Consistency of the group Lasso and multiple kernel learning

Francis R. Bach FRANCIS.BACH@MINES.ORG
Centre de Morphologie Mathématique
Ecole des Mines de Paris
35, rue Saint Honoré
77300 Fontainebleau, France

Abstract

We consider the least-square regression problem with regularization by a block 1-norm, i.e., a sum of Euclidean norms over spaces of dimensions larger than one. This problem, referred to as the group Lasso, extends the usual regularization by the 1-norm where all spaces have dimension one, where it is commonly referred to as the Lasso. In this paper, we study the asymptotic model consistency of the group Lasso. We derive necessary and sufficient conditions for the consistency of group Lasso under practical assumptions, such as model misspecification. When the linear predictors and Euclidean norms are replaced by functions and reproducing kernel Hilbert norms, the problem is usually referred to as multiple kernel learning and is commonly used for learning from heterogeneous data sources and for non linear variable selection. Using tools from functional analysis, and in particular covariance operators, we extend the consistency results to this infinite dimensional case and also propose an adaptive scheme to obtain a consistent model estimate, even when the necessary condition required for the non adaptive scheme is not satisfied.

1. Introduction

Regularization has emerged as a dominant theme in machine learning and statistics. It provides an intuitive and principled tool for learning from high-dimensional data. Regularization by squared Euclidean norms or squared Hilbertian norms has been thoroughly studied in various settings, from approximation theory to statistics, leading to efficient practical algorithms based on linear algebra and very general theoretical consistency results (Tikhonov and Arsenin, 1997, Wahba, 1990, Hastie et al., 2001, Steinwart, 2001, Cucker and Smale, 2002).

In recent years, regularization by non Hilbertian norms has generated considerable interest in linear supervised learning, where the goal is to predict a response as a linear function of covariates; in particular, regularization by the 1-norm (the sum of absolute values), a method commonly referred to as the *Lasso* (Tibshirani, 1994, Osborne et al., 2000), allows to perform variable selection. However, regularization by non Hilbertian norms cannot be solved empirically by simple linear algebra and instead leads to general convex optimization problems and much of the early effort has been dedicated to algorithms to solve the optimization problem efficiently. In particular, the *Lars* algorithm of Efron et al. (2004)

allows to find the entire regularization path (i.e. the set of solutions for all values of the regularization parameters) at the cost of a matrix inversion.

As the consequence of the optimality conditions, regularization by 1-norm leads to *sparse* solutions, i.e., loading vectors with many zeros. Recent works (Zhao and Yu, 2006, Yuan and Lin, 2007, Zou, 2006) have looked precisely at the model consistency of the Lasso, i.e., if we know that the data were generated from a sparse loading vector, does the Lasso actually recover it when the number of data points grows? In the case of a fixed number of covariates, the Lasso does recover the sparsity pattern if and only if a certain simple condition on the generating covariance matrices is verified (Yuan and Lin, 2007). In particular, in low correlation settings, the Lasso is indeed consistent. However, in presence of strong correlations, the Lasso cannot be consistent, shedding light on potential problems of such procedures for variable selection. Adaptive versions where data-dependent weights are added to the 1-norm allow to keep the consistency in all situations (Zou, 2006).

A cousin to the Lasso is the *group Lasso*, where the covariates are assumed to be clustered in groups, and instead of summing the absolute values of each individual loading, the sum of Euclidean norms of the loadings in each group is used. Intuitively, this should drive all the weights in one group to zero *together*, and thus lead to group selection (Yuan and Lin, 2006). In Section 2, we extend the consistency results of the Lasso to the group Lasso, showing that similar correlation conditions are necessary and sufficient conditions for consistency. The passage from groups of size one to groups of larger sizes leads however to a slightly weaker result as we can not get a single necessary and sufficient condition (in Section 2.4, we show that the stronger result similar to the Lasso is not true as soon as one group has dimension larger than one). In our proofs, we relax the assumptions usually made for such consistency results, i.e., that the model is completely well-specified (conditional expectation of the response which is linear in the covariates and constant conditional variance). In the context of *misspecification*, which is a common situation when applying methods such as the ones presented in this paper, we simply prove convergence to the best linear predictor (which is assumed to be sparse), both in terms of loading values and sparsity patterns.

The group Lasso essentially replaces groups of size one by groups of size larger than one. It is natural in this context to allow the size of each group to grow unbounded, i.e., to replace the sum of Euclidean norms by a sum of appropriate Hilbertian norms. When the Hilbert spaces are reproducing kernel Hilbert spaces (RKHS), this procedure turns out to be equivalent to learn the best convex combination of a set of basis kernels, where each kernel corresponds to one Hilbertian norm used for regularization (Bach et al., 2004a). This framework, referred to as *multiple kernel learning* (Bach et al., 2004a), has applications in kernel selection, data fusion from heterogeneous data sources and non linear variable selection (Lanckriet et al., 2004a). In this latter case, multiple kernel learning can exactly be seen as variable selection in a *generalized additive model* (Hastie and Tibshirani, 1990). We extend the consistency results of the group Lasso to this non parametric case, by using covariance operators and appropriate notions of functional analysis. These notions allow to carry out the analysis entirely in *"primal/input"* space, while the algorithm has to work in *"dual/feature"* space to avoid infinite dimensional optimization. Throughout the paper, we will always go back and forth between primal and dual formulations, primal formulation for analysis and dual formulation for algorithms.

The paper is organized as follows: in Section 2, we present the consistency results for the group Lasso, while in Section 3, we extend these to Hilbert spaces. Finally, we present the adaptive scheme in Section 4 and illustrate our set of results with simulations on synthetic examples in Section 5.

2. Consistency of the group-Lasso

We consider the prediction problem of $Y \in \mathbb{R}$ from $X \in \mathbb{R}^p$, where X has a block structure with m blocks: $X = (X_1, \dots, X_m)$ with each $X_j \in \mathbb{R}^{p_j}$, and $\sum_{j=1}^m p_j = p$. The only assumptions that we make on the joint distribution P_{XY} are the following:

(**A1**) X and Y have finite fourth order moments: $\mathbb{E}\|X\|^4 < \infty$ and $\mathbb{E}\|Y\|^4 < \infty$.

(**A2**) The joint matrix of second order moments $\Sigma_{XX} = \mathbb{E}XX^\top \in \mathbb{R}^{p\times p}$ is invertible.

(**A3**) We let $\mathbf{w} \in \mathbb{R}^p$ denote any minimizer of $\mathbb{E}(Y - X^\top w)^2$. We assume that $\mathbb{E}((Y - \mathbf{w}^\top X)^2|X)$ is almost surely greater than $\sigma^2_{\min} > 0$. We let denote $\mathbf{J} = \{j, \mathbf{w}_j = 0\}$ the sparsity pattern of $\mathbf{w}$.[1]

The assumption (**A3**) does not state that $\mathbb{E}(Y|X)$ is a linear function of X and that the conditional variance is constant, as is commonly done in most works dealing with consistency for linear supervised learning. We simply assume that given the best linear predictor of Y given X (defined by $\mathbf{w}$), there is a still a strictly positive amount of variance in Y. If (**A2**) is satisfied, then the loading vector $\mathbf{w}$ is uniquely defined and is equal to $\mathbf{w} = (\mathbb{E}XX^\top)^{-1}\mathbb{E}XY$. Note that throughout this paper, we do not include a constant term, but we could do so by adding a constant random variable as a group of size one. In particular, all moment matrices are never centered and we will refer to second order moment matrices as non centered covariance matrices, or simply covariance matrices.

We often use the notation $\varepsilon = Y - \mathbf{w}^\top X$. In terms of (non centered) covariance matrices, our assumption (**A3**) leads to: $\Sigma_{\varepsilon\varepsilon|X} = \mathbb{E}(\varepsilon\varepsilon|X) \geqslant \sigma^2_{\min}$ and $\Sigma_{\varepsilon X} = \mathbb{E}\varepsilon X = 0$ (but ε might not in general be independent from X).

We always assume that the number m of groups is fixed and finite. Considering cases where m is allowed to grow with the number of observed data points, in the line of Meinshausen and Yu (2006), is outside the scope of this paper.

Notations Throughout this paper, we consider block covariance matrices Σ_{XX} with m^2 blocks $\Sigma_{X_iX_j}$, $i, j = 1, \dot{;} m$. We refer to the submatrix composed of all blocks indexed by sets I, J as $\Sigma_{X_IX_J}$. Similarly, our loadings are vectors defined following block structure, $w = (w_1, \dots, w_m)$ and we denote w_I the elements indexed by I.

2.1 Group-Lasso

We consider *independent and identically distributed* (i.i.d.) data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \dots, n$, sampled from P_{XY} and the data are given in the form of matrices $\bar{Y} \in \mathbb{R}^n$ and $\bar{X} \in \mathbb{R}^{n\times p}$ and we write $\bar{X} = (\bar{X}_1, \dots, \bar{X}_m)$ where each $\bar{X}_j \in \mathbb{R}^{n\times p_j}$. Throughout this paper,

1. Note that throughout this paper, we use boldface fonts for population quantities.

we make the same i.i.d. assumption; dealing with non identically distributed or dependent data and extending our results in those situations are left for future research.

We consider the following optimization problem:

$$\min_{w \in \mathbb{R}^p} \frac{1}{2n} \|\bar{Y} - \bar{X} w\|^2 + \lambda_n \sum_{j=1}^m d_j \|w_j\|,$$

where d_j is a set of strictly positive fixed weights. Note that considering weights in the block 1-norm is important in practice as those have an influence regarding the consistency of the estimator (see Section 4 for further details). Note that we can rewrite $\frac{1}{2n}\|\bar{Y} - \bar{X}w\|^2$ as follows:

$$\frac{1}{2n}\|\bar{Y} - \bar{X}w\|^2 = \frac{1}{2}\hat{\Sigma}_{YY} - \hat{\Sigma}_{YX} w + \frac{1}{2} w^\top \hat{\Sigma}_{XX} w,$$

where $\hat{\Sigma}_{YY} = \frac{1}{n}\bar{Y}^\top \bar{Y}$, $\hat{\Sigma}_{YX} = \frac{1}{n}\bar{Y}^\top \bar{X}$ and $\hat{\Sigma}_{XX} = \frac{1}{n}\bar{X}^\top \bar{X}$ are empirical (non centered) covariance matrices. Our optimization problem is thus equivalent to:

$$\min_{w \in \mathbb{R}^p} \frac{1}{2}\hat{\Sigma}_{YY} - \hat{\Sigma}_{YX} w + \frac{1}{2} w^\top \hat{\Sigma}_{XX} w + \lambda_n \sum_{j=1}^m d_j \|w_j\|. \tag{1}$$

We denote $\hat{w}$ any minimizer of Eq. (1). We refer to $\hat{w}$ as the *group-Lasso* estimate[2]. Note that with probability tending to one, if (**A2**) is satisfied (i.e., if Σ_{XX} is invertible), there is a unique minimum.

Problem (1) is a non-differentiable convex optimization problem, for which classical tools from convex optimization (Boyd and Vandenberghe, 2003) lead to the following optimality conditions (see proof in Appendix A.1):

Proposition 1 *A vector $w \in \mathbb{R}^p$ with sparsity pattern $J = J(w) = \{j,\ w_j \neq 0\}$ is optimal for problem (1) if and only if*

$$\forall j \in J^c, \qquad \left\| \hat{\Sigma}_{X_j X} w - \hat{\Sigma}_{X_j Y} \right\| \leqslant \lambda_n d_j, \tag{2}$$

$$\forall j \in J, \qquad \hat{\Sigma}_{X_j X} w - \hat{\Sigma}_{X_j Y} = -w_j \frac{\lambda_n d_j}{\|w_j\|}. \tag{3}$$

2.2 Algorithms

Efficient *exact* algorithms exist for the regular Lasso, i.e., for the case where all group dimensions p_j are equal to one. They are based on the piecewise linearity of the set of solutions as a function of the regularization parameter λ_n (Efron et al., 2004). For the group Lasso, however, the path is only piecewise differentiable, and following such a path is not as efficient as for the Lasso. Other algorithms have been designed to solve problem (1) for a single value of λ_n, in the original group Lasso setting (Yuan and Lin, 2006) and in the multiple kernel setting (Bach et al., 2004a,b, Sonnenburg et al., 2006, Rakotomamonjy et al., 2007). In this paper, we study path consistency of the group Lasso and of multiple kernel learning, and in simulations we use the publicly available code for the algorithm of Bach et al. (2004b), that computes an approximate but entire path, by following the piecewise smooth path with predictor-corrector methods.

2. We use the convention that all "hat" notations correspond to data-dependent and thus n-dependent quantities, so we do not need the explicit dependence on n.

2.3 Consistency results

We consider the following two conditions:

$$\max_{i \in \mathbf{J}^c} \frac{1}{d_i} \left\| \Sigma_{X_i X_\mathbf{J}} \Sigma_{X_\mathbf{J} X_\mathbf{J}}^{-1} \operatorname{Diag}(d_j / \|\mathbf{w}_j\|) \mathbf{w}_\mathbf{J} \right\| < 1, \tag{4}$$

$$\max_{i \in \mathbf{J}^c} \frac{1}{d_i} \left\| \Sigma_{X_i X_\mathbf{J}} \Sigma_{X_\mathbf{J} X_\mathbf{J}}^{-1} \operatorname{Diag}(d_j / \|\mathbf{w}_j\|) \mathbf{w}_\mathbf{J} \right\| \leqslant 1, \tag{5}$$

where $\operatorname{Diag}(d_j/\|\mathbf{w}_j\|)$ denotes the block-diagonal matrix (with block sizes p_j) with $\frac{d_j}{\|\mathbf{w}_j\|} I_{p_j}$ on the diagonal, and $\mathbf{w}_\mathbf{J}$ denotes the concatenation of the loadings indexed by $\mathbf{J}$. These are conditions on both the input (through the joint covariance matrix Σ_{XX}) and on the weight vector $\mathbf{w}$. Note that, when all blocks have size 1, this corresponds to the conditions derived for the Lasso (Zhao and Yu, 2006, Yuan and Lin, 2007, Zou, 2006). Note also the difference between the *strong condition* (4) and the *weak condition* (5). For the Lasso, with our assumptions, Yuan and Lin (2007) has shown that the strong condition (4) is necessary and sufficient for path consistency of the Lasso; i.e., the path of solutions consistently contains an estimate which is both consistent for the 2-norm (regular consistency) and the 0-norm (consistency of patterns), if and only if condition (4) is satisfied.

In the case of the group Lasso, even with a finite fixed number of groups, our results are not as strong, as we can only get the strict condition as sufficient and the weak condition as necessary. In Section 2.4, we show that this cannot be improved in general. More precisely the following theorem, proved in Appendix B.1, shows that if the condition (4) is satisfied, any regularization parameter that satisfies a certain decay conditions will lead to a consistent estimator; thus the strong condition (4) is sufficient for path-consistency:

Theorem 2 *Assume (**A1**), (**A2**) and (**A3**). If condition (4) is satisfied, then for any sequence λ_n such that $\lambda_n \to 0$ and $\lambda_n n^{1/2} \to +\infty$, then the group-Lasso estimate $\hat{w}$ defined in Eq. (1) converges in probability to $\mathbf{w}$ and the sparsity pattern $J(\hat{w}) = \{j, \hat{w}_j \neq 0\}$ converges in probability to $\mathbf{J}$ (i.e., $P(J(\hat{w}) = \mathbf{J}) \to 1$).*

The following theorem, proved in Appendix B.2, states that if there is a consistent solution on the path, then the weak condition (5) must be satisfied.

Theorem 3 *Assume (**A1**), (**A2**) and (**A3**). If there exists a (possibly data-dependent) sequence λ_n such that $\hat{w}$ converges to $\mathbf{w}$ and $J(\hat{w})$ converges to $\mathbf{J}$ in probability, then condition (5) is satisfied.*

On the one hand, Theorem 2 states that under the "low correlation" condition (4), the group Lasso is indeed consistent. On the other hand, the result (and the similar one for the Lasso) is rather disappointing regarding the applicability of the group Lasso as a practical group selection method, as Theorem 3 states that if the weak correlation condition (5) is not satisfied, we cannot have consistency.

Moreover, this is to be contrasted with a thresholding procedure of the joint least-square estimator, which is also consistent with no conditions (but the invertibility of Σ_{XX}), if the threshold is properly chosen (smaller than the smallest norm $\|\mathbf{w}_j\|$ for $j \in \mathbf{J}$ or with appropriate decay conditions). However, the Lasso and group Lasso do not have to set such

a threshold, and empirical evidence shows that in the finite sample case, they do perform better (Tibshirani, 1994), in particular in the case where the number m of groups is allowed to grow (Meinshausen and Yu, 2006). In this paper we focus on the extension from uni-dimensional groups to multi-dimensional groups for finite number of groups m and leave the possibility of letting m grow with n for future research.

Finally, by looking carefully at condition (4) and (5), we can see that if we were to increase the weight d_j for $j \in \mathbf{J}^c$ and decrease the weights otherwise, we could always be consistent: this however requires the (potentially empirical) knowledge of $\mathbf{J}$ and this is exactly the idea behind the adaptive scheme that we present in Section 4.

2.4 Refinements of sufficient condition

Our current results state that the strict condition (4) is sufficient for path-consistency of the group Lasso, while the weak condition (5) is only necessary. When all groups have dimension one, then the strict condition is also necessary (Yuan and Lin, 2007). The main technical reason for this difference is that in dimension one, the set of vectors of unit norm is discrete, and thus regular squared norm consistency leads to estimates of the signs of the loadings (i.e., their normalized versions $\hat{w}_j/\|\hat{w}_j\|$) which are ultimately constant. When groups have size larger than one, then $\hat{w}_j/\|\hat{w}_j\|$ will not be ultimately constant (just consistent) and this added dependence on data leads to the following refinement of Theorem 2 (see proof in Appendix B.3):

Theorem 4 *Assume (**A1**), (**A2**) and (**A3**). Assume the weak condition (5) is satisfied and that for all $i \in \mathbf{J}^c$ such that $\frac{1}{d_i}\left\|\Sigma_{X_iX_\mathbf{J}}\Sigma_{X_\mathbf{J}X_\mathbf{J}}^{-1}\operatorname{Diag}(d_j/\|\mathbf{w}_j\|)\mathbf{w}_\mathbf{J}\right\| = 1$, we have*

$$\Delta^\top \Sigma_{X_\mathbf{J}X_i}\Sigma_{X_iX_\mathbf{J}}\Sigma_{X_\mathbf{J}X_\mathbf{J}}^{-1}\operatorname{Diag}\left[d_j/\|\mathbf{w}_j\|\left(I_{p_j} - \frac{\mathbf{w}_j\mathbf{w}_j^\top}{\mathbf{w}_j^\top\mathbf{w}_j}\right)\right]\Delta > 0, \tag{6}$$

with $\Delta = -\Sigma_{X_\mathbf{J}X_\mathbf{J}}^{-1}\operatorname{Diag}(d_j/\|\mathbf{w}_j\|)\mathbf{w}_\mathbf{J}$. Then for any sequence λ_n such that $\lambda_n \to 0$ and $\lambda_n n^{1/4} \to +\infty$, then the group-Lasso estimate $\hat{w}$ defined in Eq. (1) converges in probability to $\mathbf{w}$ and the sparsity pattern $J(\hat{w}) = \{j, \hat{w}_j \neq 0\}$ converges in probability to $\mathbf{J}$.

This theorem is of lower practical significance than Theorem 2 and Theorem 3. It merely shows that the link between strict/weak conditions and sufficient/necessary conditions are in a sense tight (as soon as there exists $j \in \mathbf{J}$ such that $p_j > 1$, it is easy to exhibit examples where Eq. (6) is or is not satisfied). The previous theorem does not contradict the fact that condition (4) is necessary for path-consistency in the Lasso case: indeed, if w_j has dimension one, then $I_{p_j} - \frac{\mathbf{w}_j\mathbf{w}_j^\top}{\mathbf{w}_j^\top\mathbf{w}_j}$ is always equal to zero, and thus Eq. (6) is never satisfied. Note that when condition (6) is an equality, we could still refine the condition by using higher orders in the asymptotic expansions presented in Appendix B.3.

2.5 Loading independent sufficient condition

Condition (4) depends on the loading vector $\mathbf{w}$ and on the sparsity pattern $\mathbf{J}$, which are both a priori unknown. In this section, we consider sufficient conditions that do not depend on the loading vector, but only on the sparsity pattern $\mathbf{J}$ and of course on the covariance

matrices. The following condition is sufficient for consistency of the group-lasso, for all possible loading vectors $\mathbf{w}$ with sparsity pattern $\mathbf{J}$:

$$C(\Sigma_{XX}, d, \mathbf{J}) = \max_{i \in \mathbf{J}^c} \quad \max_{\forall j \in \mathbf{J},\ \|u_j\|=1} \left\| \frac{1}{d_i} \Sigma_{X_i X_\mathbf{J}} \Sigma_{X_\mathbf{J} X_\mathbf{J}}^{-1} \operatorname{Diag}(d_j) u_\mathbf{J} \right\| < 1. \tag{7}$$

As opposed to the Lasso case, $C(\Sigma_{XX}, d, \mathbf{J})$ cannot be readily computed in closed form, but we have the following upper bound:

$$C(\Sigma_{XX}, d, \mathbf{J}) \leqslant \max_{i \in \mathbf{J}^c} \frac{1}{d_i} \sum_{j \in \mathbf{J}} d_j \left\| \sum_{k \in \mathbf{J}} \Sigma_{X_i X_k} \left(\Sigma_{X_\mathbf{J} X_\mathbf{J}}^{-1} \right)_{kj} \right\|,$$

where for a matrix M, $\|M\|$ denotes its maximal singular value (also known as its spectral norm). This leads to the following sufficient condition for consistency of the group-Lasso (which extends the condition of Yuan and Lin (2007)):

$$\max_{i \in \mathbf{J}^c} \frac{1}{d_i} \sum_{j \in \mathbf{J}} d_j \left\| \sum_{k \in \mathbf{J}} \Sigma_{X_i X_k} \left(\Sigma_{X_\mathbf{J} X_\mathbf{J}}^{-1} \right)_{kj} \right\| < 1. \tag{8}$$

Note that testing the existence of a set of weights d such that Eq. (8) is true is a linear programming problem. Moreover, given a set of weights d, better sufficient conditions than Eq. (8) may be obtained by solving a convex optimization problem:

Proposition 5 *The quantity* $\max_{\forall j \in \mathbf{J},\ \|u_j\|=1} \left\| \Sigma_{X_i X_\mathbf{J}} \Sigma_{X_\mathbf{J} X_\mathbf{J}}^{-1} \operatorname{Diag}(d_j) u_\mathbf{J} \right\|^2$ *is upperbounded by*

$$\max_{M \succcurlyeq 0,\ \operatorname{tr} M_{ii}=1} \operatorname{tr} M \left(\operatorname{Diag}(d_j) \Sigma_{X_\mathbf{J} X_\mathbf{J}}^{-1} \Sigma_{X_\mathbf{J} X_i} \Sigma_{X_i X_\mathbf{J}} \Sigma_{X_\mathbf{J} X_\mathbf{J}}^{-1} \operatorname{Diag}(d_j) \right), \tag{9}$$

where M is a matrix defined by blocks following the block structure of $\Sigma_{X_\mathbf{J} X_\mathbf{J}}$. Moreover, the bound is also equal to

$$\min_{\lambda \in \mathbb{R}^m,\ \operatorname{Diag}(d_j) \Sigma_{X_\mathbf{J} X_\mathbf{J}}^{-1} \Sigma_{X_\mathbf{J} X_i} \Sigma_{X_i X_\mathbf{J}} \Sigma_{X_\mathbf{J} X_\mathbf{J}}^{-1} \operatorname{Diag}(d_j) \preccurlyeq \operatorname{Diag}(\lambda)} \lambda^\top 1_m.$$

Proof We let denote $M = uu^\top \succcurlyeq 0$. Then if all u_j for $j \in \mathbf{J}$ have norm 1, then we have $\operatorname{tr} M_{jj} = 1$ for all $j \in \mathbf{J}$. This implies the convex relaxation. The second problem is easily obtained as the convex dual of the first problem (Boyd and Vandenberghe, 2003). ■

Note that for the Lasso, the convex bound in Eq. (9) is tight and leads to the bound given above in Eq. (8) and by Yuan and Lin (2007). For the Lasso, Zhao and Yu (2006) consider several particular patterns of dependencies using Eq. (8). Note that this condition (and not the condition in Eq. (7)) is independent from the dimension and thus does not readily lead to rules of thumbs allowing to set the weight d_j as a function of the dimension p_j; several rules of thumbs have been suggested, that loosely depend on the dimension on the blocks, in the context of the linear group Lasso (Yuan and Lin, 2006) or multiple kernel learning (Bach et al., 2004b); we argue in this paper, that weights should also depend on the response as well (see Section 4).

2.6 Alternative formulation to the group-Lasso

Following Bach et al. (2004a), we can instead consider regularization by the square of the block 1-norm:

$$\min_{w\in\mathbb{R}^p} \frac{1}{2n}\|\bar{Y}-\bar{X}w\|^2 + \frac{1}{2}\mu_n\left(\sum_{j=1}^m d_j\|w_j\|\right)^2.$$

This leads to the same path of solutions, but it is better behaved because each variable which is not zero is still regularized by the squared norm. The alternative version has also two advantages: (a) it has very close links to more general frameworks for learning the kernel matrix from data (Lanckriet et al., 2004b), and (b) it is essential in our proof of consistency in the functional case. We also get the equivalent formulation by using empirical covariance matrices:

$$\min_{w\in\mathbb{R}^p} \frac{1}{2}\hat{\Sigma}_{YY} - \hat{\Sigma}_{YX}w + \frac{1}{2}w^\top\hat{\Sigma}_{XX}w + \frac{1}{2}\mu_n\left(\sum_{j=1}^m d_j\|w_j\|\right)^2. \tag{10}$$

The following proposition gives the optimality conditions for the convex optimization problem defined in Eq. (10) (see proof in Appendix A.2):

Proposition 6 *A vector $w\in\mathbb{R}^p$ with sparsity pattern $J=\{j,\ w_j\neq 0\}$ is optimal for problem (10) if and only if*

$$\forall j\in J^c,\qquad \left\|\hat{\Sigma}_{X_jX}w-\hat{\Sigma}_{X_jY}\right\| \leqslant \mu_n d_j\left(\textstyle\sum_i d_i\|w_i\|\right), \tag{11}$$

$$\forall j\in J,\qquad \hat{\Sigma}_{X_jX}w-\hat{\Sigma}_{X_jY} = -\mu_n\left(\textstyle\sum_i d_i\|w_i\|\right)\frac{d_jw_j}{\|w_j\|}. \tag{12}$$

Note the correspondence at the optimum between optimal solutions of the two optimization problems in Eq. (1) and Eq. (10) through $\lambda_n=\mu_n\left(\sum_i d_i\|w_i\|\right)$. As far as consistency results are concerned, Theorem 3 immediately applies to the alternative formulation because the regularization paths are the same. For Theorem 2, it does not readily apply. But since the relationship between λ_n and μ_n at optimum is $\lambda_n=\mu_n\left(\sum_i d_i\|w_i\|\right)$ and that $\sum_i d_i\|\hat{w}_i\|$ converges to a constant whenever $\hat{w}$ is consistent, it does apply as well with minor modifications (in particular, to deal with the case where $\mathbf{J}$ is empty, which requires $\mu_n=\infty$).

3. Covariance operators and multiple kernel learning

We now extend the previous consistency results to the case of non-parametric estimation, where each group is a potentially infinite dimensional space of functions. Namely, the non parametric group Lasso aims at estimating a sparse linear combination of functions of separate random variables, and can then be seen as a variable selection in a generalized additive model (Hastie and Tibshirani, 1990). Moreover, as shown in Section 3.5, the non-parametric group Lasso may also be seen as equivalent to learning a convex combination of kernels, a framework referred to as multiple kernel learning (MKL).

In this nonparametric context, covariance operators constitute appropriate tools for the statistical analysis and are becoming standard in the theoretical analysis of kernel

methods (Fukumizu et al., 2004, Gretton et al., 2005, Fukumizu et al., 2007, Caponnetto and de Vito, 2005). The following section reviews important concepts. For more details, see Baker (1973) and Fukumizu et al. (2004).

3.1 Review of covariance operator theory

In this section, we first consider a single set $\mathcal{X}$ and a positive definite kernel k on $\mathcal{X}$, associated with the reproducing kernel Hilbert space (RKHS) $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathbb{R}$ (see, e.g., Schölkopf and Smola (2001) or Berlinet and Thomas-Agnan (2003) for an introduction to RKHS theory). The Hilbert space and its dot product $\langle \cdot, \cdot \rangle$ are such that for all $x \in \mathcal{X}$, then $k(\cdot, x) \in \mathcal{F}$ and for all $f \in \mathcal{F}$, $\langle k(\cdot, x), f \rangle = f(x)$, which leads to the *reproducing property* $\langle k(\cdot, x), k(\cdot, y) \rangle = k(x, y)$ for any $(x, y) \in \mathcal{X} \times \mathcal{X}$.

Covariance operator and norms Given a random variable X on $\mathcal{X}$ with bounded second order moment, i.e., such that $\mathbb{E}k(X, X) < \infty$, we can define the (non centered) covariance operator as the bounded linear operator Σ_{XX} from $\mathcal{F}$ to $\mathcal{F}$ such that for all $(f, g) \in \mathcal{F} \times \mathcal{F}$,

$$\langle f, \Sigma_{XX} g \rangle = \mathbb{E}(f(X)g(X)).$$

The operator Σ_{XX} is *auto-adjoint*, *non-negative* and *Hilbert-Schmidt*, i.e., for any orthonormal basis $(e_p)_{p \geqslant 1}$ of $\mathcal{F}$, then $\sum_{p=1}^{\infty} \|\Sigma_{XX} e_p\|^2$ is finite; in this case, the value does not depend on the chosen basis and is referred to as the square of the Hilbert-Schmidt norm. The norm that we use by default in this paper is the operator norm $\|\Sigma_{XX}\| = \sup_{f \in \mathcal{F},\ \|f\|=1} \|\Sigma_{XX} f\|$, which is dominated by the Hilbert-Schmidt norm. Note that in the finite dimensional case, the (non centered) covariance operator is exactly the (non centered) covariance matrix, and the Hilbert-Schmidt norm is the Frobenius norm, while the operator norm is the maximum singular value (also referred to as the spectral norm).

Empirical estimators Given data $x_i, i = 1, \ldots, n$ sampled i.i.d. from P_X, then the empirical estimate $\hat{\Sigma}_{XX}$ of Σ_{XX} is defined such that $\langle f, \hat{\Sigma}_{XX} g \rangle$ is the empirical (non centered) covariance between $f(X)$ and $g(X)$, which leads to:

$$\hat{\Sigma}_{XX} = \frac{1}{n} \sum_{i=1}^{n} k(\cdot, x_i) \otimes k(\cdot, x_i),$$

where $u \otimes v$ is the operator defined by $\langle f, (u \otimes v) g \rangle = \langle f, u \rangle \langle g, v \rangle$. If we further assume that the fourth order moment is finite, i.e., $\mathbb{E}k(X, X)^2 < \infty$, then the estimate is uniformly consistent i.e., $\|\hat{\Sigma}_{XX} - \Sigma_{XX}\| = O_p(n^{-1/2})$ (see Fukumizu et al. (2007) and Appendix C.1), which generalizes the usual result of finite dimension.[3]

Cross-covariance and joint covariance operators Covariance operator theory can be extended to cases with more than one random variables (Baker, 1973). In our situation, we have m input spaces $\mathcal{X}_1, \ldots, \mathcal{X}_m$ and m random variables $X = (X_1, \ldots, X_m)$ and m RKHS $\mathcal{F}_1, \ldots, \mathcal{F}_m$.

3. A random variable Z_n is said to be of order $O_p(a_n)$ if for any $\eta > 0$, there exists $M > 0$ such that $\sup_n P(|Z_n| > M a_n) < \eta$. See Van der Vaart (1998) for further definitions and properties of asymptotics in probability.

If we assume that $\mathbb{E}k_j(X_j, X_j) < \infty$, for all $j = 1, \dots, m$, then we can naturally define the cross-covariance operators $\Sigma_{X_i X_j}$ from $\mathcal{F}_j$ to $\mathcal{F}_i$ such that $\forall (f_i, f_j) \in \mathcal{F}_i \times \mathcal{F}_j$,

$$\langle f_i, \Sigma_{X_i X_j} f_j \rangle = \mathbb{E}(f_i(X_i) f_j(X_j)).$$

These are also Hilbert-Schmidt operators, and if we further assume that $\mathbb{E}k_j(X_j, X_j)^2 < \infty$, for all $j = 1, \dots, m$, then the natural empirical estimators converges to the population quantities in Hilbert-Schmidt and operator norms at rate $O_p(n^{-1/2})$. We can now define a joint block covariance operator on $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_m$ following the block structure of covariance matrices in Section 2. As in the finite dimensional case, it leads to a joint covariance operator Σ_{XX} and we can refer to sub-blocks as $\Sigma_{X_I X_J}$ for the blocks indexed by I and J.

Moreover, we can define the bounded (i.e., with finite operator norm) correlation operators through $\Sigma_{X_i X_j} = \Sigma_{X_i X_i}^{1/2} C_{X_i X_j} \Sigma_{X_j X_j}^{1/2}$ (Baker, 1973). Throughout this paper we will make the assumption that those operators $C_{X_i X_j}$ are *compact* for $i \neq j$: compact operators can be characterized as limits of finite rank operators or as operators that can be diagonalized on a countable basis with spectrum composed of a sequence tending to zero (see, e.e., Brezis (1980)). This implies that the joint operator C_{XX}, naturally defined on $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_m$, is of the form "identity plus compact". It thus has a minimum and a maximum eigenvalue which are both between 0 and 1 (Brezis, 1980). If those eigenvalues are strictly greater than zero, then the operator is invertible, as are all the square sub-blocks. Moreover, the joint correlation matrix is lower-bounded by a strictly positive constant times the identity operator.

Translation invariant kernels A particularly interesting ensemble of RKHS in the context of nonparametric estimation is the set of translation invariant kernels defined over $\mathcal{X} = \mathbb{R}^p$, where $p \geqslant 1$, of the form $k(x, x') = q(x' - x)$ where q is a function on $\mathbb{R}^p$ with pointwise nonnegative integrable Fourier transform (which implies that q is continuous). In this case, the associated RKHS is $\mathcal{F} = \{q_{1/2} * g,\ g \in L^2(\mathbb{R}^p)\}$, where $q_{1/2}$ denotes the inverse Fourier transform of the square root of the Fourier transform of q and $*$ denotes the convolution, and $L^2(\mathbb{R}^p)$ denotes the space of square integrable functions. The norm is equal to

$$\|f\|^2 = \int \frac{|F(\omega)|^2}{Q(\omega)} d\omega,$$

where F and Q are the Fourier transforms of f and q (Wahba, 1990, Schölkopf and Smola, 2001). Functions in the RKHS are then functions with appropriately integrable derivatives. In this paper, when using infinite dimensional kernels, we use the Gaussian kernel $k_\tau(x, x') = q_\tau(x - x') = \exp(-\frac{\|x - x'\|^2}{2\tau^2})$.

One-dimensional Hilbert spaces In this paper, we also consider real random variables Y and ε embedded in the natural Euclidean structure of real numbers (i.e., we consider the linear kernel on $\mathbb{R}$). In this setting the covariance operator $\Sigma_{X_j Y}$ from $\mathbb{R}$ to $\mathcal{F}_j$ can be canonically identified as an element of $\mathcal{F}_j$. Throughout this paper, we always use this identification.

3.2 Problem formulation

We assume in this section and in the remaining of the paper that $X_j \in \mathcal{X}_j$ where $\mathcal{X}_j$ is any set on which we have a reproducible kernel Hilbert spaces $\mathcal{F}_j$, associated with the positive kernel $k_j : \mathcal{X}_j \times \mathcal{X}_j \to \mathbb{R}$. We now make the following assumptions, that extends the assumptions (**A1**), (**A2**) and (**A3**). For each of them, we detail the main implications as well as common natural sufficient conditions. The first two conditions (**A4**) and (**A5**) depend solely on the input variables, while the two other ones, (**A6**) and (**A7**) consider the relationship between X and Y.

(**A4**) For each $j = 1 \ldots, m$, $\mathcal{F}_j$ is a separable reproducing kernel Hilbert space associated with kernel k_j, and the random variables $k_j(\cdot, X_j)$ have finite fourth-order moment, i.e., $\mathbb{E}k_j(X_j, X_j)^2 < \infty$.

This is a non restrictive assumption in many situations; for example, when (a) $\mathcal{X}_j = \mathbb{R}^{p_j}$ and the kernel function (such as the Gaussian kernel) is bounded, and when (b) $\mathcal{X}_j$ is a compact subset of $\mathbb{R}^{p_j}$ and the kernel is any continuous function such as linear or polynomial. This implies notably, as shown in Section 3.1, that we can define covariance, cross-covariance and correlation operators that all Hilbert-Schmidt (Baker, 1973, Fukumizu et al., 2007) and can all be estimated at rate $O_p(n^{-1/2})$.

(**A5**) All cross-correlation operators are compact and the joint correlation operator C_{XX} is invertible.

This is also a condition uniquely on the input spaces and not on Y. Following Fukumizu et al. (2007), a simple sufficient condition is that we have measurable spaces and distributions with joint density p_X (and marginal distributions $p_{X_i}(x_i)$ and $p_{X_iX_j}(x_i, x_j)$) and that the *mean square contingency* between all pairs of variables is finite, i.e.

$$\mathbb{E}\left\{\frac{p_{X_iX_j}(x_i, x_j)}{p_{X_i}(x_i)p_{X_j}(x_j)} - 1\right\} < \infty.$$

The contingency is a measure of statistical dependency (Renyi, 1959), and thus this sufficient condition simply states that two variables X_i and X_j cannot be too dependent. In the context of multiple kernel learning for heterogeneous data fusion, this corresponds to having sources which are heterogeneous enough. Essentially, we use this assumption to make sure that the functions $\mathbf{f}_1, \ldots, \mathbf{f}_m$ are unique. This ensures the non existence of any set of functions $f_1, \ldots, f_m$ in $\mathcal{F}_1, \ldots, \mathcal{F}_m$, such that $\mathbb{E}f_j(X_j)^2 > 0$ and a linear combination is zero on the support of the random variables. In the context of generalized additive models, this assumption is referred to as the empty *concurvity space* assumption (Hastie and Tibshirani, 1990).

(**A6**) There exists functions $\mathbf{f} = (\mathbf{f}_1, \ldots, \mathbf{f}_m) \in \mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_m$ and a function $\mathbf{h}$ of $X = (X_1, \ldots, X_m)$ such that $\mathbb{E}(Y|X) = \sum_{j=1}^m \mathbf{f}_j(X_j) + \mathbf{h}(X)$ with $\mathbb{E}h(X)^2 < \infty$ and $\mathbb{E}\mathbf{h}(X)f_j(X_j) = 0$ for all $j = 1, \ldots, m$ and $f_j \in \mathcal{F}_j$. We assume that $\mathbb{E}((Y - \mathbf{f}(X))^2|X)$ is almost surely greater than $\sigma^2_{\min} > 0$ and smaller than $\sigma^2_{\max} < \infty$. We let denote $\mathbf{J} = \{j, \mathbf{f}_j = 0\}$ the sparsity pattern of $\mathbf{f}$.

This assumption on the conditional expectation of Y given X is not the most general and follows common assumptions in approximation theory (see, e.g., Caponnetto and de Vito (2005), Cucker and Smale (2002) and references therein). It allows misspecification, but it essentially requires that the conditional expectation of Y given sums of measurable functions of X_j is attained at functions in the RKHS, and not merely measurable functions. Dealing with more general assumptions requires to consider consistency for norms weaker than the RKHS norms (Caponnetto and de Vito, 2005, Steinwart, 2001), and is left for future research. Note also, that to simplify proofs, we assume a finite upper-bound $\sigma^2_{\max}$ on the residual variance.

(**A7**) For all j, $\exists \mathbf{g}_j \in \mathcal{F}_j$ such that $\mathbf{f}_j = \Sigma^{1/2}_{X_jX_j}\mathbf{g}_j$, i.e., each $\mathbf{f}_j$ is in the range of $\Sigma^{1/2}_{X_jX_j}$.

This technical condition, already used by Caponnetto and de Vito (2005), which concerns all RKHS independently, ensures that we obtain consistency for the norm of the RKHS (and not another weaker norm) for the least-squares estimates. Note also that it implies that $\mathbb{E}f_j(X_j)^2 > 0$. For practical cases, this is not a restrictive assumption. Indeed, for finite dimensional Hilbert spaces, this is always true. For the common situation where $\mathcal{X}_j = \mathbb{R}^{p_j}$, P_{X_j} (the marginal distribution of X_j) has a density $p_{X_j}(x_j)$ with respect to the Lebesgue measure and the kernel is of the form $k_j(x_j, x'_j) = q_j(x_j - x'_j)$, we have the following proposition (proved in Appendix C.5):

Proposition 7 *Assume $\mathcal{X} = \mathbb{R}^p$ and X is a random variable on $\mathcal{X}$ with distribution P_X that has a strictly positive density $p_X(x)$ with respect to the Lebesgue measure. Assume $k(x, x') = q(x - x')$ for a function $q \in L^2(\mathbb{R}^p)$ has an integrable pointwise positive Fourier transform, with associated RKHS $\mathcal{F}$. If f can be written as $f = q * g$ (convolution of q and g) with $\int \frac{g(x)^2}{p_X(x)}dx < \infty$, then $f \in \mathcal{F}$ is in the range of the square root $\Sigma^{1/2}_{XX}$ of the covariance operator.*

The previous proposition gives natural conditions regarding f and p_X. Indeed, the condition $\int \frac{g(x)^2}{p_X(x)}dx < \infty$ corresponds to a natural support condition, i.e., f should be zero where X has no mass, otherwise, we will not be able to estimate f; note the similarity with the usual condition regarding the variance of importance sampling estimation (Brémaud, 1999).

Notations Throughout this section, we refer to functions $f = (f_1, \dots, f_m) \in \mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_m$ and the joint covariance operator Σ_{XX}. In the following, we always use the norms of the RKHS. When considering operators, we use the operator norm. We also refer to a subset of f indexed by J through f_J.

3.3 Nonparametric group Lasso

Given i.i.d data (x_{ij}, y_i), $i = 1, \dots, n$, $j = 1, \dots, m$, where each $x_{ij} \in \mathcal{X}_j$, our goal is to estimate consistently the functions $\mathbf{f}_j$ and which of them are zero. We let denote $\bar{X}$ and $\bar{Y}$ the data matrices. We consider the following problem:

$$\min_{f \in \mathcal{F}} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} f_j(x_{ji})\right)^2 + \frac{\mu_n}{2}\left(\sum_{j=1}^{m} d_j\|f_j\|\right)^2.$$

We obtain a similar formulation to Eq. (10), where empirical covariance matrices are replaced by empirical covariance operators:

$$\min_{f\in\mathcal{F}} \frac{1}{2}\hat{\Sigma}_{YY} - \langle f, \hat{\Sigma}_{XY}\rangle + \frac{1}{2}\langle f, \hat{\Sigma}_{XX} f\rangle + \frac{\mu_n}{2}\left(\sum_{j=1}^{m} d_j \|f_j\|\right)^2. \tag{13}$$

We let denote $\hat{f}$ any minimizer of Eq. (13), and we refer to it as the non parametric group Lasso estimate, or also the multiple kernel learning estimate. Note that by Proposition 11, the previous problem has indeed minimizers.

Note that formally, the finite and infinite dimensional formulation in Eq. (10) and Eq. (13) are the same, and this is the main reason why covariance operators are very practical tools for the analysis. Furthermore, we have the corresponding proposition regarding optimality conditions (see proof in Appendix A.3):

Proposition 8 *A function $f \in \mathcal{F}$ with sparsity pattern $J = J(f) = \{j,\ f_j \neq 0\}$ is optimal for problem (13) if and only if*

$$\forall j \in J^c, \qquad \left\|\hat{\Sigma}_{X_jX} f - \hat{\Sigma}_{X_jY}\right\| \leqslant \mu_n d_j \left(\textstyle\sum_i d_i \|f_i\|\right), \tag{14}$$

$$\forall j \in J, \qquad \hat{\Sigma}_{X_jX} f - \hat{\Sigma}_{X_jY} = -\mu_n \left(\textstyle\sum_i d_i \|f_i\|\right) \frac{d_j f_j}{\|f_j\|}. \tag{15}$$

A consequence (and in fact the first part of the proof) is that an optimal function f must be in the range of $\hat{\Sigma}_{XY}$ and $\hat{\Sigma}_{XX}$, i.e., an optimal f is supported by the data; that is, each f_j is a linear combination of functions $k_j(x, x_{ij})$, $i = 1, \ldots, n$. This is a rather circumvoluted way of presenting the representer theorem (Wahba, 1990), but this is the easiest for the theoretical analysis of consistency. However, to actually compute the estimate $\hat{f}$ from data, we need the usual formulation with dual parameters (see Section 3.5).

Moreover, one important conclusion is that all our optimization problems in spaces of functions can be in fact transcribed into finite-dimensional problems. In particular, all notions from multivariate differentiable calculus may be used without particular care regarding the infinite dimension.

3.4 Consistency results

We consider the following strict and weak conditions, which correspond to condition (4) and (5) in the finite dimensional case:

$$\max_{i\in\mathbf{J}^c} \frac{1}{d_i} \left\|\Sigma_{X_iX_i}^{1/2} C_{X_iX_\mathbf{J}} C_{X_\mathbf{J}X_\mathbf{J}}^{-1} \operatorname{Diag}(d_j/\|\mathbf{f}_j\|)\mathbf{g}_\mathbf{J}\right\| < 1, \tag{16}$$

$$\max_{i\in\mathbf{J}^c} \frac{1}{d_i} \left\|\Sigma_{X_iX_i}^{1/2} C_{X_iX_\mathbf{J}} C_{X_\mathbf{J}X_\mathbf{J}}^{-1} \operatorname{Diag}(d_j/\|\mathbf{f}_j\|)\mathbf{g}_\mathbf{J}\right\| \leqslant 1, \tag{17}$$

where $\operatorname{Diag}(d_j/\|\mathbf{f}_j\|)$ denotes the block-diagonal operator (with block sizes p_j) with operators $\frac{d_j}{\|\mathbf{f}_j\|} I_{\mathcal{F}_j}$ on the diagonal. Note that this is well-defined because C_{XX} is invertible and that it reduces to Eq. (4) and Eq. (5) for finite dimensional spaces. The following theorems give necessary and sufficient conditions for the path consistency of the nonparametric group Lasso (see proofs in Appendix C.2 and Appendix C.3):

Theorem 9 *Assume (**A4**), (**A5**) , (**A6**), (**A7**) and that* $\mathbf{J}$ *is not empty. If condition (16) is satisfied, then for any sequence* μ_n *such that* $\mu_n \to 0$ *and* $\mu_n n^{1/2} \to +\infty$*, then any sequence of nonparametric group Lasso estimates* $\hat{f}$ *converges in probability to* $\mathbf{f}$ *and the sparsity pattern* $J(\hat{f}) = \{j, \hat{f}_j \neq 0\}$ *converges in probability to* $\mathbf{J}$*.*

Theorem 10 *Assume (**A4**), (**A5**) , (**A6**), (**A7**) and that* $\mathbf{J}$ *is not empty. If there exists a (possibly data-dependent) sequence* μ_n *such* $\hat{f}$ *converges to* $\mathbf{f}$ *and* $\hat{J}$ *converges to* $\mathbf{J}$ *in probability, then condition (17) is satisfied.*

Essentially, the results in finite dimension also hold when groups have infinite dimensions. We leave the extensions of the refined results in Section 2.4 to future work. Condition (16) might be hard to check in practice since it involves inversion of correlation operators; see Section 3.6 for an estimate from data.

3.5 Multiple kernel learning formulation

Proposition 8 does not readily lead to an algorithm for computing the estimate $\hat{f}$. In this section, following Bach et al. (2004a), we link the group Lasso to the multiple kernel learning framework (Lanckriet et al., 2004b, Pontil and Micchelli, 2005, Rakotomamonjy et al., 2007) Problem (13) is an optimization problem on a potentially infinite dimensional space of functions. However, the following proposition shows that it reduces to a finite dimensional problem that we now precise (see proof in Appendix A.4):

Proposition 11 *The dual of problem (13) is*

$$\max_{\alpha \in \mathbb{R}^n} \left\{ -\frac{1}{2n} \|\bar{Y} - n\mu_n \alpha\|^2 - \frac{1}{2\mu_n} \max_{i=1,\dots,m} \frac{\alpha^\top K_i \alpha}{d_i^2} \right\}, \tag{18}$$

where $(K_i)_{ab} = k_i(x_a, x_b)$ *are the kernel matrices in* $\mathbb{R}^{n \times n}$*, for* $i = 1, \dots, m$*. Moreover, the dual variable* $\alpha \in \mathbb{R}^n$ *is optimal if and only if there exists* $\eta \in \mathbb{R}_+^m$ *such that* $\sum_{j=1}^m \eta_j d_j^2 = 1$ *and*

$$\left(\sum_{j=1}^m \eta_j K_j + n\mu_n I_n \right) \alpha = \bar{Y}, \tag{19}$$

$$\forall j \in \{1, \dots, m\}, \ \frac{\alpha^\top K_j \alpha}{d_j^2} < \max_{i=1,\dots,m} \frac{\alpha^\top K_i \alpha}{d_i^2} \Rightarrow \eta_j = 0. \tag{20}$$

the optimal function may then be written as $f_j = \eta_j \sum_{i=1}^n \alpha_i k_j(\cdot, x_{ij})$*.*

Since the problem in Eq. (18) is strictly convex, there is a unique dual solution α. Note that Eq. (19) corresponds to the optimality conditions for the least-square problem:

$$\min_{f \in \mathcal{F}} \frac{1}{2} \hat{\Sigma}_{YY} - \langle f, \hat{\Sigma}_{XY} \rangle + \frac{1}{2} \langle f, \hat{\Sigma}_{XX} f \rangle + \frac{1}{2} \mu_n \sum_{j=1}^m \frac{\|w_j\|^2}{\eta_i},$$

whose dual problem is:

$$\max_{\alpha \in \mathbb{R}^n} \left\{ -\frac{1}{2n} \|\bar{Y} - n\mu_n \alpha\|^2 - \frac{1}{2\mu_n} \alpha^\top \left(\sum_{j=1}^m \eta_i K_i \right) \alpha \right\},$$

and unique solution is $\alpha = (\sum_{j=1}^m \eta_j K_j + n\mu_n I_n)^{-1}\bar{Y}$. That is, the solution of the MKL problem leads to dual parameters α and set of weights $\eta \geqslant 0$ such that α is the solution to the least-square problem with kernel $K = \sum_{j=1}^m \eta_j K_j$. Bach et al. (2004a) has shown in a very similar context (hinge loss instead of the square loss) that the optimal η in Proposition 11 can be obtained as the minimizer of

$$J(\eta) = \max_{\alpha \in \mathbb{R}^n} -\frac{1}{2n}\|\bar{Y} - n\mu_n\alpha\|^2 - \frac{1}{2\mu_n}\alpha^\top \left(\sum_{j=1}^m \eta_j K_j\right)\alpha,$$

with respect to $\eta \geqslant 0$ such that $\sum_{j=1}^m \eta_j d_j^2 = 1$. This formulation allows to derive probably approximately correct error bounds (Lanckriet et al., 2004b, Bousquet and Herrmann, 2003). Besides, this formulation allows η to be negative, as long as the matrix $\sum_{j=1}^m \eta_j K_j$ is positive semi-definite. However, theoretical advantages of such a possibility still remain unclear.

3.6 Estimation of correlation condition (16)

Condition (4) is simple to compute while the non parametric condition (16) might be hard to check even if all densities are known. The following proposition shows that we can consistently estimate the quantities $\left\|\Sigma_{X_iX_i}^{1/2} C_{X_iX_\mathbf{J}} C_{X_\mathbf{J}X_\mathbf{J}}^{-1} \operatorname{Diag}(d_j/\|\mathbf{f}_j\|)\mathbf{g}_\mathbf{J}\right\|$ given i.i.d. samples (see proof in Appendix C.4):

Proposition 12 *Assume $\kappa_n \to 0$ and $\kappa_n n^{1/2} \to \infty$. Let $\alpha = \left(\sum_{j\in\mathbf{J}} K_j + n\kappa_n I_n\right)^{-1}\bar{Y}$ and $\hat{\eta}_j = \frac{1}{d_j}(\alpha^\top K_j \alpha)^{1/2}$. We Then, for all i, the norm $\left\|\Sigma_{X_iX_i}^{1/2} C_{X_iX_\mathbf{J}} C_{X_\mathbf{J}X_\mathbf{J}}^{-1} \operatorname{Diag}(d_j/\|\mathbf{f}_j\|)\mathbf{g}_\mathbf{J}\right\|$ is consistently estimated by:*

$$\left\| K_j^{1/2} \left(\sum_{j\in\mathbf{J}} K_j + n\kappa_n I_n\right)^{-1} \left(\sum_{j\in\mathbf{J}} \frac{1}{\hat{\eta}_j} K_j\right)\alpha \right\|. \tag{21}$$

In Section 5, we use this proposition with $n = 10,000$ to empirically check that the condition (17) is not satisfied.

4. Adaptive multiple kernel learning

In previous sections, we have shown that specific necessary and sufficient conditions are needed for path consistency of the group Lasso and multiple kernel learning. The following procedure, adapted from the adaptive Lasso of Zou (2006), leads to a two-step procedure that always achieves consistency, with no condition such as Eq. (16). We first begin by the consistency of the least-square estimate (see proof in Appendix C.6):

Theorem 13 *Assume (**A4**), (**A5**), (**A6**), (**A7**), (**A5**). The unique minimizer $\hat{f}_{\kappa_n}^{LS}$ of*

$$\frac{1}{2}\hat{\Sigma}_{YY} - \hat{\Sigma}_{YX} f + \frac{1}{2}\langle f, \hat{\Sigma}_{XX} f\rangle + \frac{\kappa_n}{2}\sum_{j=1}^m \|f_j\|^2,$$

converges in probability to f if $\kappa_n \to 0$ and $\kappa_n n^{1/2} \to 0$. Moreover, we have $\|\hat{f}_{\kappa_n}^{LS} - f\| = O_p(\kappa_n^{1/2} + \kappa_n^{-1} n^{-1/2})$.

Since the least-square estimate is consistent and we have an upper bound on its convergence rate, we follow Zou (2006) and use it to defined adaptive weights d_j for which we get consistency without any conditions on the value of the correlation operators.

Theorem 14 *Assume (**A4**), (**A5**), (**A6**), (**A7**), (**A5**). Let $\hat{f}^{LS}_{n^{-1/3}}$ be the least-square estimate with regularization parameter proportional to $n^{-1/3}$, as defined in Theorem 13. Let $\hat{f}$ denote any minimizer of*

$$\frac{1}{2}\hat{\Sigma}_{YY} - \hat{\Sigma}_{YX}f + \frac{1}{2}\langle f, \hat{\Sigma}_{XX}f\rangle + \frac{\mu_0 n^{-1/3}}{2}\left(\sum_{j=1}^{m}\|(\hat{f}^{LS}_{\kappa_n})_j\|^{-\gamma}\|f_j\|\right)^2.$$

For any $\gamma > 1$, $\hat{f}$ converges to $\mathbf{f}$ and $J(\hat{f})$ converges to $\mathbf{J}$ in probability.

Theorem 14 allows to set up a specific set of weights d_j. This provides a principled way to define data adaptive weights, that allows to solve (at least theoretically) the potential consistency problems of the usual MKL framework (see Section 5 for illustration on synthetic examples).

In the context of multiple kernel learning, the following proposition gives the expression for the solution of the least-square problem, necessary for the computation of adaptive weights in Theorem 14.

Proposition 15 *The solution of the least-square problem in Theorem 13 is given by*

$$\forall j,\ f_j^{LS} = \sum_{i=1}^{n}\alpha_i k_j(\cdot, x_{ij}) \text{ with } \alpha = \left(\sum_{j=1}^{m} K_j + n\kappa_n I_n\right)^{-1}\bar{Y}$$

with norms $\|f_j^{LS}\| = \left(\alpha^\top K_j \alpha\right)^{1/2}$, $j = 1, \ldots, m$.

Other weighting schemes have been suggested, based on various heuristics. A notable one is the normalization of kernel matrices by their trace (Lanckriet et al., 2004b), which leads to $d_j = (\mathrm{tr}\hat{\Sigma}_{X_jX_j})^{1/2} = (\frac{1}{n}\mathrm{tr}K_j)^{1/2}$. Bach et al. (2004b) have observed empirically that such normalization might lead to suboptimal solutions and consider weights d_j that grow with the empirical ranks of the kernel matrices. In this paper, we give theoretical arguments that indicate that weights which do depend on the data are more appropriate and work better (see Section 5 for examples).

5. Simulations

In this section, we illustrate the consistency results obtained in this paper with a few simple simulations on synthetic examples. In the finite dimensional group case, we generated a joint random covariance matrix for 6 groups of size 3, obtained by sampling the entries of its square root from independent standard normal distributions; three non zero loadings were also sampled from independent standard normal distributions. Finally, we chose a noise level of standard deviation 1/2. For cases when the correlation conditions (4) and (5) were or were not satisfied, we then tried different weighting schemes, i.e., different ways of

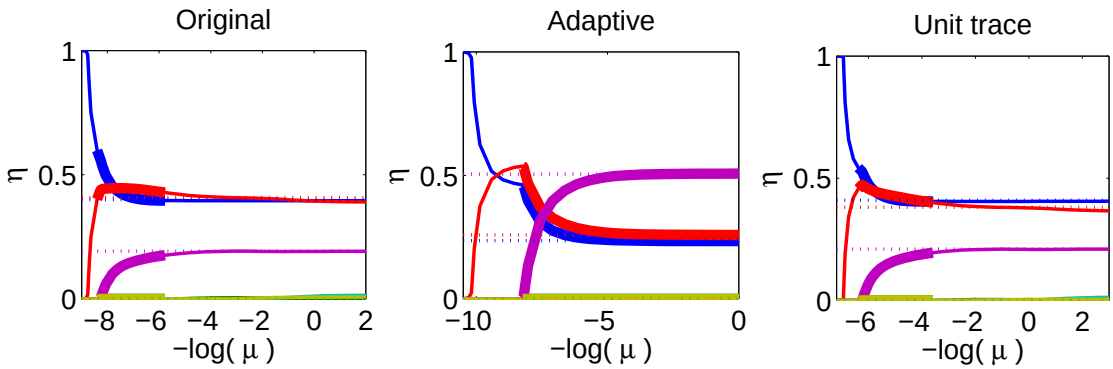

Figure 1: Regularization paths for the group Lasso for three weighting schemes (with correlation condition satisfied): (left) unit weights, (center) adaptive weights, (right) weights corresponding to unit trace constraint. For each of the three plots, plain curves correspond to values of estimated $\hat{\eta}_j$, dotted curves to population values η_j, and bold curves to model consistent estimates. As expected, all weighting schemes lead to correct model selection for at least some parts of the paths.

setting the weights d_j of the block 1-norm: unit weights, weights equal to $d_j = (\mathrm{tr}\hat{\Sigma}_{X_jX_j})^{1/2}$ (which corresponds to unit trace constraint on the kernel matrix), and adaptive weights as defined in Section 4. In Figure 1 and Figure 2, we plot the approximate regularization paths corresponding to 200 i.i.d. samples, computed by the algorithm of Bach et al. (2004b). We only plot the values of the estimated variables $\hat{\eta}_j, j = 1, \ldots, m$ for the alternative formulation in Section 2.6, which are proportional to $\|\hat{w}_j\|$ and normalized so that $\sum_{j=1}^m \hat{\eta}_j d_j^2 = 1$ (note that the values of η_j depends on d_j so that even the population values may be different for different weighting schemes, even when applied on the same data). We compare them to the population values η_j: both in terms of values, and in terms of their sparsity pattern (η_j is zero for the weights which are equal to zero).

In Figure 1, the strict correlation condition (4) was selected to be satisfied for the unit weighting scheme, thus, as expected the unit weighting scheme leads to correct model selection for at least some parts of the paths. In the example we show, the condition (4) was also satisfied for the unit trace scheme; and the adaptive weights also perform well. However, in Figure 2, where the weak condition (5) was not satisfied, none of the two response-independent schemes are selecting the correct model, while the adaptive weights do.

For the non parametric case, we followed the same approach and we generated 6 correlated Gaussian real random variables (with covariance matrix with condition number 100). We then selected three zero functions and the three functions shown in Figure 3 and a Gaussian kernel with bandwidth τ equal to one, for each dimension. We selected a covariance matrix such that the condition (17), as estimated in Section 3.6 from 10,000 samples, was not satisfied. In Figure 4, we show the regularization paths from 400 i.i.d. samples. As expected the adaptive weighting scheme manages to obtain a model consistent estimate.

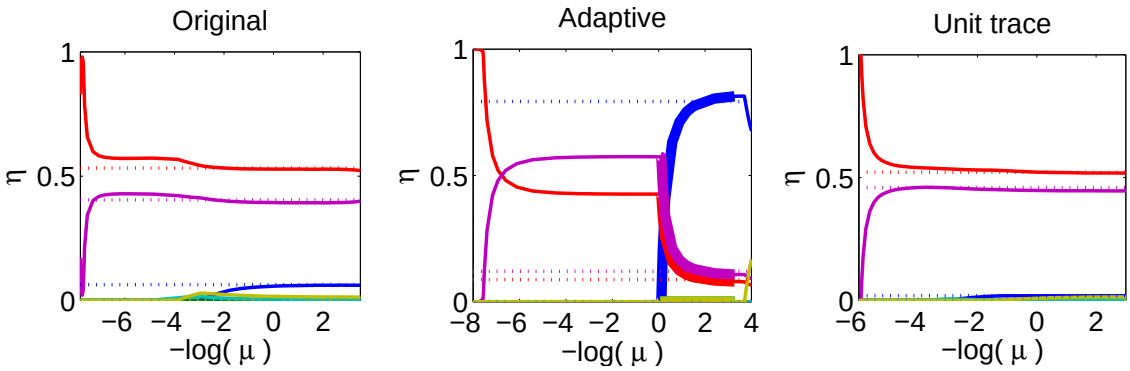

Figure 2: Regularization paths for the group Lasso for three weighting schemes (with correlation condition not satisfied): (left) unit weights, (center) adaptive weights, (right) weights corresponding to unit trace constraint. For each of the three plots, plain curves correspond to values of estimated $\hat{\eta}_j$, dotted curves to population values η_j, and bold curves to model consistent estimates. In this situation, only the adaptive weights leads to correct model selection for at least some parts of the paths.

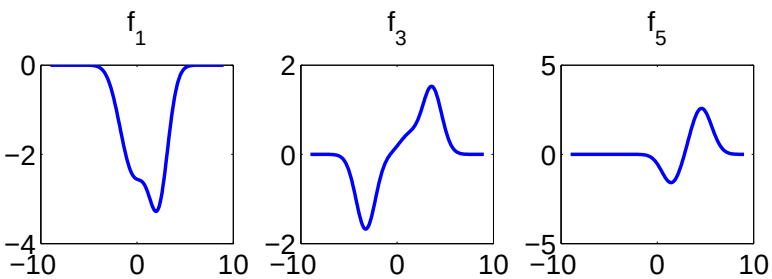

Figure 3: Functions to be estimated in the synthetic non parametric group Lasso experiments.

However, such schemes should be used with care, as there is one added free parameter (the regularization parameter κ of the least-square estimate used to define the weights): if chosen too large, all adaptive weights are equal, and thus there is no adaptation, while if chosen too small, the least-square estimate is overfit (this is the case in the right plot of Figure 4).

6. Conclusion

In this paper, we have extended some of the theoretical results of the Lasso to the group Lasso, for finite dimensional groups and infinite dimensional groups. In particular, under practical assumptions regarding the distributions the data are sampled from, we provide

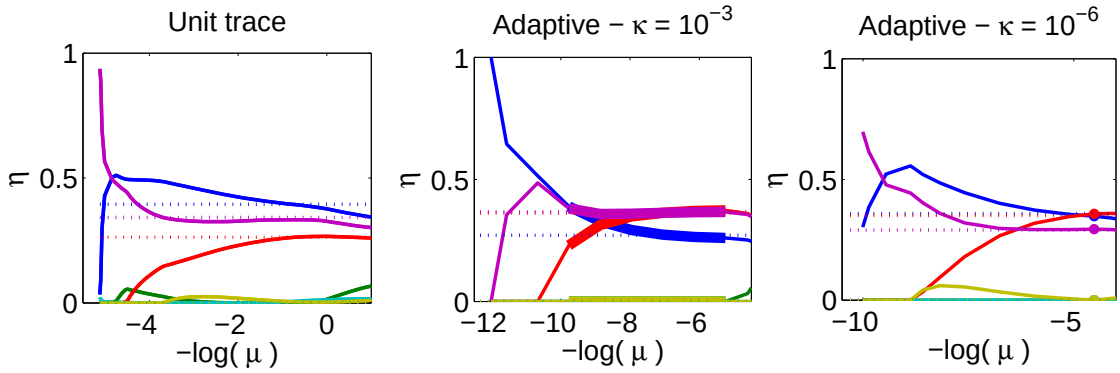

Figure 4: Regularization paths for the nonparametric group Lasso for three weighting schemes (with correlation condition not satisfied): (left) unit trace weights, (center) adaptive weights obtained with $\kappa = 10^{-3}$, (right) adaptive weights obtained with $\kappa = 10^{-6}$. For each of the three plots, plain curves correspond to values of estimated $\hat{\eta}_j$, dotted curves to population values η_j, and bold curves to model consistent estimates. In this situation, only the adaptive weights with reasonable regularization parameter κ leads to correct model selection for at least some parts of the paths.

necessary and sufficient conditions for model consistency of the group Lasso and its nonparametric version, multiple kernel learning.

The current work could be extended in several ways: first, a more detailed study of the limiting distributions of the group Lasso and adaptive group Lasso estimators could be carried and then extend the analysis of Zou (2006) or Juditsky and Nemirovski (2000), Wu et al. (2007), in particular regarding convergence rates. Second, our results should extend to generalized linear models, such as logistic regression (Meier et al., 2006). Finally, it is of interest to let the number m of groups or kernels to grow unbounded and extend the results of Zhao and Yu (2006), Meinshausen and Yu (2006) to the group Lasso.

Appendix A. Proof of optimization results

In this appendix, we give detailed proofs of the various propositions on optimality conditions and dual problems.

A.1 Proof of Proposition 1

We rewrite problem in Eq. (1), in the form

$$\min_{w \in \mathbb{R}^p,\ v \in \mathbb{R}^m} \frac{1}{2}\hat{\Sigma}_{YY} - \hat{\Sigma}_{YX} w + \frac{1}{2} w^\top \hat{\Sigma}_{XX} w + \lambda_n \sum_{j=1}^m d_j v_j,$$

with constraints $\forall j, \|w_j\| \leqslant v_j$. In order to deal with these constraints we use the tools from conic programming with the second-order cone, also known as the "ice cream" cone (Boyd and Vandenberghe, 2003). We consider the Lagrangian with dual variables $(\beta_j, \gamma_j) \in \mathbb{R}^{p_j} \times \mathbb{R}$ such that $\|\beta_j\| \leqslant \gamma_j$:

$$\mathcal{L}(w, v, \beta, \gamma) = \frac{1}{2}\hat{\Sigma}_{YY} - \hat{\Sigma}_{YX} w + \frac{1}{2} w^\top \hat{\Sigma}_{XX} w + \lambda_n d^\top v - \beta^\top w - \gamma^\top v.$$

The derivatives with respect to primal variables are

$$\begin{aligned} \nabla_w \mathcal{L}(w, v, \beta, \gamma) &= \hat{\Sigma}_{XX} w - \hat{\Sigma}_{XY} - \beta, \\ \nabla_v \mathcal{L}(w, v, \beta, \gamma) &= \lambda_n d - \gamma. \end{aligned}$$

At optimality, primal and dual variables are completely characterized by w and β. Since the dual and the primal problems are strictly feasible, strong duality holds and the KKT conditions for reduced primal/dual variables (w, β) are

$$\forall j,\ \|\beta_j\| \leqslant \lambda_n d_j \qquad \text{(dual feasibility)}\ , \tag{22}$$

$$\forall j,\ \beta_j = \hat{\Sigma}_{X_j X} w - \hat{\Sigma}_{X_j Y} \qquad \text{(stationarity)}\ , \tag{23}$$

$$\forall j,\ \beta_j^\top w_j + \|w_j\| \lambda_n d_j = 0 \qquad \text{(complementary slackness)}\ . \tag{24}$$

Complementary slackness for the second order cone has special consequences: $w_j^\top \beta_j + \|w_j\| \lambda_n d_j = 0$ if and only if (Boyd and Vandenberghe, 2003, Lobo et al., 1998), either (a) $w_j = 0$, or (b) $w_j \neq 0$, $\|\beta_j\| = \lambda_n d_j$ and $\exists \eta_j > 0$ such that $w_j = -\frac{\eta_j}{\lambda_n} \beta_j$ (anti-proportionality), which implies $\beta_j = -w_j \frac{\lambda_n d_j}{\|w_j\|}$ and $\eta_j = \|w_j\| / d_j$. This leads to the proposition.

A.2 Proof of Proposition 6

We follow the proof of Proposition 1 and of Bach et al. (2004a). We rewrite problem in Eq. (10), in the form

$$\min_{w \in \mathbb{R}^p,\ v \in \mathbb{R}^m,\ t \in \mathbb{R}} \ \frac{1}{2}\hat{\Sigma}_{YY} - \hat{\Sigma}_{YX} w + \frac{1}{2} w^\top \hat{\Sigma}_{XX} w + \frac{1}{2}\mu_n t^2,$$

with constraints that $\forall j, \|w_j\| \leqslant v_j$ and $d^\top v \leqslant t$. We consider the Lagrangian with dual variables $(\beta_j, \gamma_j) \in \mathbb{R}^{p_j} \times \mathbb{R}$ and $\delta \in \mathbb{R}_+$ such that $\|\beta_j\| \leqslant \gamma_j$:

$$\mathcal{L}(w, v, \beta, \gamma, \delta) = \frac{1}{2}\hat{\Sigma}_{YY} - \hat{\Sigma}_{YX} w + \frac{1}{2} w^\top \hat{\Sigma}_{XX} w + \frac{1}{2}\mu_n t^2 - \beta^\top w - \gamma^\top v + \delta(d^\top v - t).$$

The derivatives with respect to primal variables are

$$\begin{aligned} \nabla_w \mathcal{L}(w, v, \beta, \gamma) &= \hat{\Sigma}_{XX} w - \hat{\Sigma}_{XY} - \beta, \\ \nabla_v \mathcal{L}(w, v, \beta, \gamma) &= \delta d - \gamma, \\ \nabla_t \mathcal{L}(w, v, \beta, \gamma) &= \mu_n t - \delta. \end{aligned}$$

At optimality, primal and dual variables are completely characterized by w and β. Since the dual and the primal problems are strictly feasible, strong duality holds and the KKT conditions for reduced primal/dual variables (w, β) are

$$\forall j, \beta_j = \hat{\Sigma}_{X_j X} w - \hat{\Sigma}_{X_j Y} \qquad \text{(stationarity - 1)} , \tag{25}$$

$$\forall j, \sum_{j=1}^{m} d_j \|w_j\| = \frac{1}{\mu_n} \max_{i=1,\dots,m} \frac{\|\beta_i\|}{d_i} \qquad \text{(stationarity - 2)} , \tag{26}$$

$$\forall j, \left(\frac{\beta_j}{d_j}\right)^\top w_j + \|w_j\| \max_{i=1,\dots,m} \frac{\|\beta_i\|}{d_i} = 0 \qquad \text{(complementary slackness)} . \tag{27}$$

Complementary slackness for the second order cone implies that:

$$\left(\frac{\beta_j}{d_j}\right)^\top w_j + \|w_j\| \max_{i=1,\dots,m} \frac{\|\beta_i\|}{d_i} = 0,$$

if and only if, either (a) $w_j = 0$, or (b) $w_j \neq 0$ and $\frac{\|\beta_j\|}{d_j} = \max_{i=1,\dots,m} \frac{\|\beta_i\|}{d_i}$, and $\exists \eta_j \geqslant 0$ such that $w_j = -\eta_j \beta_j / \mu_n$, which implies $\|w_j\| = \frac{\eta_j d_j}{\mu_n} \max_{i=1,\dots,m} \frac{\|\beta_i\|}{d_i}$.

By writing $\eta_j = 0$ if $w_j = 0$ (i.e., in order to cover all cases), we have from Eq. (26) $\sum_{j=1}^m d_j \|w_j\| = \frac{1}{\mu_n} \max_{i=1,\dots,m} \frac{\|\beta_i\|}{d_i}$, which implies $\sum_{j=1}^m d_j^2 \eta_j = 1$ and thus $\forall j,\ \eta_j \propto \frac{\|w_j\|/d_j}{\sum_i d_i \|w_i\|}$. This leads to $\beta_j = -w_j \mu_n / \eta_j = -\frac{w_j}{\|w_j\|} \sum_{i=1}^n d_i \|w_i\|$. The proposition follows.

A.3 Proof of Proposition 8

By following the usual proof of the representer theorem (Wahba, 1990), we obtain that each optimal function f_j must be supported by the data points, i.e., there exists $\alpha \in \mathbb{R}^{m \times n}$ such that for all $j = 1, \dots, m$, $f_j = \sum_{i=1}^n \alpha_{ji} k(\cdot, x_{ij})$. When using this representation back into Eq. (13), we obtain an optimization problem that only depends on $\phi_j = G_j^\top \alpha_j$ for $j = 1, \dots, n$ where G_j denotes any square root of the kernel matrix K_j , i.e. $K_j = G_j G_j^\top$. This problem is exactly the finite dimensional problem in Eq. (10), where $\bar{X}_j$ is replaced by G_j and w_j by ϕ_j. Thus Proposition 6 applies and we can easily derive the current proposition by expressing all terms through the functions f_j. Note that in this proposition, we do not show that the α_j are all proportional to the same vector, as is done in Appendix A.4.

A.4 Proof of Proposition 11

We prove the proposition in the linear case. Going to the general case, can be done in the same way as done in Appendix A.3. We simply need to add a new variable $u = \bar{X} w$ and to "dualize" it. That is, we rewrite problem in Eq. (10), in the form

$$\min_{w \in \mathbb{R}^p,\ v \in \mathbb{R}^m,\ t \in \mathbb{R},\ u \in \mathbb{R}^n} \frac{1}{2n} \|\bar{Y} - u\|^2 + \frac{1}{2} \mu_n t^2,$$

with constraints that $\forall j, \|w_j\| \leqslant v_j$ and $d^\top v \leqslant t$ and $\bar{X}w = u$. We consider the Lagrangian with dual variables $(\beta_j, \gamma_j) \in \mathbb{R}^{p_j} \times \mathbb{R}$ and $\delta \in \mathbb{R}_+$ such that $\|\beta_j\| \leqslant \gamma_j$, and $\alpha \in \mathbb{R}^n$:

$$\mathcal{L}(w, v, u, \beta, \gamma, \alpha, \delta) = \frac{1}{2n}\|\bar{Y} - u\|^2 + \mu_n \alpha^\top (u - \bar{X}w) + \frac{1}{2}\mu_n t^2 - \sum_{j=1}^m \left\{ \beta_j^\top w_j + \gamma_j v_j \right\} + \delta(d^\top v - t).$$

The derivatives with respect to primal variables are

$$\begin{aligned}
\nabla_w \mathcal{L}(w, v, u, \beta, \gamma, \alpha) &= -\mu_n \bar{X}^\top \alpha - \beta \\
\nabla_v \mathcal{L}(w, v, u, \beta, \gamma, \alpha) &= \delta d - \gamma \\
\nabla_t \mathcal{L}(w, v, u, \beta, \gamma, \alpha) &= \mu_n t - \delta \\
\nabla_u \mathcal{L}(w, v, u, \beta, \gamma, \alpha) &= \frac{1}{n}(u - \bar{Y} + \mu_n n \alpha).
\end{aligned}$$

Equating them to zero, we get the dual problem in Eq. (18). Since the dual and the primal problems are strictly feasible, strong duality holds and the KKT conditions for reduced primal/dual variables (w, α) are

$$\forall j, \bar{X}w - \bar{Y} + \mu_n n \alpha = 0 \qquad \text{(stationarity - 1)}\ , \tag{28}$$

$$\forall j, \sum_{j=1}^m d_j \|w_j\| = \max_{i=1,\dots,m} \frac{(\alpha^\top K_i \alpha)^{1/2}}{d_i} \qquad \text{(stationarity - 2)}\ , \tag{29}$$

$$\forall j, \left(\frac{-\bar{X}_j^\top \alpha}{d_j} \right)^\top w_j + \|w_j\| \max_{i=1,\dots,m} \frac{(\alpha^\top K_i \alpha)^{1/2}}{d_i} = 0 \qquad \text{(complementary slackness)}\ . \tag{30}$$

Complementary slackness for the second order cone goes leads to:

$$\left(\frac{-\bar{X}_j^\top \alpha}{d_j} \right)^\top w_j + \|w_j\| \max_{i=1,\dots,m} \frac{(\alpha^\top K_i \alpha)^{1/2}}{d_i} = 0,$$

if and only if, either (a) $w_j = 0$, or (b) $w_j \neq 0$ and $\frac{(\alpha^\top K_j \alpha)^{1/2}}{d_j} = \max_{i=1,\dots,m} \frac{(\alpha^\top K_i \alpha)^{1/2}}{d_i}$, and $\exists \eta_j \geqslant 0$ such that $w_j = -\eta_j \left(-\bar{X}_j^\top \alpha \right)$, which implies $\|w_j\| = \eta_j d_j \max_{i=1,\dots,m} \frac{(\alpha^\top K_i \alpha)^{1/2}}{d_i}$.

By writing $\eta_j = 0$ if $w_j = 0$ (to cover all cases), we have from Eq. (29), $\sum_{j=1}^m d_j \|w_j\| = \max_{i=1,\dots,m} \frac{(\alpha^\top K_i \alpha)^{1/2}}{d_i}$, which implies $\sum_{j=1}^m d_j^2 \eta_j = 1$. The proposition follows from the fact that at optimality, $w_j = -\eta_j - \bar{X}_j^\top \alpha$

Appendix B. Detailed proofs for the group Lasso

In this appendix, detailed proofs of the consistency results for the finite dimensional case (Theorems 2 and 3) are presented. Some of the results presented in this appendix are corollaries of the more general results in Appendix C, but their proofs in the finite dimensional case are much simpler.

B.1 Proof of Theorem 2

We begin with a lemma, which states that if we restrict ourselves to the covariates which we are after (i.e., indexed by $\mathbf{J}$), we get a consistent estimate as soon as λ_n tends to zero:

Lemma 16 *Let $\tilde{w}_{\mathbf{J}}$ any minimizer of*

$$\frac{1}{2n}\|\bar{Y} - \bar{X}_{\mathbf{J}} w_{\mathbf{J}}\|^2 + \lambda_n \sum_{j \in \mathbf{J}} d_j \|w_j\| = \frac{1}{2}\hat{\Sigma}_{YY} - \hat{\Sigma}_{YX_{\mathbf{J}}} w_{\mathbf{J}} + \frac{1}{2} w_{\mathbf{J}}^\top \hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}} w_{\mathbf{J}} + \lambda_n \sum_{j \in \mathbf{J}} d_j \|w_j\|.$$

If $\lambda_n \to 0$, then $\tilde{w}_{\mathbf{J}}$ converges to $\mathbf{w}_{\mathbf{J}}$ in probability.

Proof If λ_n tends to zero, then the cost function defining $\tilde{w}_{\mathbf{J}}$ converges to $F_n(w_{\mathbf{J}}) = \frac{1}{2}\Sigma_{YY} - \Sigma_{YX_{\mathbf{J}}} w_{\mathbf{J}} + \frac{1}{2} w_{\mathbf{J}}^\top \Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}} w_{\mathbf{J}}$ whose unique (because $\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}}$ is positive definite) global minimum is $\mathbf{w}_{\mathbf{J}}$ (true generating value). The convergence of $\tilde{w}_{\mathbf{J}}$ is thus a simple consequence of standard results in M-estimation (Van der Vaart, 1998, Fu and Knight, 2000). ■

We now prove Theorem 2. Let $\tilde{w}_{\mathbf{J}}$ be defined as in Lemma 16. We extend it by zeros on $\mathbf{J}^c$. We already know from Lemma 16 that we have consistency in squared norm. We now need to prove that the probability that $\tilde{w}$ is optimal for problem in Eq. (1) is tending to one.

By definition of $\tilde{w}_{\mathbf{J}}$, the optimality condition (3) is satisfied. We now need to verify optimality condition (2). Denoting $\varepsilon = Y - \mathbf{w}^\top X$, we have:

$$\hat{\Sigma}_{XY} = \hat{\Sigma}_{XX}\mathbf{w} + \hat{\Sigma}_{X\varepsilon} = \left(\Sigma_{XX} + O_p(n^{-1/2})\right)\mathbf{w} + O_p(n^{-1/2}) = \Sigma_{XX_{\mathbf{J}}}\mathbf{w}_{\mathbf{J}} + O_p(n^{-1/2}),$$

because of classical results on convergence of empirical covariances to covariances (Van der Vaart, 1998), which are applicable because we have the fourth order moment condition (**A1**). We thus have:

$$\hat{\Sigma}_{XY} - \hat{\Sigma}_{XX_{\mathbf{J}}}\tilde{w}_{\mathbf{J}} = \Sigma_{XX_{\mathbf{J}}}(\mathbf{w}_{\mathbf{J}} - \tilde{w}_{\mathbf{J}}) + O_p(n^{-1/2}). \tag{31}$$

From the optimality condition $\hat{\Sigma}_{X_{\mathbf{J}}Y} - \hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}}\tilde{w}_{\mathbf{J}} = \lambda_n \operatorname{Diag}(d_j/\|\tilde{w}_j\|)\tilde{w}_{\mathbf{J}}$ defining $\tilde{w}_{\mathbf{J}}$ and Eq. (31), we obtain:

$$\tilde{w}_{\mathbf{J}} - \mathbf{w}_{\mathbf{J}} = -\lambda_n \Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}}^{-1} \operatorname{Diag}(d_j/\|\tilde{w}_j\|)\tilde{w}_{\mathbf{J}} + O_p(n^{-1/2}). \tag{32}$$

Therefore,

$$\begin{aligned}
\hat{\Sigma}_{X_{\mathbf{J}^c}Y} - \hat{\Sigma}_{X_{\mathbf{J}^c}X_{\mathbf{J}}}\tilde{w}_{\mathbf{J}} &= \Sigma_{X_{\mathbf{J}^c}X_{\mathbf{J}}}(\mathbf{w}_{\mathbf{J}} - \tilde{w}_{\mathbf{J}}) + O_p(n^{-1/2}) \text{ by Eq. (31)} , \\
&= \lambda_n \Sigma_{X_{\mathbf{J}^c}X_{\mathbf{J}}}\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}}^{-1} \operatorname{Diag}(d_j/\|\tilde{w}_j\|)\tilde{w}_{\mathbf{J}} + O_p(n^{-1/2}) \text{ by Eq. (32)}.
\end{aligned}$$

Since $\tilde{w}$ is consistent, and $\lambda_n n^{1/2} \to +\infty$, then for each $i \in J^c$,

$$\frac{1}{d_i \lambda_n}\left(\hat{\Sigma}_{X_iY} - \hat{\Sigma}_{X_iX_{\mathbf{J}}}\tilde{w}_{\mathbf{J}}\right)$$

converges in probability to $\Sigma_{X_iX_\mathbf{J}}\Sigma_{X_\mathbf{J}X_\mathbf{J}}^{-1}\operatorname{Diag}(d_j/\|\mathbf{w}_j\|)\mathbf{w}_\mathbf{J}$ which is of norm strictly smaller than one because condition (4) is satisfied. Thus the probability that $\tilde{w}$ is indeed optimal, which is equal to

$$P\left\{\forall i \in \mathbf{J}^c, \frac{1}{d_i\lambda_n}\left\|\hat{\Sigma}_{X_iY} - \hat{\Sigma}_{X_iX_\mathbf{J}}\tilde{w}_\mathbf{J}\right\| \leqslant 1\right\} \geqslant \prod_{i\in\mathbf{J}^c} P\left\{\frac{1}{d_i\lambda_n}\left\|\hat{\Sigma}_{X_iY} - \hat{\Sigma}_{X_iX_\mathbf{J}}\tilde{w}_\mathbf{J}\right\| \leqslant 1\right\}$$

is tending to 1, which implies the theorem.

B.2 Proof of theorem 3

We prove the theorem by contradiction, by assuming that there exists $i \in \mathbf{J}^c$ such that

$$\frac{1}{d_i}\left\|\Sigma_{X_iX_\mathbf{J}}\Sigma_{X_\mathbf{J}X_\mathbf{J}}^{-1}\operatorname{Diag}(d_j/\|\mathbf{w}_j\|)\mathbf{w}_\mathbf{J}\right\| > 1.$$

Since with probability tending to one $J(\hat{w}) = \mathbf{J}$, with probability tending to one, we have from optimality condition (3):

$$\hat{w}_\mathbf{J} = \hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}}^{-1}\left(\hat{\Sigma}_{X_\mathbf{J}Y} - \lambda_n\operatorname{Diag}(d_j/\|\hat{w}_j\|)\hat{w}_\mathbf{J}\right),$$

and thus

$$\begin{aligned}\hat{\Sigma}_{X_iY} - \hat{\Sigma}_{X_iX_\mathbf{J}}\hat{w}_\mathbf{J} &= \hat{\Sigma}_{X_iY} - \hat{\Sigma}_{X_iX_\mathbf{J}}\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}}^{-1}\hat{\Sigma}_{X_\mathbf{J}Y} + \lambda_n\hat{\Sigma}_{X_iX_\mathbf{J}}\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}}^{-1}\operatorname{Diag}(d_j/\|\hat{w}_j\|)\hat{w}_\mathbf{J}\\ &= A_n + B_n.\end{aligned}$$

The second term B_n in the last expression (divided by λ_n) converges to

$$v = \Sigma_{X_iX_\mathbf{J}}\Sigma_{X_\mathbf{J}X_\mathbf{J}}^{-1}\operatorname{Diag}(d_j/\|\mathbf{w}_j\|)\mathbf{w}_\mathbf{J} \in \mathbb{R}^{p_j},$$

because $\hat{w}$ is assumed to converge in probability to $\mathbf{w}$ and empirical covariance matrices converge to population covariance matrices. By assumption $\|v\| > d_i$, which implies that the probability $\mathbb{P}\left\{\left(\frac{v}{\|v\|}\right)^\top(B_n/\lambda_n) \geqslant (d_i + \|v\|)/2)\right\}$ converges to one.

The first term is equal to:

$$\begin{aligned}A_n &= \hat{\Sigma}_{X_iY} - \hat{\Sigma}_{X_iX_\mathbf{J}}\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}}^{-1}\hat{\Sigma}_{X_\mathbf{J}Y}\\ &= \hat{\Sigma}_{X_iX_\mathbf{J}}\mathbf{w} - \hat{\Sigma}_{X_iX_\mathbf{J}}\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}}^{-1}\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}}\mathbf{w} + \hat{\Sigma}_{X_i\varepsilon} - \hat{\Sigma}_{X_iX_\mathbf{J}}\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}}^{-1}\hat{\Sigma}_{X_\mathbf{J}\varepsilon}\\ &= \hat{\Sigma}_{X_i\varepsilon} - \hat{\Sigma}_{X_iX_\mathbf{J}}\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}}^{-1}\hat{\Sigma}_{X_\mathbf{J}\varepsilon}\\ &= \hat{\Sigma}_{X_i\varepsilon} - \Sigma_{X_iX_\mathbf{J}}\Sigma_{X_\mathbf{J}X_\mathbf{J}}^{-1}\hat{\Sigma}_{X_\mathbf{J}\varepsilon} + o_p(n^{-1/2})\\ &= \frac{1}{n}\sum_{k=1}^n \varepsilon_k\left(x_{ki} - \Sigma_{X_iX_\mathbf{J}}\Sigma_{X_\mathbf{J}X_\mathbf{J}}^{-1}x_{k\mathbf{J}}\right) + o_p(n^{-1/2}) = C_n + o_p(n^{-1/2})\end{aligned}$$

The random variable C_n is a mean of i.i.d. random variables with finite variance. Thus, by the central limit theorem, it is asymptotically normal (Van der Vaart, 1998). We thus

simply need to compute the mean and the variance of C_n. We have $\mathbb{E}C_n = 0$ because $\mathbb{E}(X\varepsilon) = \Sigma_{X\varepsilon} = 0$, and

$$\begin{aligned}
\mathrm{var}(C_n) &= \mathbb{E}C_n^2 = \mathbb{E}(\mathbb{E}(C_n^2|\bar{X})) \\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{k=1}^n E(\varepsilon_k^2|\bar{X})\left(x_{ki} - \Sigma_{X_iX_\mathbf{J}}\Sigma_{X_\mathbf{J}X_\mathbf{J}}^{-1}x_{k\mathbf{J}}\right)^2\right] \\
&\succcurlyeq \mathbb{E}\left[\frac{1}{n^2}\sum_{k=1}^n \sigma_{\min}^2\left(x_{ki} - \Sigma_{X_iX_\mathbf{J}}\Sigma_{X_\mathbf{J}X_\mathbf{J}}^{-1}x_{k\mathbf{J}}\right)^2\right] \\
&= \frac{1}{n}\sigma_{\min}^2\left(\Sigma_{X_iX_i} - \Sigma_{X_iX_\mathbf{J}}\Sigma_{X_\mathbf{J}X_\mathbf{J}}^{-1}\Sigma_{X_\mathbf{J}X_i}\right).
\end{aligned}$$

Thus $n^{1/2}C_n$ is asymptotically normal with mean 0 and covariance matrix larger than $\sigma_{\min}^2\Sigma_{X_i|X_\mathbf{J}} = \sigma_{\min}^2 \times (\Sigma_{X_iX_i} - \Sigma_{X_iX_\mathbf{J}}\Sigma_{X_\mathbf{J}X_\mathbf{J}}^{-1}\Sigma_{X_\mathbf{J}X_i})$ which is positive definite (because this is the conditional covariance of X_i given $X_\mathbf{J}$ and Σ_{XX} is assumed invertible). Therefore $P(n^{1/2}v^\top A_n > 0)$ converges to $1/2$, which implies that $P(\frac{v}{\|v\|}^\top(A_n + B_n)/\lambda_n \geqslant (d_i + \|v\|)/2)$ is asymptotically bounded below by $1/2$. Thus since $\|(A_n + B_n)/\lambda_n\| \geqslant \frac{v}{\|v\|}^\top(A_n + B_n)/\lambda_n \geqslant (d_i + \|v\|)/2 > d_i$ implies that $\hat{w}$ is not optimal, we get a contradiction, which concludes the proof.

B.3 Proof of Theorem 4

We first prove the following refinement of Lemma 16:

Lemma 17 *Let $\tilde{w}_\mathbf{J}$ any minimizer of*

$$\frac{1}{2n}\|\bar{Y} - \bar{X}_\mathbf{J}w_\mathbf{J}\|^2 + \lambda_n\sum_{j\in\mathbf{J}} d_j\|w_j\| = \frac{1}{2}\hat{\Sigma}_{YY} - \hat{\Sigma}_{YX_\mathbf{J}}w_\mathbf{J} + \frac{1}{2}w_\mathbf{J}^\top\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}}w_\mathbf{J} + \lambda_n\sum_{j\in\mathbf{J}} d_j\|w_j\|.$$

If $\lambda_n \to 0$ and $\lambda_n n^{1/2} \to \infty$, then $\frac{1}{\lambda_n}(\tilde{w}_\mathbf{J} - \mathbf{w}_\mathbf{J})$ converges in probability to

$$\Delta = -\Sigma_{X_\mathbf{J}X_\mathbf{J}}^{-1}\,\mathrm{Diag}(d_j/\|\mathbf{w}_j\|)\mathbf{w}_\mathbf{J}.$$

Proof We write $\tilde{w}_\mathbf{J} = \mathbf{w}_\mathbf{J} + \lambda_n\tilde{\Delta}$. $\tilde{\Delta}$ is the minimizer of the following function:

$$\begin{aligned}
F(\Delta) &= \hat{\Sigma}_{YX_\mathbf{J}}(\mathbf{w}_\mathbf{J} + \lambda_n\Delta) + \frac{1}{2}(\mathbf{w}_\mathbf{J} + \lambda_n\Delta)^\top\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}}(\mathbf{w}_\mathbf{J} + \lambda_n\Delta) + \lambda_n\sum_{j\in\mathbf{J}} d_j\|\mathbf{w}_j + \lambda_n\Delta_j\| \\
&= \lambda_n\hat{\Sigma}_{YX_\mathbf{J}}\Delta + \frac{\lambda_n^2}{2}\Delta^\top\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}}\Delta + \lambda_n\mathbf{w}_\mathbf{J}^\top\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}}\Delta + \lambda_n\sum_{j\in\mathbf{J}} d_j\left(\|\mathbf{w}_j + \lambda_n\Delta_j\| - \|\mathbf{w}_j\|\right) + \text{ cst} \\
&= \lambda_n\hat{\Sigma}_{\varepsilon X_\mathbf{J}}\Delta + \frac{\lambda_n^2}{2}\Delta^\top\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}}\Delta + \lambda_n\sum_{j\in\mathbf{J}} d_j\left(\|\mathbf{w}_j + \lambda_n\Delta_j\| - \|\mathbf{w}_j\|\right) + \text{ cst},
\end{aligned}$$

by using $\hat{\Sigma}_{YX_\mathbf{J}} = \mathbf{w}_\mathbf{J}^\top\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}} + \hat{\Sigma}_{\varepsilon X_\mathbf{J}}$. The first term is $O_p(n^{-1/2}\lambda_n) = o_p(\lambda_n^2)$, while the last ones are equal to $\|\mathbf{w}_j + \lambda_n\Delta_j\| - \|\mathbf{w}_j\| = \lambda_n\frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}^\top\Delta_j + o_p(\lambda_n)$. Thus

$$F(\Delta)/\lambda_n^2 = \frac{1}{2}\Delta^\top\Sigma_{X_\mathbf{J}X_\mathbf{J}}\Delta + \sum_{j\in\mathbf{J}}\frac{d_j\mathbf{w}_j}{\|\mathbf{w}_j\|}^\top\Delta_j + o_p(1).$$

By Lemma 16, $\hat{w}_{\mathbf{J}}$ is $O_p(1)$ and the limiting function has an unique minimum; standard results in M-estimation shows that $\tilde{\Delta}$ converges in probability to the minimum of the last expression which is exactly $\Delta = -\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}}^{-1} \operatorname{Diag}(d_j/\|\mathbf{w}_j\|)\mathbf{w}_{\mathbf{J}}$. ■

We now turn to the proof of Theorem 4. We follow the proof of Theorem 2. Given $\tilde{w}$ defined through Lemma 16 and 17, we need to satisfy optimality condition (3) for all $i \in \mathbf{J}^c$, with probability tending to one. For all those i such that $\frac{1}{d_i}\left\|\Sigma_{X_iX_{\mathbf{J}}}\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}}^{-1} \operatorname{Diag}(d_j/\|\mathbf{w}_j\|)\mathbf{w}_{\mathbf{J}}\right\| < 1$, then we know from Appendix B.1, that the optimality condition is indeed satisfied. We now focus on those i such that $\frac{1}{d_i}\left\|\Sigma_{X_iX_{\mathbf{J}}}\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}}^{-1} \operatorname{Diag}(d_j/\|\mathbf{w}_j\|)\mathbf{w}_{\mathbf{J}}\right\| = 1$, and for which we have the condition in Eq. (6). From Eq. (32) and the few arguments that follow, we get that

$$\hat{\Sigma}_{X_iY} - \hat{\Sigma}_{X_iX_{\mathbf{J}}}\tilde{w}_{\mathbf{J}} = \lambda_n \Sigma_{X_iX_{\mathbf{J}}}\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}}^{-1} \operatorname{Diag}(d_j/\|\tilde{w}_j\|)\tilde{w}_{\mathbf{J}} + O_p(n^{-1/2}) \tag{33}$$

Moreover, we have from Lemma 17 and standard differential calculus, i.e., the gradient and the Hessian of the function $v \mapsto \|v\|$ are $v/\|v\|$ and $\frac{1}{\|v\|}\left(I - \frac{vv^\top}{v^\top v}\right)$:

$$\frac{\tilde{w}_j}{\|\tilde{w}_j\|} = \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} + \frac{\lambda_n}{\|\mathbf{w}_j\|}\left(I_{p_j} - \frac{\mathbf{w}_j\mathbf{w}_j^\top}{\mathbf{w}_j^\top\mathbf{w}_j}\right)\Delta_j + o_p(\lambda_n). \tag{34}$$

From Eq. (33) and Eq. (34), we get:

$$\begin{aligned}\frac{1}{\lambda_n}(\hat{\Sigma}_{X_iY} - \hat{\Sigma}_{X_iX_{\mathbf{J}}}\tilde{w}_{\mathbf{J}}) &= \Sigma_{X_iX_{\mathbf{J}}}\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}}^{-1} \operatorname{Diag}(d_j/\|\mathbf{w}_j\|)\mathbf{w}_{\mathbf{J}} + o_p(\lambda_n) \\ &\quad + \lambda_n \Sigma_{X_iX_{\mathbf{J}}}\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}}^{-1} \operatorname{Diag}\left[d_j/\|\mathbf{w}_j\|\left(I_{p_j} - \frac{\mathbf{w}_j\mathbf{w}_j^\top}{\mathbf{w}_j^\top\mathbf{w}_j}\right)\right]\Delta + O_p(n^{-1/2}\lambda_n^{-1}).\end{aligned}$$

Since $\lambda_n = o_p(n^{-1/4})$, we have $O_p(n^{-1/2}\lambda_n^{-1}) = o_p(\lambda_n)$. Thus, since we assumed that $\|\Sigma_{X_iX_{\mathbf{J}}}\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}}^{-1} \operatorname{Diag}(d_j/\|\mathbf{w}_j\|)\mathbf{w}_{\mathbf{J}}\| = d_i$, we have:

$$\begin{aligned}\left\|\frac{1}{\lambda_n}(\hat{\Sigma}_{X_iY} - \hat{\Sigma}_{X_iX_{\mathbf{J}}}\tilde{w}_{\mathbf{J}})\right\|^2 &= d_i^2 + o_p(\lambda_n) \\ &\quad - 2\lambda_n \Delta^\top \Sigma_{X_{\mathbf{J}}X_i}\Sigma_{X_iX_{\mathbf{J}}}\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}}^{-1} \operatorname{Diag}\left(d_j/\|\mathbf{w}_j\|(I_{p_j} - \frac{\mathbf{w}_j\mathbf{w}_j^\top}{\mathbf{w}_j^\top\mathbf{w}_j})\right)\Delta,\end{aligned}$$

which is asymptotically strictyl smaller than d_i^2 if Eq. (6) is satisfied, which proves optimality and concludes the proof.

Appendix C. Detailed proofs for the non parametric formulation

We first prove lemmas that will be useful for further proofs, and then prove the consistency results for the non parametric case.

C.1 Useful lemmas on empirical covariance operators

We have the following lemma, proved by Fukumizu et al. (2007), which states that the empirical covariance estimator converges in probability at rate $O_p(n^{-1/2})$ to the population covariance operators:

Lemma 18 *Assume (**A4**) and (**A6**). Then* $\|\hat{\Sigma}_{XX} - \Sigma_{XX}\| = O_p(n^{-1/2})$ *(for the operator norm),* $\|\hat{\Sigma}_{XY} - \Sigma_{XY}\| = O_p(n^{-1/2})$ *and* $\|\hat{\Sigma}_{X\varepsilon}\| = O_p(n^{-1/2})$.

The following lemma is useful in several proofs:

Lemma 19 *Assume (**A4**). Then* $\left\| \left(\hat{\Sigma}_{XX} + \mu_n I\right)^{-1} \Sigma_{XX} - (\Sigma_{XX} + \mu_n I)^{-1} \Sigma_{XX} \right\| = O_p(n^{-1/2}\mu_n)$, *and* $\left\| \left(\hat{\Sigma} + \mu_n I\right)^{-1} \hat{\Sigma}_{XX} - (\Sigma_{XX} + \mu_n I)^{-1} \Sigma_{XX} \right\| = O_p(n^{-1/2}\mu_n)$.

Proof We have:

$$\begin{aligned} & \left(\hat{\Sigma}_{XX} + \mu_n I\right)^{-1} \Sigma_{XX} - (\Sigma_{XX} + \mu_n I)^{-1} \Sigma_{XX} \\ = \; & \left(\hat{\Sigma}_{XX} + \mu_n I\right)^{-1} (\Sigma_{XX} - \hat{\Sigma}_{XX}) (\Sigma_{XX} + \mu_n I)^{-1} \Sigma_{XX} \end{aligned}$$

This is the product of operators whose norms are respectively upper bounded by μ_n^{-1},$O_p(n^{-1/2})$ and 1, which leads to the first inequality (we use $\|AB\| \leqslant \|A\|\|B\|$). The second inequality follows along similar lines. ■

Note that the two previous lemma also hold for any suboperator of Σ_{XX}, i.e., for $\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}}$.

Lemma 20 *Assume (**A4**), (**A7**) and (**A5**). There exists* $\mathbf{h_J} \in \mathcal{F}_{\mathbf{J}}$ *such that* $\mathbf{f_J} = \Sigma^{1/2}_{X_{\mathbf{J}}X_{\mathbf{J}}} \mathbf{h}_J$.

Proof The range condition implies that

$$\mathbf{f_J} \;=\; \mathrm{Diag}(\Sigma^{1/2}_{X_jX_j})\mathbf{g_J} = \mathrm{Diag}(\Sigma^{1/2}_{X_jX_j}) C^{1/2}_{X_{\mathbf{J}}X_{\mathbf{J}}} C^{-1/2}_{X_{\mathbf{J}}X_{\mathbf{J}}} \mathbf{g_J}.$$

The result follows from the identity $\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}} = \mathrm{Diag}(\Sigma^{1/2}_{X_jX_j}) C^{1/2}_{X_{\mathbf{J}}X_{\mathbf{J}}} (\mathrm{Diag}(\Sigma^{1/2}_{X_jX_j}) C^{1/2}_{X_{\mathbf{J}}X_{\mathbf{J}}})^*$ and the fact that if $\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}} = UU^*$ and $f = U\alpha$ then there exists β such that $f = \Sigma^{1/2}_{X_{\mathbf{J}}X_{\mathbf{J}}}\beta$ (Baker, 1973). ■

C.2 Proof of Theorem 9

We now extend Lemma 16 to covariance operators, which requires to use the alternative formulation and a slower rate of decrease for the regularization parameter:

Lemma 21 *Let* $\tilde{f}_{\mathbf{J}}$ *be any minimizer of*

$$\frac{1}{2}\hat{\Sigma}_{YY} - \hat{\Sigma}_{YX_{\mathbf{J}}} f_{\mathbf{J}} + \frac{1}{2}\langle f_{\mathbf{J}}, \hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}} f_{\mathbf{J}}\rangle + \frac{\mu_n}{2}\left(\sum_{j\in\mathbf{J}} d_j \|f_j\|\right)^2.$$

If $\mu_n \to 0$ *and* $\mu_n n^{1/2} \to +\infty$, *then* $\|\tilde{f}_{\mathbf{J}} - \mathbf{f_J}\|$ *converges to zero in probability. Moreover for any* η_n *such that* $\eta_n \gg \mu_n^{1/2} + \mu_n^{-1} n^{-1/2}$ *then* $\|\tilde{f}_{\mathbf{J}} - \mathbf{f_J}\| = O_p(\eta_n)$.

Proof Note that from Pontil et al. (2007), we have:

$$\left(\sum_{j\in\mathbf{J}} d_j\|f_j\|\right)^2 \leqslant \left(\sum_{j\in\mathbf{J}} d_j\|\mathbf{f}_j\|\right)\sum_{j\in\mathbf{J}}\frac{d_j\|f_j\|^2}{\|\mathbf{f}_j\|},$$

with equality if and only if $\|f_j\| = \|\mathbf{f}_j\|$ for all $j \in \mathbf{J}$. We consider the unique minimizer $\bar{f}_{\mathbf{J}}$ of the following cost function, built by replacing the regularization by its upperbound,

$$F(f_{\mathbf{J}}) = \frac{1}{2}\hat{\Sigma}_{YY} - \hat{\Sigma}_{YX_{\mathbf{J}}}f_{\mathbf{J}} + \frac{1}{2}\langle f_{\mathbf{J}}, \hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}}f_{\mathbf{J}}\rangle + \frac{\mu_n}{2}\left(\sum_{j\in\mathbf{J}} d_j\|\mathbf{f}_j\|\right)\sum_{j\in\mathbf{J}}\frac{d_j\|f_j\|^2}{\|\mathbf{f}_j\|}.$$

Since it is a regularized least-squares problem, we have:

$$\bar{f}_{\mathbf{J}} = \left(\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D\right)^{-1}\left(\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}}\mathbf{f}_{\mathbf{J}} + \hat{\Sigma}_{X_{\mathbf{J}}\varepsilon}\right),$$

where $D = \left(\sum_{j\in\mathbf{J}} d_j\|\mathbf{f}_j\|\right)\mathrm{Diag}(d_j/\|\mathbf{f}_j\|)$. Note that D is upperbounded and lowerbounded (as an auto-adjoint operator) by *strictly positive* constants times the identity operator, i.e., $D_{\max}I \succcurlyeq D \succcurlyeq D_{\min}I$. We now prove that $\bar{f}_{\mathbf{J}} - \mathbf{f}_{\mathbf{J}}$ is converging to zero in probability. We have:

$$\left(\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D\right)^{-1}\hat{\Sigma}_{X_{\mathbf{J}}\varepsilon} = O_p(n^{-1/2}\mu_n^{-1}), \tag{35}$$

because of Lemma 18 and $\left\|\left(\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D\right)^{-1}\right\| \leqslant D_{\min}^{-1}\mu_n^{-1}$. Moreover, similarly, we have

$$\left(\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D\right)^{-1}\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}}f_{\mathbf{J}} - \left(\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D\right)^{-1}\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}}\mathbf{f}_{\mathbf{J}} = O_p(n^{-1/2}\mu_n^{-1}). \tag{36}$$

Besides, by Lemma 19,

$$\left(\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D\right)^{-1}\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}}f_{\mathbf{J}} - \left(\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D\right)^{-1}\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}}f_{\mathbf{J}} = O_p(n^{-1/2}\mu_n^{-1}). \tag{37}$$

Thus $\bar{f}_{\mathbf{J}} - \mathbf{f}_{\mathbf{J}} = V + O_p(n^{-1/2}\mu_n^{-1})$, where

$$V = \left[\left(\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D\right)^{-1}\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}} - I\right]\mathbf{f}_{\mathbf{J}} = -\left(\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D\right)^{-1}\mu_n D\mathbf{f}_{\mathbf{J}}.$$

We have

$$\begin{aligned}
\|V\|^2 &= \mu_n^2\langle \mathbf{f}_{\mathbf{J}}, D\left(\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D\right)^{-2} D\mathbf{f}_{\mathbf{J}}\rangle \\
&\leqslant D_{\max}^2\mu_n^2\langle \mathbf{f}_{\mathbf{J}}, \left(\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D_{\min}I\right)^{-2}\mathbf{f}_{\mathbf{J}}\rangle \\
&\leqslant D_{\max}^2\mu_n\langle \mathbf{f}_{\mathbf{J}}, \left(\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D_{\min}I\right)^{-1}\mathbf{f}_{\mathbf{J}}\rangle \\
&\leqslant D_{\max}^2\mu_n\langle \mathbf{h}_{\mathbf{J}}, \Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}}\left(\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D_{\min}I\right)^{-1}\mathbf{h}_{\mathbf{J}}\rangle \text{ by Lemma 20,} \\
&\leqslant D_{\max}^2\mu_n\|\mathbf{h}_{\mathbf{J}}\|^2.
\end{aligned}$$

Finally we obtain $\|\bar{f}_{\mathbf{J}} - \mathbf{f}_{\mathbf{J}}\| = O_p(\mu_n^{1/2} + n^{-1/2}\mu_n^{-1})$.

We now consider the cost function defining $\tilde{f}_{\mathbf{J}}$:

$$F_n(f_{\mathbf{J}}) = \frac{1}{2}\hat{\Sigma}_{YY} - \hat{\Sigma}_{YX_{\mathbf{J}}} f_{\mathbf{J}} + \frac{1}{2}\langle f_{\mathbf{J}}, \hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}} f_{\mathbf{J}}\rangle + \frac{\mu_n}{2}\left(\sum_{j\in\mathbf{J}} d_j\|f_j\|\right)^2.$$

We have (note that although we seem to take infinite dimensional derivatives, everything can be done in the finite subspace spanned by the data):

$$\begin{aligned} F_n(f_{\mathbf{J}}) - F(f_{\mathbf{J}}) &= \frac{\mu_n}{2}\left[\left(\sum_{j\in\mathbf{J}} d_j\|f_j\|\right)^2 - \left(\sum_{j\in\mathbf{J}} d_j\|\mathbf{f}_j\|\right)\sum_{j\in\mathbf{J}}\frac{d_j\|f_j\|^2}{\|\mathbf{f}_j\|}\right], \\ \nabla_{f_i}F_n(f_{\mathbf{J}}) - \nabla_{f_i}F(f_{\mathbf{J}}) &= \mu_n\left[\left(\sum_{j\in\mathbf{J}} d_j\|f_j\|\right)\frac{d_i f_i}{\|f_i\|} - \left(\sum_{j\in\mathbf{J}} d_j\|\mathbf{f}_j\|\right)\frac{d_i f_i}{\|\mathbf{f}_i\|}\right]. \end{aligned}$$

Since the right hand side of the previous equation corresponds to a continuously differentiable function of $f_{\mathbf{J}}$ around $\mathbf{f}_{\mathbf{J}}$ (with upper-bounded derivatives around $\mathbf{f}_{\mathbf{J}}$), we have:

$$\|\nabla_{f_i}F_n(\bar{f}_{\mathbf{J}}) - 0\| \leqslant C\mu_n\|\mathbf{f}_{\mathbf{J}} - \bar{f}_{\mathbf{J}}\| = \mu_n O_p(\mu_n^{1/2} + n^{-1/2}\mu_n).$$

Moreover, on the ball of center $\bar{f}_{\mathbf{J}}$ and radius η_n such that $\eta_n \gg \mu_n^{1/2} + \mu_n^{-1}n^{-1/2}$ (to make sure that it asymptotically contains $\bar{f}_{\mathbf{J}}$, which implies that on the ball each f_j, $j \in \mathbf{J}$ are bounded away from zero), and $\eta_n \ll 1$ (so that we get consistency), we have a lower bound on the second derivative of $\left(\sum_{j\in\mathbf{J}} d_j\|f_j\|\right)$. Thus for any element of the ball,

$$F_n(f_{\mathbf{J}}) \geqslant F_n(\bar{f}_{\mathbf{J}}) + \langle\nabla_{f_{\mathbf{J}}}F_n(\bar{f}_{\mathbf{J}}), (f_{\mathbf{J}} - \bar{f}_{\mathbf{J}})\rangle + C\mu_n\|f_{\mathbf{J}} - \bar{f}_{\mathbf{J}}\|^2,$$

where C is a constant > 0. This implies that the value of $F_n(f_{\mathbf{J}})$ on the edge of the ball is larger than

$$F_n(\bar{f}_{\mathbf{J}}) + \eta_n\mu_n O_p(\mu_n^{1/2} + n^{-1/2}\mu_n^{-1}) + C\eta_n^2\mu_n,$$

Thus if $\eta_n^2\mu_n \gg \eta_n\mu_n^{3/2}$ and $\eta_n^2\mu_n \gg n^{-1/2}\eta_n$, then we must have all minima inside the ball of radius η_n (because with probability tending to one, the value on the edge is greater than one value inside and the function is convex) which implies that the global minimum of F_n is at most η_n away from $\bar{f}_{\mathbf{J}}$ and thus since $\bar{f}_{\mathbf{J}}$ is $O(\mu_n^{1/2})$ away from $\mathbf{f}_{\mathbf{J}}$, we have the consistency if

$$\eta_n \ll 1 \text{ and } \eta_n \gg \mu_n^{1/2} + n^{-1/2}\mu_n^{-1},$$

which concludes the proof of the lemma. ■

We now prove Theorem 9. Let $\tilde{f}_{\mathbf{J}}$ be defined as in Lemma 16. We extend it by zeros on $\mathbf{J}^c$. We already know the squared norm consistency by Lemma 16. We need to prove that

with probability tending to one $\tilde{f}$ is optimal for problem in Eq. (13). We have by the first optimality condition for $\tilde{f}_{\mathbf{J}}$:

$$\hat{\Sigma}_{X_{\mathbf{J}}Y} - \hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}}\tilde{f}_{\mathbf{J}} = \mu_n \|\tilde{f}\|_d \operatorname{Diag}(d_j/\|\tilde{f}_j\|)\tilde{f}_{\mathbf{J}},$$

where we use the notation $\|f\|_d = \sum_{j=1}^m d_j \|f_j\|$. We thus have by solving for $\tilde{f}_{\mathbf{J}}$ and using $\hat{\Sigma}_{X_{\mathbf{J}}Y} = \hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}}\mathbf{f}_{\mathbf{J}} + \hat{\Sigma}_{X_{\mathbf{J}}\varepsilon}$:

$$\tilde{f}_{\mathbf{J}} = \left(\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D_n\right)^{-1}\left(\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}}\mathbf{f}_{\mathbf{J}} + \hat{\Sigma}_{X_{\mathbf{J}}\varepsilon}\right),$$

with the notation $D_n = \|\tilde{f}\|_d \operatorname{Diag}(d_j/\|\tilde{f}_j\|)$. We can now put that back into $\hat{\Sigma}_{X_{\mathbf{J}^c}Y} - \hat{\Sigma}_{X_{\mathbf{J}^c}X_{\mathbf{J}}}\tilde{f}_{\mathbf{J}}$ and show that this will have small enough norm with probability tending to one. We have for all $i \in \mathbf{J}^c$:

$$\begin{aligned}
\hat{\Sigma}_{X_iY} - \hat{\Sigma}_{X_iX_{\mathbf{J}}}\tilde{f}_{\mathbf{J}} &= \hat{\Sigma}_{X_iY} - \hat{\Sigma}_{X_iX_{\mathbf{J}}}\left(\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D_n\right)^{-1}\left(\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}}\mathbf{f}_{\mathbf{J}} + \hat{\Sigma}_{X_{\mathbf{J}}\varepsilon}\right) \\
&= -\hat{\Sigma}_{X_iX_{\mathbf{J}}}\left(\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D_n\right)^{-1}\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}}\mathbf{f}_{\mathbf{J}} \\
&\quad + \hat{\Sigma}_{X_iY} - \hat{\Sigma}_{X_iX_{\mathbf{J}}}\left(\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D_n\right)^{-1}\hat{\Sigma}_{X_{\mathbf{J}}\varepsilon} \\
&= -\hat{\Sigma}_{X_iX_{\mathbf{J}}}\mathbf{f}_{\mathbf{J}} + \hat{\Sigma}_{X_iX_{\mathbf{J}}}\left(\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D_n\right)^{-1}\mu_n D_n \mathbf{f}_{\mathbf{J}} \\
&\quad + \hat{\Sigma}_{X_iY} - \hat{\Sigma}_{X_iX_{\mathbf{J}}}\left(\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D_n\right)^{-1}\hat{\Sigma}_{X_{\mathbf{J}}\varepsilon} \\
&= \hat{\Sigma}_{X_iX_{\mathbf{J}}}\left(\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D_n\right)^{-1}\mu_n D_n \mathbf{f}_{\mathbf{J}} \\
&\quad + \hat{\Sigma}_{X_i\varepsilon} - \hat{\Sigma}_{X_iX_{\mathbf{J}}}\left(\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D_n\right)^{-1}\hat{\Sigma}_{X_{\mathbf{J}}\varepsilon} \\
&= A_n + B_n.
\end{aligned}$$

The first term A_n (divided by μ_n) is equal to

$$\frac{A_n}{\mu_n} = \hat{\Sigma}_{X_iX_{\mathbf{J}}}\left(\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D_n\right)^{-1} D_n \mathbf{f}_{\mathbf{J}}.$$

We can replace $\hat{\Sigma}_{X_iX_{\mathbf{J}}}$ in $\frac{A_n}{\mu_n}$ by $\Sigma_{X_iX_{\mathbf{J}}}$ at cost $O_p(n^{-1/2}\mu_n^{-1/2})$ because $\langle \mathbf{f}_{\mathbf{J}}, \Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}}^{-1}\mathbf{f}_{\mathbf{J}}\rangle < \infty$ (by Lemma 20). Also, we can replace $\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}}$ in $\frac{A_n}{\mu_n}$ by $\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}}$ at cost $O_p(n^{1/2}\mu_n^{-1})$ as a consequence of Lemma 19. Those two are $o_p(1)$ by assumptions on μ_n. Thus,

$$\frac{A_n}{\mu_n} = \Sigma_{X_iX_{\mathbf{J}}}\left(\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D_n\right)^{-1} D_n \mathbf{f}_{\mathbf{J}} + o_p(1).$$

Furthermore, we let denote $D = \|\mathbf{f}\|_d \operatorname{Diag}(d_j/\|\mathbf{f}_j\|)$. From Lemma 21, we know that $D_n - D = o_p(1)$. Thus we can replace D_n by D at cost $o_p(1)$ to get:

$$\frac{A_n}{\mu_n} = \Sigma_{X_iX_{\mathbf{J}}}\left(\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}} + \mu_n D\right)^{-1} D\mathbf{f}_{\mathbf{J}} + o_p(1) = C_n + o_p(1).$$

We now show that this last deterministic term $C_n \in \mathcal{F}_i$ converges to:

$$C = \Sigma^{1/2}_{X_iX_i} C_{X_iX_\mathbf{J}} C^{-1}_{X_\mathbf{J}X_\mathbf{J}} D\mathbf{g}_\mathbf{J},$$

where, from (**A7**), $\mathbf{f}_j = \Sigma^{1/2}_{X_jX_j}\mathbf{g}_j$. We have

$$\begin{aligned} C_n - C &= \Sigma^{1/2}_{X_iX_i} C_{X_iX_\mathbf{J}} \left[\mathrm{Diag}(\Sigma^{1/2}_{X_jX_j})\,(\Sigma_{X_\mathbf{J}X_\mathbf{J}} + \mu_n D)^{-1}\,\mathrm{Diag}(\Sigma^{1/2}_{X_jX_j}) - C^{-1}_{X_\mathbf{J}X_\mathbf{J}}\right] D\mathbf{g}_\mathbf{J} \\ &= \Sigma^{1/2}_{X_iX_i} C_{X_iX_\mathbf{J}} K_n D\mathbf{g}_\mathbf{J}. \end{aligned}$$

where $K_n = \mathrm{Diag}(\Sigma^{1/2}_{X_jX_j})\,(\Sigma_{X_\mathbf{J}X_\mathbf{J}} + \mu_n D)^{-1}\,\mathrm{Diag}(\Sigma^{1/2}_{X_jX_j}) - C^{-1}_{X_\mathbf{J}X_\mathbf{J}}$. Moreover, we have:

$$\mathrm{Diag}(\Sigma^{1/2}_{X_jX_j}) C_{X_\mathbf{J}X_\mathbf{J}} K_n = -\mu_n D\,(\Sigma_{X_\mathbf{J}X_\mathbf{J}} + \mu_n D)^{-1}\,\mathrm{Diag}(\Sigma^{1/2}_{X_jX_j}).$$

Following Fukumizu et al. (2007), the range of the operator $\left(\Sigma^{1/2}_{X_iX_i} C_{X_iX_\mathbf{J}}\right)^* = C_{X_\mathbf{J}X_i}\Sigma^{1/2}_{X_iX_i}$ is included in the closure of the range of $\mathrm{Diag}(\Sigma_{X_jX_j})$ (which is equal to the range of $\Sigma_{X_\mathbf{J}X_\mathbf{J}}$ by Lemma 20). For any $v_\mathbf{J} \in \mathcal{F}_\mathbf{J}$ in the intersection of two ranges, we have $v_\mathbf{J} = C_{X_\mathbf{J}X_\mathbf{J}}\,\mathrm{Diag}(\Sigma_{X_jX_j})u_\mathbf{J}$ (note that $C_{X_\mathbf{J}X_\mathbf{J}}$ is invertible), and thus

$$\begin{aligned} \langle K_n D\mathbf{g}_\mathbf{J}, v_\mathbf{J}\rangle &= \langle K_n D\mathbf{g}_\mathbf{J}, C_{X_\mathbf{J}X_\mathbf{J}}\,\mathrm{Diag}(\Sigma_{X_jX_j})u_\mathbf{J}\rangle \\ &= \langle -\mu_n D\,(\Sigma_{X_\mathbf{J}X_\mathbf{J}} + \mu_n D)^{-1}\,\mathrm{Diag}(\Sigma^{1/2}_{X_jX_j}) D\mathbf{g}_\mathbf{J}, u_\mathbf{J}\rangle \end{aligned}$$

which $O_p(\mu_n^{1/2})$ and thus tends to zero. Since this holds for all elements in the intersection of the ranges, Lemma 9 by Fukumizu et al. (2007) implies that $\|C_n - C\|$ converges to zero.

We now simply need to show that the second term B_n is dominated by μ_n. We have: $\|\hat{\Sigma}_{X_i\varepsilon}\| = O_p(n^{-1/2})$ and $\|\hat{\Sigma}_{X_iX_\mathbf{J}}\left(\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}} + \mu_n D_n\right)^{-1}\hat{\Sigma}_{X_\mathbf{J}\varepsilon}\| \leqslant \|\hat{\Sigma}_{X_i\varepsilon}\|$, thus, because condition (4) is satisfied and $\mu_n n^{1/2} \to +\infty$, then $B_n = o_p(\mu_n)$ and thus for for each $i \in J^c$,

$$\frac{1}{d_i\mu_n\|\mathbf{f}\|_d}\left(\hat{\Sigma}_{X_iY} - \hat{\Sigma}_{X_iX_\mathbf{J}}\tilde{f}_\mathbf{J}\right)$$

converges in probability to a limit with norm strictly smaller than one. Thus

$$\mathbb{P}\left\{\frac{1}{d_i\mu_n\|\mathbf{f}\|_d}\left\|\hat{\Sigma}_{X_iY} - \hat{\Sigma}_{X_iX_\mathbf{J}}\tilde{f}_\mathbf{J}\right\| \leqslant 1\right\}$$

is tending to 1, which implies the theorem (using the same arguments than in the proof of Theorem 2 in Appendix B.1).

C.3 Proof of Theorem 10

Before proving the analog of the second group-Lasso theorem, we need the following additional proposition, which states that consistency of the patterns can only be achieved if $\mu_n n^{1/2} \to \infty$ (even if chosen in a data dependent way).

Proposition 22 *Assume (**A4**), (**A5**), (**A6**), (**A7**) and that* $\mathbf{J}$ *is not empty. If* $\hat{f}$ *is converging in probability to* $\mathbf{f}$ *and* $J(\hat{f})$ *converges in probability to* $\mathbf{J}$*, then* $\mu_n n^{1/2} \to \infty$ *in probability.*

Proof We give a proof by contradiction, and we thus assume that there exists $M > 0$ such that $\liminf_{n\to\infty} \mathbb{P}(\mu_n n^{1/2} < M) > 0$. This imposes that there exists a subsequence which is almost surely bounded by M (Durrett, 2004). Thus, we can take a further subsequence which converges to a limit $\mu_0 \in [0, \infty)$. We now consider such a subsequence (and still use the notation of the original sequence for simplicity).

With probability tending to one, we have the optimality condition (15):

$$\hat{\Sigma}_{X_{\mathbf{J}}\varepsilon} + \hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}}\mathbf{f}_{\mathbf{J}} = \hat{\Sigma}_{X_{\mathbf{J}}Y} = \hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}}\hat{f}_{\mathbf{J}} + \mu_n \|\hat{f}\|_d \operatorname{Diag}(d_j/\|\hat{f}_j\|)\hat{f}_{\mathbf{J}}.$$

If we let denote $D_n = n^{1/2}\mu_n\|\hat{f}\|_d \operatorname{Diag}(d_j/\|\hat{f}_j\|)$, we get:

$$D_n\mathbf{f}_{\mathbf{J}} = \left[\hat{\Sigma}_{X_{\mathbf{J}}X_{\mathbf{J}}} + D_n n^{-1/2}\right] n^{1/2}\left[\mathbf{f}_{\mathbf{J}} - \hat{f}_{\mathbf{J}}\right] + n^{1/2}\hat{\Sigma}_{X_{\mathbf{J}}\varepsilon},$$

which can be approximated as follows (we denote $D = \|\mathbf{f}\|_d \operatorname{Diag}(d_j/\|\mathbf{f}_j\|)$):

$$\mu_0 D\mathbf{f}_{\mathbf{J}} + o_p(1) = \Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}} n^{1/2}\left[\mathbf{f}_{\mathbf{J}} - \hat{f}_{\mathbf{J}}\right] + o_p(1) + n^{1/2}\hat{\Sigma}_{X_{\mathbf{J}}\varepsilon},$$

We can now write for $i \in \mathbf{J}^c$:

$$\begin{aligned} n^{1/2}\left(\hat{\Sigma}_{X_iY} - \hat{\Sigma}_{X_iX_{\mathbf{J}}}\hat{f}_{\mathbf{J}}\right) &= n^{1/2}\hat{\Sigma}_{X_i\varepsilon} + \hat{\Sigma}_{X_iX_{\mathbf{J}}} n^{1/2}(\mathbf{f}_{\mathbf{J}} - \hat{f}_{\mathbf{J}}) \\ &= n^{1/2}\hat{\Sigma}_{X_i\varepsilon} + \Sigma_{X_iX_{\mathbf{J}}} n^{1/2}(\mathbf{f}_{\mathbf{J}} - \hat{f}_{\mathbf{J}}) + o_p(1). \end{aligned}$$

We now consider an arbitrary vector $w_{\mathbf{J}} \in \mathcal{F}_{\mathbf{J}}$, such that $\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}}w_{\mathbf{J}}$ is different from zero (such vector exists because $\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}} \neq 0$). Since the range of $\Sigma_{X_{\mathbf{J}}X_i}$ is included in the range of $\Sigma_{X_{\mathbf{J}}X_i}$ (Baker, 1973), there exists $v_i \in \mathcal{F}_i$ such that $\Sigma_{X_{\mathbf{J}}X_i}v_i = \Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}}w_{\mathbf{J}}$. Note that since $\Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}}w_{\mathbf{J}}$ is different from zero, we must have $\Sigma^{1/2}_{X_iX_i}v_i \neq 0$. We have:

$$\begin{aligned} n^{1/2}\langle v_i, \hat{\Sigma}_{X_iY} - \hat{\Sigma}_{X_iX_{\mathbf{J}}}\hat{f}_{\mathbf{J}}\rangle &= n^{1/2}\langle v_i, \hat{\Sigma}_{X_i\varepsilon}\rangle + \langle w_{\mathbf{J}}, \Sigma_{X_{\mathbf{J}}X_{\mathbf{J}}} n^{1/2}(\mathbf{f}_{\mathbf{J}} - \hat{f}_{\mathbf{J}}) + o_p(1) \\ &= n^{1/2}\langle v_i, \hat{\Sigma}_{X_i\varepsilon}\rangle + \langle w_{\mathbf{J}}, \mu_0 Df_{\mathbf{J}} - n^{1/2}\hat{\Sigma}_{X_{\mathbf{J}}\varepsilon}\rangle + o_p(1) \\ &= \langle w_{\mathbf{J}}, \mu_0 Df_{\mathbf{J}}\rangle + n^{1/2}\langle v_i, \hat{\Sigma}_{X_i\varepsilon}\rangle - n^{1/2}\langle w_{\mathbf{J}}, \hat{\Sigma}_{X_{\mathbf{J}}\varepsilon}\rangle + o_p(1) \end{aligned}$$

The random variable $E_n = n^{1/2}\langle v_i, \hat{\Sigma}_{X_i\varepsilon}\rangle - n^{1/2}\langle w_{\mathbf{J}}, \hat{\Sigma}_{X_{\mathbf{J}}\varepsilon}\rangle$ is the sum of i.i.d random variables with finite second moment. It is thus asymptotically normal, and we simply need to

compute its mean and variance. The mean is zero and we have:

$$\begin{aligned}
E_n &= \frac{1}{n^{1/2}} \sum_{k=1}^{n} \left(v_i(x_{ki}) - \sum_{j \in \mathbf{J}} w_j(x_{kj}) \right) \varepsilon_k, \\
\mathbb{E}(E_n^2 | \bar{X}) &= \frac{1}{n} \sum_{k=1}^{n} \left(v_i(x_{ki}) - \sum_{j \in \mathbf{J}} w_j(x_{kj}) \right)^2 \mathbb{E}(\varepsilon_k^2 | \bar{X}) \\
&\geqslant \frac{1}{n} \sum_{k=1}^{n} \left(v_i(x_{ki}) - \sum_{j \in \mathbf{J}} w_j(x_{kj}) \right)^2 \sigma_{\min}^2 \\
&= \sigma_{\min}^2 \langle v_i, \hat{\Sigma}_{X_i X_i} v_i \rangle + \sigma_{\min}^2 \langle w_{\mathbf{J}}, \hat{\Sigma}_{X_{\mathbf{J}} X_{\mathbf{J}}} w_{\mathbf{J}} \rangle - 2\sigma_{\min}^2 \langle v_i, \hat{\Sigma}_{X_i X_{\mathbf{J}}} w_{\mathbf{J}} \rangle, \text{ and thus} \\
\mathbb{E} E_n^2 &\geqslant \sigma_{\min}^2 \langle v_i, \Sigma_{X_i X_i} v_i \rangle + \sigma_{\min}^2 \langle w_{\mathbf{J}}, \Sigma_{X_{\mathbf{J}} X_{\mathbf{J}}} w_{\mathbf{J}} \rangle - 2\sigma_{\min}^2 \langle v_i, \Sigma_{X_i X_{\mathbf{J}}} w_{\mathbf{J}} \rangle \\
&= \sigma_{\min}^2 \langle v_i, \Sigma_{X_i X_i} v_i \rangle - \sigma_{\min}^2 \langle v_i, \Sigma_{X_i X_{\mathbf{J}}} w_{\mathbf{J}} \rangle.
\end{aligned}$$

The operator $C_{X_{\mathbf{J}} X_{\mathbf{J}}}^{-1} C_{X_{\mathbf{J}} X_i}$ has the same range as $C_{X_{\mathbf{J}} X_{\mathbf{J}}}$ (because C is invertible), and is thus included in the closure of the range of $\mathrm{Diag}(\Sigma_{X_j X_j}^{1/2})$ (Baker, 1973). Thus for any $u \in \mathcal{F}_i$, $C_{X_{\mathbf{J}} X_{\mathbf{J}}}^{-1} C_{X_{\mathbf{J}} X_i} u$ can be expressed as a limit of terms of the form $\mathrm{Diag}(\Sigma_{X_j X_j}^{1/2}) t$. We thus have that

$$\langle u, C_{X_i X_{\mathbf{J}}} \mathrm{Diag}(\Sigma_{X_j X_j}^{1/2}) w_{\mathbf{J}} \rangle = \langle u, C_{X_i X_{\mathbf{J}}} C_{X_{\mathbf{J}} X_{\mathbf{J}}}^{-1} C_{X_{\mathbf{J}} X_{\mathbf{J}}} \mathrm{Diag}(\Sigma_{X_j X_j}^{1/2}) w_{\mathbf{J}} \rangle$$

can be expressed as a limit of terms of the form

$$\begin{aligned}
\langle t, \mathrm{Diag}(\Sigma_{X_j X_j}^{1/2}) C_{X_{\mathbf{J}} X_{\mathbf{J}}} \mathrm{Diag}(\Sigma_{X_j X_j}^{1/2}) w_{\mathbf{J}} \rangle &= \langle t, \Sigma_{X_{\mathbf{J}} X_{\mathbf{J}}} w_{\mathbf{J}} \rangle = \langle t, \Sigma_{X_{\mathbf{J}} X_i} v_i \rangle \\
&= \langle t, \mathrm{Diag}(\Sigma_{X_j X_j}^{1/2}) C_{X_{\mathbf{J}} X_i} \Sigma_{X_i X_i}^{1/2} v_i \rangle \to \langle u, C_{X_i X_{\mathbf{J}}} C_{X_{\mathbf{J}} X_{\mathbf{J}}}^{-1} C_{X_{\mathbf{J}} X_i} \Sigma_{X_i X_i}^{1/2} v_i \rangle.
\end{aligned}$$

This implies that $C_{X_i X_{\mathbf{J}}} \mathrm{Diag}(\Sigma_{X_j X_j}^{1/2}) w_{\mathbf{J}} = C_{X_i X_{\mathbf{J}}} C_{X_{\mathbf{J}} X_{\mathbf{J}}}^{-1} C_{X_{\mathbf{J}} X_i} \Sigma_{X_i X_i}^{1/2} v_i$, and thus we have:

$$\begin{aligned}
\mathbb{E} E_n^2 &\geqslant \sigma_{\min}^2 \langle v_i, \Sigma_{X_i X_i} v_i \rangle - \sigma_{\min}^2 \langle v_i, \Sigma_{X_i X_i}^{1/2} C_{X_i X_{\mathbf{J}}} \mathrm{Diag}(\Sigma_{X_j X_j}^{1/2}) w_{\mathbf{J}} \rangle \\
&= \sigma_{\min}^2 \langle v_i, \Sigma_{X_i X_i} v_i \rangle - \sigma_{\min}^2 \langle v_i, \Sigma_{X_i X_i}^{1/2} C_{X_i X_{\mathbf{J}}} C_{X_{\mathbf{J}} X_{\mathbf{J}}}^{-1} C_{X_{\mathbf{J}} X_i} \Sigma_{X_i X_i}^{1/2} v_i \rangle \\
&= \sigma_{\min}^2 \langle \Sigma_{X_i X_i}^{1/2} v_i, (I - C_{X_i X_{\mathbf{J}}} C_{X_{\mathbf{J}} X_{\mathbf{J}}}^{-1} C_{X_{\mathbf{J}} X_i}) \Sigma_{X_i X_i}^{1/2} v_i \rangle.
\end{aligned}$$

By assumption (**A5**), the operator $I - C_{X_i X_{\mathbf{J}}} C_{X_{\mathbf{J}} X_{\mathbf{J}}}^{-1} C_{X_{\mathbf{J}} X_i}$ is lower bounded by a strictly positive constant times the identity matrix, and thus, since $\Sigma_{X_i X_i}^{1/2} v_i \neq 0$, we have $\mathbb{E} E_n^2 > 0$. This implies that $n^{1/2} \langle v_i, \hat{\Sigma}_{X_i Y} - \hat{\Sigma}_{X_i X_{\mathbf{J}}} \hat{f}_{\mathbf{J}} \rangle$ converges to a normal distribution with strictly positive variance. Thus the probability $\mathbb{P}\left(n^{1/2} \langle v_i, \hat{\Sigma}_{X_i Y} - \hat{\Sigma}_{X_i X_{\mathbf{J}}} \hat{f}_{\mathbf{J}} \rangle \geqslant d_i \|\hat{f}\|_d \|v_i\| + 1 \right)$ converges to a strictly positive limit (note that $\|\hat{f}\|_d$ can be replaced by $\|\mathbf{f}\|_d$ without changing the result). Since $\mu_n n^{1/2} \to \mu_0 < \infty$, this implies that

$$\mathbb{P}\left(\mu_n^{-1} \langle v_i, \hat{\Sigma}_{X_i Y} - \hat{\Sigma}_{X_i X_{\mathbf{J}}} \hat{f}_{\mathbf{J}} \rangle > d_i \|\hat{f}\|_d \|v_i\| \right)$$

is asymptotically strictly positive (i.e. has a strictly positive $\liminf$). Thus the optimality condition (14) is not satisfied with non vanishing probability, which is a contradiction and proves the proposition.

■

We now go back to the proof of Theorem 10. We prove by contradiction, by assuming that there exists $i \in \mathbf{J}^c$ such that

$$\frac{1}{d_i}\left\|\Sigma_{X_iX_i}^{1/2} C_{X_iX_\mathbf{J}} C_{X_\mathbf{J}X_\mathbf{J}}^{-1} \operatorname{Diag}(d_j/\|\mathbf{f}_j\|)\mathbf{g}_\mathbf{J}\right\| > 1.$$

Since with probability tending to one $J(\hat{f}) = \mathbf{J}$, with probability tending to one, we have from optimality condition (15), and the usual line of arguments:

$$\begin{aligned}\hat{\Sigma}_{X_iY} - \hat{\Sigma}_{X_iX_\mathbf{J}}\hat{f}_\mathbf{J} &= \mu_n\hat{\Sigma}_{X_iX_\mathbf{J}}\left(\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}} + \mu_n\|\hat{f}\|_d \operatorname{Diag}(d_j/\|\hat{f}_j\|)\right)^{-1}\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}}\mathbf{f} \\ &\quad + \hat{\Sigma}_{X_i\varepsilon} - \hat{\Sigma}_{X_iX_\mathbf{J}}\left(\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}} + \|\hat{f}\|_d\mu_n \operatorname{Diag}(d_j/\|\hat{f}_j\|)\right)^{-1}\hat{\Sigma}_{X_\mathbf{J}\varepsilon}.\end{aligned}$$

Following the same argument as in the proof of Theorem 9, (and because $\mu_n n^{1/2} \to +\infty$ as a consequence of Proposition 22), the first term in the last expression (divided by μ_n) converges to

$$v_i = \Sigma_{X_iX_i}^{1/2} C_{X_iX_\mathbf{J}} C_{X_\mathbf{J}X_\mathbf{J}}^{-1}\|\mathbf{f}\|_d \operatorname{Diag}(d_j/\|\mathbf{f}_j\|)\mathbf{g}_\mathbf{J}$$

By assumption $\|v_i\| > d_i\|\mathbf{f}\|_f$. We have the second term:

$$\begin{aligned}&\hat{\Sigma}_{X_i\varepsilon} - \hat{\Sigma}_{X_iX_\mathbf{J}}\left(\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}} + \mu_n\|\hat{f}\|_d \operatorname{Diag}(d_j/\|\hat{f}_j\|)\right)^{-1}\hat{\Sigma}_{X_\mathbf{J}\varepsilon} \\ =\ & O_p(n^{-1/2}) - \hat{\Sigma}_{X_iX_\mathbf{J}}\left(\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}} + \mu_n\|\mathbf{f}\|_d \operatorname{Diag}(d_j/\|\mathbf{f}_j\|)\right)^{-1}\hat{\Sigma}_{X_\mathbf{J}\varepsilon} + O_p(n^{-1/2}).\end{aligned}$$

The remaining term can be bounded as follows:

$$\begin{aligned}&\mathbb{E}\left(\left\|\hat{\Sigma}_{X_iX_\mathbf{J}}\left(\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}} + \mu_n\|\mathbf{f}\|_d \operatorname{Diag}(d_j/\|\mathbf{f}_j\|)\right)^{-1}\hat{\Sigma}_{X_\mathbf{J}\varepsilon}\right\|^2 |\bar{X}\right) \\ \leqslant\ & \frac{\sigma_{\max}^2}{n}\left\|\hat{\Sigma}_{X_iX_\mathbf{J}}\left(\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}} + \mu_n\|\mathbf{f}\|_d \operatorname{Diag}(d_j/\|\mathbf{f}_j\|)\right)^{-1}\hat{\Sigma}_{X_\mathbf{J}X_\mathbf{J}}^{1/2}\right\|^2 \\ \leqslant\ & \frac{\sigma_{\max}^2}{n}\operatorname{tr}\hat{\Sigma}_{X_iX_i},\end{aligned}$$

which implies that the full expectation is $O(n^{-1})$ (because our operators are trace-class, i.e. have finite trace). Thus the remaining term is $O_p(n^{-1/2})$ and thus negligible compared to μ_n, therefore $\frac{1}{\mu_n\|\hat{f}\|_d}\left(\hat{\Sigma}_{X_iY} - \hat{\Sigma}_{X_iX_\mathbf{J}}\hat{f}_\mathbf{J}\right)$ converges in probability to a limit which is of norm strictly greater than d_i. Thus there is a non vanishing probability of being strictly larger than d_i, which implies that with non vanishing probability, the optimality condition (14) is not satisfied, which is a contradiction. This concludes the proof.

C.4 Proof of Proposition 12

Note that the estimator defined in Eq. (21) is exactly equal to

$$\left\| \hat{\Sigma}_{X_i X_{\mathbf{J}}} (\hat{\Sigma}_{X_{\mathbf{J}} X_{\mathbf{J}}} + \kappa_n I)^{-1} \operatorname{Diag}(d_j / \|(\hat{f}^{LS}_{\kappa_n})_j\|)(\hat{f}^{LS}_{\kappa_n})_{\mathbf{J}} \right\| .$$

Using Theorem 13 and the arguments from Appendix C.2, we get the consistency result.

C.5 Range condition of covariance operators

We let denote $C(q)$ the convolution operator by q on the space of real functions on $\mathbb{R}^p$ and $T(p)$ the pointwise multiplication by $p(x)$. In this appendix, since we look at different Hilbertian products of functions on $\mathbb{R}^p$, we use the notations $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ and $\langle \cdot, \cdot \rangle_{L^2(p_X)}$ and $\langle \cdot, \cdot \rangle_{L^2(\mathbb{R}^p)}$ for the dot products in the RKHS $\mathcal{F}$, the space $L^2(p_X)$ of square integrable functions with respect to $p(x)dx$, and the space $L^2(\mathbb{R}^p)$ of square integrable functions with respect to the Lebesgue measure. With our assumptions, for all $\tilde{f}, \tilde{g} \in L^2(\mathbb{R}^p)$, we have

$$\langle \tilde{f}, \tilde{g} \rangle_{L^2} = \langle C(q)^{1/2} \tilde{f}, C(q)^{1/2} \tilde{g} \rangle_{\mathcal{F}}.$$

Denote by $\{\lambda_k\}_{k \geq 1}$ and $\{e_k\}_{k \geq 1}$ the eigenvalues and the eigenvectors of the covariance operator Σ_{XX}, respectively. Since $p_X(x)$ was assumed to be strictly positive, all eigenvalues are strictly positive. For $k \geqslant 1$, set $f_k = \lambda_k^{-1/2} e_k$. By construction, for any $k, \ell \geqslant 1$,

$$\begin{aligned} \lambda_k \delta_{k,\ell} = \langle e_k, \Sigma e_\ell \rangle_{\mathcal{F}} &= \int p(x) e_k(x) e_\ell(x) dx \\ &= \lambda_k^{1/2} \lambda_\ell^{1/2} \int p_X(x) f_k(x) f_\ell(x) dx = \lambda_k^{1/2} \lambda_\ell^{1/2} \langle f_k, f_\ell \rangle_{L^2(p_X)} . \end{aligned}$$

Thus $\{f_k\}_{k \geqslant 1}$ is an orthonormal sequence in $L^2(p_X)$. Let $f = C(q)g$ for $g \in L^2(\mathbb{R}^p)$. Note that f is in the range of $\Sigma_{XX}^{1/2}$ if and only if $\left\langle f, \Sigma^{-1} f \right\rangle$ is finite. We have:

$$\begin{aligned} \left\langle f, \Sigma^{-1} f \right\rangle = \sum_{p=1}^{\infty} \lambda_p^{-1} \langle e_p, f \rangle_{\mathcal{F}}^2 = \sum_{p=1}^{\infty} \lambda_p^{-1} \langle e_p, g \rangle_{L^2(\mathbb{R}^p)}^2 &= \sum_{p=1}^{\infty} \lambda_p^{-1} \left(\int g(x) e_p(x) dx \right)^2 \\ &= \sum_{p=1}^{\infty} \left\langle p_X^{-1} g, f_p \right\rangle_{L^2(p_X)}^2 \leqslant \|p_X^{-1} g\|_{L^2(p_X)}^2 = \int \frac{g^2(x)}{p_X(x)} dx, \end{aligned}$$

which concludes the proof.

C.6 Proof of Theorem 13

We have:

$$\hat{f}^{LS}_{\kappa_n} = \left(\hat{\Sigma}_{XX} + \kappa_n I \right)^{-1} \hat{\Sigma}_{XY}$$

and thus:

$$\begin{aligned} \hat{f}^{LS}_{\kappa_n} - \mathbf{f} &= \left(\hat{\Sigma}_{XX} + \kappa_n I \right)^{-1} \hat{\Sigma}_{XX} \mathbf{f} - \mathbf{f} + \left(\hat{\Sigma}_{XX} + \kappa_n I \right)^{-1} \hat{\Sigma}_{X\varepsilon} \\ &= (\Sigma_{XX} + \kappa_n I)^{-1} \Sigma_{XX} \mathbf{f} - \mathbf{f} + O_p(n^{-1/2} \kappa_n) \text{ from Lemma 19} \\ &= -(\Sigma_{XX} + \kappa_n I)^{-1} \kappa_n \mathbf{f} + O_p(n^{-1/2} \kappa_n). \end{aligned}$$

Since $\mathbf{f} = \Sigma_{XX}^{1/2} \mathbf{g}$, we have $\| -(\Sigma_{XX} + \kappa_n I)^{-1} \kappa_n \mathbf{f} \|^2 \leqslant C \kappa_n \|\mathbf{g}\|^2$, which concludes the proof.

C.7 Proof of Theorem 14

We define $\tilde{f}$ as the minimizer of the same cost function restricted to $f_{\mathbf{J}^c} = 0$. Because $\hat{f}^{LS}_{n^{-1/3}}$ is consistent, the norms of $(\hat{f}^{LS}_{n^{-1/3}})_j$ for $j \in \mathbf{J}$ are bounded away from zero, and Lemma 21 applies with $\mu_n = \mu_0 n^{-1/3}$, i.e., $\tilde{f}$ converges in probability to $\mathbf{f}$ and so are the patterns of zeros (which is obvious by construction of $\tilde{f}$). Moreover, for any $\eta > 0$, from Lemma 21, we have $\|\tilde{f}_{\mathbf{J}} - f_{\mathbf{J}}\| = O_p(n^{-1/6+\eta})$ (because $\mu_n^{-1/2} + n^{-1/2}\mu_n^{-1} = O_p(n^{-1/6})$).

What remains to be shown is that with probability tending to one, $\tilde{f}$ is optimal for the full problem. We just need to show that with probability tending to one, for all $i \in \mathbf{J}^c$,

$$\|\hat{\Sigma}_{X_i\varepsilon} - \hat{\Sigma}_{X_i X_{\mathbf{J}}}(\tilde{f}_{\mathbf{J}} - f_{\mathbf{J}})\| \leqslant \mu_n \|\tilde{f}\|_d \|(\hat{f}^{LS}_{n^{-1/3}})_i\|^{-\gamma}. \tag{38}$$

Note that $\|\tilde{f}\|_d$ converges in probability to $\|\mathbf{f}\|_d > 0$. Moreover, by Theorem 13, $\|(\hat{f}^{LS}_{n^{-1/3}})_i - \mathbf{f}_i\| = O_p(n^{-1/6})$. Thus, if $i \in \mathbf{J}^c$, i.e., if $\mathbf{f}_i = 0$, then $\|(\hat{f}^{LS}_{n^{-1/3}})_i\| = O_p(n^{-1/6})$. The left hand side in Eq. (38) is thus upper bounded by $O_p(n^{-1/2} + n^{-1/6+\eta})$ while the right hand side is lower bounded asymptotically by $n^{-1/3}n^{\gamma/6}$. Thus if $-1/6 + \eta < -1/3 + \gamma/6$, then with probability tending to one we get the correct optimality condition. As soon as $\gamma > 1$, we can find η small enough and strictly positive, which concludes the proof.

References

F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the 21st International Conference on Machine Learning*, 2004a.

F. R. Bach, R. Thibaux, and M. I. Jordan. Computing regularization paths for learning multiple kernels. In *Advances in Neural Information Processing Systems 17*, 2004b.

C. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2003.

O. Bousquet and D. J. L. Herrmann. On the complexity of learning the kernel matrix. In *Advances in Neural Information Processing Systems 17*, 2003.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge Univ. Press, 2003.

P. Brémaud. *Markov chains, Gibbs fields, Monte Carlo simulation, and queues*. Springer-Verlag, New York, 1999.

H. Brezis. *Analyse Fonctionelle*. Paris: Masson, 1980.

A. Caponnetto and E. de Vito. Fast rates for regularized least-squares algorithm. Technical Report 248/AI Memo 2005-013, CBCL, Massachusetts Institute of Technology, 2005.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1), 2002.

R. Durrett. *Probability: theory and examples.* Duxbury Press, third edition, 2004.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407, 2004.

W. Fu and K. Knight. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28(5): 1356–1378, 2000.

K. Fukumizu, F. Bach, and A. Gretton. Statistical convergence of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(8), 2007.

K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schlkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 12 2005.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer-Verlag, 2001.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models.* Chapman & Hall, 1990.

A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric regression. *Annals of Statistics*, 28(3):681–712, 2000.

G. R. G. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble. A statistical framework for genomic data fusion. *Bioinf.*, 20:2626–2635, 2004a.

G. R. G. Lanckriet, N. Cristianini, L. El Ghaoui, P. Bartlett, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5: 27–72, 2004b.

M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lébret. Applications of second-order cone programming. *Linear Algebra and its Applications*, 284:193–228, 1998.

L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. Technical Report 131, Eidgenöossische Technische Hochschule (ETH), Zürich, Switzerland, 2006.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. Technical Report 720, Departement of Statisics, UC Berkeley, 2006.

M. R. Osborne, B. Presnell, and B. A. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000.

M. Pontil, A. Argyriou, and T. Evgeniou. Multi-task feature learning. In *Advances in Neural Information Processing Systems*, 2007.

M. Pontil and C.A. Micchelli. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6:1099–1125, 2005.

A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet. More efficiency in multiple kernel learning. In *Proceedings of the Twenty-fourth International Conference on Machine Learning*, 2007.

A. Renyi. On Measures of Dependence. *Acta Mathematica Academy Sciences Hungary*, 10: 441–451, 1959.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2001.

S. Sonnenburg, G. Rtsch, C. Schfer, and B. Schlkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565, 07 2006.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal Royal Statististics*, 58(1):267–288, 1994.

A. N. Tikhonov and V. Y. Arsenin. *Solutions of ill-posed problems*. V. H. Winston and Sons, 1997.

A. W. Van der Vaart. *Asymptotic Statistics*. Cambridge Univ. Press, 1998.

G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

Q. Wu, Y. Ying, and D.-X. Zhou. Multi-kernel regularized classifiers. *Journal of Complexity*, 23(1):108–134, 2007.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, 68(1):49–67, 2006.

M. Yuan and Y. Lin. On the non-negative garrotte estimator. *Journal of The Royal Statistical Society Series B*, 69(2):143–161, 2007.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, December 2006.