Multi-task Regression using Minimal Penalties

Matthieu Solnon, Sylvain Arlot, Francis Bach

To cite this version:

Matthieu Solnon, Sylvain Arlot, Francis Bach. Multi-task Regression using Minimal Penalties. Journal of Machine Learning Research, Microtome Publishing, 2012, 13, pp.2773-2812. hal-00610534v3

HAL Id: hal-00610534

https://hal.archives-ouvertes.fr/hal-00610534v3

Submitted on 23 Oct 2012

Multi-task Regression using Minimal Penalties

Matthieu Solnon MATTHIEU.SOLNON@ENS.FR
ENS; Sierra Project-team
Département 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
Département 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
Département d'Informatique de l'École Normale Supérieure
(CNRS/ENS/INRIA UMR 8548)
23, avenue d'Italie, CS 81321
75214 Paris Cedex 13, France

Editor: Tong Zhang

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 the sample size, for instance because of experimental limitations. Hopefully, in many situations practicioners can find many related and similar problems, and might use these problems as if more observations were available for the 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 $f(X_i) = \mathbb{E}[Y_i|X_i]$, by minimizing the quadratic risk of the estimator regularized by a certain norm. A practically important task is to calibrate a regularization parameter, that is, 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, for example, 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 empirically show that it can lead to performance improvement (Thrun and O’Sullivan, 1996; Caruana, 1997; Bakker and Heskes, 2003). 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 consider a setting where $p \gg n$, and formulate a sparsity assumption which enables to use the group Lasso, assuming all the different functions have a small set of common active covariates (see for instance Obozinski et al., 2011; Lounici et al., 2010). We exclude this setting from our analysis, because of the Hilbertian nature of our problem, and thus will not consider the similarity between the tasks in terms of sparsity, but rather in terms of an Euclidean similarity. Another theoretical approach has also been 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 M (typically of size p), which characterizes the relationship between the tasks and extends the regularization parameter from 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 for the estimation of the task similarity matrix M, both with high probability and in expectation. Finally we give some simulation experiment results and show that our technique correctly deals with the multi-task settings with a low sample-size.

1.1 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

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

  that is, the vector in which the columns $Y^j := (Y_{i,j})_{1 \le i \le n}$ 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.
- $\lVert\cdot\rVert_2$ is the usual Euclidean norm on $\mathbb{R}^k$ for any $k \in \mathbb{N}$: $\forall u \in \mathbb{R}^k$, $\lVert u \rVert_2^2 := \sum_{i=1}^k u_i^2$.

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.

2.1 Multi-task with a Fixed Kernel

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 $\mathcal{D}_n = (X_i, Y_i^1, \dots, Y_i^p)_{i=1}^n \in (\mathcal{X} \times \mathbb{R}^p)^n$, which gives us the positive semidefinite kernel matrix $K = (k(X_i, X_\ell))_{1 \le i, \ell \le n} \in \mathcal{S}_n^+(\mathbb{R})$. For each task $j \in \{1, \dots, p\}$, $\mathcal{D}_n^j = (X_i, y_i^j)_{i=1}^n$ 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 carried out similarly by defining $X_i = (X_i^1, \dots, X_i^p)$, but the notations would be more complicated.

We now define the model. We assume $(f^1, \dots, f^p) \in \mathcal{F}^p$, 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

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

This means that, while the observations are independent, the outputs of the different tasks can be 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 $\big(f^1(X_i),\dots,f^p(X_i)\big)_{i=1}^n$. 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 $(f^j(X_i))_{1\le i\le n\,,\,1\le j\le p}$ and introduce the vector $f := \mathrm{vec}(F) = (f^1(X_1),\dots,f^1(X_n),\dots,f^p(X_1),\dots,f^p(X_n)) \in \mathbb{R}^{np}$, 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. Generalizing the work of Evgeniou et al. (2005), we estimate $(f^1,\dots,f^p) \in \mathcal{F}^p$ by

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

Although M could have a general unconstrained form we may restrict M to certain forms, for either computational or statistical reasons.

Remark 2 *Requiring that $M \succeq 0$ implies that Equation (2) is a convex optimization problem, which 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 3 *The case where the p tasks are treated independently can be considered in this setting: taking $M = M_{\mathrm{ind}}(\lambda) := \frac{1}{p}\mathrm{Diag}(\lambda_1,\dots,\lambda_p)$ for any $\lambda \in \mathbb{R}^p$ leads to the criterion*

$$\frac{1}{p}\sum_{j=1}^p \left[\frac{1}{n}\sum_{i=1}^n (y_i^j - g^j(X_i))^2 + \lambda_j \|g^j\|_{\mathcal{F}}^2 \right] \ , \tag{3}$$

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

Example 4 *As done by Evgeniou et al. (2005), for every $\lambda,\mu \in (0,+\infty)^2$, define*

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

Taking $M = M_{\mathrm{similar}}(\lambda,\mu)$ in Equation (2) leads to the criterion

$$\frac{1}{np}\sum_{i=1}^n\sum_{j=1}^p (y_i^j - g^j(X_i))^2 + \lambda\sum_{j=1}^p \left\|g^j\right\|_{\mathcal{F}}^2 + \frac{\mu}{2}\sum_{j=1}^p\sum_{k=1}^p \left\|g^j - g^k\right\|_{\mathcal{F}}^2 \ . \tag{4}$$

Minimizing Equation (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 contributions of the paper is to go beyond this case and learn from data a more general similarity matrix* M *between tasks.*

Example 5 *We extend Example 4 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* $\{1, \ldots, p\}$, $\mathbf{1}_I$ *the vector* v *with components* $v_i = \mathbf{1}_{i \in I}$, *and* $\operatorname{Diag}(I)$ *the diagonal matrix* d *with components* $d_{i,i} = \mathbf{1}_{i \in I}$. *We then define*

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

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:

$$\frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} (y_i^j - g^j(X_i))^2 + \lambda \sum_{j=1}^{p} \left\| g^j \right\|_{\mathcal{F}}^2 + \frac{\mu}{2k} \sum_{j \in I} \sum_{k \in I} \left\| g^j - g^k \right\|_{\mathcal{F}}^2 + \frac{\nu}{2(p-k)} \sum_{j \in I^c} \sum_{k \in I^c} \left\| g^j - g^k \right\|_{\mathcal{F}}^2 \enspace . \tag{5}$$

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

Remark 6 *Since* I_p *and* $\mathbf{1}\mathbf{1}^\top$ *can be diagonalized simultaneously, minimizing Equation* (4) *and Equation* (5) *is quite easy: it only demands optimization over two independent parameters, which can be done with the procedure of Arlot and Bach (2011).*

Remark 7 *As stated below (Proposition 8),* 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 $\mathcal{X} \times \{1, \ldots, p\}$ by

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

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

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

which is a scalar product as soon as M is positive semi-definite (see proof in Appendix A) and leads to a RKHS (see proof in Appendix B):

Proposition 8 *With the preceding notations* $\langle \cdot, \cdot \rangle_{\mathcal{G}}$ *is a scalar product on* $\mathcal{G}$.

Corollary 9 $(\mathcal{G}, \langle \cdot, \cdot \rangle_{\mathcal{G}})$ *is a RKHS.*

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

Definition 10 (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 $p \times q$ blocks, the block of index (i,j) being $A_{i,j} \cdot B$:*

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

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

Proposition 11 *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 $\widetilde{K}_M := M^{-1} \otimes K$.*

Proposition 11 is proved in Appendix C. We can then apply the representer's theorem (Schölkopf and Smola, 2002) to the minimization problem (2) and deduce that $\widehat{f}_M = A_M y$ with

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

2.2 Optimal Choice of the Kernel

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 oracle risk is defined as

$$\inf_{M \in \mathcal{M}} \left\{ \left\| \widehat{f}_M - f \right\|_2^2 \right\} . \tag{6}$$

The ideal choice, called the oracle, is any matrix

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

Nothing here ensures the oracle exists. However in some special cases (see for instance Example 12) the infimum of $\| \widehat{f}_M - f \|^2$ over the set $\{ \widehat{f}_M, \ M \in \mathcal{M} \}$ may be attained by a function $f^* \in \mathcal{F}^p$—which we will call "oracle" by a slight abuse of notation—while the former problem does not have a solution.

From now on we always suppose that the infimum of $\{ \| \widehat{f}_M - f \|^2 \}$ over $\mathcal{M}$ is attained by some function $f^\star \in \mathcal{F}^p$. However the oracle $M^\star$ is not an estimator, since it depends on f.

Example 12 (Partial computation of the oracle in a simple setting) *It is possible in certain simple settings to exactly compute the oracle (or, at least, some part of it).*

Consider for instance the set-up where the p functions are taken to be equal (that is, $f^1 = \cdots = f^p$). In this setting it is natural to use the set

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

Using the estimator $\widehat{f}_M = A_M y$ we can then compute the quadratic risk using the bias-variance decomposition given in Equation (36):

$$\mathbb{E}\left[\left\| \widehat{f}_M - f \right\|_2^2 \right] = \| (A_M - I_{np}) f \|_2^2 + \operatorname{tr}(A_M^\top A_M \cdot (\Sigma \otimes I_n)) \ .$$

Computations (*reported in Appendix D*) *show that, with the change of variables $\widetilde{\mu} = \lambda + p\mu$, the bias does not depend on $\widetilde{\mu}$ and the variance is a decreasing function of $\widetilde{\mu}$. Thus the oracle is obtained when $\widetilde{\mu} = +\infty$, leading to a situation where the oracle functions $f^{1,\star}, \ldots, f^{p,\star}$ verify $f^{1,\star} = \cdots = f^{p,\star}$. It is also noticeable that, if one assumes the maximal eigenvalue of Σ stays bounded with respect to p, the variance is of order $\mathcal{O}(p^{-1})$ while the bias is bounded with respect to p.*

As explained by Arlot and Bach (2011), we choose

$$\widehat{M} \in \operatorname*{argmin}_{M \in \mathcal{M}} \{ \operatorname{crit}(M) \} \quad \text{with} \quad \operatorname{crit}(M) = \frac{1}{np} \left\| y - \widehat{f}_M \right\|_2^2 + \operatorname{pen}(M) \ ,$$

where the penalty term $\operatorname{pen}(M)$ has to be chosen appropriately.

Remark 13 *Our model* (1) *does not constrain the functions $f^1, \ldots, f^p$. Our way to express the similarities between the tasks (that is, between the f^j) is via the set $\mathcal{M}$, which represents the a priori knowledge the statistician has about the problem. Our goal is to build an estimator whose risk is the closest possible to the oracle risk. Of course using an inappropriate set $\mathcal{M}$ (with respect to the target functions $f^1, \ldots, f^p$) may lead to bad overall performances. Explicit multi-task settings are given in Examples 3, 4 and 5 and through simulations in Section 6.*

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*

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

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,

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

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

$$\mathrm{pen}_{\mathrm{id}}(M) = \frac{2\,\mathrm{tr}\left(A_M \cdot (\Sigma \otimes I_n)\right)}{np} \enspace . \tag{7}$$

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.

Note that estimating Σ is a mean towards estimating M. The technique we develop later for this purpose is not purely a multi-task technique, and may also be used in a different context.

3. Single Task Framework: Estimating a Single Variance

This section recalls some of the main results from Arlot and Bach (2011) which can be considered as solving 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

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

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

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

By analogy with the case where A_λ is an orthogonal projection matrix, $\mathrm{df}(\lambda) := \mathrm{tr}(A_\lambda)$ is called the effective degree of freedom, first introduced by Mallows (1973); see also the work 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

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

We shall see now that it is a *minimal penalty* in the following sense. If for every $C > 0$

$$\widehat{\lambda}_0(C) \in \operatorname*{argmin}_{\lambda \in [0,+\infty]} \left\{ \frac{1}{n} \left\lVert A_{\lambda,K} Y - Y \right\rVert_2^2 + C \,\mathrm{pen}_{\min}(\lambda, K) \right\} \enspace ,$$

then—up to concentration inequalities—$\widehat{\lambda}_0(C)$ acts as a mimimizer of

$$g_C(\lambda) = \mathbb{E}\left[\frac{1}{n} \left\lVert A_\lambda Y - Y \right\rVert_2^2 + C \,\mathrm{pen}_{\min}(\lambda) \right] - \sigma^2 = \frac{1}{n} \left\lVert (A_\lambda - I_n) f \right\rVert_2^2 + (C - \sigma^2) \,\mathrm{pen}_{\min}(\lambda) \enspace .$$

The former theoretical arguments show that

- if $C < \sigma^2$, $g_C(\lambda)$ decreases with $\mathrm{df}(\lambda)$ so that $\mathrm{df}(\widehat{\lambda}_0(C))$ is huge: the procedure overfits;
- if $C > \sigma^2$, $g_C(\lambda)$ increases with $\mathrm{df}(\lambda)$ when $\mathrm{df}(\lambda)$ is large enough so that $\mathrm{df}(\widehat{\lambda}_0(C))$ is much smaller than when $C < \sigma^2$.

The following algorithm was introduced by Arlot and Bach (2011) and uses this fact to estimate σ^2.

Algorithm 14 **Input:** $Y \in \mathbb{R}^n$, $K \in \mathcal{S}_n^{++}(\mathbb{R})$

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

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

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 14 is detailed by Arlot and Massart (2009), and we discuss the way we implemented Algorithm 14 in Section 6. The output $\widehat{C}$ of Algorithm 14 is a provably consistent estimator of σ^2, as stated in the following theorem.

Theorem 15 (Corollary of Theorem 1 of Arlot and Bach, 2011) *Let* $\beta = 150$. *Suppose*
$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ *with* $\sigma^2 > 0$, *and that* $\lambda_0 \in (0, +\infty)$ *and* $d_n \geq 1$ *exist such that*

$$\mathrm{df}(\lambda_0) \leq \sqrt{n} \text{ and } \frac{1}{n} \left\| (A_{\lambda_0} - I_n) F \right\|_2^2 \leq d_n \sigma^2 \sqrt{\frac{\ln n}{n}} \ . \tag{8}$$

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

$$\left(1 - \beta(2+\delta) \sqrt{\frac{\ln n}{n}} \right) \sigma^2 \leq \widehat{C} \leq \left(1 + \beta(2+\delta) d_n \sqrt{\frac{\ln(n)}{n}} \right) \sigma^2 \ . \tag{9}$$

Remark 16 *The values* $n/10$ *and* $n/3$ *in Algorithm 14 have no particular meaning and can be replaced by* n/k, n/k', *with* $k > k' > 2$. *Only* β *depends on* k *and* k'. *Also the bounds required in Assumption* (8) *only impact the right hand side of Equation* (9) *and are chosen to match the left hand side. See Proposition 10 of Arlot and Bach (2011) for more details.*

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

$$Y_z := Y \cdot z = F \cdot z + E \cdot z = F_z + \varepsilon_z \ , \tag{10}$$

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 17 *Let $a(z)$ be the output $\widehat{C}$ of Algorithm 14 applied to problem* (10)*, that is, with inputs $Y_z \in \mathbb{R}^n$ and $K \in \mathcal{S}_n^{++}(\mathbb{R})$.*

The idea is to apply Algorithm 14 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 $\mathcal{Z} = \{e_i,\ i \in \{1,\dots,p\}\} \cup \{e_i + e_j,\ 1 \le i < j \le p\}$. 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 18 *Let $J : \mathbb{R}^{\frac{p(p+1)}{2}} \to \mathcal{S}_p(\mathbb{R})$ be defined by*

$$J(a_1,\dots,a_p,a_{1,2},\dots,a_{1,p},\dots,a_{p-1,p})_{i,i} = a_i \text{ if } 1 \le i \le p \ ,$$

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

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:

$$\widehat{\Sigma} := J\left(a(e_1),\dots,a(e_p),a(e_1+e_2),\dots,a(e_1+e_p),\dots,a(e_{p-1}+e_p)\right) \ . \tag{11}$$

Remark 19 *If a diagonalization basis $(e_1',\dots,e_p')$ (whose basis matrix is P) of Σ is known, or if Σ is diagonal, then a simplified version of the algorithm defined by Equation* (11) *is*

$$\widehat{\Sigma}_{\text{simplified}} = P^\top \operatorname{Diag}(a(e_1'),\dots,a(e_p'))P \ . \tag{12}$$

This algorithm has a smaller computational cost and leads to better theoretical bounds (see Remark 24 and Section 5.2).

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:

$$\left.\begin{aligned} &\forall j \in \{1,\dots,p\}\,,\ \exists \lambda_{0,j} \in (0,+\infty)\,, \\ &\qquad \operatorname{df}(\lambda_{0,j}) \le \sqrt{n} \quad \text{and} \quad \frac{1}{n}\left\|(A_{\lambda_{0,j}} - I_n)F_{e_j}\right\|_2^2 \le \Sigma_{j,j}\sqrt{\frac{\ln n}{n}} \end{aligned}\right\} \tag{13}$$

We can now state the first main result of the paper.

Theorem 20 *Let $\widehat{\Sigma}$ be defined by Equation* (11)*, $\alpha = 2$ and assume* (13) *holds. For every $\delta \ge 2$, a constant $n_0(\delta)$, an absolute constant $L_1 > 0$ and an event Ω exist such that $\mathbb{P}(\Omega) \ge 1 - p(p+1)/2 \times n^{-\delta}$ and if $n \ge n_0(\delta)$, on Ω,*

$$(1-\eta)\Sigma \preceq \widehat{\Sigma} \preceq (1+\eta)\Sigma \tag{14}$$

$$\text{where} \qquad \eta := L_1(2+\delta)p\sqrt{\frac{\ln(n)}{n}}c(\Sigma)^2 \ .$$

Theorem 20 is proved in Section E. 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 Equation (7) is precisely estimated when Σ is replaced by $\widehat{\Sigma}$.

An important feature of Theorem 20 is that it holds under very mild assumptions on the mean f of the data (see Remark 22). 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 21 (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 22 (On assumption (13)) *Assumption (13) 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 15), at the price of multiplying η by d_n in the upper bound of Equation (14). More generally the bounds on the degree of freedom and the bias in (13) only influence the upper bound of Equation (14). The rates are chosen here to match the lower bound, see Proposition 10 of Arlot and Bach (2011) for more details.*

Assumption (13) is rather classical in model selection, see Arlot and Bach (2011) for instance. In particular, (a weakened version of) (13) 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 23 (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 20, since it comes from an union bound over the one-dimensional estimations.*

Remark 24 *When $\widehat{\Sigma} = \widehat{\Sigma}_{\text{simplified}}$ as defined by Equation (12), that is, when a diagonalization basis of Σ is known, Theorem 20 still holds on a set of larger probability $1 - \kappa p n^{-\delta}$ with a reduced error $\eta = L_1(\alpha+\delta)\sqrt{\ln(n)/n}$. Then, a consistent estimation of Σ is possible whenever $p = O(n^\delta)$ for some $\delta \geq 0$.*

5. Oracle Inequality

This section aims at proving "oracle inequalities", as usually done in a model selection setting: given a set of models or of estimators, the goal is to upper bound the risk of the selected estimator by the oracle risk (defined by Equation (6)), up to an additive term and a multiplicative factor. We show two oracle inequalities (Theorems 26 and 29) that correspond to two possible definitions of $\widehat{\Sigma}$.

Note that "oracle inequality" sometimes has a different meaning in the literature (see for instance Lounici et al., 2011) when the risk of the proposed estimator is controlled by the risk of an estimator using information coming from the true parameter (that is, available only if provided by an oracle).

5.1 A General Result for Discrete Matrix Sets $\mathcal{M}$

We first show that the estimator introduced in Equation (11) is precise enough to derive an oracle inequality when plugged in the penalty defined in Equation (7) in the case where $\mathcal{M}$ is finite.

Definition 25 *Let $\widehat{\Sigma}$ be the estimator of Σ defined by Equation* (11)*. We define*

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

We assume now the following holds true:

$$\exists (C, \alpha_{\mathcal{M}}) \in (0, +\infty)^2, \quad \operatorname{card}(\mathcal{M}) < C n^{\alpha_{\mathcal{M}}} \enspace . \tag{15}$$

Theorem 26 *Let $\alpha = \max(\alpha_{\mathcal{M}}, 2)$, $\delta \geq 2$ and assume* (13) *and* (15) *hold true. Absolute constants $L_2, \kappa' > 0$, a constant $n_1(\delta)$ and an event $\widetilde{\Omega}$ exist such that $\mathbb{P}(\widetilde{\Omega}) \geq 1 - \kappa' p(p + C) n^{-\delta}$ and the following holds as soon as $n \geq n_1(\delta)$. First, on $\widetilde{\Omega}$,*

$$\frac{1}{np} \left\| \widehat{f}_{\widehat{M}} - f \right\|_2^2 \leq \left(1 + \frac{1}{\ln(n)} \right)^2 \inf_{M \in \mathcal{M}} \left\{ \frac{1}{np} \left\| \widehat{f}_M - f \right\|_2^2 \right\} + L_2 c(\Sigma)^4 \operatorname{tr}(\Sigma) (\alpha + \delta)^2 \frac{p^3 \ln(n)^3}{np} \enspace . \tag{16}$$

Second, an absolute constant L_3 exists such that

$$\begin{aligned} \mathbb{E}\left[\frac{1}{np} \left\| \widehat{f}_{\widehat{M}} - f \right\|_2^2 \right] \leq & \left(1 + \frac{1}{\ln(n)} \right)^2 \mathbb{E}\left[\inf_{M \in \mathcal{M}} \left\{ \frac{1}{np} \left\| \widehat{f}_M - f \right\|_2^2 \right\} \right] \\ & + L_2 c(\Sigma)^4 \operatorname{tr}(\Sigma) (\alpha + \delta)^2 \frac{p^3 \ln(n)^3}{np} + L_3 \frac{\sqrt{p(p+C)}}{n^{\delta/2}} \left(\left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \Sigma \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert + \frac{\| f \|_2^2}{np} \right) \enspace . \end{aligned} \tag{17}$$

Theorem 26 is proved in Section F.

Remark 27 *If $\widehat{\Sigma} = \widehat{\Sigma}_{\text{simplified}}$ is defined by Equation* (12) *the result still holds on a set of larger probability $1 - \kappa' p(1 + C) n^{-\delta}$ with a reduced error, similar to the one in Theorem 29.*

5.2 A Result for a Continuous Set of Jointly Diagonalizable Matrices

We now show a similar result when matrices in $\mathcal{M}$ can be jointly diagonalized. It turns out a faster algorithm can be used instead of Equation (11) with a reduced error and a larger probability event in the oracle inequality. Note that we no longer assume $\mathcal{M}$ is finite, so it can be parametrized by continuous parameters.

Suppose now the following holds, which means the matrices of $\mathcal{M}$ are jointly diagonalizable:

$$\exists P \in O_p(\mathbb{R}) , \quad \mathcal{M} \subseteq \left\{ P^\top \operatorname{Diag}(d_1, \ldots, d_p) P \, , \, (d_i)_{i=1}^p \in (0, +\infty)^p \right\} \ . \tag{18}$$

Let P be the matrix defined in Assumption (18), $\widetilde{\Sigma} = P \Sigma P^\top$ and recall that $A_\lambda = K(K + n\lambda I_n)^{-1}$. Computations detailed in Appendix D show that the ideal penalty introduced in Equation (7) can be written as

$$\begin{aligned} &\forall M = P^\top \operatorname{Diag}(d_1, \ldots, d_p) P \in \mathcal{M}, \\ &\qquad \operatorname{pen}_{\mathrm{id}}(M) = \frac{2 \operatorname{tr}\left(A_M \cdot (\Sigma \otimes I_n)\right)}{np} = \frac{2}{np} \left(\sum_{j=1}^p \operatorname{tr}(A_{pd_j}) \widetilde{\Sigma}_{j,j} \right) \ . \end{aligned} \tag{19}$$

Equation (19) shows that under Assumption (18), we do not need to estimate the entire matrix Σ in order to have a good penalization procedure, but only to estimate the variance of the noise in p directions.

Definition 28 *Let $(e_1, \ldots, e_p)$ be the canonical basis of $\mathbb{R}^p$, $(u_1, \ldots, u_p)$ be the orthogonal basis defined by $\forall j \in \{1, \ldots, p\}, \ u_j = P^\top e_j$. We then define*

$$\widehat{\Sigma}_{HM} = P \operatorname{Diag}(a(u_1), \ldots, a(u_p)) P^\top \ ,$$

*where for every $j \in \{1, \ldots, p\}$, $a(u_j)$ denotes the output of Algorithm 14 applied to Problem (**Pu$_\mathbf{j}$**), and*

$$\widehat{M}_{HM} \in \operatorname*{argmin}_{M \in \mathcal{M}} \left\{ \left\| \widehat{f}_M - y \right\|_2^2 + 2 \operatorname{tr}\left(A_M \cdot (\widehat{\Sigma}_{HM} \otimes I_n) \right) \right\} \ . \tag{20}$$

Theorem 29 *Let $\alpha = 2$, $\delta \geq 2$ and assume* (13) *and* (18) *hold true. Absolute constants $L_2 > 0$, and κ'', a constant $n_1(\delta)$ and an event $\widetilde{\Omega}$ exist such that $\mathbb{P}(\widetilde{\Omega}) \geq 1 - \kappa'' p n^{-\delta}$ and the following holds as soon as $n \geq n_1(\delta)$. First, on $\widetilde{\Omega}$,*

$$\frac{1}{np} \left\| \widehat{f}_{\widehat{M}_{HM}} - f \right\|_2^2 \leq \left(1 + \frac{1}{\ln(n)} \right)^2 \inf_{M \in \mathcal{M}} \left\{ \frac{1}{np} \left\| \widehat{f}_M - f \right\|_2^2 \right\} + L_2 \operatorname{tr}(\Sigma) (2 + \delta)^2 \frac{\ln(n)^3}{n} \ . \tag{21}$$

Second, an absolute constant L_4 exists such that

$$\begin{aligned} \mathbb{E}\left[\frac{1}{np} \left\| \widehat{f}_{\widehat{M}_{HM}} - f \right\|_2^2 \right] \leq \left(1 + \frac{1}{\ln(n)} \right)^2 \mathbb{E}\left[\inf_{M \in \mathcal{M}} \left\{ \frac{1}{np} \left\| \widehat{f}_M - f \right\|_2^2 \right\} \right] \\ + L_4 \operatorname{tr}(\Sigma) (2 + \delta)^2 \frac{\ln(n)^3}{n} + \frac{p}{n^{\delta/2}} \frac{\|f\|_2^2}{np} \ . \end{aligned} \tag{22}$$

Theorem 29 is proved in Section F.

5.3 Comments on Theorems 26 and 29

Remark 30 *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 of Theorem 26.*

Remark 31 (Scaling of (n, p)**)** *When assumption* (15) *holds, Equation* (16) *implies the asymptotic optimality of the estimator* $\widehat{f}_{\widehat{M}}$ *when*

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

In particular, only (n, p) *such that* $p^3 \ll n/(\ln(n))^3$ *are admissible. When assumption* (18) *holds, the scalings required to ensure optimality in Equation* (21) *are more favorable:*

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

It is to be noted that p *still influences the left hand side via* $\operatorname{tr}\Sigma$*.*

Remark 32 *Theorems 26 and 29 are non asymptotic oracle inequalities, with a multiplicative 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 Equations* (16)*,* (21)*,* (17) *and* (22) *can be further diminished, but this yields a greater remainder term as a consequence.*

Remark 33 (On assumption (18)**)** *Assumption* (18) *actually means all matrices in* $\mathcal{M}$ *can be diagonalized in a unique orthogonal basis, and thus can be parametrized by their eigenvalues as in Examples 3, 4 and 5.*

In that case the optimization problem is quite easy to solve, as detailed in Remark 36. If not, solving (20) *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* $\mathcal{M}$ *or, in practice, to use gradient descent.*

Compared to the setting of Theorem 26, assumption (18) *allows a simpler estimator for the penalty* (19)*, with an increased probability and a reduced error in the oracle inequality.*

The main theoretical limitation comes from the fact that the probabilistic concentration tools used apply to discrete sets $\mathcal{M}$ *(through union bounds). The structure of kernel ridge regression allows us to have a uniform control over a continuous set for the single-task estimators at the "cost" of* n *pointwise controls, which can then be extended to the multi-task setting via* (18)*. We conjecture Theorem 29 still holds without* (18) *as long as* $\mathcal{M}$ *is not "too large", which could be proved similarly up to some uniform concentration inequalities.*

Note also that if $\mathcal{M}_1, \ldots, \mathcal{M}_K$ *all satisfy* (18) *(with different matrices* P_k*), then Theorem 29 still holds for* $\mathcal{M} = \bigcup_{k=1}^{K} \mathcal{M}_k$ *with the penalty defined by Equation* (20) *with* $P = P_k$ *when* $M \in \mathcal{M}_k$*, and* $\mathbb{P}(\widetilde{\Omega}) \geq 1 - 9Kp^2 n^{-\delta}$*, by applying the union bound in the proof.*

Remark 34 (Relationship with the trace norm) *Our approach relies on the minimization of Equation* (2) *with respect to* f*. Argyriou et al. (2008) has shown that if we also* *minimize Equation* (2) *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.*

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 an 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 $X_1, \ldots, X_n \in \mathbb{R}^d$ are drawn (repeatedly and independently for each sample) independently from the multivariate standard Gaussian distribution. For every $j \in \{1, \ldots, p\}$, $f^j(\cdot) = \sum_{i=1}^m \alpha_i^j k(\cdot, z_i)$ where $m = 4$ and $z_1, \ldots, z_m \in \mathbb{R}^d$ are drawn (once for all experiments except in Experiment D) independently from the multivariate standard Gaussian distribution, independently 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. Matlab code is available online.[1]

6.1 Experiments

Five experimental settings are considered:

A⌋ Various numbers of tasks: $n = 10$ 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: $p \in \{2k \,/\, k = 1, \ldots, 25\}$. The covariance matrix is $\Sigma = 10 \cdot I_p$.

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 $n \in \{50k \,/\, k = 1, \ldots, 20\}$.

C⌋ Various noise levels: $n = 100$, $p = 5$ and $\forall j$, $f^j = f_A$. The varying parameter is $\Sigma = \Sigma_{C,t} := 5t \cdot I_5$ with $t \in \{0.2k \,/\, k = 1, \ldots, 50\}$. We also ran the experiments for $t = 0.01$ and $t = 100$.

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

[1] Matlab code can be found at `http://www.di.ens.fr/~solnon/multitask_minpen_en.html`.

number of Σ_E is $c(\Sigma_E) \approx 24.95$. We pick the function $f_D := \sum_{i=1}^{m} \alpha_i k(\cdot, z_i)$ by drawing $(\alpha_1, \dots, \alpha_m)$ and $(z_1, \dots, z_m)$ from standard multivariate normal distribution (independently in each replication) and finally $f^1 = \dots = f^5 = f_D$, $f^6 = \dots = f^{10} = -f_D$.

E⌋ Comparison to cross-validation parameter selection: $p = 5$, $\Sigma = 10 \cdot I_5$, $\forall j$, $f^j = f_A$. The sample size is taken in $\{10, 50, 100, 250\}$.

6.2 Collections of Matrices

Two different sets of matrices $\mathcal{M}$ are considered in the Experiments A–C, following Examples 3 and 4:

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

$$\text{and} \quad \mathcal{M}_{\text{ind}} := \{ M_{\text{ind}}(\lambda) = \operatorname{Diag}(\lambda_1, \dots, \lambda_p) \,/\, \lambda \in (0, +\infty)^p \} \ .$$

In Experiment D, we also use two different sets of matrices, following Example 5:

$$\mathcal{M}_{\text{clus}} := \bigcup_{I \subset \{1,\dots,p\}, I \notin \{\{1,\dots,p\}, \emptyset\}} \left\{ M_I(\lambda, \mu, \mu) \,/\, (\lambda, \mu) \in (0, +\infty)^2 \right\} \cup \mathcal{M}_{\text{similar}}$$

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

Remark 35 *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 is adapted to the structure of the Experiment D. We call this setting "segmentation into intervals".*

6.3 Estimators

In Experiments A–C, we consider four estimators obtained by combining two collections $\mathcal{M}$ of matrices with two formulas for Σ which are plugged into the penalty (7) (that is, either Σ known or estimated by $\widehat{\Sigma}$):

$$\forall \alpha \in \{\text{similar}, \text{ind}\} \ , \ \forall S \in \left\{ \Sigma, \widehat{\Sigma}_{\text{HM}} \right\} \ , \quad \widehat{f}_{\alpha,S} := \widehat{f}_{\widehat{M}_{\alpha,S}} = A_{\widehat{M}_{\alpha,S}} y$$

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

and $\widehat{\Sigma}_{\text{HM}}$ is defined in Section 5.2. As detailed in Examples 3–4, $\widehat{f}_{\text{ind},\widehat{\Sigma}_{\text{HM}}}$ and $\widehat{f}_{\text{ind},\Sigma}$ are concatenations of single-task estimators, whereas $\widehat{f}_{\text{similar},\widehat{\Sigma}_{\text{HM}}}$ 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 $\sum_{j,k} \|f^j - f^k\|_{\mathcal{F}}^2$. In Experiment D we consider the following three estimators, that depend on the choice of the collection $\mathcal{M}$:

$$\forall \beta \in \{\text{clus}, \text{interval}, \text{ind}\} \ , \quad \widehat{f}_\beta := \widehat{f}_{\widehat{M}_\beta} = A_{\widehat{M}_\beta} y$$

$$\text{where} \quad \widehat{M}_\beta \in \operatorname*{argmin}_{M \in \mathcal{M}_\beta} \left\{ \frac{1}{np} \left\| y - \widehat{f}_M \right\|_2^2 + \frac{2}{np} \operatorname{tr}\left(A_M \cdot (\widehat{\Sigma} \otimes I_n) \right) \right\}$$

and $\widehat{\Sigma}$ is defined by Equation (11).

In Experiment E we consider the estimator $\widehat{f}_{\text{similar},\widehat{\Sigma}_{\text{HM}}}$. As explained in the following remark the parameters of the former estimator are chosen by optimizing (20), in practice by choosing a grid. We also consider the estimator $\widehat{f}_{\text{similar,CV}}$ where the parameters are selected by performing 5-fold cross-validation on the mentionned grid.

Remark 36 (Optimization of (20)) *Thanks to Assumption (18) the optimization problem (20) can be solved easily. It suffices to diagonalize in a common basis the elements of $\mathcal{M}$ and the problem splits into several multi-task problems, each with one real parameter. The optimization was then done by using a grid on the real parameters, chosen such that the degree of freedom takes all integer values from 0 to n.*

Remark 37 (Finding the jump in Algorithm 14) *Algorithm 14 raises the question of how to detect the jump of $\mathrm{df}(\lambda)$, which happens around $C = \sigma^2$. We chose to select an estimator $\widehat{C}$ of σ^2 corresponding to the smallest index such that $\mathrm{df}(\widehat{\lambda}_0(\widehat{C})) < n/2$. Another approach is to choose the index corresponding to the largest instantaneous jump of $\mathrm{df}(\widehat{\lambda}_0(C))$ (which is piece-wise constant and non-increasing). This approach has a major drawback, because it sometimes selects a jump far away from the "real" jump around σ^2, when the real jump consists of several small jumps. Both approaches gave similar results in terms of prediction error, and we chose the first one because of its direct link to the theoretical criterion given in Theorem 15.*

6.4 Results

In each experiment, $N = 1000$ independent samples $y \in \mathbb{R}^{np}$ have been generated. Expectations are estimated thanks to empirical means over the N samples. Error bars correspond to the classical Gaussian 95% confidence interval (that is, empirical standard-deviation over the N samples multiplied by $1.96/\sqrt{N}$). The results of Experiments A–C are reported in Figures 2–8. The results of Experiments C–E are reported in Tables 1–3. The p-values correspond to the classical Gaussian difference test, where the hypotheses tested are of the shape $\mathbb{H}_0 = \{q > 1\}$ against the hypotheses $\mathbb{H}_1 = \{q \leq 1\}$, where the different quantities q are detailed in Tables 2–3.

t	0.01	100
$\mathbb{E}[\lVert\widehat{f}_{\text{similar},\widehat{\Sigma}} - f\rVert^2 / \lVert\widehat{f}_{\text{ind},\widehat{\Sigma}} - f\rVert^2]$	1.80 ± 0.02	0.300 ± 0.003
$\mathbb{E}[\lVert\widehat{f}_{\text{similar},\widehat{\Sigma}} - f\rVert^2]$	$(2.27 \pm 0.38) \times 10^{-2}$	0.357 ± 0.048
$\mathbb{E}[\lVert\widehat{f}_{\text{similar},\Sigma} - f\rVert^2]$	$(1.20 \pm 0.28) \times 10^{-2}$	0.823 ± 0.080
$\mathbb{E}[\lVert\widehat{f}_{\text{ind},\widehat{\Sigma}} - f\rVert^2]$	$(1.26 \pm 0.26) \times 10^{-2}$	1.51 ± 0.07
$\mathbb{E}[\lVert\widehat{f}_{\text{ind},\Sigma} - f\rVert^2]$	$(1.20 \pm 0.24) \times 10^{-2}$	4.47 ± 0.13

Table 1: Results of Experiment C for the extreme values of t.

6.5 Comments

As expected, multi-task learning significantly helps when all f^j are equal, as soon as p is large enough (Figure 1), especially for small n (Figure 6) and large noise-levels (Figure 8 and

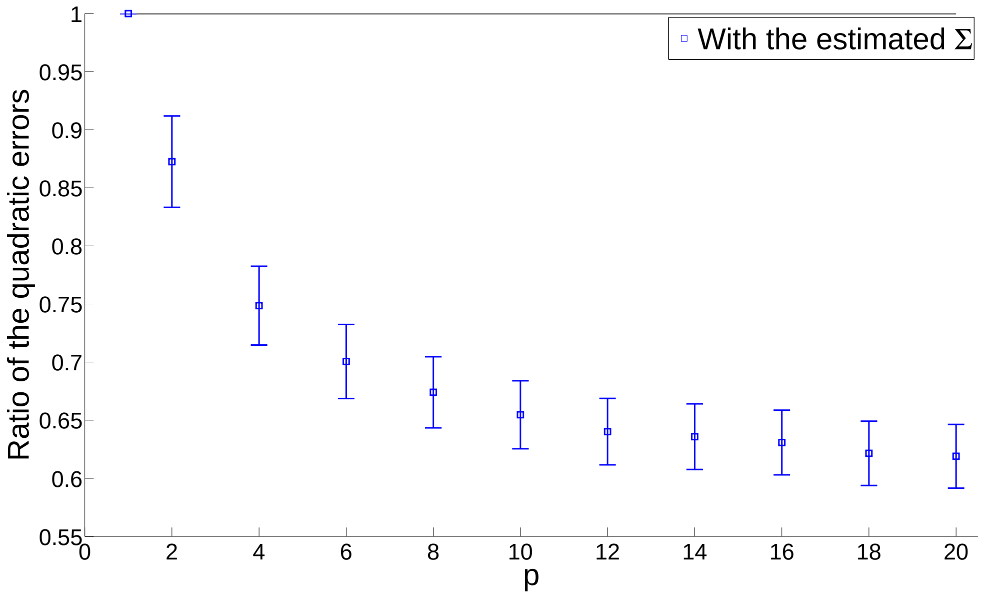

Figure 1: Increasing the number of tasks p (Experiment A), improvement of multi-task compared to single-task: $\mathbb{E}[\|\widehat{f}_{\text{similar},\widehat{\Sigma}} - f\|^2 / \|\widehat{f}_{\text{ind},\widehat{\Sigma}} - f\|^2]$.

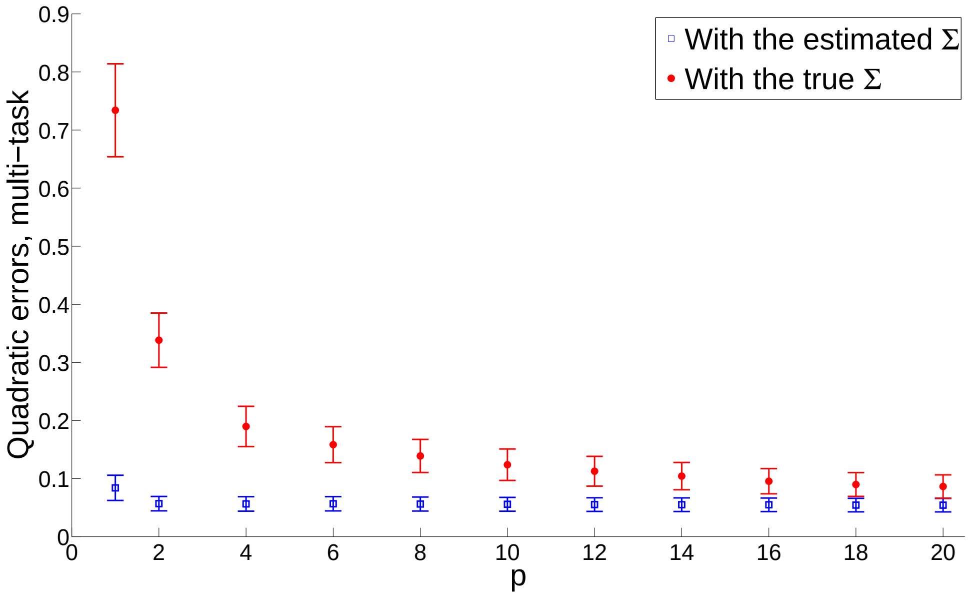

Figure 2: Increasing the number of tasks p (Experiment A), quadratic errors of multi-task estimators $(np)^{-1}\mathbb{E}[\|\widehat{f}_{\text{similar},S} - f\|^2]$. Blue: $S = \widehat{\Sigma}$. Red: $S = \Sigma$.

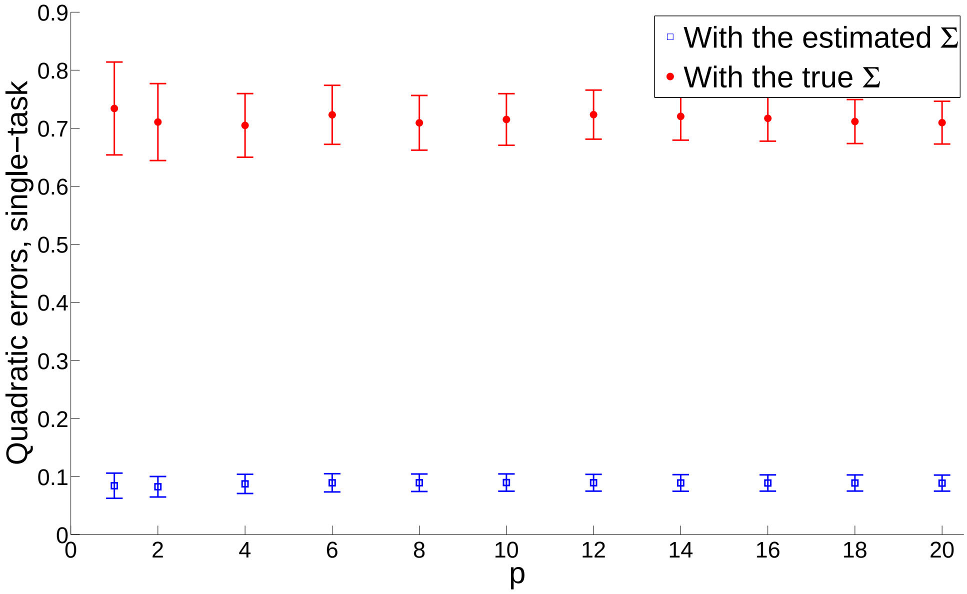

Figure 3: Increasing the number of tasks p (Experiment A), quadratic errors of single-task estimators $(np)^{-1}\mathbb{E}[\|\widehat{f}_{\mathrm{ind},S} - f\|^2]$. Blue: $S = \widehat{\Sigma}$. Red: $S = \Sigma$.

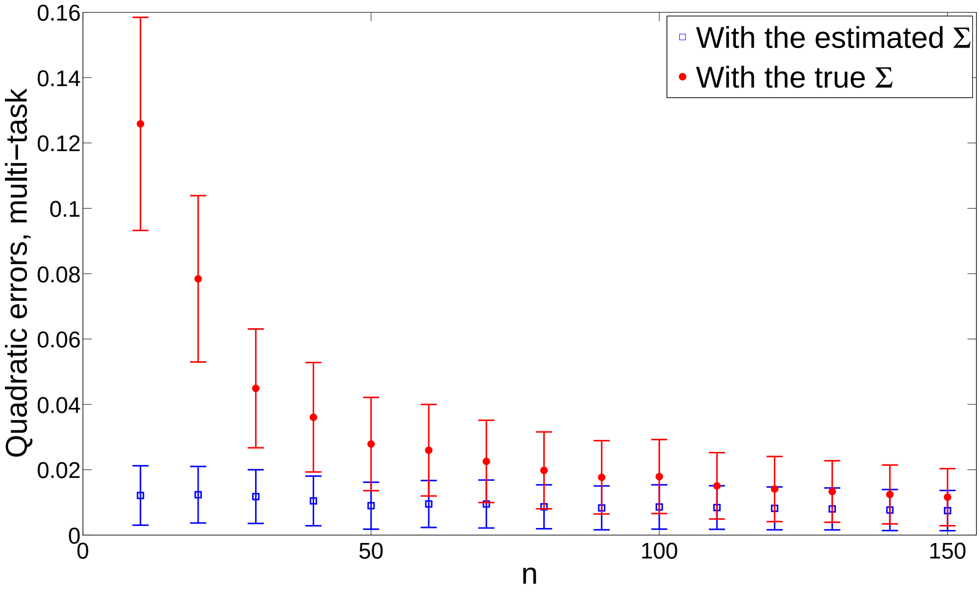

Figure 4: Increasing the sample size n (Experiment B), quadratic errors of multi-task estimators $(np)^{-1}\mathbb{E}[\|\widehat{f}_{\mathrm{similar},S} - f\|^2]$. Blue: $S = \widehat{\Sigma}$. Red: $S = \Sigma$.

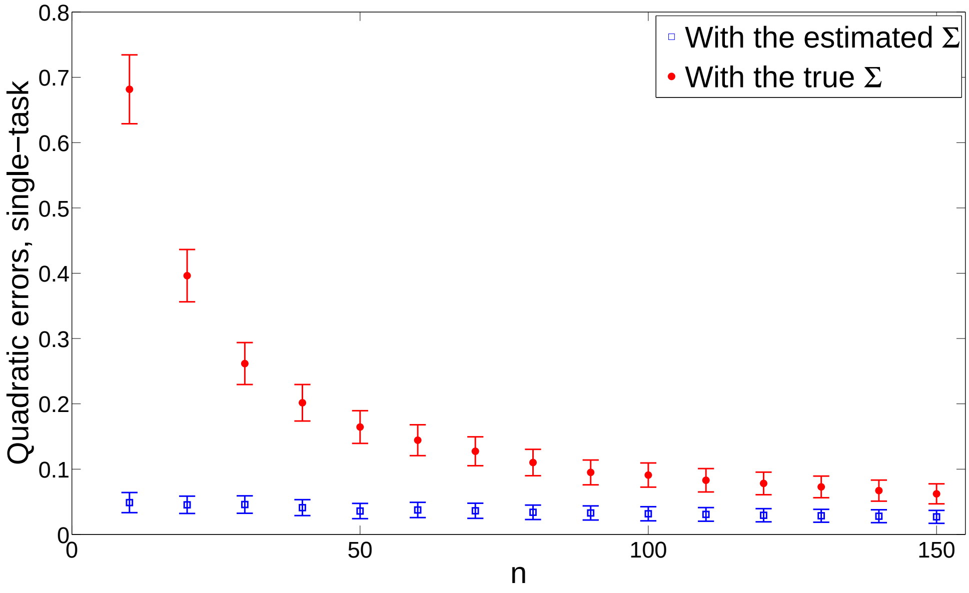

Figure 5: Increasing the sample size n (Experiment B), quadratic errors of single-task estimators $(np)^{-1}\mathbb{E}[\|\widehat{f}_{\mathrm{ind},S} - f\|^2]$. Blue: $S = \widehat{\Sigma}$. Red: $S = \Sigma$.

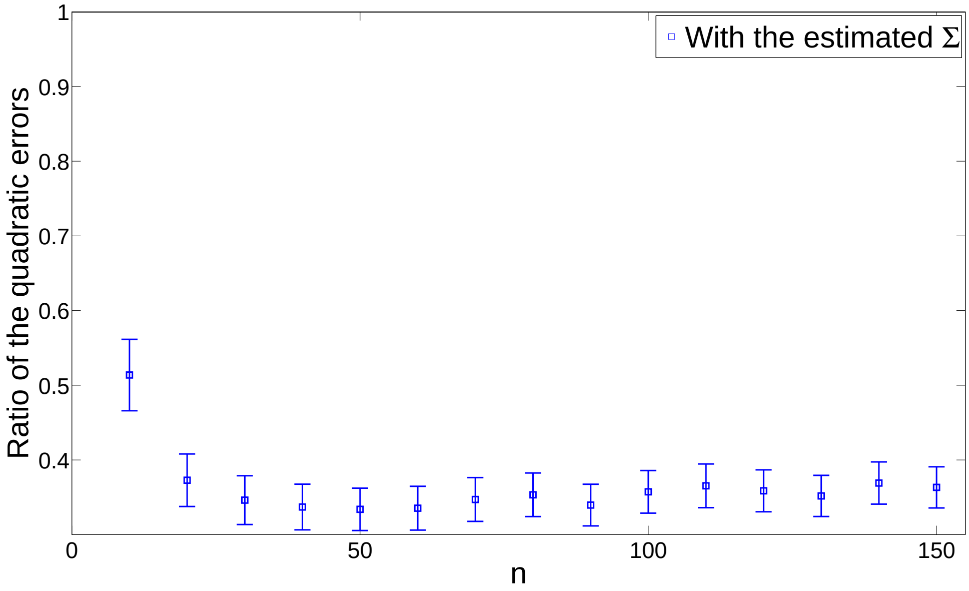

Figure 6: Increasing the sample size n (Experiment B), improvement of multi-task compared to single-task: $\mathbb{E}[\|\widehat{f}_{\mathrm{similar},\widehat{\Sigma}} - f\|^2 / \|\widehat{f}_{\mathrm{ind},\widehat{\Sigma}} - f\|^2]$.

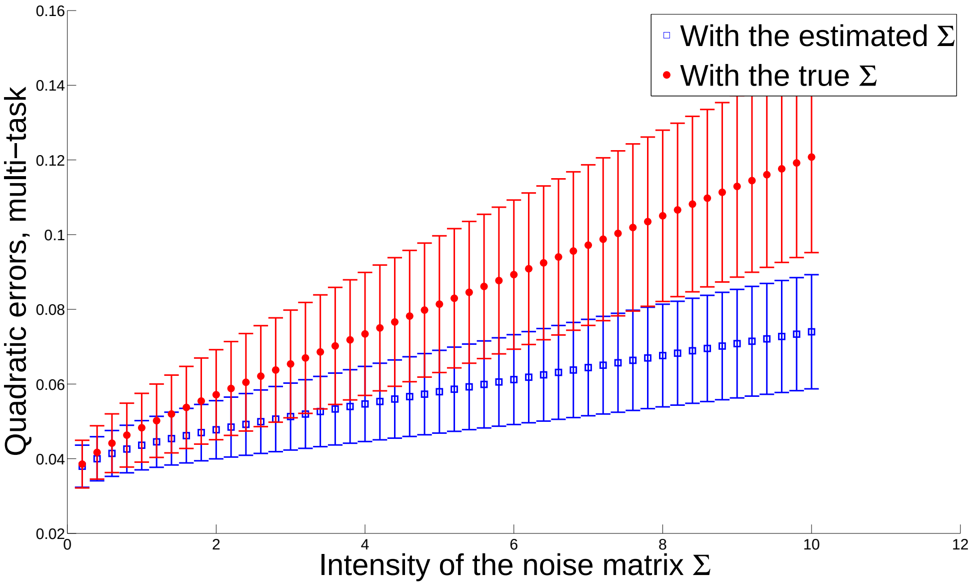

Figure 7: Increasing the signal-to-noise ratio (Experiment C), quadratic errors of multi-task estimators $(np)^{-1}\mathbb{E}[\|\widehat{f}_{\mathrm{similar},S} - f\|^2]$. Blue: $S = \widehat{\Sigma}$. Red: $S = \Sigma$.

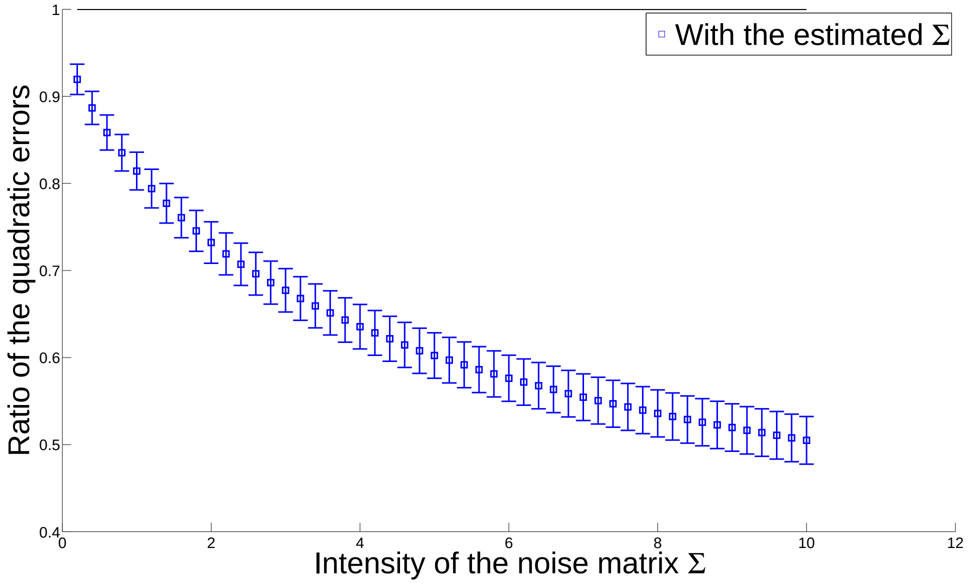

Figure 8: Increasing the signal-to-noise ratio (Experiment C), improvement of multi-task compared to single-task: $\mathbb{E}[\|\widehat{f}_{\mathrm{similar},\widehat{\Sigma}} - f\|^2 / \|\widehat{f}_{\mathrm{ind},\widehat{\Sigma}} - f\|^2]$.

q	$\mathbb{E}[q]$	$\mathrm{Std}[q]$	p-value for $\mathbb{H}_0 = \{q > 1\}$
$\lVert \widehat{f}_{\mathrm{clus}} - f \rVert^2 / \lVert \widehat{f}_{\mathrm{ind}} - f \rVert^2$	0.668	0.294	$< 10^{-15}$
$\lVert \widehat{f}_{\mathrm{interval}} - f \rVert^2 / \lVert \widehat{f}_{\mathrm{ind}} - f \rVert^2$	0.660	0.270	$< 10^{-15}$
$\lVert \widehat{f}_{\mathrm{interval}} - f \rVert^2 / \lVert \widehat{f}_{\mathrm{clus}} - f \rVert^2$	1.00	0.165	0.50

Table 2: Clustering and segmentation (Experiment D).

q	n	$\mathbb{E}[q]$	$\mathrm{Std}[q]$	p-value for $\mathbb{H}_0 = \{q > 1\}$
$\lVert \widehat{f}_{\mathrm{similar},\widehat{\Sigma}_{\mathrm{HM}}} - f \rVert^2 / \lVert \widehat{f}_{\mathrm{similar,CV}} - f \rVert^2$	10	0.35	0.46	$< 10^{-15}$
$\lVert \widehat{f}_{\mathrm{similar},\widehat{\Sigma}_{\mathrm{HM}}} - f \rVert^2 / \lVert \widehat{f}_{\mathrm{similar,CV}} - f \rVert^2$	50	0.56	0.42	$< 10^{-15}$
$\lVert \widehat{f}_{\mathrm{similar},\widehat{\Sigma}_{\mathrm{HM}}} - f \rVert^2 / \lVert \widehat{f}_{\mathrm{similar,CV}} - f \rVert^2$	100	0.71	0.34	$< 10^{-15}$
$\lVert \widehat{f}_{\mathrm{similar},\widehat{\Sigma}_{\mathrm{HM}}} - f \rVert^2 / \lVert \widehat{f}_{\mathrm{similar,CV}} - f \rVert^2$	250	0.87	0.19	$< 10^{-15}$

Table 3: Comparison of our method to 5-fold cross-validation (Experiment E).

Table 1). Increasing the number of tasks rapidly reduces the quadratic error with multi-task estimators (Figure 2) contrary to what happens with single-task estimators (Figure 3).

A noticeable phenomenon also occurs in Figure 2 and even more in Figure 3: the estimator $\widehat{f}_{\mathrm{ind},\Sigma}$ (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 14 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 2) 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".

Figures 4 and 5 show 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 over-penalization phenomenon discussed above disappears.

Table 2 shows that using the multitask procedure improves the estimation accuracy, both in the clustering setting and in the segmentation setting. The last line of Table 2 does not show that the clustering setting improves over the "segmentation into intervals" one, which was awaited if a model close to the oracle is selected in both cases.

Table 3 finally shows that our parameter tuning procedure outperforms 5-fold cross-validation.

7. Conclusion and Future Work

This paper shows that taking into account the unknown similarity between p regression tasks can be done optimally (Theorem 26). The crucial point is to estimate the $p \times p$ covariance matrix Σ of the noise (covariance between tasks), in order to learn the task similarity matrix

M. Our main contributions are twofold. First, an estimator of Σ is defined in Section 4, where non-asymptotic bounds on its error are provided under mild assumptions on the mean of the sample (Theorem 20). Second, we show an oracle inequality (Theorem 26), more particularly with a simplified estimation of Σ and increased performances when the matrices of $\mathcal{M}$ are jointly diagonalizable (which often corresponds to cases where we have a prior knowledge of what the relations between the tasks would be). We do plan to expand our results to larger sets $\mathcal{M}$, which may require new concentration inequalities and new optimization algorithms.

Simulation experiments show that our algorithm works with reasonable sample sizes, and that our multi-task estimator often performs 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.

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 E and F.

Appendix A. Proof of Proposition 8

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 $\forall i \in \{1 \ldots p\}$, $g(\cdot, i) =$

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

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

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 9

Proof If $(x,j) \in \mathcal{X} \times \{1,\dots,p\}$, the application $(f^1,\dots,f^p) \mapsto f^j(x)$ 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 $\forall n \in \mathbb{N}$, $\forall i \in \{1\dots p\}$, $g_n(\cdot,i) = \sum_{k=1}^p T_{i,k} f_n(\cdot,k)$. 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 11

Proof We define

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

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:

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

■

Appendix D. Computation of the Quadratic Risk in Example 12

We consider here that $f^1 = \cdots = f^p$. We use the set $\mathcal{M}_{\text{similar}}$:

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

Using the estimator $\widehat{f}_M = A_M y$ we can then compute the quadratic risk using the bias-variance decomposition given in Equation (36):

$$
\mathbb{E}\left[\left\| \widehat{f}_M - f \right\|_2^2 \right] = \|(A_M - I_{np})f\|_2^2 + \operatorname{tr}(A_M^\top A_M \cdot (\Sigma \otimes I_n)) \ .
$$

Les us denote by $(e_1,\dots,e_p)$ the canonical basis of $\mathbb{R}^p$. The eigenspaces of $p^{-1}\mathbf{1}\mathbf{1}^\top$ are:

- span $\{e_1 + \cdots + e_p\}$ corresponding to eigenvalue p,

- span $\{e_2 - e_1, \ldots, e_p - e_1\}$ corresponding to eigenvalue 0.

Thus, with $\widetilde{\mu} = \lambda + p\mu$ we can diagonalize in an orthonormal basis any matrix $M_{\lambda,\mu} \in \mathcal{M}$ as $M = P^\top D_{\lambda,\widetilde{\mu}} P$, with $D = D_{\lambda,\widetilde{\mu}} = \mathrm{Diag}\{\lambda, \widetilde{\mu}, \ldots, \widetilde{\mu}\}$. Les us also diagonalise in an orthonormal basis K: $K = Q^\top \Delta Q$, $\Delta = \mathrm{Diag}\{\mu_1, \ldots, \mu_n\}$. Thus we can write (see Properties 38 and 39 for basic properties of the Kronecker product):

$$A_M = A_{M_{\lambda,\mu}} = (P^\top \otimes Q^\top) \left[(D^{-1} \otimes \Delta) \left((D^{-1} \otimes \Delta) + npI_{np} \right)^{-1} \right] (P \otimes Q) \ .$$

We can then note that $(D^{-1} \otimes \Delta) \left((D^{-1} \otimes \Delta) + npI_{np} \right)^{-1}$ is a diagonal matrix, whose diagonal entry of index $(j-1)n + i$ ($i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, p\}$) is

$$\begin{cases} \frac{\mu_i}{\mu_i + np\lambda} \text{ if } j = 1 \ , \\ \frac{\mu_i}{\mu_i + np\widetilde{\mu}} \text{ if } j > 1 \ . \end{cases}$$

We can now compute both bias and variance.

Bias: We can first remark that $(P^\top \otimes Q^\top) = (P \otimes Q)^\top$ is an orthogonal matrix and that $P \times \mathbf{1} = (1, 0, \ldots, 0)^\top$. Thus, as in this setting $f^1 = \cdots = f^p$, we have $f = \mathbf{1} \otimes (f^1(X_1), \ldots, f^1(X_n))^\top$ and $(P^\top \otimes Q^\top) f = (1, 0, \ldots, 0)^\top \otimes Q(f^1(X_1), \ldots, f^1(X_n))^\top$. To keep notations simple we note $Q(f^1(X_1), \ldots, f^1(X_n))^\top := (g_1, \ldots, g_n)^\top$. Thus

$$\begin{aligned} \|(A_M - I_{np}) f\|_2^2 &= \|(P \otimes Q)^\top \left[(D^{-1} \otimes K) \left((D^{-1} \otimes K) + npI_{np} \right)^{-1} - I_{np} \right] (P \otimes Q) f\|_2^2 \\ &= \| \left[(D^{-1} \otimes \Delta) \left((D^{-1} \otimes \Delta) + npI_{np} \right)^{-1} - I_{np} \right] \\ &\qquad \times (1, 0, \ldots, 0)^\top \otimes (g_1, \ldots, g_n)^\top \|_2^2 \ . \end{aligned}$$

As only the first n terms of $(P \otimes Q) f$ are non-zero we can finally write

$$\|(A_M - I_{np}) f\|_2^2 = \sum_{i=1}^n \left(\frac{np\lambda}{\mu_i + np\lambda} \right)^2 g_i^2 \ .$$

Variance: First note that

$$(P \otimes Q)(\Sigma \otimes I_n)(P \otimes Q)^\top = (P \Sigma P^\top \otimes I_n) \ .$$

We can also note that $\widetilde{\Sigma} := P\Sigma P^\top$ is a symmetric positive definite matrix, with positive diagonal coefficients. Thus we can finally write

$$\begin{aligned}\operatorname{tr}(A_M^\top A_M \cdot (\Sigma \otimes I_n)) &= \operatorname{tr}\Bigg(P \otimes Q)^\top \left[(D^{-1} \otimes \Delta) \left((D^{-1} \otimes \Delta) + npI_{np} \right)^{-1} \right]^2 \\ &\qquad \times (P \otimes Q)(\Sigma \otimes I_n) \Bigg) \\ &= \operatorname{tr}\Bigg(\left[(D^{-1} \otimes \Delta) \left((D^{-1} \otimes \Delta) + npI_{np} \right)^{-1} \right]^2 \\ &\qquad \times (P \otimes Q)(\Sigma \otimes I_n)(P \otimes Q)^\top \Bigg) \\ &= \sum_{i=1}^n \left[\left(\frac{\mu_i}{\mu_i + np\lambda} \right)^2 \widetilde{\Sigma}_{1,1} + \left(\frac{\mu_i}{\mu_i + np\widetilde{\mu}} \right)^2 \sum_{j=2}^p \widetilde{\Sigma}_{j,j} \right] \enspace .\end{aligned}$$

As noted at the end of Example 12 this leads to an oracle which has all its p functions equal.

D.1 Proof of Equation (19) in Section 5.2

Let $M \in \mathcal{S}_p^{++}(\mathbb{R})$, $P \in \mathcal{O}_p(\mathbb{R})$ such that $M = P^\top \operatorname{Diag}(d_1, \dots, d_p) P$ and $\widetilde{\Sigma} = P\Sigma P^\top$. We recall that $A_\lambda = K(K + n\lambda I_n)^{-1}$. The computations detailed above also show that the ideal penalty introduced in Equation (7) can be written as

$$\operatorname{pen}_{\mathrm{id}}(M) = \frac{2 \operatorname{tr}\left(A_M \cdot (\Sigma \otimes I_n) \right)}{np} = \frac{2}{np} \left(\sum_{j=1}^p \operatorname{tr}(A_{pd_j}) \widetilde{\Sigma}_{j,j} \right) \enspace .$$

Appendix E. Proof of Theorem 20

Theorem 20 is proved in this section, after stating some classical linear algebra results (Section E.1).

E.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 20.

Property 38 *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 39 *Let $A \in \mathcal{M}_n(\mathbb{R})$, $B \in \mathcal{M}_B(\mathbb{R})$, $(A \otimes B)^\top = (A^\top \otimes B^\top)$.*

Definition 40 *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*

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

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

Property 41 *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.

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

We also have:

Proposition 42

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

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

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

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

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

■

This norm can also be written in other forms:

Property 43 *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 $\|\!|S|\!\| = \|S\|_2$*

E.2 The Proof

We now give a proof of Theorem 20, using Lemmas 46, 48 and 49, which are stated and proved in Section E.3. The outline of the proof is the following:

1. Apply Theorem 15 to problem (10) 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

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

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

Proof *1. Apply Theorem 15:* We start by noticing that Assumption (13) actually holds true with all $\lambda_{0,j}$ equal. Indeed, let $(\lambda_{0,j})_{1 \le j \le p}$ be given by Assumption (13) and define $\lambda_0 := \min_{j=1,\dots,p} \lambda_{0,j}$. 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 $j \in \{1, \dots, p\}$, $\lambda_0 \le \lambda_{0,j}$ 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

$$\frac{1}{n} \left\|(A_{\lambda_0} - I_n)F_{e_j}\right\|_2^2 \le \frac{1}{n} \left\|(A_{\lambda_{0,j}} - I_n)F_{e_j}\right\|_2^2 \le \Sigma_{j,j} \sqrt{\frac{\ln(n)}{n}} \ . \tag{23}$$

In particular, Equation (8) holds with $d_n = 1$ for problem (10) whatever $z \in \{e_1, \dots, e_p\}$.

Let us now consider the case $z = e_i + e_j$ with $i \neq j \in \{1, \dots, p\}$. Using Equation (23) and that $F_{e_i + e_j} = F_{e_i} + F_{e_j}$, we have

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

The last term is bounded as follows:

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

because Lemma 46 shows

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

Therefore, Equation (8) holds with $d_n = 1 + c(\Sigma)$ for problem (10) whatever $z \in \mathcal{Z}$.

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

$$\eta_1 := \beta(2 + \delta)(1 + c(\Sigma)) \sqrt{\frac{\ln(n)}{n}} \ .$$

By Theorem 15, for every $z \in \mathcal{Z}$, an event Ω_z of probability greater than $1 - n^{-\delta}$ exists on which, if $n \ge n_0(\delta)$,

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

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

$$\|\zeta - s\|_\infty \le \eta_1 \|s\|_\infty \ , \tag{24}$$

and $\mathbb{P}(\Omega) \ge 1 - p(p+1)/2 \times n^{-\delta}$ 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 $\|s\|_\infty \le 4\|\Sigma\|_\infty$ and $C_1(p) = 1$ by Lemma 48, Equation (24) implies that on Ω,

$$\|\zeta - s\|_\infty \le 4\eta_1 \|\Sigma\|_\infty \le 4\eta_1 |||\Sigma||| \ . \tag{25}$$

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 49, $C_2(p) \le \frac{3}{2}p$. By Equation (25), on Ω,

$$\| \widehat{\Sigma} - \Sigma \| = \| J(\zeta) - J(s) \| \le C_2(p) \left\| \zeta - s \right\|_\infty \le 4\eta_1 C_2(p) \| \Sigma \| \ . \tag{26}$$

Since

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

and $\| \Sigma \| \| \Sigma^{-1} \| = c(\Sigma)$, Equation (26) implies that on Ω,

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

To conclude, Equation (14) holds on Ω with

$$\eta = 6pc(\Sigma)\beta(2+\delta)(1+c(\Sigma))\sqrt{\frac{\ln(n)}{n}} \le L_1(2+\delta)p\sqrt{\frac{\ln(n)}{n}} c(\Sigma)^2 \tag{27}$$

for some numerical constant L_1. ■

Remark 44 *As stated in Arlot and Bach (2011), we need* $\sqrt{n_0(\delta)/\ln(n_0(\delta))} \ge 504$ *and* $\sqrt{n_0(\delta)}/\ln(n_0(\delta)) \ge 24(290+\delta)$*.*

Remark 45 *To ensure that the estimated matrix* $\widehat{\Sigma}$ *is positive-definite we need that* $\eta < 1$*, that is,*

$$\sqrt{\frac{n}{\ln(n)}} > 6\beta(2+\delta)pc(\Sigma)\left(1+c(\Sigma)\right) \ .$$

E.3 Useful Lemmas

Lemma 46 *Let* $p \ge 1$*,* $\Sigma \in \mathcal{S}_p^{++}(\mathbb{R})$ *and* $c(\Sigma)$ *its condition number. Then,*

$$\forall 1 \le i < j \le p \, , \quad \Sigma_{i,j} \ge -\frac{c(\Sigma)-1}{c(\Sigma)+1} \frac{\Sigma_{i,i} + \Sigma_{j,j}}{2} \ , \tag{28}$$

Remark 47 *The proof of Lemma 46 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, (28) only involves 2×2 submatrices $\widetilde{\Sigma}(i,j) \in \mathcal{S}_2^{++}(\mathbb{R})$ for which

$$1 \le c(\widetilde{\Sigma}) \le c\left(\Sigma\right) \quad \text{hence} \quad 0 \le \frac{c(\widetilde{\Sigma})-1}{c(\widetilde{\Sigma})+1} \le \frac{c(\Sigma)-1}{c(\Sigma)+1} \ .$$

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

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

Therefore,

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

So, Equation (28) is equivalent to

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

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

Lemma 48 *For every $p \geq 1$, $C_1(p) := \sup_{\Sigma \in \mathcal{S}_p(\mathbb{R})} \frac{\|\Sigma\|_\infty}{\vert\!\vert\!\vert \Sigma \vert\!\vert\!\vert} = 1$.*

Proof With $\Sigma = I_p$ we have $\|\Sigma\|_\infty = \vert\!\vert\!\vert \Sigma \vert\!\vert\!\vert = 1$, 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 $\mathbb{R}^p$,

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

∎

Lemma 49 *For every $p \geq 1$, let $C_2(p) := \sup_{\zeta \in \mathbb{R}^{p(p+1)/2}} \frac{\vert\!\vert\!\vert J(\zeta) \vert\!\vert\!\vert}{\|\zeta\|_\infty}$. Then,*

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

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 $\vert\!\vert\!\vert J(\zeta) \vert\!\vert\!\vert = p$ 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 $\|z\|_2 = 1$

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

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

Appendix F. Proof of Theorem 26

The proof of Theorem 26 is similar to the proof of Theorem 3 in Arlot and Bach (2011). We give it here for the sake of completeness. We also show how to adapt its proof to demonstrate Theorem 29. The two main mathematical results used here are Theorem 20 and a gaussian concentration inequality from Arlot and Bach (2011).

F.1 Key Quantities and their Concentration Around their Means

Definition 50 *We introduce, for $S \in \mathcal{S}_p^{++}(\mathbb{R})$,*

$$\widehat{M}_o(S) \in \operatorname*{argmin}_{M \in \mathcal{M}} \left\{ \left\| \widehat{F}_M - Y \right\|_2 + 2 \operatorname{tr}\left(A_M \cdot (S \otimes I_n) \right) \right\} \tag{29}$$

Definition 51 *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 $D = \operatorname{Diag}(d_1, \ldots, d_p)$, then*

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

Definition 52 *For every $M \in \mathcal{M}$, we define*

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

Definition 53 *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 $\theta_1, \theta_2, \theta_3, \theta_4 \in (0, 1]$:

$$|\delta_1(M)| \leq \theta_1 \operatorname{tr}\left(A_M^\top A_M \cdot (\Sigma \otimes I_n) \right) + (C_A + C_B \theta_1^{-1}) x \|\Sigma\| \tag{30}$$

$$|\delta_2(M)| \leq \theta_2 \operatorname{tr}\left(A_M^\top A_M \cdot (\Sigma \otimes I_n) \right) + (C_C + C_D \theta_2^{-1}) x \|\Sigma\| \tag{31}$$

$$|\delta_3(M)| \leq \theta_3 \left\| (I_{np} - A_M) f \right\|_2^2 + C_E \theta_3^{-1} x \|\Sigma\| \tag{32}$$

$$|\delta_4(M)| \leq \theta_4 \left\| (I_{np} - A_M) f \right\|_2^2 + C_F \theta_4^{-1} x \|\Sigma\| \tag{33}$$

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 54 *Let*

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

Suppose that (18) *holds. Then* $\mathbb{P}(\Omega_x(\mathcal{M}, C_A, C_B, C_C, C_D, C_E, C_F)) \geq 1 - pe^{1027+\ln(n)}e^{-x}$. *Suppose that* (15) *holds. Then* $\mathbb{P}(\Omega_x(\mathcal{M}, C_A, C_B, C_C, C_D, C_E, C_F)) \geq 1 - 6p\,\mathrm{card}(\mathcal{M})e^{-x}$.

.

Proof

First common step. Let $M \in \mathcal{M}$, $P_M \in \mathcal{O}_p(\mathbb{R})$ such that $M = P_M^\top D P_M$, with $D = \mathrm{Diag}(d_1, \dots, d_p)$. We can write:

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

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

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

We can see that the quantities δ_i decouple, therefore

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

Supposing (18)**.** Assumption (18) implies that the matrix P used above is the same for all the matrices M of $\mathcal{M}$. 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 uniformly on $\mathcal{M}$

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

Supposing (15)**.** We can use Lemma 8 of Arlot and Bach (2011) where we have p concentration results on the sets $\widetilde{\Omega}_{j,M}$, each of probability at least $1 - 6e^{-x}$ we can state that, on the set $\bigcap_{j=1}^{p} \bigcap_{M \in \mathcal{M}} \widetilde{\Omega}_i$, we have uniformly on $\mathcal{M}$ the same inequalities written above.

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

■

F.2 Intermediate Result

We first prove a general oracle inequality, under the assumption that the penalty we use (with an estimator of Σ) does not underestimate the ideal penalty (involving Σ) too much.

Proposition 55 *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*

$$\begin{aligned}
&\operatorname{tr}\left(A_{\widehat{M}_o(S)} \cdot ((S - \Sigma) \otimes I_n) \right) \\
&\qquad \geq -\theta_S \operatorname{tr}\left(A_{\widehat{M}_o(S)} \cdot (\Sigma \otimes I_n) \right) \inf_{M \in \mathcal{M}} \left\{ \frac{b(M) + v_2(M) + K_S \ln(n) \|\!|\Sigma|\!\|}{v_1(M)} \right\}
\end{aligned} \tag{34}$$

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

$$\begin{aligned}
\frac{1}{np} \left\| \widehat{f}_{\widehat{M}_o(S)} - f \right\|_2^2 &\leq \frac{1 + 2\theta}{1 - 2\theta - 4\theta_S} \inf_{M \in \mathcal{M}} \left\{ \frac{1}{np} \left\| \widehat{F}_M - F \right\|_2^2 + \frac{2 \operatorname{tr}\left(A_M \cdot ((S - \Sigma)_+ \otimes I_n) \right)}{np} \right\} \\
&+ \frac{1}{1 - 2\theta - 4\theta_S} \left[(2C_A + 3C_C + 6C_D + 6C_E + \frac{2}{\theta}(C_B + C_F))\gamma + \frac{\theta_S K_S}{4} \right] \frac{\ln(n) \|\!|\Sigma|\!\|}{np}
\end{aligned} \tag{35}$$

Proof The proof of Proposition 55 is very similar to the one of Proposition 5 in Arlot and Bach (2011). First, we have

$$\left\|\widehat{f}_M - f\right\|_2^2 = b(M) + v_2(M) + \delta_2(M) + \delta_3(M) \ , \tag{36}$$

$$\left\|\widehat{f}_M - y\right\|_2^2 = \|\widehat{f}_M - f\|_2^2 - 2v_1(M) - 2\delta_1(M) + \delta_4(M) + \|\varepsilon\|_2^2 \ . \tag{37}$$

Combining Equation (29) and (37), we get:

$$\begin{aligned}&\left\|\widehat{f}_{\widehat{M_o}(S)} - f\right\|_2^2 + 2\operatorname{tr}\left(A_{\widehat{M_o}(S)} \cdot ((S-\Sigma)_+ \otimes I_n)\right) + \widehat{\Delta}(\widehat{M_o}(S)) \\ &\leq \inf_{M\in\mathcal{M}} \left\{ \left\|\widehat{f}_M - f\right\|_2^2 + 2\operatorname{tr}\left(A_M \cdot ((S-\Sigma)\otimes I_n)\right) + \widehat{\Delta}(M) \right\} \ .\end{aligned} \tag{38}$$

On the event $\Omega_{\gamma \ln(n)}$, for every $\theta \in (0,1]$ and $M \in \mathcal{M}$, using Equation (30) and (33) with $\theta = \theta_1 = \theta_4$,

$$|\widehat{\Delta}(M)| \leq \theta(b(M) + v_2(M)) + (C_A + \frac{1}{\theta}(C_B + C_F))\gamma \ln(n) |||\Sigma||| \ . \tag{39}$$

Using Equation (31) and (32) with $\theta_2 = \theta_3 = 1/2$ we get that for every $M \in \mathcal{M}$ Equation

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

which is equivalent to

$$b(M) + v_2(M) \leq 2\left\|\widehat{F}_M - F\right\|_2^2 + 2(C_C + 2C_D + 2C_E)\gamma \ln(n) |||\Sigma||| \ . \tag{40}$$

Combining Equation (39) and (40), 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 Equation (38), and with $C_1 = C_A$, $C_2 = 2C_C + 4C_D + 4C_E$ and $C_3 = C_B + C_F$ we get

$$\begin{aligned}(1-2\theta)\left\|\widehat{f}_{\widehat{M_o}(S)} - f\right\|_2^2 + 2\operatorname{tr}\left(A_{\widehat{M_o}(S)} \cdot ((S-\Sigma)_+ \otimes I_n)\right) \leq \\ \inf_{M\in\mathcal{M}} \left\{ \left\|\widehat{f}_M - f\right\|_2^2 + 2\operatorname{tr}\left(A_M \cdot ((S-\Sigma)\otimes I_n)\right) \right\} + \left(C_1 + C_2\theta + \frac{C_3}{\theta}\right)\gamma \ln(n) |||\Sigma||| \ .\end{aligned} \tag{41}$$

Using Equation (34) we can state that

$$\operatorname{tr}\left(A_{\widehat{M_o}(S)} \cdot ((S-\Sigma)\otimes I_n)\right) \geq \frac{b(\widehat{M_o}(S)) + v_2(\widehat{M_o}(S)) + K_S \ln(n) |||\Sigma|||}{v_1(\widehat{M_o}(S))} \operatorname{tr}\left(A_{\widehat{M_o}(S)} \cdot (\Sigma \otimes I_n)\right)$$

so 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 Equation (35) using Equation (40) and (41). ■

F.3 The Proof Itself

We now show Theorem 26 as a consequence of Proposition 55. It actually suffices to show that $\widehat{\Sigma}$ does not underestimate Σ too much, and that the second term in the infimum of Equation (35) is negligible in front of the quadratic error $(np)^{-1}\|\widehat{f}_M - f\|^2$.

Proof On the event Ω introduced in Theorem 20, Equation (14) holds. Let

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

By Lemma 56 below, we have:

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

We supposed Assumption (15) holds. Using elementary algebra it is easy to show that, for every symmetric positive definite matrices A, M and N of size p, $M \succeq N$ implies that $\operatorname{tr}(AM) \geq \operatorname{tr}(AN)$. In order to have $\widehat{M_o}(\widehat{\Sigma})$ satisfying Equation (34), Theorem 20 shows that it suffices to have, for every $\theta_S > 0$,

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

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 $\theta_S = \theta = (9\ln(n))^{-1}$. Let Ω be the set given by Theorem 20. Using Equation (35) and requiring that $\ln(n) \geq 6$ we get, on the set $\widetilde{\Omega} = \Omega \cap \Omega_{(\alpha+\delta)\ln(n)}(\mathcal{M}, C_A, C_B, C_C, C_D, C_E, C_F)$ of probability $1 - (p(p+1)/2 + 6pC)n^{-\delta}$, using that $\alpha \geq 2$:

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

Using Equation (27) and defining

$$\eta_2 := 12\beta(\alpha+\delta)\gamma\sqrt{\frac{\ln(n)}{n}} \quad ,$$

we get

$$\frac{1}{np}\left\|\widehat{f}_{\widehat{M}}-f\right\|_2 \leq \left(1+\frac{1}{\ln(n)}\right)\inf_{M\in\mathcal{M}}\left\{\frac{1}{np}\left\|\widehat{f}_M-f\right\|_2^2+\eta_2\frac{\operatorname{tr}(A_M\cdot(\Sigma\otimes I_n))}{np}\right\}$$
$$+\left(1-\frac{2}{3\ln(n)}\right)^{-1}\left[2C_A+3C_C+6C_D+6C_E+\ln(n)\left(18C_B+18C_F+\frac{729\beta^2\gamma^2\operatorname{tr}(\Sigma)^2}{4\|\Sigma\|^2}\right)\right]$$
$$\times(\alpha+\delta)^2\frac{\ln(n)^2\|\Sigma\|}{np}\ . \tag{42}$$

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 56 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 $72\beta^2\ln(n)\gamma^2\operatorname{tr}(\Sigma)/(C_n\|\Sigma\|)$ leads to

$$\eta_2 v_1(M)\leq 2C_n v_2(M)+\frac{72\beta^2\ln(n)\gamma^2\operatorname{tr}(\Sigma)}{C_n}\ . \tag{43}$$

Then, since $v_2(M)\leq v_2(M)+b(M)$ and using also Equation (36), 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 Equation (31) and (32),

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

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 Equation (43), we get

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

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

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

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 54 and we get, for some constant L_2,

$$\left(1-\frac{2}{3\ln(n)}\right)^{-1}\left[1851.5+\ln(n)\left(5530.5+\frac{729\beta^2\gamma^2}{4|\!|\!|\Sigma|\!|\!|^2}\right)+616.5\left(1+\frac{1}{\ln(n)}\right)\frac{1}{\ln(n)}\right] \\ +\frac{72\beta^2\ln(n)\gamma^2\operatorname{tr}(\Sigma)}{C_n} \le L_2\ln(n)\gamma^2\frac{\operatorname{tr}(\Sigma)^2}{|\!|\!|\Sigma|\!|\!|^2}$$

From this we can deduce Equation (16) by noting that $\gamma \le 2pc(\Sigma)^2$.

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

$$\begin{aligned}\mathbb{E}\left[\frac{1}{np}\left\|\widehat{f}_{\widehat{M}}-f\right\|_2^2\right] &= \mathbb{E}\left[\frac{\mathbf{1}_{\widetilde{\Omega}}}{np}\left\|\widehat{f}_{\widehat{M}}-f\right\|_2^2\right]+\mathbb{E}\left[\frac{\mathbf{1}_{\widetilde{\Omega}^c}}{np}\left\|\widehat{f}_{\widehat{M}}-f\right\|_2^2\right] \\ &\le \mathbb{E}\left[R_{n,\delta}\right]+\frac{1}{np}\sqrt{\frac{4p(p+1)+6pC}{n^\delta}}\sqrt{\mathbb{E}\left[\left\|\widehat{f}_{\widehat{M}}-f\right\|_2^4\right]} \ . \qquad (44)\end{aligned}$$

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

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

So

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

together with Equation (42) and Equation (44), induces Equation (17), using that for some constant $L_3>0$,

$$12\sqrt{\frac{p(p+1)/2+6pC}{n^\delta}}\left(|\!|\!|\Sigma|\!|\!|+\frac{4}{np}\left\|f\right\|_2^2\right) \le L_3\frac{\sqrt{p(p+C)}}{n^{\delta/2}}\left(|\!|\!|\Sigma|\!|\!|+\frac{1}{np}\left\|f\right\|_2^2\right) \ .$$

■

Lemma 56 *Let $n,p \ge 1$ be two integers, $x \ge 0$ and $\Sigma \in \mathcal{S}_p^{++}(\mathbb{R})$. Then,*

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

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 $A\in\mathcal{M}_{np}(\mathbb{R})$,

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

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

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

Therefore

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

■

F.4 Proof of Theorem 29

We now prove Theorem 29, first by proving that $\widehat{\Sigma}_{\mathrm{HM}}$ leads to a sharp enough approximation of the penalty.

Lemma 57 *Let $\widehat{\Sigma}_{HM}$ be defined as in Definition 28, $\alpha=2$, $\kappa>0$ be the numerical constant defined in Theorem 15 and assume* (13) *and* (18) *hold. For every $\delta\geq 2$, a constant $n_0(\delta)$, an absolute constant $L_1>0$ and an event Ω exist such that $\mathbb{P}(\Omega_{HM})\geq 1-pn^{-\delta}$ and for every $n\geq n_0(\delta)$, on Ω_{HM}, for every M in $\mathcal{M}$*

$$(1-\eta)\operatorname{tr}(A_M\cdot(\Sigma\otimes I_n))\leq\operatorname{tr}(A_M\cdot(\widehat{\Sigma}_{HM}\otimes I_n))\leq(1+\eta)\operatorname{tr}(A_M\cdot(\Sigma\otimes I_n))\enspace , \quad (45)$$

where $\qquad \eta:=L_1(\alpha+\delta)\sqrt{\frac{\ln(n)}{n}}\enspace .$

Proof Let P be defined by (18). Let $M\in\mathcal{M}$, and $(d_1,\dots,d_p)\in(0,+\infty)^p$ such that $M=P^\top\operatorname{Diag}(d_1,\dots,d_p)P$. Thus, as shown in Section D, we have with $\widetilde{\Sigma}=P\Sigma P^\top$:

$$\operatorname{tr}(A_M\cdot(\Sigma\otimes I_n))=\sum_{j=1}^p\operatorname{tr}(A_{pd_j})\widetilde{\Sigma}_{j,j}\enspace .$$

let $\widetilde{\sigma}_j$ be defined as in Definition 28 (and thus $\widehat{\Sigma}_{\mathrm{HM}}=P\operatorname{Diag}(\widetilde{\sigma}_1,\dots,\widetilde{\sigma}_p)P^\top$), we then have by Theorem 15 that for every $j\in\{1,\dots,p\}$ an event Ω^j of probability $1-\kappa n^{-\delta}$ exists such that on Ω^j $|\widetilde{\Sigma}_{j,j}-\widetilde{\sigma}_j|\leq\eta\widetilde{\Sigma}_{j,j}$. Since

$$\operatorname{tr}(A_M\cdot(\widehat{\Sigma}_{\mathrm{HM}}\otimes I_n))=\sum_{j=1}^p\operatorname{tr}(A_{pd_j})\widetilde{\sigma}_j\enspace ,$$

taking $\Omega_{\mathrm{HM}}=\cap_{j=1}^p\Omega^j$ suffices to conclude. ■

Proof [**of Theorem 26**] Adapting the proof of Theorem 26 to Assumption (18) first requires to take $\gamma=1$ as Lemma 57 allows us. It then suffices to take the set $\widetilde{\Omega}=\Omega_{\mathrm{HM}}\cap\Omega_{(2+\delta)\ln(n)}(\mathcal{M},C_A,C_B,C_C,C_D,C_E,C_F)$ (thus replacing α by 2) of probability $1-(p(p+1)/2+p)n^{-\delta}\geq 1-p^2n^{-\delta}$—supposing $p\geq 2$—if we require that $2\ln(n)\geq 1027$.

To get to the oracle inequality in expectation we use the same technique than above, but we note that $\sqrt{\mathbb{P}(\widetilde{\Omega}^c)} \leq \widetilde{L_4} \times p/n^{\delta/2}$. We can finally define the constant L_4 by:

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

■

References

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.

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.

Karim Lounici, Massimiliano Pontil, Sarah van de Geer, and Alexandre Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4): 2164–2204, 2011.

Colin L. Mallows. Some comments on C_P. *Technometrics*, pages 661–675, 1973.

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.