)| &= 2\langle \widetilde{A}_{d_1,\ldots,d_p}\widetilde{\varepsilon}, (\widetilde{A}_{d_1,\ldots,d_p} - I_{np})\widetilde{f}\rangle\ , \\ |\delta_4(M)| &= 2\langle \widetilde{\varepsilon}, (I_{np} - \widetilde{A}_{d_1,\ldots,d_p})\widetilde{f}\rangle\ . \end{aligned}$$

We can see that the quantities δ_i decouple, therefore

$$
\begin{aligned}
|\delta_1(M)| &= \sum_{i=1}^{p} \langle \widetilde{\varepsilon}_i, B_{d_i} \widetilde{\varepsilon}_i \rangle - \mathbb{E}\left[\langle \widetilde{\varepsilon}_i, B_{d_i} \widetilde{\varepsilon} \rangle \right] \ , \\
|\delta_2(M)| &= \sum_{i=1}^{p} \|B_{d_i} \widetilde{\varepsilon}_i\|_2^2 - \mathbb{E}\left[\|B_{d_i} \widetilde{\varepsilon}_i\|_2^2 \right] \ , \\
|\delta_3(M)| &= \sum_{i=1}^{p} 2 \langle B_{d_i} \widetilde{\varepsilon}_i, (B_{d_i} - I_n) \widetilde{f}_i \rangle \ , \\
|\delta_4(M)| &= \sum_{i=1}^{p} 2 \langle \widetilde{\varepsilon}_i, (I_n - B_{d_i}) \widetilde{f}_i \rangle \ .
\end{aligned}
$$

Using Lemma 9 of Arlot and Bach (2011), where we have p concentration results on the sets $\widetilde{\Omega}_i$, each of probability at least $1 - e^{1027+\ln(n)} e^{-x}$ we can state that, on the set $\bigcap_{i=1}^{p} \widetilde{\Omega}_i$, we have

$$
\begin{aligned}
|\delta_1(M)| &\leq \sum_{i=1}^{p} \theta_1 \operatorname{Var}[\widetilde{\varepsilon}_i] \operatorname{tr}(B_{d_i}^\top B_{d_i}) + (C_A + C_B \theta_1^{-1}) x \operatorname{Var}[\widetilde{\varepsilon}_i] \ , \\
|\delta_2(M)| &\leq \sum_{i=1}^{p} \theta_2 \operatorname{Var}[\widetilde{\varepsilon}_i] \operatorname{tr}(B_{d_i}^\top B_{d_i}) + (C_C + C_D \theta_2^{-1}) x \operatorname{Var}[\widetilde{\varepsilon}_i] \ , \\
|\delta_3(M)| &\leq \sum_{i=1}^{p} \theta_3 \left\| (I_n - B_{d_i}) \widetilde{f}_i \right\|_2^2 + C_E \theta_3^{-1} x \operatorname{Var}[\widetilde{\varepsilon}_i] \ , \\
|\delta_4(M)| &\leq \sum_{i=1}^{p} \theta_4 \left\| (I_n - B_{d_i}) \widetilde{f}_i \right\|_2^2 + C_F \theta_4^{-1} x \operatorname{Var}[\widetilde{\varepsilon}_i] \ .
\end{aligned}
$$

To conclude, it suffices to see that for every $i \in \{1, \ldots, p\}$, $\operatorname{Var}[\widetilde{\varepsilon}_i] \leq \|\Sigma\|$. ■

E.2 Intermediate result

We first prove a general oracle inequality, under the assumption that we use inside the penalty an estimation of Σ which does not underestimate Σ too much.

Proposition 21. *Let $C_A, C_B, C_C, C_D, C_E \geq 0$ be fixed constants, $\gamma > 0$, $\theta_S \in [0, 1/4)$ and $K_S \geq 0$. On $\Omega_{\gamma \ln(n)}(\mathcal{M}, C_A, C_B, C_C, C_D, C_E)$, for every $S \in \mathcal{S}_p^{++}(\mathbb{R})$ such that*

$$
S \succeq \Sigma \left(1 - \theta_S \inf_{M \in \mathcal{M}} \left\{ \frac{b(M) + v_2(M) + K_S \ln(n) \|\Sigma\|}{v_1(M)} \right\} \right) \tag{E.6}
$$

and for every $\theta \in (0, (1 - 4\theta_S)/2)$, we have:

$$\frac{1}{np}\left\|\widehat{f}_{M_o(S)}-f\right\|_2^2 \leq \frac{1+2\theta}{1-2\theta-4\theta_S}\inf_{M\in\mathcal{M}}\left\{\frac{1}{np}\left\|\widehat{F}_M-F\right\|_2^2+\frac{2\operatorname{tr}\left(A_M\cdot((S-\Sigma)_+\otimes I_n)\right)}{np}\right\}$$
$$+\frac{1}{1-2\theta-4\theta_S}\left[(2C_A+3C_C+6C_D+6C_E+\frac{2}{\theta}(C_B+C_F))\gamma+\frac{\theta_S K_S}{4}\right]\frac{\ln(n)|||\Sigma|||}{np} \quad \text{(E.7)}$$

Proof The proof of Proposition 21 is very similar to the one of Proposition 5 in Arlot and Bach (2011). First, we have

$$\left\|\widehat{f}_M-f\right\|_2^2 = b(M)+v_2(M)+\delta_2(M)+\delta_3(M)\ , \quad \text{(E.8)}$$
$$\left\|\widehat{f}_M-y\right\|_2^2 = \|\widehat{f}_M-f\|_2^2-2v_1(M)-2\delta_1(M)+\delta_4(M)+\|\varepsilon\|_2^2\ . \quad \text{(E.9)}$$

Combining Eq. (E.1) and (E.9), we get:

$$\begin{aligned}&\left\|\widehat{f}_{\widehat{M}_o(S)}-f\right\|_2^2+2\operatorname{tr}\left(A_{\widehat{M}_o(S)}\cdot((S-\Sigma)_+\otimes I_n)\right)+\widehat{\Delta}(\widehat{M}_o(S))\\ &\leq \inf_{M\in\mathcal{M}}\left\{\left\|\widehat{f}_M-f\right\|_2^2+2\operatorname{tr}\left(A_M\cdot((S-\Sigma)\otimes I_n)\right)+\widehat{\Delta}(M)\right\}\ .\end{aligned} \quad \text{(E.10)}$$

On the event $\Omega_{\gamma\ln(n)}$, for every $\theta\in(0,1]$ and $M\in\mathcal{M}$, using Eq. (E.2) and (E.5) with $\theta=\theta_1=\theta_4$,

$$|\widehat{\Delta}(M)| \leq \theta(b(M)+v_2(M))+(C_A+\frac{1}{\theta}(C_B+C_F))\gamma\ln(n)|||\Sigma|||\ . \quad \text{(E.11)}$$

Using Eq. (E.3) and (E.4) with $\theta_2=\theta_3=1/2$ we get that for every $M\in\mathcal{M}$ Eq.

$$\left\|\widehat{F}_M-F\right\|_2^2 \geq \frac{1}{2}(b(M)+v_2(M))-(C_C+2C_D+2C_E)\gamma\ln(n)|||\Sigma|||\ ,$$

which is equivalent to

$$b(M)+v_2(M) \leq 2\left\|\widehat{F}_M-F\right\|_2^2+2(C_C+2C_D+2C_E)\gamma\ln(n)|||\Sigma|||\ . \quad \text{(E.12)}$$

Combining Eq. (E.11) and (E.12), we get

$$|\widehat{\Delta}(M)| \leq 2\theta\left\|\widehat{F}_M-F\right\|_2^2+\left(C_A+(2C_C+4C_D+4C_E)\theta+(C_B+C_F)\frac{1}{\theta}\right)\gamma\ln(n)|||\Sigma|||\ .$$

With Eq. (E.10), and with $C_1=C_A$, $C_2=2C_C+4C_D+4C_E$ and $C_3=C_B+C_F$ we get

$$\begin{aligned}&(1-2\theta)\left\|\widehat{f}_{\widehat{M}_o(S)}-f\right\|_2^2+2\operatorname{tr}\left(A_{\widehat{M}_o(S)}\cdot((S-\Sigma)_+\otimes I_n)\right)\leq\\ &\inf_{M\in\mathcal{M}}\left\{\left\|\widehat{f}_M-f\right\|_2^2+2\operatorname{tr}\left(A_M\cdot((S-\Sigma)\otimes I_n)\right)\right\}+\left(C_1+C_2\theta+\frac{C_3}{\theta}\right)\gamma\ln(n)|||\Sigma|||\ .\end{aligned} \quad \text{(E.13)}$$

Using Eq. (E.6) we can state that

$$\operatorname{tr}\left(A_{\widehat{M}_o(S)}\cdot((S-\Sigma)\otimes I_n)\right) \geq -\theta_S\left((b(\widehat{M}_o(S))+v_2(\widehat{M}_o(S))+K_S\ln(n)|||S|||\right)\ ,$$

which then leads to Eq. (E.7) using Eq. (E.12) and (E.13). ■

E.3 The proof itself

We now show Theorem 10 as a consequence of Proposition 21. It actually suffices to show that $\widehat{\Sigma}$ does not underestimate Σ too much, and that the second term in the infimum of Eq. (E.7) is negligible in front of the quadratic error $(np)^{-1}\|\widehat{f}_M - f\|^2$.

Proof On the event Ω introduced in Theorem 8, Eq. (4.2) holds. Let

$$\gamma = c(\Sigma)\left(1 + c(\Sigma)\right) \quad .$$

By Lemma 22 below, we have:

$$\inf_{M\in\mathcal{M}} \left\{ \frac{b(M) + v_2(M) + K_S \ln(n) |\!|\!| \Sigma |\!|\!|}{v_1(M)} \right\} \geq 2\sqrt{\frac{K_S \ln(n) |\!|\!| \Sigma |\!|\!|}{n \operatorname{tr}(\Sigma)}} \quad .$$

In order to have $\widehat{M}_o(\widehat{\Sigma})$ satisfying Eq. (E.6), it suffices to have, for every $\theta_S > 0$,

$$2\theta_S \sqrt{\frac{K_S \ln(n) |\!|\!| \Sigma |\!|\!|}{n \operatorname{tr}(\Sigma)}} = 6\beta(\alpha+\delta) p \gamma \sqrt{\frac{\ln(n)}{n}} \quad ,$$

which leads to the choice

$$K_S = \left(\frac{3\beta(\alpha+\delta)\gamma \operatorname{tr}(\Sigma)}{\theta_S |\!|\!| \Sigma |\!|\!|} \right)^2 \quad .$$

We now take $\theta_S = \theta = (9\ln(n))^{-1}$. Using Eq. (E.7) and requiring that $\ln(n) \geq 6$, we get on $\widetilde{\Omega} = \Omega \cap \Omega_{(\alpha+\delta)\ln(n)}(\mathcal{M}, C_A, C_B, C_C, C_D, C_E, C_F)$:

$$\begin{aligned}
\frac{1}{np}\left\| \widehat{f}_{\widehat{M}} - f \right\|_2 \leq \left(1 + \frac{1}{\ln(n)}\right) \inf_{M\in\mathcal{M}} \left\{ \frac{1}{np}\left\| \widehat{f}_M - f \right\|_2^2 + \frac{2\operatorname{tr}\left(A_M \cdot ((\widehat{\Sigma} - \Sigma)_+ \otimes I_n) \right)}{np} \right\} \\
+ \left(1 - \frac{2}{3\ln(n)}\right)^{-1} \left[2C_A + 3C_C + 6C_D + 6C_E + \ln(n)\left(18C_B + 18C_F + \frac{729\beta^2 p^2 \gamma^2 \operatorname{tr}(\Sigma)^2}{4|\!|\!|\Sigma|\!|\!|^2} \right) \right] \\
\times (\alpha+\delta)^2 \frac{\ln(n)^2 |\!|\!| \Sigma |\!|\!|}{np}
\end{aligned}$$

Using Eq. (D.5) and defining

$$\eta_2 := 12\beta(\alpha+\delta) p \sqrt{\frac{\ln(n)}{n}} c(\Sigma)\left(1 + c(\Sigma)\right) \quad , \tag{E.14}$$

we get

$$\begin{aligned}
\frac{1}{np}\left\| \widehat{f}_{\widehat{M}} - f \right\|_2 \leq \left(1 + \frac{1}{\ln(n)}\right) \inf_{M\in\mathcal{M}} \left\{ \frac{1}{np}\left\| \widehat{f}_M - f \right\|_2^2 + \eta_2 \frac{\operatorname{tr}(A_M \cdot (\Sigma \otimes I_n))}{np} \right\} \\
+ \left(1 - \frac{2}{3\ln(n)}\right)^{-1} \left[2C_A + 3C_C + 6C_D + 6C_E + \ln(n)\left(18C_B + 18C_F + \frac{729\beta^2 p^2 \gamma^2 \operatorname{tr}(\Sigma)^2}{4|\!|\!|\Sigma|\!|\!|^2} \right) \right] \\
\times (\alpha+\delta)^2 \frac{\ln(n)^2 |\!|\!| \Sigma |\!|\!|}{np} \quad .
\end{aligned} \tag{E.15}$$

Now, to get a classical oracle inequality, we have to show that $\eta_2 v_1(M) = \eta_2 \operatorname{tr}(A_M \cdot (\Sigma \otimes I_n))$ is negligible in front of $\|\widehat{f}_M - f\|^2$. Lemma 22 ensures that:

$$\forall M \in \mathcal{M}\,, \forall x \geq 0\,, \quad 2\sqrt{\frac{x|||\Sigma|||}{n \operatorname{tr}(\Sigma)}} v_1(M) \leq v_2(M) + x|||\Sigma||| \ .$$

With $0 < C_n < 1$, taking x to be equal to $72\beta^2 p^2 \ln(n) c(\Sigma)^2 (1+c(\Sigma))^2 \operatorname{tr}(\Sigma)/(C_n|||\Sigma|||)$ leads to

$$\eta_2 v_1(M) \leq 2C_n v_2(M) + \frac{72\beta^2 p^2 \ln(n) c(\Sigma)^2 (1+c(\Sigma))^2 \operatorname{tr}(\Sigma)}{C_n} \ . \tag{E.16}$$

Then, since $v_2(M) \leq v_2(M) + b(M)$ and using also Eq. (E.8), we get

$$v_2(M) \leq \left\| \widehat{f}_M - f \right\|_2^2 + |\delta_2(m)| + |\delta_3(M)| \ .$$

On $\widetilde{\Omega}$ we have that for every $\theta \in (0,1)$, using Eq. (E.3) and (E.4),

$$|\delta_2(M)| + |\delta_3(M)| \leq 2\theta \left(\left\| \widehat{f}_M - f \right\|_2^2 - |\delta_2(M)| - |\delta_3(M)| \right) + (C_C + (C_D + C_E)\theta^{-1})(\alpha+\delta)\ln(n)|||\Sigma||| \ ,$$

which leads to

$$|\delta_2(M)| + |\delta_3(M)| \leq \frac{2\theta}{1+2\theta} \left\| \widehat{f}_M - f \right\|_2^2 + \frac{C_C + (C_D + C_E)\theta^{-1}}{1+2\theta} (\alpha+\delta)\ln(n)|||\Sigma||| \ .$$

Now, combining this equation with Eq. (E.16), we get

$$\eta_2 v_1(M) \leq \left(1 + \frac{4C_n\theta}{1+2\theta}\right) \left\| \widehat{f}_M - f \right\|_2^2 + 2C_n \frac{C_C + (C_D + C_E)\theta^{-1}}{1+2\theta} (\alpha+\delta)\ln(n)|||\Sigma||| + \frac{72\beta^2 p^2 \ln(n) c(\Sigma)^2 (1+c(\Sigma))^2 \operatorname{tr}(\Sigma)}{C_n} \ .$$

Taking $\theta = 1/2$ then leads to

$$\eta_2 v_1(M) \leq (1+C_n) \left\| \widehat{f}_M - f \right\|_2^2 + C_n (C_C + 2(C_D + C_E))(\alpha+\delta)\ln(n)|||\Sigma||| + \frac{72\beta^2 p^2 \ln(n) c(\Sigma)^2 (1+c(\Sigma))^2 \operatorname{tr}(\Sigma)}{C_n} \ .$$

We now take $C_n = 1/\ln(n)$. We now replace the constants C_A, C_B, C_C, C_D, C_E, C_F by their values in Lemma 20, and if we require that $2\ln(n) \geq 1027$, we get, for some constant L_3,

$$\left(1 - \frac{2}{3\ln(n)}\right)^{-1} \left[1851.5 + \ln(n)\left(5530.5 + \frac{729\beta^2 p^2 \gamma^2 \operatorname{tr}(\Sigma)^2}{4|||\Sigma|||^2}\right) + 616.5\left(1 + \frac{1}{\ln(n)}\right)\frac{1}{\ln(n)}\right] + \frac{72\beta^2 p^2 \ln(n) c(\Sigma)^2 (1+c(\Sigma))^2 \operatorname{tr}(\Sigma)}{C_n} \leq L_3 \ln(n) p^2 c(\Sigma)^4 \frac{\operatorname{tr}(\Sigma)^2}{|||\Sigma|||^2} \tag{E.17}$$

From this we can deduce Eq. (5.2).

Finally we deduce an oracle inequality in expectation by noting that if $n^{-1}\|f_{\widehat{M}} - f\|^2 \leq R_{n,\delta}$ on $\widetilde{\Omega}$, using Cauchy-Schwarz inequality

$$\mathbb{E}\left[\frac{1}{np}\left\|\widehat{f}_{\widehat{M}} - f\right\|_2^2\right] = \mathbb{E}\left[\frac{\mathbf{1}_{\widetilde{\Omega}}}{np}\left\|\widehat{f}_{\widehat{M}} - f\right\|_2^2\right] + \mathbb{E}\left[\frac{\mathbf{1}_{\widetilde{\Omega}^c}}{np}\left\|\widehat{f}_{\widehat{M}} - f\right\|_2^2\right]$$
$$\leq \mathbb{E}\left[R_{n,\delta}\right] + \frac{1}{np}\sqrt{\frac{4p(p+1)+p}{n^\delta}}\sqrt{\mathbb{E}\left[\left\|\widehat{f}_{\widehat{M}} - f\right\|_2^4\right]} \quad . \tag{E.18}$$

We can remark that, since $\|\!|A_M|\!\| \leq 1$,

$$\left\|\widehat{f}_M - f\right\|_2^2 \leq 2\left\|A_M\varepsilon\right\|_2^2 + 2\left\|(I_{np} - A_M)f\right\|_2^2 \leq 2\left\|\varepsilon\right\|_2^2 + 8\left\|f\right\|_2^2 \quad .$$

So

$$\mathbb{E}\left[\left\|\widehat{f}_{\widehat{M}} - f\right\|_2^4\right] \leq 12\left(np\|\!|\Sigma|\!\| + 4\left\|f\right\|_2^2\right)^2 \quad ,$$

together with Eq. (E.15) and Eq. (E.18), induces Eq. (5.3), using that for some constant $L_4 > 0$,

$$12\sqrt{\frac{4p(p+1)+p}{n^\delta}}\left(\|\!|\Sigma|\!\| + \frac{4}{np}\left\|f\right\|_2^2\right) \leq L_4\frac{p}{n^{\delta/2}}\left(\|\!|\Sigma|\!\| + \frac{1}{np}\left\|f\right\|_2^2\right) \quad .$$

We can finally define the constant L_2 by:

$$L_3 c(\Sigma)^4 \operatorname{tr}(\Sigma)(\alpha+\delta)^2\frac{p^3\ln(n)^3}{np} + L_4\frac{p}{n^{\delta/2}}\|\!|\Sigma|\!\| \leq L_2 c(\Sigma)^4\operatorname{tr}(\Sigma)(\alpha+\delta)^2\frac{p^3\ln(n)^3}{np}$$

■

Lemma 22. *Let $n, p \geq 1$ be two integers, $x \geq 0$ and $\Sigma \in \mathcal{S}_p^{++}(\mathbb{R})$. Then,*

$$\inf_{A\in\mathcal{M}_{np}(\mathbb{R}),\|\!|A|\!\|\leq 1}\left\{\frac{\operatorname{tr}(A^\top A\cdot(\Sigma\otimes I_n)) + x\|\!|\Sigma|\!\|}{\operatorname{tr}(A\cdot(\Sigma\otimes I_n))}\right\} \geq 2\sqrt{\frac{x\|\!|\Sigma|\!\|}{n\operatorname{tr}(\Sigma)}}$$

Proof First note that the bilinear form on $\mathcal{M}_{np}(\mathbb{R})$, $(A,B)\mapsto \operatorname{tr}(A^\top B\cdot(\Sigma\otimes I_n))$ is a scalar product. By Cauchy-Schwarz inequality, for every $A\in\mathcal{M}_{np}(\mathbb{R})$,

$$\operatorname{tr}(A\cdot(\Sigma\otimes I_n))^2 \leq \operatorname{tr}(\Sigma\otimes I_n)\operatorname{tr}(A^\top A\cdot(\Sigma\otimes I_n)) \quad .$$

Thus, since $\operatorname{tr}(\Sigma\otimes I_n) = n\operatorname{tr}(\Sigma)$, if $c = \operatorname{tr}(A\cdot(\Sigma\otimes I_n)) > 0$,

$$\operatorname{tr}(A^\top A\cdot(\Sigma\otimes I_n)) \geq \frac{c^2}{n\operatorname{tr}(\Sigma)} \quad .$$

Therefore

$$\inf_{A\in\mathcal{M}_{np}(\mathbb{R}),\|A\|\leq 1}\left\{\frac{\operatorname{tr}(A^\top A\cdot(\Sigma\otimes I_n))+x\|\Sigma\|}{\operatorname{tr}(A\cdot(\Sigma\otimes I_n))}\right\}\geq\inf_{c>0}\left\{\frac{c}{n\operatorname{tr}(\Sigma)}+\frac{x\|\Sigma\|}{c}\right\}$$
$$\geq 2\sqrt{\frac{x\|\Sigma\|}{n\operatorname{tr}(\Sigma)}}\ .$$

■

Bibliography

Hirotogu Akaike. Statistical predictor identification. *Annals of the Institute of Statistical Mathematics*, 22:203–217, 1970.

Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, December 2005. ISSN 1532-4435.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

Sylvain Arlot. Model selection by resampling penalization. *Electron. J. Stat.*, 3:557–624 (electronic), 2009. ISSN 1935-7524. doi: 10.1214/08-EJS196.

Sylvain Arlot and Francis Bach. Data-driven calibration of linear estimators with minimal penalties, July 2011. arXiv:0909.1884v2.

Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10:245–279 (electronic), 2009.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, May 1950.

Bart Bakker and Tom Heskes. Task clustering and gating for bayesian multitask learning. *Journal of Machine Learning Research*, 4:83–99, December 2003. ISSN 1532-4435. doi: http://dx.doi.org/10.1162/153244304322765658.

Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138:33–73, 2007.

Philip J. Brown and James V. Zidek. Adaptive multivariate ridge regression. *The Annals of Statistics*, 8(1):pp. 64–74, 1980. ISSN 00905364.

Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, July 1997. ISSN 0885-6125. doi: 10.1023/A:1007379606734.

Theodoros Evgeniou, Charles A. Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

Gilles Gasso, Alain Rakotomamonjy, and Stéphane Canu. Recovering sparse signals with non-convex penalties and dc programming. *IEEE Trans. Signal Processing*, 57(12):4686–4698, 2009.

Trevor J. Hastie and Robert J. Tibshirani. *Generalized Additive Models*. Taylor and Francis, 1990. ISBN 9780412343902.

Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991. ISBN 9780521467131.

Laurent Jacob, Francis Bach, and Jean-Philippe Vert. Clustered multi-task learning: A convex formulation. *Computing Research Repository*, pages –1–1, 2008.

Matthieu Lerasle. Optimal model selection in density estimation. *Ann. Inst. H. Poincaré Probab. Statist.*, 2011. ISSN 0246-0203. Accepted. arXiv:0910.1654.

Percy Liang, Francis Bach, Guillaume Bouchard, and Michael I. Jordan. Asymptotically optimal regularization in smooth parametric models. In *Advances in Neural Information Processing Systems*, 2010.

Karim Lounici, Massimiliano Pontil, Alexandre B. Tsybakov, and Sara van de Geer. Oracle inequalities and optimal inference under group sparsity. Technical Report arXiv:1007.1771, Jul 2010. Comments: 37 pages.

Guillaume Obozinski, Martin J. Wainwright, and Michael I. Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1):1–17, 2011.

Carl E. Rasmussen and Christopher K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, 12 2002.

Sebastian Thrun and Joseph O'Sullivan. Discovering structure in multiple learning tasks: The TC algorithm. *Proceedings of the 13th International Conference on Machine Learning*, 1996.

Grace Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990. ISBN 0-89871-244-0.

Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098, 2005.