ESAIM: PS 25 (2021) 298–324
https://doi.org/10.1051/ps/2021010

ESAIM: Probability and Statistics
www.esaim-ps.org

A PARTIAL GRAPHICAL MODEL WITH A STRUCTURAL PRIOR ON THE DIRECT LINKS BETWEEN PREDICTORS AND RESPONSES

Eunice Okome Obiang[1], Pascal Jézéquel[2,3,4] and Frédéric Proïa[1,*]

Abstract. This paper is devoted to the estimation of a partial graphical model with a structural Bayesian penalization. Precisely, we are interested in the linear regression setting where the estimation is made through the direct links between potentially high-dimensional predictors and multiple responses, since it is known that Gaussian graphical models enable to exhibit direct links only, whereas coefficients in linear regressions contain both direct and indirect relations (due *e.g.* to strong correlations among the variables). A smooth penalty reflecting a generalized Gaussian Bayesian prior on the covariates is added, either enforcing patterns (like row structures) in the direct links or regulating the joint influence of predictors. We give a theoretical guarantee for our method, taking the form of an upper bound on the estimation error arising with high probability, provided that the model is suitably regularized. Empirical studies on synthetic data and a real dataset are conducted.

Mathematics Subject Classification. 62A09, 62F30, 62J05.

Received September 17, 2020. Accepted June 7, 2021.

1. Introduction

We are interested in the recovery and estimation of direct links between high-dimensional predictors and a set of responses. Whereas the graphical models seem a natural way to go, we propose to take account of a prior knowledge on the predictors, when possible. This is typically the case when dealing with genetic markers whose joint influence may be anticipated thanks to some kind of genetic distance, or when the predictors are supposed to represent a continuous phenomenon so that consecutive covariates probably act together. In this regard, while taking up the graphical approach, we introduce some Bayesian information in a structural regularization of the estimation procedure, although the inference remains frequentist, thereby following the idea of Chiquet *et al.* [7]. This strategy also enables to affect the amount of shrinkage by playing with some hyperparametrization in the prior, while sparsity may be obtained *via* usual penalty-based patterns. Regarding the mathematical formalization of the graphical models that we will just briefly discuss in this introduction, we refer the reader

Keywords and phrases: High-dimensional linear regression, partial graphical model, structural penalization, sparsity, convex optimization.

[1] Univ Angers, CNRS, LAREMA, SFR MATHSTIC, 49000 Angers, France.
[2] Unité de Bioinformatique, Institut de Cancérologie de l'Ouest, Bd Jacques Monod, 44805 Saint Herblain Cedex, France.
[3] SIRIC ILIAD, Nantes, Angers, France.
[4] CRCINA, INSERM, CNRS, University of Nantes, University of Angers, Health Research Institute-University of Nantes, 8 Quai Moncousu - BP 70721, 44007, Nantes Cedex 1, France.

* Corresponding author: frederic.proia@univ-angers.fr

© The authors. Published by EDP Sciences, SMAI 2021

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

to the very complete handbook recently edited by Maathuis *et al.* [16]. We also refer the reader to the book of Hastie *et al.* [11] and to the one of Giraud [10], both related to the standard high-dimensional statistical methods. Before introducing the model and the organization of this work, let us describe the notation used throughout the paper.

1.1. Notation

For any matrix A, $|A|_* = \|\mathrm{vec}(A)\|_*$ is the elementwise ℓ_* norm of A and $|A|_*^-$ is $|A|_*$ deprived of the diagonal terms of A. We also note $\|A\|_F = |A|_2$ the Frobenius norm of A and $\|A\|_2$ the spectral norm of A. The Frobenius inner product between any matrices A and B of same dimensions is $\langle\langle A, B\rangle\rangle = \langle \mathrm{vec}(A), \mathrm{vec}(B)\rangle = \mathrm{tr}(A^t\, B)$ whereas $\langle u, v\rangle = u^t\, v$ is the inner product of the Euclidean real space. For any vector u, $|u|_0$ is the number of non-zero values in u. For a matrix A, $[A]_C$ is to be understood as the matrix A whose elements outside of the set of coordinates C are set to zero and $\mathrm{vec}(A)$ is the vectorization of A into a column vector. The eigenvalues of a square matrix A of size d with spectrum $\mathrm{sp}(A)$ are $\lambda_i(A)$ taken in decreasing order (from $\lambda_1(A) = \lambda_{\max}(A)$ to $\lambda_d(A) = \lambda_{\min}(A)$). The cones of symmetric positive semi-definite and definite matrices of dimension d are $\mathbb{S}_+^d$ and $\mathbb{S}_{++}^d$ respectively.

1.2. The partial graphical model

In the classic Gaussian graphical model (GGM) setting, we aim at estimating the precision matrix $\Omega = \Sigma^{-1}$ of jointly normally distributed random vectors $Y \in \mathbb{R}^q$ and $X \in \mathbb{R}^p$ with zero mean and covariance Σ. The point is that it induces a graphical structure among the variables and the support of Ω is closely related to the conditional interdependences between them. Let us consider, now and in all the study, the sample covariances of n independent observations (Y_i, X_i), denoted by

$$S_{yy}^{(n)} = \frac{1}{n}\sum_{i=1}^n Y_i\, Y_i^t, \qquad S_{yx}^{(n)} = \frac{1}{n}\sum_{i=1}^n Y_i\, X_i^t \qquad \text{and} \qquad S_{xx}^{(n)} = \frac{1}{n}\sum_{i=1}^n X_i\, X_i^t. \tag{1.1}$$

Maximizing the penalized likelihood of a GGM boils down to finding $\Omega \in \mathbb{S}_{++}^{p+q}$ that minimizes the convex objective

$$L_n(\Omega) = -\ln\det(\Omega) + \langle\langle S^{(n)}, \Omega\rangle\rangle + \lambda\,\mathrm{pen}(\Omega) \tag{1.2}$$

where $S^{(n)}$ is the full sample covariance built from the blocks (1.1). The penalty function $\mathrm{pen}(\Omega)$ is usually $|\Omega|_1$ or even $|\Omega|_1^-$. Efficient algorithms exist to get solutions for (1.2), see *e.g.* Banerjee *et al.* [2], Yuan and Lin [28], Lu [15] or the graphical Lasso of Friedman *et al.* [9]. The reader may also look at the theoretical guarantees of Ravikumar *et al.* [21]. However, thinking of X_i as a predictor of size p associated with a response Y_i of size q, the partial Gaussian graphical model (PGGM), developped *e.g.* by Sohn and Kim [26] or Yuan and Zhang [29], appears as a powerful tool to exhibit direct relationships between the predictors and the responses. To understand this, consider the decomposition into blocks

$$\Omega = \begin{pmatrix} \Omega_{yy} & \Omega_{yx} \\ \Omega_{yx}^t & \Omega_{xx} \end{pmatrix} \qquad \text{and} \qquad \Sigma = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{yx}^t & \Sigma_{xx} \end{pmatrix}$$

where $\Omega_{yy} \in \mathbb{S}_{++}^q$, $\Omega_{yx} \in \mathbb{R}^{q\times p}$ and $\Omega_{xx} \in \mathbb{S}_{++}^p$ and where the same goes for Σ_{xx}, Σ_{yx} and Σ_{xx}. The precision matrix $\Omega = \Sigma^{-1}$ satisfies, by blockwise inversion,

$$\Omega_{yy}^{-1} = \Sigma_{yy} - \Sigma_{yx}\,\Sigma_{xx}^{-1}\,\Sigma_{yx}^t \qquad \text{and} \qquad \Omega_{yx} = -(\Sigma_{yy} - \Sigma_{yx}\,\Sigma_{xx}^{-1}\,\Sigma_{yx}^t)^{-1}\,\Sigma_{yx}\,\Sigma_{xx}^{-1}. \tag{1.3}$$

The conditional distribution peculiar to Gaussian vectors

$$Y_i \,|\, X_i \;\sim\; \mathcal{N}(-\Omega_{yy}^{-1}\,\Omega_{yx}\,X_i,\,\Omega_{yy}^{-1})$$

gives a new light on the multiple-output regression $Y_i = B^{\,t}\,X_i + E_i$ with Gaussian noise $E_i \sim \mathcal{N}(0, R)$, through the reparametrization $B = -\Omega_{yx}^{t}\,\Omega_{yy}^{-1}$ and $R = \Omega_{yy}^{-1}$. Whereas B contains direct and indirect links between the predictors and the responses (due *e.g.* to strong correlations among the variables), Ω_{yx} only contains direct links, as it is shown by the graphical models theory. In other words, the direct links are closely related to the concept of partial correlations between X and Y (see Meinshausen and Bülmann [17] or Peng *et al.* [19], for the univariate case). For example, the direct link between predictor k and response ℓ may be evaluated through the partial correlation $\mathbb{C}\mathrm{orr}(Y_\ell, X_k \,|\, Y_{\neq\,\ell}, X_{\neq\,k})$ contained, apart from a multiplicative coefficient, in the ℓ-th row and k-th column of Ω_{yx} (see *e.g.* Cor. A.6 in [10]) with the particularly interesting consequence that the support of Ω_{yx} is sufficient to identify direct relationships between X and Y. Hence, in the partial setting, the objective reduces to the estimation of the direct links Ω_{yx} together with the conditional precision matrix of the responses Ω_{yy}. Maximizing the penalized conditional log-likelihood of the model now comes down to minimizing the new convex objective

$$\begin{aligned} L_n(\Omega_{yy}, \Omega_{yx}) \;=\; & -\ln\det(\Omega_{yy}) + \langle\!\langle S_{yy}^{(n)}, \Omega_{yy}\rangle\!\rangle + 2\,\langle\!\langle S_{yx}^{(n)}, \Omega_{yx}\rangle\!\rangle \\ & + \langle\!\langle S_{xx}^{(n)}, \Omega_{yx}^{t}\,\Omega_{yy}^{-1}\,\Omega_{yx}\rangle\!\rangle + \lambda\,\mathrm{pen}(\Omega_{yy}) + \mu\,\mathrm{pen}(\Omega_{yx}) \end{aligned} \tag{1.4}$$

over $(\Omega_{yy}, \Omega_{yx}) \in \mathbb{S}_{++}^{q} \times \mathbb{R}^{q\times p}$ for some usual penalty functions. It is worth noting that $\mathrm{pen}(\Omega_{yx})$ often plays a crucial role in modern statistics dealing with high-dimensional predictors (and the natural choice is $|\Omega_{yx}|_1$ to get sparsity) while we may choose $\lambda = 0$, because the number of responses is generally small. In the seminal papers [26, 29], the authors consider $|\Omega_{yy}|_1$ and $|\Omega_{yy}|_1^{-}$ for $\mathrm{pen}(\Omega_{yy})$, respectively. Yuan and Zhang [29] also point out that no estimation of Ω_{xx} is needed anymore. In a graphical model, the estimation of Ω_{yx} and Ω_{yy} depends on the accuracy of the estimation of Ω which, in turn, is strongly affected by the one of Ω_{xx}, especially in a high-dimensional setting. The partial model overrides this issue, the focus is on Ω_{yx} and Ω_{yy} while Ω_{xx} has disappeared from the objective function (1.4). The latter is obtained either by considering the multiple-output Gaussian regression scheme, or, as it is done in [29], by eliminating Ω_{xx} thanks to a first optimization step in (1.2). In this paper, we will consider the penalties

$$\mathrm{pen}(\Omega_{yy}) = |\Omega_{yy}|_1^{-} \quad \text{and} \quad \mathrm{pen}(\Omega_{yx}) = |\Omega_{yx}|_1 \tag{1.5}$$

which correspond to the PGGM (Gm) of [29]. The Spring (Spr) of [7] can also be seen as a PGGM but with no penalty on Ω_{yy} (replaced with an additional structuring one on Ω_{yx}, we will come back to this point thereafter), so for Spr we may consider $\lambda = 0$. The generalized procedure (GenGm) at the heart of the study relies on a combination between these two approaches. We will see in due time that we keep both the penalties of Gm and the structuring one of Spr on Ω_{yx}. Finally, the intermediate solution consisting in estimating Ω_{yy} and B through the conditional distribution $Y_i \,|\, X_i \sim \mathcal{N}(B^{\,t}\,X_i,\,\Omega_{yy}^{-1})$ with penalizations both on B and Ω_{yy}, presented and analyzed by Rothman *et al.* [23] and by Lee and Liu [14], is better known as a multivariate regression with covariance estimation (MRCE). However, it has been shown that the objective function suffers from a lack of convexity and that the optimization procedure may be debatable, in addition to the less convenient setup for statistical interpretation (B contains both direct and indirect influences) compared to PGGM. Without claiming to be exhaustive, let us conclude this quick introduction by citing some related works, like the structural generalization of the Elastic-Net of Slawski *et al.* [25], the Dantzig approach of Cai *et al.* [6] put in practice on genomic data [5], the greedy research of the non-zero pattern in Ω of Johnson *et al.* [13], the approach of Fan *et al.* [8] using a non-convex SCAD penalty to reduce the bias of the Lasso in the estimation of Ω, the eQTL data analysis of Yin and Li [27] which makes use of a sparse conditional GGM, and so on. All the references inside will complete this concise list.

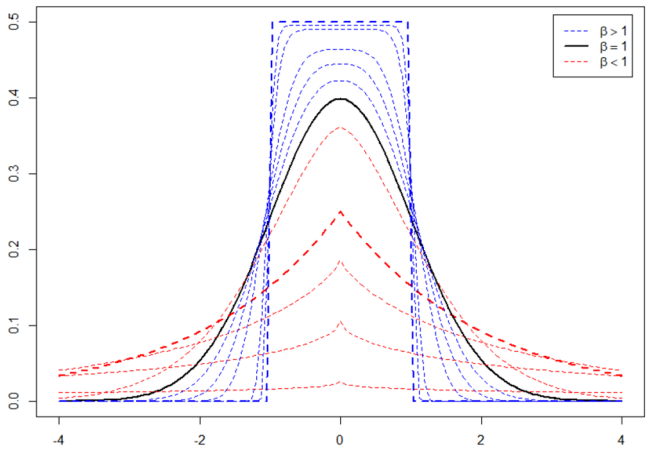

FIGURE 1. Marginal shape of the generalized Gaussian distribution ($d = 1$ and $V = 1$) for some $\beta < 1$ (dotted red), $\beta = 1$ (black) and some $\beta > 1$ (dotted blue). The noteworthy cases $\beta = 1/2$ (Laplace), $\beta = 1$ (Gaussian) and $\beta = +\infty$ (uniform) are highlighted.

1.3. Organization of the paper

To sum up, we have two goals in this paper:

1. Give some theoretical guarantees to the (slightly modified) model introduced in Chiquet *et al.* [7].
2. Generalize the result of Yuan and Zhang [29] to the case where a structural penalization is added in the estimation step.

In Section 2, we introduce the model, consisting in putting a generalized Gaussian prior on the direct links before the procedure of estimation of Ω_{yy} and Ω_{yx}, and we detail the new convex objective. Then we provide some error bounds for our estimates, useful as theoretical guarantees of performance. Section 3 is devoted to empirical considerations. We explain how we deal with the minimization of the new objective and we test the method on simulations first, and next on a real dataset (a Canadian average annual weather cycle, see *e.g.* [20]). After a short conclusion in Section 4, we finally prove our results in Section 5. The numerous constants appearing in the results and the proofs are gathered in the Appendix, for the sake of readability.

2. A generalized Gaussian prior on the direct links

We use the definition given in formulas (1)–(2) of [18] for the so-called d-dimensional multivariate generalized Gaussian $\mathcal{GN}(0, 1, V, \beta)$ distribution with mean 0, scale 1, scatter parameter $V \in \mathbb{S}^d_{++}$ and shape parameter $\beta > 0$. According to the authors, the density takes the form of

$$\forall\, z \in \mathbb{R}^d, \qquad f_{V,\,\beta}(z) = \frac{\beta\,\Gamma(\frac{d}{2})}{\pi^{\frac{d}{2}}\,\Gamma(\frac{d}{2\beta})\,2^{\frac{d}{2\beta}}\sqrt{\det(V)}}\,\exp\left(-\frac{\langle z, V^{-1}z\rangle^{\beta}}{2}\right)$$

where Γ is the Euler Gamma function.

We clearly recognize the Gaussian $\mathcal{N}(0, V)$ setting for $\beta = 1$. Moreover, for $\beta = 1/2$, it can be seen as a multivariate Laplace distribution whereas it is known to converge to some uniform distribution as $\beta \to +\infty$. The marginal shapes ($d = 1$ and $V = 1$) of the distribution are represented in Figure 1, depending on whether $\beta < 1$, $\beta = 1$ or $\beta > 1$. Our results hold for all $\beta \geqslant 1$ but, as will be explained in due course, we shall not theoretically deviate too much from the Gaussianity in the prior (even if we will allow ourselves some exceptions in the practical works). The usual Bayesian approach for multiple-output Gaussian regression having B as matrix of

coefficients and R as noise variance consists in a conjugate prior $\mathrm{vec}(B) \sim \mathcal{N}(b, R \otimes L^{-1})$ for some information matrix $L \in \mathbb{S}^p_{++}$ and a centering value b (see *e.g.* Sect. 2.8.5 of [22]). In the PGGM reformulation, we have $R = \Omega_{yy}^{-1}$ and $B = -\Omega_{yx}^t\,\Omega_{yy}^{-1}$ as explained in Section 1, and of course we shall choose $b = 0$ to meet our purposes. Thus,

$$\mathrm{vec}(\Omega_{yx}^t) = -(\Omega_{yy} \otimes I_p)\,\mathrm{vec}(B) \;\sim\; \mathcal{N}(0, \Omega_{yy} \otimes L^{-1})$$

is a natural prior for the direct links (this is in particular the choice of the authors of [7]). Following the same logic, let us choose $\Omega_{yy} \otimes L^{-1}$ for scatter parameter and suppose that

$$\mathrm{vec}(\Omega_{yx}^t) \;\sim\; \mathcal{GN}(0, 1, \Omega_{yy} \otimes L^{-1}, \beta). \tag{2.1}$$

In this way, we can play on the intensity of the constraint we want to bring on Ω_{yx}, from a non-informative prior to quasi-boundedness through Laplace and Gaussian distributions. This prior entails an additional smooth term acting as a structural penalization in the objective (1.4) that becomes

$$\begin{aligned} L_n(\Omega_{yy}, \Omega_{yx}) = &-\ln\det(\Omega_{yy}) + \langle\!\langle S_{yy}^{(n)}, \Omega_{yy}\rangle\!\rangle + 2\,\langle\!\langle S_{yx}^{(n)}, \Omega_{yx}\rangle\!\rangle \\ &+\langle\!\langle S_{xx}^{(n)}, \Omega_{yx}^t\,\Omega_{yy}^{-1}\,\Omega_{yx}\rangle\!\rangle + \eta\,\langle\!\langle L, \Omega_{yx}^t\,\Omega_{yy}^{-1}\,\Omega_{yx}\rangle\!\rangle^{\beta} + \lambda\,|\Omega_{yy}|_1^- + \mu\,|\Omega_{yx}|_1 \end{aligned} \tag{2.2}$$

with three regularization parameters (λ, μ, η). The smooth penalization lends weight to the prior on Ω_{yx} and thereby plays on the extent of shrinkage and structuring through β, whereas $|\Omega_{yx}|_1$ and $|\Omega_{yy}|_1^-$ are designed to induce sparsity. One can note that this is closely related to the log-likelihood of a hierarchical model of the form

$$\begin{cases} Y_i \,|\, X_i, \Omega_{yx} \;\sim\; \mathcal{N}(-\Omega_{yy}^{-1}\,\Omega_{yx}\,X_i,\, \Omega_{yy}^{-1}) \\ \mathrm{vec}(\Omega_{yx}^t) \;\sim\; \mathcal{GN}(0, 1, \Omega_{yy} \otimes L^{-1}, \beta) \end{cases}$$

where the emphasis is on Ω_{yx} in the prior and Ω_{yy} remains a fixed parameter, although it is important to see that, in this work, the estimation step does not rely on a posterior distribution. The following proposition is related to the existence of a global minimum for our objective (2.2) with respect to $(\Omega_{yy}, \Omega_{yx})$ as soon as $\beta \geqslant 1$.

Proposition 2.1. *Assume that $\beta \geqslant 1$. Then, $L_n(\Omega_{yy}, \Omega_{yx})$ defined in (2.2) is jointly convex with respect to $(\Omega_{yy}, \Omega_{yx})$.*

Proof. See Section 5.2. □

Now and throughout the rest of the paper, denote by $\theta = (\Omega_{yy}, \Omega_{yx}) \in \Theta = \mathbb{S}^q_{++} \times \mathbb{R}^{q\times p}$ the $(q \times (q+p))$-matrix of parameters of the model, with true value $\theta^* = (\Omega^*_{yy}, \Omega^*_{yx})$. As it is usually done in studies implying sparsity, we will also consider S of cardinality $|S|$, the true active set of θ^* defined as $S = \{(i,j),\, \theta^*_{i,j} \neq 0\}$, and its complement $\bar{S}$. Our results also depend on some basic assumptions related to the true covariances of the Gaussian observations, and we will assume that the following holds.

$$\Sigma^*_{xx} \in \mathbb{S}^p_{++}, \qquad \Omega^*_{yy} \in \mathbb{S}^q_{++}, \qquad B \neq 0 \text{ (that is, } \Omega^*_{yx} \neq 0) \qquad \text{and} \qquad \Omega^*_{yx}\,L\,\Omega^{*\,t}_{yx} \in \mathbb{S}^q_{++}. \tag{H$_1$}$$

This is a natural hypothesis in our framework, in particular we suppose that there is at least a link between X and Y.

Remark 2.2 (Null model)**.** Even if it is of less interest, our study does not exclude the case where $\Omega^*_{yx} = 0$. Indeed, we might as well consider that $\Omega^*_{yx} = 0$ and get the same results, but some constants should be refined. On the other hand, $\Sigma^*_{xx} \in \mathbb{S}^p_{++}$ and $\Omega^*_{yy} \in \mathbb{S}^q_{++}$ are crucial.

Under (H_1), the random matrices

$$A_n = (S_{yy}^{(n)} - \Sigma_{yy}^*) - \Omega_{yy}^{*\,-1}\,\Omega_{yx}^*\,(S_{xx}^{(n)} - \Sigma_{xx}^*)\,\Omega_{yx}^{*\,t}\,\Omega_{yy}^{*\,-1} \quad \text{with} \quad h_a = |A_n|_\infty \tag{2.3}$$

and

$$B_n = 2\,((S_{yx}^{(n)} - \Sigma_{yx}^*) + \Omega_{yy}^{*\,-1}\,\Omega_{yx}^*\,(S_{xx}^{(n)} - \Sigma_{xx}^*)) \quad \text{with} \quad h_b = |B_n|_\infty \tag{2.4}$$

are going to play a fundamental role, especially h_a and h_b. Let us now provide some theoretical guarantees for the estimation of θ in our model, provided that the regularization parameters are located in a particular area $(\lambda, \mu, \eta) \in \Lambda$. Consider the penalized likelihood $\ell_{\lambda,\mu,\eta}(\theta)$ given in (2.2), and estimate θ by the global minimum

$$\widehat{\theta} = \arg\min_{\Theta}\, \ell_{\lambda,\mu,\eta}(\theta) \tag{2.5}$$

obtained for $\beta \geqslant 1$. To facilitate reading, we postpone the precise definition of the numerous constants to the Appendix. We recall that p is the number of predictors, q is the number of responses and $|S|$ is the size of the true active set.

Theorem 2.3. *Fix $d_\lambda > c_\lambda > 1$, $d_\mu > c_\mu > 1$, $e_\lambda > 0$ and $e_\mu > 0$, and assume that the regularization parameters satisfy $(\lambda, \mu, \eta) \in \Lambda = [c_\lambda\, h_a,\, d_\lambda\, h_a] \times [c_\mu\, h_b,\, d_\mu\, h_b] \times [0,\, \overline{\eta}]$, where*

$$\overline{\eta} = \frac{\min\left\{\frac{(c_\lambda - 1)\,\lambda}{c_\lambda\,\ell_a},\ \frac{(c_\mu - 1)\,\mu}{c_\mu\,\ell_b},\ \frac{e_\lambda\,h_a}{\ell_a},\ \frac{e_\mu\,h_b}{\ell_b}\right\}}{\beta\, s_L^{\beta-1}}$$

for some non-random constants s_L, ℓ_a and ℓ_b defined in (A.2) and (A.3), and the random constants h_a and h_b given above. Then, under (H_1), there exists absolute constants $b_1 > 0$ and $b_2 > 0$ such that, for any $0 < b_3 < 1$ and as soon as $n > n_0$, with probability no less that $1 - \mathrm{e}^{-b_2 n} - b_3$, the estimator (2.5) satisfies

$$\|\widehat{\theta} - \theta^*\|_F \leqslant \frac{16\, m^*\, c_{\lambda,\mu}\sqrt{|S|}}{\gamma_{r,\eta,\beta,p}} \sqrt{\frac{\ln(10(p+q)^2) - \ln(b_3)}{n}}$$

where $\gamma_{r,\eta,\beta,p}$, $c_{\lambda,\mu}$ and m^* are technical constants defined in (A.7), (A.8) and (A.9), respectively, and where the minimal number of observations is given by

$$n_0 = \max\left\{\frac{(\ln(10(p+q)^2) - \ln(b_3))\, c_{\lambda,\mu}^2\, |S|\, (16\, m^*)^2}{r^{*\,2}\, \gamma_{r,\eta,\beta,p}^2},\right.$$
$$\left. b_1\,(q + \lceil s_\alpha \rceil \ln(p+q)), \ln(10(p+q)^2) - \ln(b_3)\right\} \tag{2.6}$$

with s_α defined in (A.5) and r^* in (A.6).

Proof. See Section 5.3. □

Among all these constants, we can note that s_L, ℓ_a, ℓ_b, h_a and h_b are useful to properly describe and restrict Λ, the domain of validity of (λ, μ, η) for the theorem to hold. Once Λ is fixed, the other constants take part in the upper bound of the estimation error. However, as it stands, the theorem is very difficult to interpret. The next two remarks seem essential to have an overview of the orders of magnitude involved for the number of observations, for p and q, for the estimation error and for the regularization parameters.

Remark 2.4 (Validity band)**.** Of course the degree of sparsity $|S|$ is crucial in the estimation error, but it also plays an indirect role in the probability associated with the theorem and in the numerous constants. In virtue of Lemma 5.12, we can hope that λ and μ have a wide validity band, by adjusting c_λ, c_μ, d_λ and d_μ. In turn, η also has a non-negligible area of validity, provided of course that ℓ_a, ℓ_b and s_L, all depending on combinations between Ω^*_{yx}, $\Omega^{*\,-1}_{yy}$ and L, are small enough. Accordingly, it would be to our advantage if L was both sparse and not chosen with too large elements. As it always appears together with η, we may as well take a normalized version of L (*e.g.* $|L|_\infty \leqslant 1$).

Remark 2.5 (Order of magnitude)**.** Even if the result holds for any $\beta \geqslant 1$, the terms $\propto p^{\beta-1}$ appearing in some upper bounds of the proof clearly argue in favor of a moderate choice $\beta \in [1, 1+\epsilon]$ for a small $\epsilon > 0$, depending on p. In other words, we cannot deviate too much from the Gaussianity in the prior on the direct links. For example in a very high-dimensional setting ($p \sim 10^7$), choosing $\epsilon = 0.1$ leads to $p^{\beta-1} \approx 5$ whereas we may try larger values of ϵ for the more common high-dimensional settings $p \sim 10^3$ or $p \sim 10^4$. By contrast, we can see that n_0 must (at least) grow like q for the theorem to hold, so high-dimensional responses are excluded. However in multiple-output regressions, even when p is extremely large, q generally remains small. According to all these considerations, we may roughly say that, in a high-dimensional setting with respect to p,

$$\|\widehat{\theta} - \theta^*\|_F \lesssim \sqrt{\frac{|S| \ln p}{n}}$$

with a large probability, under a suitable regularization of the model. We recognize the usual terms appearing in the error bounds of regressions with high-dimensional covariates, like the ℓ_2 error of the Lasso (see *e.g.* Chap. 11 of [11]). This is the same bound as in [29], but our additional structural penalty restricts Λ.

3. Simulations and real dataset

The minimization problem (2.5) is solved using a coordinate descent procedure, alternating between the computations of

$$\widehat{\Omega}_{yy} = \arg\min_{\mathbb{S}^q_{++}} \ell_{\lambda,\mu,\eta}(\Omega_{yy}, \widehat{\Omega}_{yx}) \quad \text{and} \quad \widehat{\Omega}_{yx} = \arg\min_{\mathbb{R}^{q\times p}} \ell_{\lambda,\mu,\eta}(\widehat{\Omega}_{yy}, \Omega_{yx}).$$

Each step is done by an Orthant-Wise Limited-Memory Quasi-Newton (OWL-QN) algorithm (see *e.g.* [1]). The first subproblem is performed through half-vectorization (vech) to ensure symmetry and we set the objective to $+\infty$ on $\bar{\mathbb{S}}^q_{++}$ to ensure positive definiteness of the solution. The coordinate descent is stopped when

$$\|\widehat{\Omega}^{(t)}_{yy} - \widehat{\Omega}^{(t-1)}_{yy}\|_2 \leqslant \epsilon \max(1, \|\widehat{\Omega}^{(t-1)}_{yy}\|_2) \quad \text{and} \quad \|\widehat{\Omega}^{(t)}_{yx} - \widehat{\Omega}^{(t-1)}_{yx}\|_2 \leqslant \epsilon \max(1, \|\widehat{\Omega}^{(t-1)}_{yx}\|_2)$$

following two consecutive iterations $t-1$ and t, where $\epsilon > 0$ is a small threshold depending on the desired precision. We are now going to try our method on synthetic data first, and then on a real dataset. We will pay attention to the role played by β, in particular we will see that it can be useful as well as counterproductive, depending on the situations.

3.1. Simulations

For each scenario, we first generate i.i.d. standard Gaussian vectors $X_i \in \mathbb{R}^p$, then $Y_i \in \mathbb{R}^q$ is simulated according to the setting and we estimate Ω_{yy} and Ω_{yx}. From the relations detailed in Section 1, we recall that $Y_i = B^t X_i + E_i$ with $E_i \sim \mathcal{N}(0, R)$ is an equivalent formulation, provided that $B = -\Omega^t_{yx}\,\Omega^{-1}_{yy}$ and $R = \Omega^{-1}_{yy}$. In a compact form, we may also write

$$Y = XB + E \quad \text{or} \quad \mathrm{vec}(Y) = (I_q \otimes X)\,\mathrm{vec}(B) + \mathrm{vec}(E)$$

where the i-th row of Y is ${Y_i}^t$ and the i-th row of X is ${X_i}^t$. Thus, we can estimate B using the Lasso (Las) and the Group-Lasso (GLas) in the vectorized form, to provide a basis for comparison between our method and the usual penalized methods. The Lasso penalty is obviously $\|\mathrm{vec}(B)\|_1$ to promote coordinate sparsity while, for the Group-Lasso, we use the penalty $\|B_1\|_2 + \ldots + \|B_p\|_2$ where B_i is the i-th row of B, to promote row sparsity and exclude altogether some predictors from the model. We also implement some variants of our generalized graphical model (GenGm). The case where $\Omega_{yy} = R^{-1}$ is known and does not need to be estimated is the Oracle (Or) and the case where $\eta = 0$ so that β has no influence is the classic PGGM (Gm). The case where $\lambda = 0$ and $\beta = 1$ is called the Spring (Spr) by the authors of [7]. We will focus on structured scenarios. With no structure in Ω_{yx}, there is no reason why our method should outperform the usual PGGM. In a completely random setting, we have observed that all PGGM procedures perform identically. In fact, a slight gain can be obtained compared to Spr and Gm simply due to the flexibility induced by the additional parameter (Spr and Gm are particular cases of GenGm). However, that clearly cannot counterbalance the extended computational times, and GenGm should not be used for such situations. The calibration of the regularization parameters is made using a cross-validation on a training set of size $n_t = 150$ and the accuracy is evaluated thanks to the mean squared prediction error (MSPE) on a validation set of size $n_v = 1000$,

$$\mathrm{MSPE} = \frac{\left\|Y + X\,\widehat{\Omega}_{yx}^t\,\widehat{\Omega}_{yy}^{-1}\right\|_F^2}{q\,n_v}. \tag{3.1}$$

Due to the large amount of treatments, the grids for cross-validation are not very sharp here but they will be carefully refined for the real dataset of the next section. The covariance between the outputs is $R = (r^{|i-j|})_{1 \leqslant i,\, j \leqslant q}$ for $r = \frac{1}{2}$ and we work with $p = 100$. Each scenario is repeated $N = 500$ times and GenGm is evaluated with numerous values of β, from 0.25 to 2 with a step of 0.25. The results of the following scenarios are summarized in Figures 2–4, respectively.

- Scenario 1 ($q = 1$). We draw $\omega_i = \pm\frac{1}{2}$ for $i = 1, \ldots, 10$ and we fill 10 randomly selected sections of size 3 in Ω_{yx} with ω_i. The remaining part of Ω_{yx} is 0.
- Scenario 2 ($q = 2$). We draw $\omega = \pm\frac{1}{2}$ and one randomly selected row of Ω_{yx} is filled with ω while the other is identically 0.
- Scenario 3 ($q = 3$). We draw $\omega_i = \pm\frac{1}{2}$ and we fill a randomly selected section of size 30 on the i-th row of Ω_{yx} with ω_i, for $i = 1, 2, 3$. The remaining part of Ω_{yx} is 0.

The row structure is promoted by a normalized first finite difference operator

$$L = \frac{1}{2}\begin{pmatrix} 1 & -1 & 0 & \ldots & 0 \\ -1 & 2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 2 & -1 \\ 0 & \ldots & 0 & -1 & 1 \end{pmatrix} \tag{3.2}$$

which, through $\Omega_{yx}\,L\,\Omega_{yx}^t$, tends to penalize the difference between two consecutive values on a same row (as does Fused-Lasso with ℓ_1 penalty). Yet, the Fused-Lasso is not a suitable alternative to GLas and Las in this precise context because $B = -\Omega_{yx}^t\,\Omega_{yy}^{-1}$ is not supposed to have a row structure even if Ω_{yx} has one. For this choice of L, one can note that, in the particular case where $R = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_q^2)$,

$$\langle\!\langle L, \Omega_{yx}^t\,\Omega_{yy}^{-1}\,\Omega_{yx}\rangle\!\rangle^\beta = \left(\sum_{i=1}^q {\sigma_i}^2 \sum_{j=2}^p (\omega_{i,j} - \omega_{i,j-1})^2\right)^\beta \geqslant \sum_{i=1}^q {\sigma_i}^{2\beta} \sum_{j=2}^p \left|\omega_{i,j} - \omega_{i,j-1}\right|^{2\beta}$$

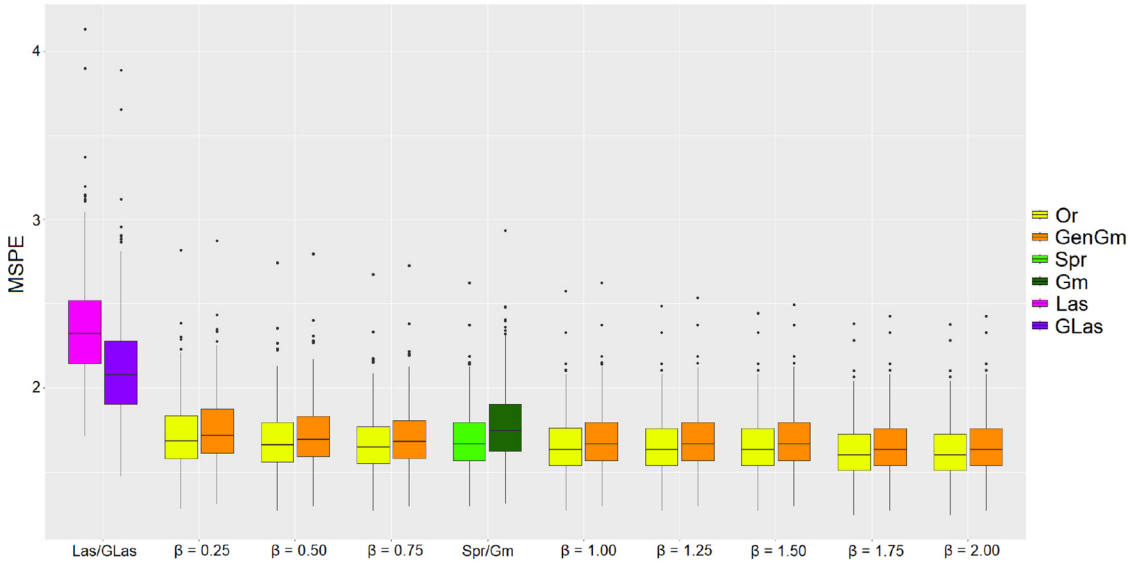

FIGURE 2. Mean squared prediction error for $N = 500$ repetitions of the weakly structured Scenario 1.

where $\omega_{i,j}$ is the (i, j)-th element of Ω_{yx}, so we may fairly expect that $\beta \geqslant 1$ is going to strengthen the smoothness of the estimation and to enforce all the more the structuring.

Remark 3.1 (Validity of the hypotheses)**.** We could as well add a small diagonal element in the matrix L defined above, positive semi-definite but not invertible. The resulting effect would be a negligible ridge-like penalization on the elements of Ω_{yx}. This is not required for the estimation procedure but useful for Theorem 2.3 to hold (see *e.g.* (H_1)). Likewise, it seemed interesting to test some settings with $\beta < 1$ even if the theory developped in the paper does not give any guarantee for them, as a basis for comparison.

First of all, one can observe that Las and GLas are left behind in all our simulations. This is not surprising since the covariance between the outputs cannot be recovered with the standard Lasso, at least for $q \geqslant 2$. Generally, GLas remains more robust compared to Las, probably due to the high level of sparsity in Ω_{yx} approximately passed to B (provided that the covariances in R are small enough), and exploited by the grouping effect. In the weakly structured setting (Scenario 1), we also observe that, as expected, all PGGM procedures perform almost identically, with obviously an advantage for Or (although small, illustrating the accuracy of the estimation). In the strongly structured settings (Scenarios 2 and 3), Gm gives results below the expected level, because it is not designed to promote such layouts. On the contrary, thanks to this choice of L showing here great efficiency, GenGm and Spr are doing pretty well. Note that, in this context, GenGm with $\beta = 1$ is almost the same as Spr since, q being small, λ does not play a crucial role. However, some empirical facts draw our attention: the prediction error decreases with β to some extent, but the most interesting fact seems to be the simultaneous decrease of its variance. It is likely that the increasing pressure exerted by β on the estimation procedure leads to a higher homogeneity in the numerical results, despite the repetitions of random experiments under random settings. In other words, the structuring seems to be strengthened and we also observe that the convergence of the algorithm is faster, which logically follows from the latter remarks (especially clear when we compare $\beta = 0.25$ and $\beta = 2$). On the other hand, for the opposite reason, we notice that the predictions are hardly better than Gm (even worse in some cases), both on average and in terms of variability, for $\beta < 1$, and these simulations tend to undermine such values of the hyperparameter. On the whole, GenGm with $\beta > 1$ might be a sound approach for practitioners who place a high priority on structuring the estimations, even if Remark 3.2 below should probably temper this statement. To conclude, let us consider the strongly structured scenarios

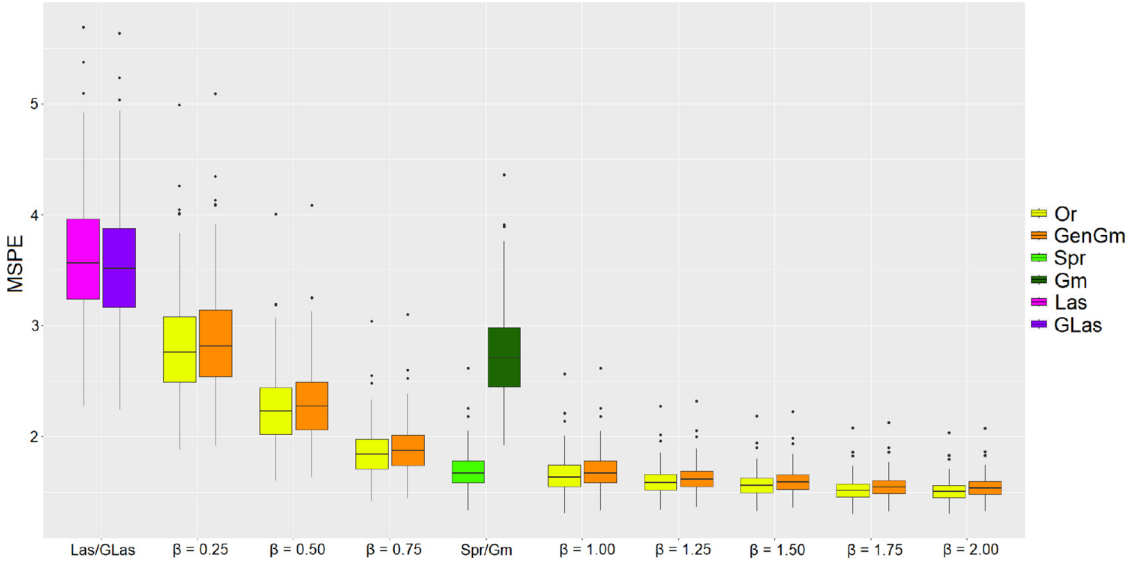

FIGURE 3. Mean squared prediction error for $N = 500$ repetitions of the strongly structured Scenario 2.

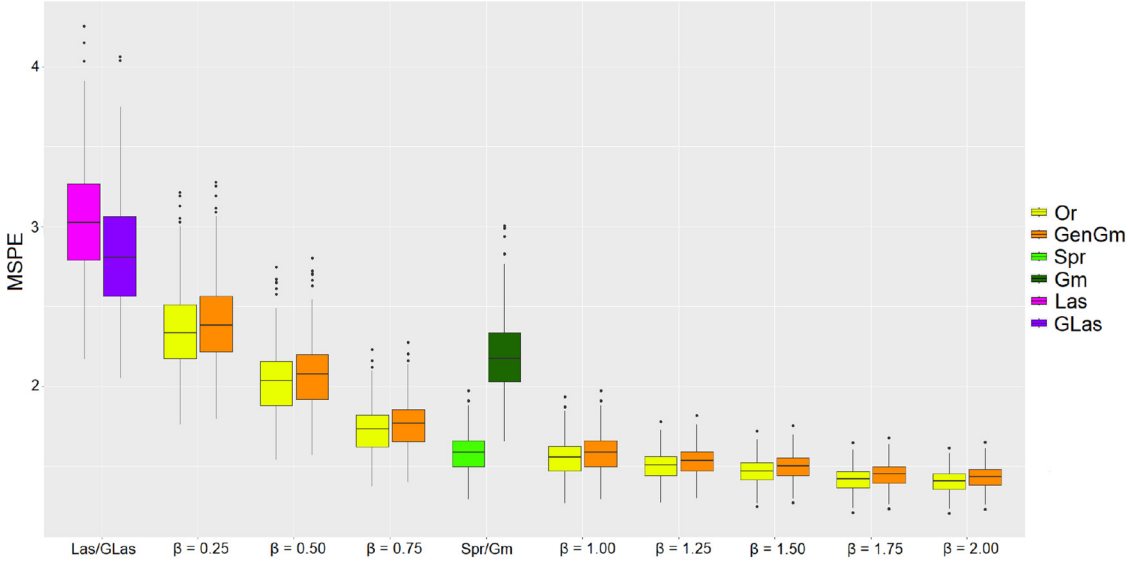

FIGURE 4. Mean squared prediction error for $N = 500$ repetitions of the strongly structured Scenario 3.

with $L = I_p$ (without structuring) in the Oracle setting with $\beta = 2$, and let us compare the results with those of Figures 3 and 4, obtained with the correct version of L given in (3.2). The results are displayed in Figure 5 where we can see that the benefit of structuring is manifest. Unsurprisingly, the results without structuring are close to those of Gm since $L = I_p$ only strengthens the shrinkage effect with ridge-like additional penalties.

Remark 3.2 (Computational time)**.** To estimate $(\Omega_{yy}, \Omega_{yx})$ in the model Spr, the authors of [7] use a very judicious and efficient method relying, in each step of the coordinate descent procedure, on a direct computation

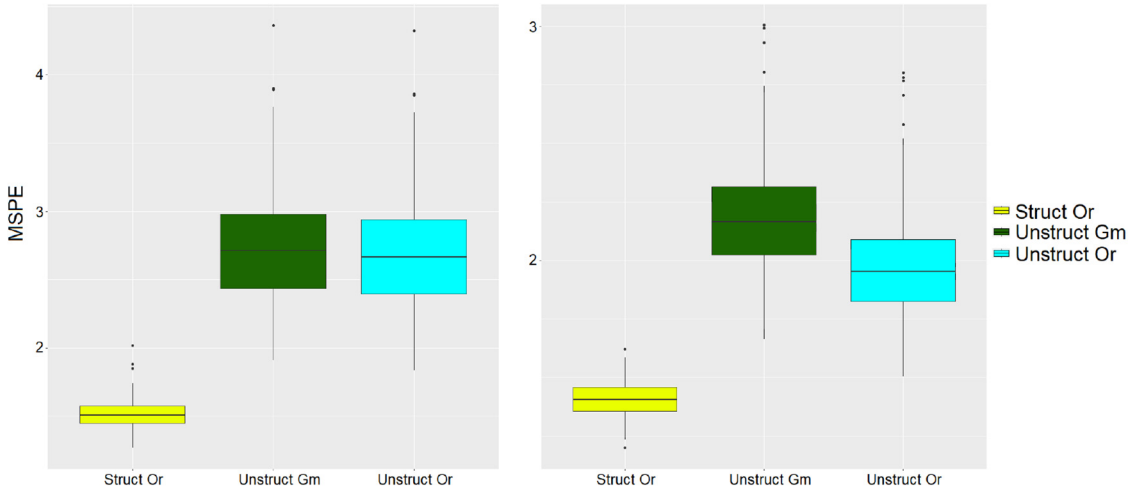

FIGURE 5. Mean squared prediction error for $N = 500$ repetitions of the strongly structured Scenario 2 (*left*) and Scenario 3 (*right*) for Or, Gm and the unstructured Or ($L = I_p$), with $\beta = 2$.

of the estimation of Ω_{yy} together with an Elastic-Net estimation of Ω_{yx}. This is possible for $\lambda = 0$ and $\beta = 1$, but unfortunately cannot be implemented in the general setting. As a result, computational times remain an issue that should be paid attention to.

Remark 3.3 (Oracle-type errors)**.** The mean value of the estimation errors $\|\widehat{\Omega}_{yx} - \Omega_{yx}\|_F^2$ leads to the same kind of observations for the models being compared in the simulations. But the minimal prediction error does not always coincide with an optimal support recovery due to the shrinkage effect on the estimation of Ω_{yx}, when the coefficients or the covariates are not very contrasting. The so-called F-score is given by

$$F = \frac{2\, p_r\, r_e}{p_r + r_e} \quad \text{where} \quad p_r = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad \text{and} \quad r_e = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

are the *precision* and the *recall*, respectively, and where T/F and P/N stand for true/false and positive/negative. In the strongly structured scenarios, F is generally located between 0.60 and 0.65, and a deeper analysis shows that a proportion of more than 0.99 of true non-zero values are recovered (that is, the part of the true active set S related to Ω_{yx}). If the models are not calibrated to reach the best prediction error but the best F-score, F regularly exceeds 0.90, at least for the structured procedures.

Nevertheless, Scenarios 2 and 3 are very strongly structured, more than one would expect from an unknown underlying generating process, and the real dataset of the next section is going to highlight the fact that the improvement may be hardly noticeable with respect to β. But we will see that β can still be useful for variable selection.

3.2. A real dataset

The dataset available as `CanadianWeather` in the `R` package `fda` contains daily temperature and precipitation at 35 different locations in Canada, averaged over annual reports starting in 1960 and ending in 1994 (see *e.g.* [20]). We intend to look at the direct links between the minimal and maximal rainfall (on the $\log_{10}$ scale) and the temperature pattern in the 35 weather stations, so as to identify the times of the year that have a strong effect on rainfall (positive as well as negative). In this context, $n = 35$, $q = 2$ and $p = 365$. Figure 6 shows temperature and log-precipitation measured over a year in Montreal, chosen as an example, together with the

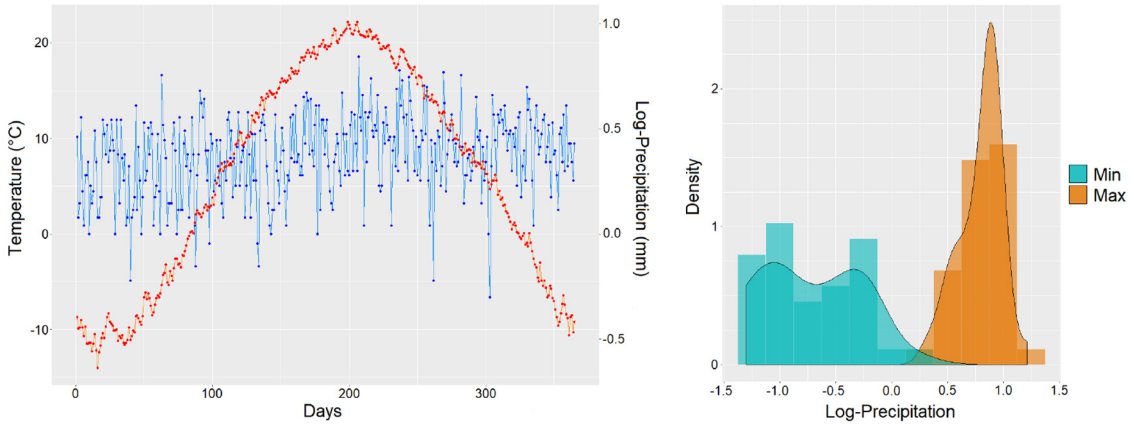

FIGURE 6. Temperature and log-precipitation measured over a year in Montreal (*left*). Empirical distribution of the minimal and maximal log-precipitation for the 35 weather stations (*right*).

empirical distribution of the minimal and maximal log-precipitation for the 35 weather stations. We can note that, since the data are averaged over numerous years, outliers are unlikely even for the extremes (min and max).

Some authors (see *e.g.* [24]) have already highlighted the pertinence of using the matrix L defined in (3.2) in this dataset, because the predictors are ordered temporally so that the selection of isolated days instead of relevant sequences of days seems an unreliable procedure for statistical interpretation. To assess the models, we repeat $N = 100$ times the following experiment: $n_t = 25$ observations are randomly selected for calibration (*via* 2-fold cross-validation) and estimation, the remaining $n_v = 10$ observations are used to compute the MSPE (3.1) related to the prediction of the minimum ($\min_{\mathrm{p}}$) and maximum ($\max_{\mathrm{p}}$) precipitation. We can see in Figure 7 that all structured PGGM perform almost identically, with the phenomenon described in the previous section still visible but to a lesser extent. We can even notice that structuring is hardly beneficial for this dataset, from a purely numerical point of view. This conclusion can also be found in [24], where the author compares the structured Elastic-Net with unstructured alternatives to predict the 0.25-, 0.50- and 0.75-quantiles of the log-precipitation, through independent regressions. But we will see that, in terms of variable selection and statistical interpretation, L and β still have a substancial role to play.

The point is that we have observed that the best prediction error does not usually coincide with a sparse solution (see Rem. 3.3) when the coefficients or the covariates are not very contrasting. In particular, this was the case of our simulation study with $\pm\frac{1}{2}$ coefficients and $\mathcal{N}(0,1)$ covariates. So, just as they look at the Lasso's regularization paths, practitioners may choose the desired degree of sparsity, depending on p/n, by adjusting the hyperparameters. Here, on the basis of the MSPE, most of the time we must retain $\mu \ll 10^{-2}$ and only a few direct links are set to zero. To look for sequences of days directly related to $\min_{\mathrm{p}}$ and $\max_{\mathrm{p}}$, we decided to constraint $\mu \geqslant 10^{-2}$ and focus on variable selection. The active set of Ω_{yx} is evaluated on the basis of $n_t = 25$ randomly chosen observations. The experiment is repeated $N = 100$ times, and the locations having a frequency of occurrence that exceeds 0.5 are retained (or, equivalently, those whose estimates have a non-zero median). This can be seen as a measure of variable importance. The results are given in Figures 8 and 9 for $\min_{\mathrm{p}}$ and $\max_{\mathrm{p}}$, respectively, with a fixed set of regularization parameters and increasing values of β. The objective is to show the influence of the latter, all other things being equal. The colored areas highlight the days having a frequency of occurrence, represented by gray crosses, that exceeds 0.5 in the $N = 100$ repetitions of the experiment. Note that, since we retain $\lambda = 0$ in these experiments, GenGm for $\beta = 1$ coincides with Spr. We can see that the increasing pressure exerted by β on the estimation procedure tends to refine the selection by giving priority to the most important variables and by dropping the others much more easily, at the cost of

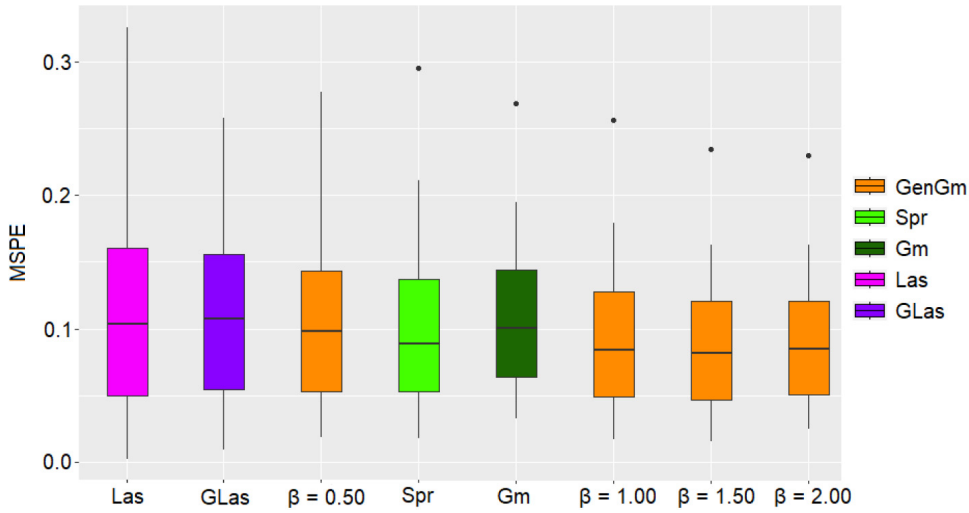

FIGURE 7. Mean squared prediction error for $N = 100$ repetitions of the experiment. GenGm for $\beta \in \{0.5, 1, 1.5, 2\}$ is compared with Spr, Gm, Las and GLas.

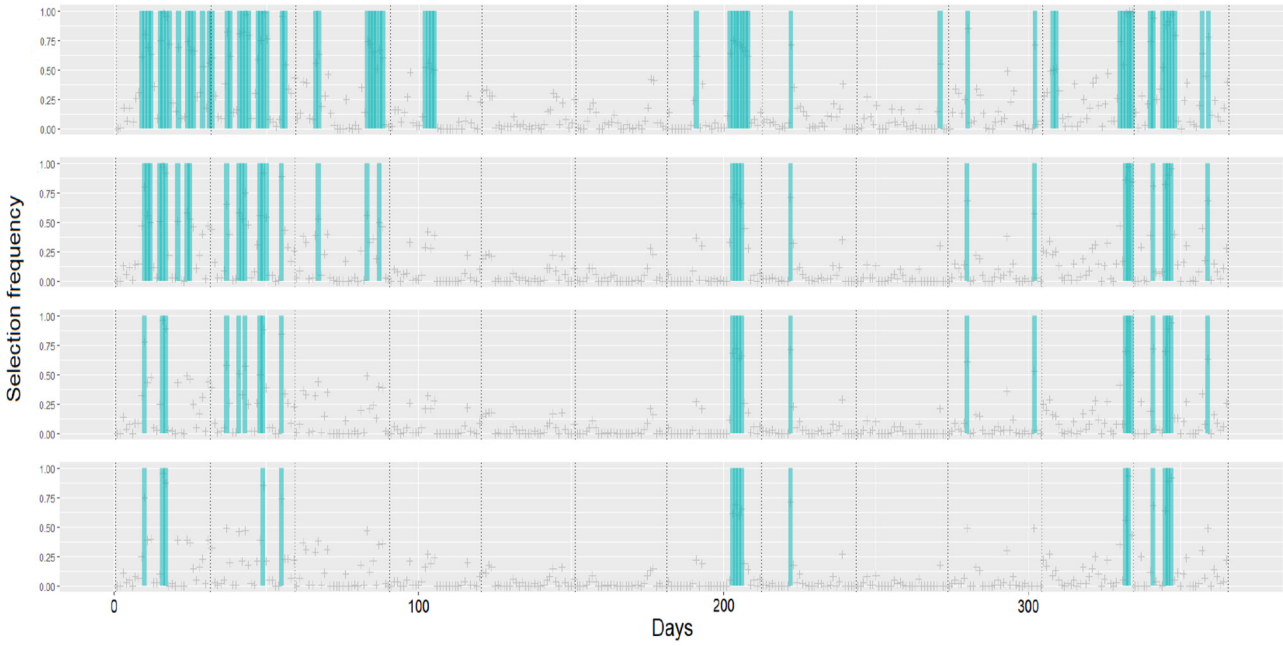

FIGURE 8. Variable selection for $\min_{\mathrm{p}}$ by GenGm with $(\lambda, \mu, \eta) = (0, 0.05, 1)$ and, from *top to bottom*, $\beta \in \{0.5, 1, 1.5, 2\}$.

prediction results: we are undoubtedly in a selection process. The sequence of inclusions

$$\widehat{S}_{\beta_2} \subset \widehat{S}_{\beta_1} \qquad \text{for } \beta_1 < \beta_2$$

that we observe for the estimated active sets is clearly a guarantee of quality for the selected variables.

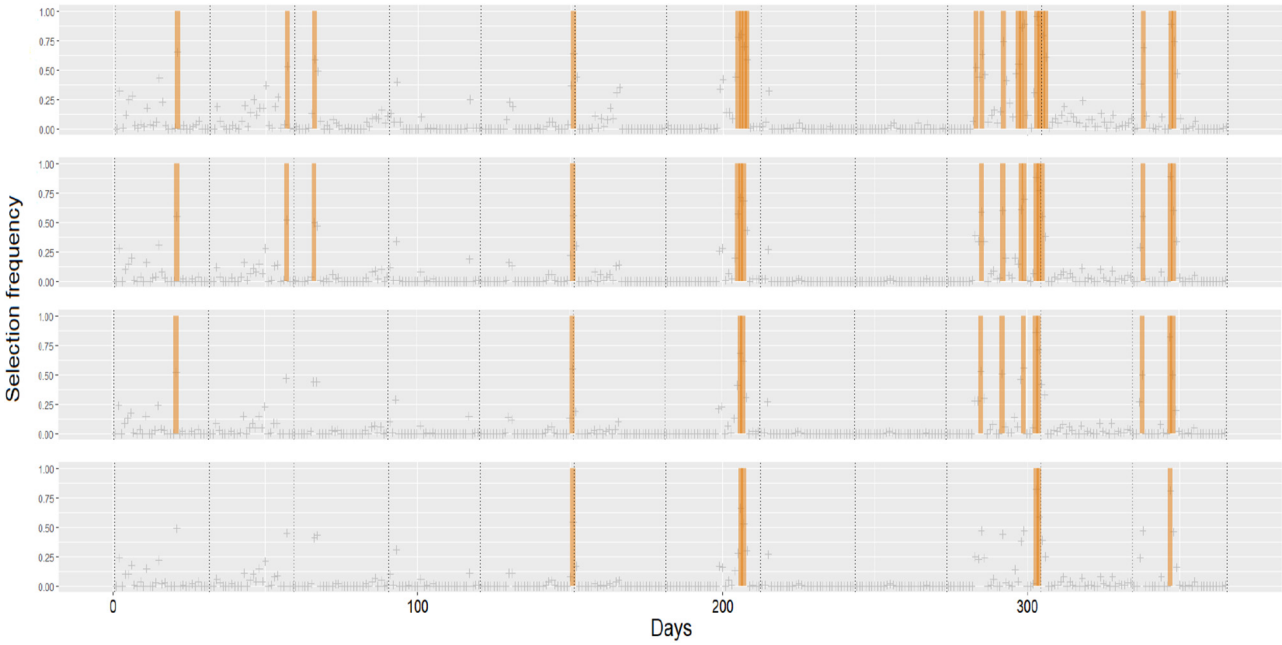

FIGURE 9. Variable selection for $\max_{\mathrm{p}}$ by GenGm with $(\lambda, \mu, \eta) = (0, 0.05, 1)$ and, from *top to bottom*, $\beta \in \{0.5, 1, 1.5, 2\}$.

The median values of the estimated direct links between the temperature of the days and the pair $(\min_{\mathrm{p}}, \max_{\mathrm{p}})$ are represented in Figure 10 together with the estimated regression coefficients, for $\beta = 2$. We recall that the relation $B = -\Omega_{yx}^t \, \Omega_{yy}^{-1}$ simply leads to

$$\widehat{B} = -\widehat{\Omega}_{yx}^t \, \widehat{\Omega}_{yy}^{-1}.$$

We detect sequences of influent days in November, December, January and February, especially related to $\min_{\mathrm{p}}$, positively at the end of the year and negatively at the beggining. This is broadly consistent with the analysis of [24] – even if the responses are not extremes but quantiles in it – with however two differences: the regression coefficients associated with $\max_{\mathrm{p}}$ are much lower compared to $\min_{\mathrm{p}}$ whereas it is not that clear in the reference, and an activity is also detected between July and August. The main explanation, at least for the first of them, probably lies in the use of graphical models that take into account the correlation between responses. Indeed, as can be seen in Figure 11 which gives an overview of the estimation of R obtained from the repeated experiments, a non-zero correlation is detected between the responses (≈ 0.32). The influence of November and December on all quantiles and that of January and February on the 0.75-quantile in [24] might actually be an artificial effect of the correlation with the 0.25-quantile. This is what our study suggests by highlighting $\min_{\mathrm{p}}$ compared to $\max_{\mathrm{p}}$: the 'real' effect appears to be on $\min_{\mathrm{p}}$ whereas $\max_{\mathrm{p}}$ seems to react only through a phenomenon of correlation with $\min_{\mathrm{p}}$. From this point of view, the interest of graphical models instead of independent regressions is particularly obvious.

Let us also mention that, interestingly enough, we notice that the role of η tends to depreciate for the large values of β. For example, for the same regularization parameters $(\lambda, \mu) = (0, 0.05)$ and $\beta = 2$, the difference between the estimated active sets for $\eta = 0.1$ and $\eta = 1$ is almost negligible (depending on the experiments, between 1 and 3 days are concerned, on average). Based on these studies and observations, we might conclude that β is insignificant when we are interested in the best prediction error on a validation set (even counterproductive with respect to computational times, *e.g.* compared to Spr), whereas it seems to have a substancial role

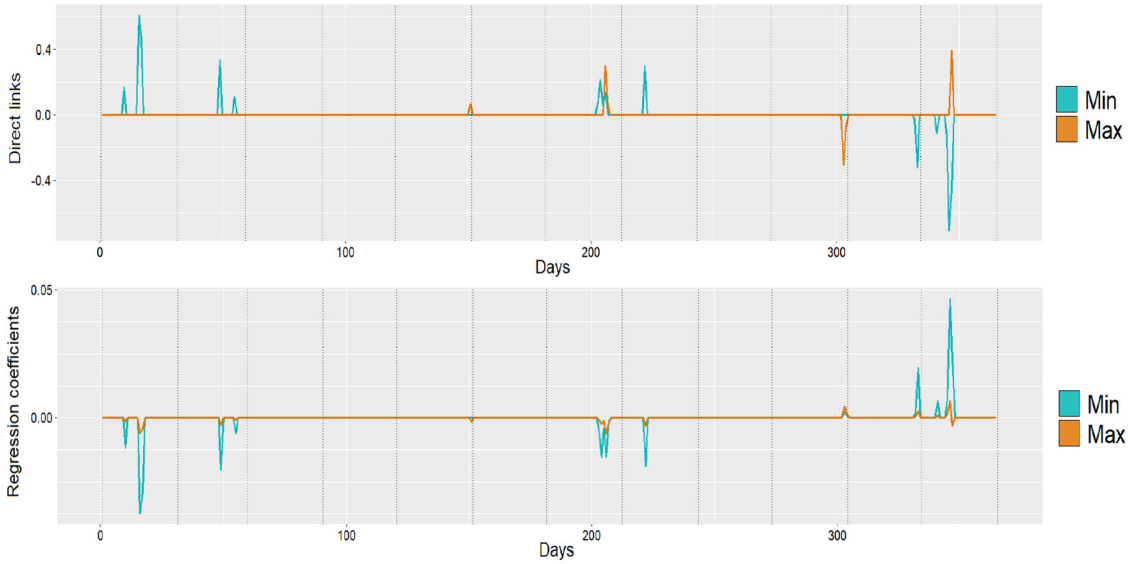

FIGURE 10. Estimated direct links (*top*) and regression coefficients (*bottom*) for the pair $(\min_{\mathrm{p}}, \max_{\mathrm{p}})$ by GenGm with $(\lambda, \mu, \eta) = (0, 0.05, 1)$ and $\beta = 2$, after the $N = 100$ experiments. Dotted lines divide the panel into months.

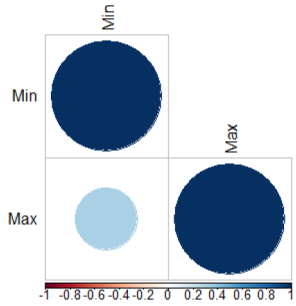

FIGURE 11. Estimated correlation between $\min_{\mathrm{p}}$ and $\max_{\mathrm{p}}$ by GenGm with $(\lambda, \mu, \eta) = (0, 0.05, 1)$ and $\beta = 2$, after the $N = 100$ experiments. The off-diagonal entry is approximately 0.32.

to play when focusing on selection, by accelerating the discrimination of variables. In the first case, η has to be carefully adjusted while in the second case, β will quickly help to reach the desired sparsity.

Remark 3.4 (Structure matrix)**.** For the simulations and the real dataset, we have used the popular first finite difference operator given in (3.2). Other examples can be found in the literature, like the promotion of a genetic distance for genomic selection in *Brassica napus* [7] or the bidimensional discretization of the Laplacian to work on handwritten digit recognition [24]. More generally, L can be used in a classic Bayesian prior supposed to promote some covariance structure on the direct links, with no 'physical' structuring in mind (like temporal, spatial or genetic proximity).

4. CONCLUSION

In conclusion, our work is a generalization of [29], using the same technical tools to establish an upper bound on the estimation error when a prior on the direct links generates an additional structural penalty in the

objective, provided that the model is suitably regularized. Our work is also an improvement of [7] since, while being inspired by the methodology of the authors, we generalize the prior and give some theoretical guarantees. The empirical study shows that the hyperparametrization in the prior, although more expensive in adjusting the parameters, is likely to refine the selection results but clearly, this does not appear as a crucial improvement compared to the two previous points. Let us conclude the paper by highlighting two weaknesses that might be trails for future studies. On the one hand, the Laplace distribution is often used as a prior in the Bayesian Lasso (see *e.g.* Sect. 6.1 of [11]). However, our reasonings do not allow $\beta = 1/2$, which may correspond to a multivariate Laplace distribution on the direct links. Combined with the first finite difference operator L, the choice $\beta = 1/2$ could generate a Fused-Lasso-type penalty. In this regard, it would be challenging and interesting to obtain some theoretical guarantees for $\beta \geqslant 1/2$ and not only for $\beta \geqslant 1$, even if our probably too brief simulation study does not encourage the choice of $\beta < 1$. On the other hand, $\lambda = 0$ is a natural choice when q is small (this is in particular the configuration of [7]), not to mention that it is computionally faster. But, the proof of our theorem needs $\lambda > c_\lambda\, h_a > 0$ to hold. We think that a reasoning enabling to deal with $\lambda = 0$ should also be beneficial to the study. More generally, it would be instructive to consider a very high-dimensional setting ($p \gg n$ and not only $p \sim 10^2$ although always larger than n, as in our experiments). Such studies should follow with omic data.

5. Technical proofs

We start in a first part by some useful linear algebra lemmas that will be repeatedly used subsequently, well-known for most of them. In a second part, we prove the joint convexity of the objective and our main result.

5.1. Linear algebra

Lemma 5.1. *Let $A \in \mathbb{S}_+^d$ and $U \in \mathbb{R}^{d\times\ell}$. Then, $U^t A U \in \mathbb{S}_+^\ell$.*

Proof. Since A is symmetric with non-negative eigenvalues, there is an orthogonal matrix P such that $A = PDP^t$ with $D = \mathrm{diag}(\mathrm{sp}(A)) \in \mathbb{S}_+^d$. Thus, for all $v \in \mathbb{R}^\ell$, it follows that $\langle v, U^t A\, U\, v\rangle = \|D^{1/2}\, P^t\, U\, v\|^2 \geqslant 0$. □

Lemma 5.2. *Let $A \in \mathbb{S}_{++}^d$ and $B \in \mathbb{S}_+^d$. Then for all i, $\lambda_i(AB) \geqslant 0$.*

Proof. The equality $AB = A^{1/2}\,(A^{1/2}\,B\,A^{1/2})\,A^{-1/2}$ shows that AB and $A^{1/2}\,B\,A^{1/2}$ are similar, so they must share the same eigenvalues. From Lemma 5.1, $\lambda_i(A^{1/2}\,B\,A^{1/2}) \geqslant 0$. □

Lemma 5.3. *Let $A \in \mathbb{S}_+^d$ and $B \in \mathbb{S}_+^d$. Then,*

$$\lambda_{\min}(A)\,\mathrm{tr}(B) \;\leqslant\; \mathrm{tr}(AB) \;\leqslant\; \lambda_{\max}(A)\,\mathrm{tr}(B).$$

Proof. Since $A - \lambda_{\min}(A) I_d \in \mathbb{S}_+^d$ and $B \in \mathbb{S}_+^d$,

$$\mathrm{tr}((A - \lambda_{\min}(A) I_d)\, B) = \mathrm{tr}(B^{1/2}\,(A - \lambda_{\min}(A) I_d)\, B^{1/2}) \geqslant 0$$

from Lemma 5.1, thus $\mathrm{tr}(AB) \geqslant \lambda_{\min}(A)\,\mathrm{tr}(B)$. The other inequality is obtained through $\lambda_{\max}(A) I_d - A \in \mathbb{S}_+^d$. □

Lemma 5.4. *Let $A \in \mathbb{S}_{++}^d$ and $B \in \mathbb{S}_+^d$. Then,*

$$\lambda_{\min}(A)\,\lambda_{\min}(B) \;\leqslant\; \lambda_{\min}(AB) \quad \textit{and} \quad \lambda_{\max}(AB) \;\leqslant\; \lambda_{\max}(A)\,\lambda_{\max}(B).$$

Proof. On the one hand, $\lambda_{\max}(AB) \leqslant \|AB\|_2 \leqslant \|A\|_2\,\|B\|_2 = \lambda_{\max}(A)\,\lambda_{\max}(B)$, since A and B are symmetric and since, from Lemma 5.2 and by hypothesis, all eigenvalues appearing in the relation are non-negative. Suppose now that B is invertible so that both A^{-1} and B^{-1} belong to $\mathbb{S}_{++}^d$. Then, $\lambda_{\max}((AB)^{-1}) \leqslant \lambda_{\max}(A^{-1})\,\lambda_{\max}(B^{-1})$

and this immediately gives $\lambda_{\min}(AB) \geqslant \lambda_{\min}(A)\,\lambda_{\min}(B)$. If B is not invertible, the relation trivially holds since we still have $\lambda_{\min}(AB) \geqslant 0$ from Lemma 5.2. □

Lemma 5.5. *Let $A \in \mathbb{S}_+^d$ and $U \in \mathbb{R}^{d\times\ell}$. Then,*

$$\lambda_{\min}(A)\,\|U\|_F^2 \;\leqslant\; \mathrm{tr}(U^{\,t}AU) \;\leqslant\; \lambda_{\max}(A)\,\|U\|_F^2.$$

Proof. Denote by u_i the i-th column of U. It is not hard to see that the i-th diagonal element of $U^{\,t}AU$ is $u_i^t A\, u_i \geqslant \lambda_{\min}(A)\,\|u_i\|^2 \geqslant 0$. Thus,

$$\mathrm{tr}(U^{\,t}AU) = \sum_{i=1}^{\ell} u_i^t A\, u_i \geqslant \lambda_{\min}(A) \sum_{i=1}^{\ell} \|u_i\|^2 = \lambda_{\min}(A)\,\|U\|_F^2.$$

The upper bound stems from $0 \leqslant u_i^t A\, u_i \leqslant \lambda_{\max}(A)\,\|u_i\|^2$. □

Lemma 5.6. *Let A and B be symmetric matrices of same dimensions. Then,*

$$\lambda_{\min}(A) + \lambda_{\min}(B) \;\leqslant\; \lambda_{\min}(A+B) \qquad \textit{and} \qquad \lambda_{\max}(A+B) \;\leqslant\; \lambda_{\max}(A) + \lambda_{\max}(B).$$

Proof. These are just two special cases of Weyl inequalities. We refer the reader to Theorem 4.3.1 of [12], for example. □

5.2. Convexity of the objective

We know from Proposition 1 of [29] and the convexity of the elementwise ℓ_1 norm that $L_n(\Omega_{yy}, \Omega_{yx}) - \eta\,\langle\!\langle L, \Omega_{yx}^t\,\Omega_{yy}^{-1}\,\Omega_{yx}\rangle\!\rangle^{\beta}$ is itself convex, but it remains to show that this is still the case with the additional smooth penalty.

Proof of Proposition 2.1

Recall that $\Theta = \mathbb{S}_{++}^q \times \mathbb{R}^{q\times p}$ and consider the mapping $\Phi : \Theta \to \mathbb{S}_+^p$ defined as

$$\forall\,(A,B) \in \Theta, \qquad \Phi(A,B) = B^{\,t}A^{-1}B.$$

We can already note from Lemma 5.1 that $\mathrm{tr}(\Phi(A,B)) \geqslant 0$. Moreover, for all $0 \leqslant h \leqslant 1$ and all $Z_i = (A_i, B_i) \in \Theta$, $i = 1, 2$, it is easy to see that

$$S_h(Z_1, Z_2) = h\,\Phi(Z_1) + (1-h)\,\Phi(Z_2) - \Phi(hZ_1 + (1-h)Z_2) \tag{5.1}$$

is the Schur complement of $hA_1 + (1-h)A_2$ in the matrix

$$M_h(Z_1, Z_2) = h\begin{pmatrix} A_1 & B_1 \\ B_1^t & B_1^t\,A_1^{-1}B_1 \end{pmatrix} + (1-h)\begin{pmatrix} A_2 & B_2 \\ B_2^t & B_2^t\,A_2^{-1}B_2 \end{pmatrix}. \tag{5.2}$$

But the decomposition

$$\begin{pmatrix} A^{1/2} & A^{-1/2}\,B \\ 0 & 0 \end{pmatrix}^t \begin{pmatrix} A^{1/2} & A^{-1/2}\,B \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} A & B \\ B^{\,t} & B^{\,t}\,A^{-1}B \end{pmatrix}$$

directly shows that $M_h(Z_1, Z_2)$ in (5.2) is symmetric and positive semi-definite. It is well-known (see *e.g.* Appendix A.5.5 of [4]) that in that case, the Schur complement (5.1) must also be positive semi-definite. Consequently, for $Z_i = (\Omega_{i,yy}, \Omega_{i,yx} L^{1/2})$, $i = 1, 2$, taking the trace of $S_h(Z_1, Z_2)$ and considering $\beta \geqslant 1$,

$$\begin{aligned}
\langle\!\langle L, P_h^t\, Q_h^{-1} P_h \rangle\!\rangle^\beta &= (\mathrm{tr}(\Phi(hZ_1 + (1-h)Z_2)))^\beta \\
&\leqslant (h\,\mathrm{tr}(\Phi(Z_1)) + (1-h)\,\mathrm{tr}(\Phi(Z_2)))^\beta \\
&= (h\,\langle\!\langle L, \Omega_{1,yx}^t\, \Omega_{1,yy}^{-1}\, \Omega_{1,yx} \rangle\!\rangle + (1-h)\,\langle\!\langle L, \Omega_{2,yx}^t\, \Omega_{2,yy}^{-1}\, \Omega_{2,yx} \rangle\!\rangle)^\beta \\
&\leqslant h\,\langle\!\langle L, \Omega_{1,yx}^t\, \Omega_{1,yy}^{-1}\, \Omega_{1,yx} \rangle\!\rangle^\beta + (1-h)\,\langle\!\langle L, \Omega_{2,yx}^t\, \Omega_{2,yy}^{-1}\, \Omega_{2,yx} \rangle\!\rangle^\beta
\end{aligned}$$

where $P_h = h\,\Omega_{1,yx} + (1-h)\,\Omega_{2,yx}$ and $Q_h = h\,\Omega_{1,yy} + (1-h)\,\Omega_{2,yy}$. This convexity inequality concludes the proof. □

5.3. Theoretical guarantees

Proof of Theorem 2.3

Let $R_n(\theta)$ be the the smooth part of the objective (2.2),

$$\begin{aligned}
R_n(\theta) = -\ln\det(\Omega_{yy}) &+ \langle\!\langle S_{yy}^{(n)}, \Omega_{yy} \rangle\!\rangle + 2\,\langle\!\langle S_{yx}^{(n)}, \Omega_{yx} \rangle\!\rangle \\
&+ \langle\!\langle S_{xx}^{(n)}, \Omega_{yx}^t\, \Omega_{yy}^{-1}\, \Omega_{yx} \rangle\!\rangle + \eta\,\langle\!\langle L, \Omega_{yx}^t\, \Omega_{yy}^{-1}\, \Omega_{yx} \rangle\!\rangle^\beta.
\end{aligned} \tag{5.3}$$

For any $\theta \in \Theta$ and $t \in \mathbb{R}$, by a Taylor expansion,

$$R_n(\theta^* + t\,(\theta - \theta^*)) = R_n(\theta^*) + t\,\langle\!\langle \nabla R_n(\theta^*), \theta - \theta^* \rangle\!\rangle + e_t(\theta, \theta^*) \tag{5.4}$$

for some second-order error term $e_t(\theta, \theta^*)$. Consider the reparametrization

$$\phi(t) = R_n(\theta^* + t\,(\theta - \theta^*)) \tag{5.5}$$

so that $\phi'(0) = \langle\!\langle \nabla R_n(\theta^*), \theta - \theta^* \rangle\!\rangle$. Let $\delta\theta_{yy} = \Omega_{yy} - \Omega_{yy}^*$ and $\delta\theta_{yx} = \Omega_{yx} - \Omega_{yx}^*$, let also $\delta\theta = \theta - \theta^*$ in a compact form. The estimation error is denoted

$$\delta\vartheta = \widehat{\theta} - \theta^* = (\widehat{\Omega}_{yy} - \Omega_{yy}^*, \widehat{\Omega}_{yx} - \Omega_{yx}^*) = (\delta\vartheta_{yy}, \delta\vartheta_{yx}). \tag{5.6}$$

Before we start the actual proof, some additional lemmas are needed. They constitute a local study in a sort of r^*-neighborhood of θ^* that we define as

$$N_{r,\alpha}(\theta^*) = \big\{\theta \in \Theta,\ \|\delta\theta\|_F \leqslant r^* \text{ and } |[\delta\theta]_{\bar{S}}|_1 \leqslant \alpha |[\delta\theta]_S|_1\big\}. \tag{5.7}$$

Our strategy can be summarized as follows:

- (Lem. 5.9) Show that there exists a configuration for the regularization parameters (λ, μ, η) so that the estimation error satisfies $|[\delta\vartheta]_{\bar{S}}|_1 \leqslant \alpha |[\delta\vartheta]_S|_1$ for some $\alpha > 0$.
- (Lem. 5.10) Find some $r^* > 0$ and $\gamma_{r,\eta,\beta,p} > 0$ such that $e_1(\theta, \theta^*) > \gamma_{r,\eta,\beta,p} \|\delta\theta\|_F^2$ as soon as $\theta \in N_{r,\alpha}(\theta^*)$.
- (Lem. 5.11) Exploit this result to show that the estimation error must also satisfy $\|\delta\vartheta\|_F \leqslant r^*$ provided that $\max\{h_a, h_b\}$ is small enough.
- (Lem. 5.12) Conclude that the theorem holds with high probability, provided that n is large enough.

For the sake of readability, we refer the reader to the Appendix for the numerous constants that are about to appear in the following lemmas and proofs. Thereafter, $N_{r,\alpha}(\theta^*)$ will always refer to α in (A.4) and r^*

in (A.6), while the second hypothesis (H_2) given below is to be assumed with the smallest integer greater than s_α in (A.5). This is a random hypothesis, which will be controlled with a probability, related to the proximity between the empirical covariance and the true covariance of the predictors, since we recall that $S^{(n)}$ has no reason to be an excellent approximation of Σ^* when $p \gg n$. This is also assumed by the authors of [29], it is a kind of restricted isometry propertie (RIP), well-known in high-dimensional studies. In particular, we will see through Lemma 5.12 that it is satisfied with high probability provided that n is large enough.

$$\forall\, u \neq 0 \text{ such that } |u|_0 \leqslant \lceil s_\alpha \rceil, \qquad \frac{1}{2}\, u^{\,t}\, \Sigma^*_{xx}\, u \;\leqslant\; u^{\,t}\, S^{(n)}_{xx}\, u \;\leqslant\; \frac{3}{2}\, u^{\,t}\, \Sigma^*_{xx}\, u. \tag{H_2}$$

$$\text{In addition,} \qquad \lambda_{\max}(\Omega^*_{yx}\, S^{(n)}_{xx}\, \Omega^{*\,t}_{yx}) \;\leqslant\; \frac{7}{5}\, \lambda_{\max}(\Omega^*_{yx}\, \Sigma^*_{xx}\, \Omega^{*\,t}_{yx}).$$

The next two lemmas give some bounds for expressions that will appear repeatedly.

Lemma 5.7. *Under* (H_1) *and* (H_2)*, for all* $\theta \in N_{r,\alpha}(\theta^*)$*, we have the bound*

$$\lambda_{\max}(\Omega^{-1}_{yy}\, \Omega_{yx}\, S^{(n)}_{xx}\, \Omega^t_{yx}) \;\leqslant\; \overline{\omega}_S$$

where $\overline{\omega}_S$ *is given in* (A.1)*. In addition,*

$$\mathrm{tr}(\delta\theta_{yx}\, S^{(n)}_{xx}\, \delta\theta^t_{yx}) \;\geqslant\; \frac{\lambda_{\min}(\Sigma^*_{xx})}{10}\, \|\delta\theta_{yx}\|^2_F.$$

Proof. Similar reasonings may be found in the proofs of Lemmas 1-2 of [29]. We simply reworked the constants to make them stick to our study. □

Lemma 5.8. *Under* (H_1)*, for all* $\theta \in N_{r,\alpha}(\theta^*)$*, we have the bounds*

$$\lambda_{\min}(\Omega^{-1}_{yy}\, \Omega_{yx}\, L\, \Omega^t_{yx}) \;\geqslant\; \underline{\omega}_L \qquad \textit{and} \qquad \lambda_{\max}(\Omega^{-1}_{yy}\, \Omega_{yx}\, L\, \Omega^t_{yx}) \;\leqslant\; \overline{\omega}_L$$

where $\underline{\omega}_L$ *and* $\overline{\omega}_L$ *are given in* (A.1)*. As a corollary,*

$$p\, \underline{\omega}_L \;\leqslant\; \langle\!\langle L, \Omega^t_{yx}\, \Omega^{-1}_{yy}\, \Omega_{yx} \rangle\!\rangle \;\leqslant\; p\, \overline{\omega}_L.$$

Proof. From Lemmas 5.1 and 5.6,

$$\begin{aligned}
2\, \lambda_{\min}(\Omega_{yx}\, L\, \Omega^t_{yx}) &\geqslant 2\, \big(\lambda_{\min}(\Omega^*_{yx}\, L\, \Omega^{*\,t}_{yx}) + \lambda_{\min}(\delta\theta_{yx}\, L\, \Omega^{*\,t}_{yx} + \Omega^*_{yx}\, L\, \delta\theta^t_{yx})\big) \\
&\geqslant 2\, \big(\lambda_{\min}(\Omega^*_{yx}\, L\, \Omega^{*\,t}_{yx}) - \|\delta\theta_{yx}\, L\, \Omega^{*\,t}_{yx} + \Omega^*_{yx}\, L\, \delta\theta^t_{yx}\|_2\big) \\
&\geqslant 2\, \big(\lambda_{\min}(\Omega^*_{yx}\, L\, \Omega^{*\,t}_{yx}) - 2\, \|\delta\theta_{yx}\|_F\, \|L\, \Omega^{*\,t}_{yx}\|_2\big) \;\;\geqslant\;\; \lambda_{\min}(\Omega^*_{yx}\, L\, \Omega^{*\,t}_{yx})
\end{aligned}$$

as soon as $\|\delta\theta_{yx}\|_F \leqslant r^*$. From Lemma 5.4, we get

$$\lambda_{\min}(\Omega^{-1}_{yy}\, \Omega_{yx}\, L\, \Omega^t_{yx}) \geqslant \frac{\lambda_{\min}(\Omega_{yx}\, L\, \Omega^t_{yx})}{\lambda_{\max}(\Omega_{yy})} \geqslant \frac{\lambda_{\min}(\Omega^*_{yx}\, L\, \Omega^{*\,t}_{yx})}{4\, \lambda_{\max}(\Omega^*_{yy})}$$

where the inequality in the denominator comes from $\lambda_{\max}(\Omega_{yy}) \leqslant \lambda_{\max}(\Omega^*_{yy}) + \lambda_{\max}(\delta\theta_{yy})$, *via* Lemma 5.6, and the fact that $\lambda_{\max}(\delta\theta_{yy}) \leqslant \|\delta\theta_{yy}\|_F \leqslant r^* \leqslant \lambda_{\max}(\Omega^*_{yy})$. For the upper bound, a similar logic gives, with

Lemma 5.5,

$$\begin{aligned}\sqrt{\lambda_{\max}(\Omega_{yx}\,L\,\Omega_{yx}^{t})} &\leqslant \sqrt{\lambda_{\max}(\Omega_{yx}^{*}\,L\,\Omega_{yx}^{*\,t})} + \sqrt{\mathrm{tr}(\delta\theta_{yx}\,L\,\delta\theta_{yx}^{t})}\\ &\leqslant \sqrt{\lambda_{\max}(\Omega_{yx}^{*}\,L\,\Omega_{yx}^{*\,t})} + \|\delta\theta_{yx}\|_F\,\sqrt{\lambda_{\max}(L)} \;\leqslant\; \sqrt{2\,\lambda_{\max}(\Omega_{yx}^{*}\,L\,\Omega_{yx}^{*\,t})}\end{aligned}$$

for $\|\delta\theta_{yx}\|_F \leqslant r^*$. It follows from Lemma 5.4 that

$$\lambda_{\max}(\Omega_{yy}^{-1}\,\Omega_{yx}\,L\,\Omega_{yx}^{t}) \leqslant \frac{\lambda_{\max}(\Omega_{yx}\,L\,\Omega_{yx}^{t})}{\lambda_{\min}(\Omega_{yy})} \leqslant \frac{4\,\lambda_{\max}(\Omega_{yx}^{*}\,L\,\Omega_{yx}^{*\,t})}{\lambda_{\min}(\Omega_{yy}^{*})}$$

where the inequality in the denominator comes from $\lambda_{\min}(\Omega_{yy}) \geqslant \lambda_{\min}(\Omega_{yy}^{*}) + \lambda_{\min}(\delta\theta_{yy})$, *via* Lemma 5.6, and the fact that $2\,\lambda_{\min}(\delta\theta_{yy}) \geqslant -2\,\|\delta\theta_{yy}\|_F \geqslant -2\,r^* \geqslant -\lambda_{\min}(\Omega_{yy}^{*})$. The corollary that concludes the lemma is now immediate. □

Lemma 5.9. *Assume that λ, μ and η are chosen according to the configuration of the theorem. Then, under* (H_1)*, the estimation error satisfies*

$$|[\delta\vartheta]_{\bar{S}}|_1 \leqslant \alpha\,|[\delta\vartheta]_S|_1$$

where $\alpha > 0$ is given in (A.4)*.*

Proof. Taking $t = 1$ in the Taylor expansion (5.4) with $\theta = \widehat{\theta}$ and considering the definition of ϕ in (5.5), by convexity,

$$R_n(\widehat{\theta}) - R_n(\theta^*) \geqslant \phi'(0).$$

The first derivative of ϕ will be explicitely computed in (5.11). For $t = 0$, we find

$$\begin{aligned}\phi'(0) &= -\langle\!\langle \Omega_{yy}^{*\,-1}, \delta\vartheta_{yy}\rangle\!\rangle + \langle\!\langle S_{yy}^{(n)}, \delta\vartheta_{yy}\rangle\!\rangle + 2\,\langle\!\langle S_{yx}^{(n)}, \delta\vartheta_{yx}\rangle\!\rangle\\ &\qquad + 2\,\langle\!\langle S_{xx}^{(n)}, \Omega_{yx}^{*\,t}\,\Omega_{yy}^{*\,-1}\,\delta\vartheta_{yx}\rangle\!\rangle - \langle\!\langle S_{xx}^{(n)}, \Omega_{yx}^{*\,t}\,\Omega_{yy}^{*\,-1}\,\delta\vartheta_{yy}\,\Omega_{yy}^{*\,-1}\,\Omega_{yx}^{*}\rangle\!\rangle\\ &\qquad + \eta\beta\,s_L^{\beta-1}\big[2\,\langle\!\langle L, \Omega_{yx}^{*\,t}\,\Omega_{yy}^{*\,-1}\,\delta\vartheta_{yx}\rangle\!\rangle - \langle\!\langle L, \Omega_{yx}^{*\,t}\,\Omega_{yy}^{*\,-1}\,\delta\vartheta_{yy}\,\Omega_{yy}^{*\,-1}\,\Omega_{yx}^{*}\rangle\!\rangle\big]\\ &= \langle\!\langle A_n + \eta\beta\,s_L^{\beta-1}C_A, \delta\vartheta_{yy}\rangle\!\rangle + \langle\!\langle B_n + \eta\beta\,s_L^{\beta-1}C_B, \delta\vartheta_{yx}\rangle\!\rangle\end{aligned}$$

where s_L is given in (A.3), where, through the blockwise relations (1.3), we recognize the random matrices A_n (with max norm h_a) and B_n (with max norm h_b) defined in (2.3) and (2.4), and where, coming from the structural regularization term,

$$C_A = -\Omega_{yy}^{*\,-1}\,\Omega_{yx}^{*}\,L\,\Omega_{yx}^{*\,t}\,\Omega_{yy}^{*\,-1} \qquad \text{and} \qquad C_B = 2\,\Omega_{yy}^{*\,-1}\,\Omega_{yx}^{*}\,L.$$

Whence it follows from the well-known relation $|\mathrm{tr}(M_1M_2)| \leqslant |M_1|_\infty\,|M_2|_1$, where M_1 and M_2 are compatible matrices, that

$$\phi'(0) \geqslant -\frac{\lambda}{c_\lambda}\,|\delta\vartheta_{yy}|_1 - \eta\beta\,s_L^{\beta-1}\ell_a\,|\delta\vartheta_{yy}|_1 - \frac{\mu}{c_\mu}\,|\delta\vartheta_{yx}|_1 - \eta\beta\,s_L^{\beta-1}\ell_b\,|\delta\vartheta_{yx}|_1$$

making use of the constants (A.3), $\lambda \geqslant c_\lambda\,h_a$ and $\mu \geqslant c_\mu\,h_b$. For the sake of clarity, let

$$\Delta_n(\theta, \theta^*) = R_n(\theta) + \lambda\,|\Omega_{yy}|_1^- + \mu\,|\Omega_{yx}|_1 - R_n(\theta^*) - \lambda\,|\Omega_{yy}^{*}|_1^- - \mu\,|\Omega_{yx}^{*}|_1.$$

For all $\theta \in \Theta$,

$$\begin{aligned}|\Omega_{yy}|_1^- - |\Omega_{yy}^*|_1^- &= |[\Omega_{yy}^* + \delta\theta_{yy}]_S|_1^- + |[\delta\theta_{yy}]_{\bar{S}}|_1^- - |[\Omega_{yy}^*]_S|_1^- \\ &\geqslant \big||[\Omega_{yy}^*]_S|_1^- - |[\delta\theta_{yy}]_S|_1^-\big| + |[\delta\theta_{yy}]_{\bar{S}}|_1^- - |[\Omega_{yy}^*]_S|_1^- \\ &\geqslant |[\delta\theta_{yy}]_{\bar{S}}|_1 - |[\delta\theta_{yy}]_S|_1\end{aligned}$$

from the triangle inequality and the fact that, as Ω_{yy}^* is positive definite, the diagonal must belong to S, *i.e.* $(j,j) \in S$ for all $1 \leqslant j \leqslant q$ so that any square matrix M of size q is such that $[M]_{\bar{S}}$ has diagonal elements all equal to zero. A similar bound obviously holds for $|\Omega_{yx}|_1 - |\Omega_{yx}^*|_1$. Now, a straightforward calculation shows that

$$\Delta_n(\widehat{\theta}, \theta^*) \geqslant \underline{c}\left(|[\delta\vartheta_{yy}]_{\bar{S}}|_1 + |[\delta\vartheta_{yx}]_{\bar{S}}|_1\right) - \overline{c}\left(|[\delta\vartheta_{yy}]_S|_1 + |[\delta\vartheta_{yx}]_S|_1\right) \tag{5.8}$$

where

$$\overline{c} = \max\left\{\frac{(c_\lambda+1)\lambda}{c_\lambda} + \eta\beta\, s_L^{\beta-1}\ell_a, \frac{(c_\mu+1)\mu}{c_\mu} + \eta\beta\, s_L^{\beta-1}\ell_b\right\}$$

and

$$\underline{c} = \min\left\{\frac{(c_\lambda-1)\lambda}{c_\lambda} - \eta\beta\, s_L^{\beta-1}\ell_a, \frac{(c_\mu-1)\mu}{c_\mu} - \eta\beta\, s_L^{\beta-1}\ell_b\right\}.$$

Thus, provided that $\underline{c} > 0$, which is stated in the configuration of the theorem, it only remains to note that, necessarily,

$$\Delta_n(\widehat{\theta}, \theta^*) \leqslant 0$$

since $\widehat{\theta}$ is the global minimizer of $\theta \mapsto R_n(\theta) + \lambda\,|\Omega_{yy}|_1^- + \mu\,|\Omega_{yx}|_1$. The identification of α given in (A.4) easily follows. □

Lemma 5.10. *Under* (H_1) *and* (H_2), *the second-order error term of* (5.4) *satisfies, for* $t = 1$ *and all* $\theta \in N_{r,\alpha}(\theta^*)$,

$$e_1(\theta, \theta^*) > \gamma_{r,\eta,\beta,p}\,\|\delta\theta\|_F^2$$

where $\gamma_{r,\eta,\beta,p} > 0$ *is given in* (A.7).

Proof. From the definition of ϕ in (5.5) and the fact that $\phi'(0) = \langle\!\langle \nabla R_n(\theta^*), \theta - \theta^* \rangle\!\rangle$, there exists $h \in\,]0,1[$ satisfying

$$e_1(\theta, \theta^*) = \frac{1}{2}\,\phi''(h). \tag{5.9}$$

To simplify the calculations, let

$$u_L = \langle\!\langle L, \Omega_{yx}^t\,\Omega_{yy}^{-1}\,\Omega_{yx} \rangle\!\rangle. \tag{5.10}$$

We are going to study the behavior of $R_n(\Omega_{yy}, \Omega_{yx})$ in the directions $\Omega_{yy} = \Omega_{yy}^* + t\,\delta\theta_{yy}$ and $\Omega_{yx} = \Omega_{yx}^* + t\,\delta\theta_{yx}$ through $\phi(t)$, where we recall that $\delta\theta_{yy} = \Omega_{yy} - \Omega_{yy}^*$ and $\delta\theta_{yx} = \Omega_{yx} - \Omega_{yx}^*$. One can see that $\phi(t)$ moves from

$R_n(\Omega_{yy}, \Omega_{yx})$ to $R_n(\Omega^*_{yy}, \Omega^*_{yx})$ as t decreases from 1 to 0. The first derivative is

$$\begin{aligned}\phi'(t) = &-\langle\!\langle \Omega_{yy}^{-1}, \delta\theta_{yy}\rangle\!\rangle + \langle\!\langle S_{yy}^{(n)}, \delta\theta_{yy}\rangle\!\rangle + 2\,\langle\!\langle S_{yx}^{(n)}, \delta\theta_{yx}\rangle\!\rangle \\ &+ 2\,\langle\!\langle S_{xx}^{(n)}, \Omega_{yx}^t\,\Omega_{yy}^{-1}\,\delta\theta_{yx}\rangle\!\rangle - \langle\!\langle S_{xx}^{(n)}, \Omega_{yx}^t\,\Omega_{yy}^{-1}\,\delta\theta_{yy}\,\Omega_{yy}^{-1}\,\Omega_{yx}\rangle\!\rangle \\ &+ \eta\beta\,u_L^{\beta-1}\big[2\,\langle\!\langle L, \Omega_{yx}^t\,\Omega_{yy}^{-1}\,\delta\theta_{yx}\rangle\!\rangle - \langle\!\langle L, \Omega_{yx}^t\,\Omega_{yy}^{-1}\,\delta\theta_{yy}\,\Omega_{yy}^{-1}\,\Omega_{yx}\rangle\!\rangle\big]. \end{aligned} \tag{5.11}$$

The second derivative is tedious to write but straightforward to establish,

$$\begin{aligned}\phi''(t) = &\langle\!\langle \Omega_{yy}^{-1}, \delta\theta_{yy}\,\Omega_{yy}^{-1}\,\delta\theta_{yy}\rangle\!\rangle + 2\,\big[\langle\!\langle S_{xx}^{(n)}, \delta\theta_{yx}^t\,\Omega_{yy}^{-1}\,\delta\theta_{yx}\rangle\!\rangle - 2\,\langle\!\langle S_{xx}^{(n)}, \Omega_{yx}^t\,\Omega_{yy}^{-1}\,\delta\theta_{yy}\,\Omega_{yy}^{-1}\,\delta\theta_{yx}\rangle\!\rangle \\ &+ \langle\!\langle S_{xx}^{(n)}, \Omega_{yx}^t\,\Omega_{yy}^{-1}\,\delta\theta_{yy}\,\Omega_{yy}^{-1}\,\delta\theta_{yy}\,\Omega_{yy}^{-1}\,\Omega_{yx}\rangle\!\rangle\big] \\ &+ 2\,\eta\beta\,u_L^{\beta-1}\big[\langle\!\langle L, \delta\theta_{yx}^t\,\Omega_{yy}^{-1}\,\delta\theta_{yx}\rangle\!\rangle - 2\,\langle\!\langle L, \Omega_{yx}^t\,\Omega_{yy}^{-1}\,\delta\theta_{yy}\,\Omega_{yy}^{-1}\,\delta\theta_{yx}\rangle\!\rangle \\ &+ \langle\!\langle L, \Omega_{yx}^t\,\Omega_{yy}^{-1}\,\delta\theta_{yy}\,\Omega_{yy}^{-1}\,\delta\theta_{yy}\,\Omega_{yy}^{-1}\,\Omega_{yx}\rangle\!\rangle\big] \\ &+ \eta\beta(\beta-1)\,u_L^{\beta-2}\big[2\,\langle\!\langle L, \Omega_{yx}^t\,\Omega_{yy}^{-1}\,\delta\theta_{yx}\rangle\!\rangle - \langle\!\langle L, \Omega_{yx}^t\,\Omega_{yy}^{-1}\,\delta\theta_{yy}\,\Omega_{yy}^{-1}\,\Omega_{yx}\rangle\!\rangle\big]^2. \end{aligned} \tag{5.12}$$

First, from the combination of Lemmas 5.1 and 5.8, we clearly have $u_L \geqslant 0$. We also note that $0 \leqslant \|\frac{2}{c}M_1 - cM_2\|_F^2 = \frac{4}{c^2}\,\|M_1\|_F^2 - 4\,\langle\!\langle M_1, M_2\rangle\!\rangle + c^2\,\|M_2\|_F^2$ for any $c \neq 0$ and any matrices M_1 and M_2 of same dimensions. It follows, after some reorganizations, that for any $c \neq 0$ and $d \neq 0$,

$$\begin{aligned}\phi''(t) \geqslant &\langle\!\langle \Omega_{yy}^{-1}, \delta\theta_{yy}\,\Omega_{yy}^{-1}\,\delta\theta_{yy}\rangle\!\rangle \\ &+ c_1\,\langle\!\langle \Omega_{yy}^{-1}, \delta\theta_{yx}\,S_{xx}^{(n)}\,\delta\theta_{yx}^t\rangle\!\rangle + c_2\,\langle\!\langle S_{xx}^{(n)}, \Omega_{yx}^t\,\Omega_{yy}^{-1}\,\delta\theta_{yy}\,\Omega_{yy}^{-1}\,\delta\theta_{yy}\,\Omega_{yy}^{-1}\,\Omega_{yx}\rangle\!\rangle \\ &+ \eta\beta\,u_L^{\beta-1}\big[d_1\,\langle\!\langle \Omega_{yy}^{-1}, \delta\theta_{yx}\,L\,\delta\theta_{yx}^t\rangle\!\rangle + d_2\,\langle\!\langle L, \Omega_{yx}^t\,\Omega_{yy}^{-1}\,\delta\theta_{yy}\,\Omega_{yy}^{-1}\,\delta\theta_{yy}\,\Omega_{yy}^{-1}\,\Omega_{yx}\rangle\!\rangle\big] \end{aligned}$$

where $c_1 = 2 - \frac{4}{c^2}$, $c_2 = 2 - c^2$, $d_1 = 2 - \frac{4}{d^2}$ and $d_2 = 2 - d^2$. Here we exploited the previous inequality twice, $u_L \geqslant 0$ and $\beta \geqslant 1$. From Lemmas 5.1, 5.3, 5.7 and 5.8, using $\mathrm{sp}(M_1\,M_2) = \mathrm{sp}(M_2\,M_1)$ for square matrices M_1 and M_2, we obtain

$$\langle\!\langle L, \Omega_{yx}^t\,\Omega_{yy}^{-1}\,\delta\theta_{yy}\,\Omega_{yy}^{-1}\,\delta\theta_{yy}\,\Omega_{yy}^{-1}\,\Omega_{yx}\rangle\!\rangle \leqslant \overline{\omega}_L\,\langle\!\langle \Omega_{yy}^{-1}, \delta\theta_{yy}\,\Omega_{yy}^{-1}\,\delta\theta_{yy}\rangle\!\rangle$$

where $\overline{\omega}_L$ is defined in (A.1). Replacing L by $S_{xx}^{(n)}$ and $\overline{\omega}_L$ by $\overline{\omega}_S$, a similar bound obviously holds. Suppose that c and d are chosen so that $c_1 > 0$, $d_1 > 0$, $c_2 < 0$ and $d_2 < 0$. Then,

$$\begin{aligned}\phi''(t) &\geqslant \langle\!\langle \Omega_{yy}^{-1}, \delta\theta_{yy}\,\Omega_{yy}^{-1}\,\delta\theta_{yy}\rangle\!\rangle\,\big[1 - |c_2|\,\overline{\omega}_S - \eta\beta\,u_L^{\beta-1}|d_2|\,\overline{\omega}_L\big] \\ &\qquad + c_1\,\langle\!\langle \Omega_{yy}^{-1}, \delta\theta_{yx}\,S_{xx}^{(n)}\,\delta\theta_{yx}^t\rangle\!\rangle + \eta\beta\,u_L^{\beta-1}d_1\,\langle\!\langle \Omega_{yy}^{-1}, \delta\theta_{yx}\,L\,\delta\theta_{yx}^t\rangle\!\rangle \\ &\geqslant \langle\!\langle \Omega_{yy}^{-1}, \delta\theta_{yy}\,\Omega_{yy}^{-1}\,\delta\theta_{yy}\rangle\!\rangle\,\big[1 - |c_2|\,\overline{\omega}_S - \eta\beta\,(p\,\overline{\omega}_L)^{\beta-1}|d_2|\,\overline{\omega}_L\big] \\ &\qquad + c_1\,\langle\!\langle \Omega_{yy}^{-1}, \delta\theta_{yx}\,S_{xx}^{(n)}\,\delta\theta_{yx}^t\rangle\!\rangle + \eta\beta\,(p\,\underline{\omega}_L)^{\beta-1}d_1\,\langle\!\langle \Omega_{yy}^{-1}, \delta\theta_{yx}\,L\,\delta\theta_{yx}^t\rangle\!\rangle. \end{aligned}$$

Now choose $\epsilon_S > 0$ and $\epsilon_L > 0$ small enough so that $\epsilon_S\,\overline{\omega}_S + \eta\beta\,p^{\beta-1}\,\overline{\omega}_L^{\beta}\,\epsilon_L < 1$ and fix $c = \sqrt{2+\epsilon_S}$ and $d = \sqrt{2+\epsilon_L}$. We finally obtain

$$\phi''(t) \geqslant a_1\,\langle\!\langle \Omega_{yy}^{-1}, \delta\theta_{yy}\,\Omega_{yy}^{-1}\,\delta\theta_{yy}\rangle\!\rangle + a_2\,\langle\!\langle \Omega_{yy}^{-1}, \delta\theta_{yx}\,S_{xx}^{(n)}\,\delta\theta_{yx}^t\rangle\!\rangle + a_3\,\langle\!\langle \Omega_{yy}^{-1}, \delta\theta_{yx}\,L\,\delta\theta_{yx}^t\rangle\!\rangle \tag{5.13}$$

where these positive constants are respectively given by

$$a_1 = 1 - \epsilon_S\,\overline{\omega}_S - \eta\beta\,p^{\beta-1}\,\overline{\omega}_L^{\beta}\,\epsilon_L, \qquad a_2 = \frac{2\,\epsilon_S}{2+\epsilon_S} \qquad \text{and} \qquad a_3 = \eta\beta\,(p\,\underline{\omega}_L)^{\beta-1}\,\frac{2\,\epsilon_L}{2+\epsilon_L}.$$

The combination of Lemmas 5.1, 5.3 and 5.5 gives, uniformly in $t \in [0,1]$,

$$\langle\!\langle \Omega_{yy}^{-1}, \delta\theta_{yy}\, \Omega_{yy}^{-1}\, \delta\theta_{yy} \rangle\!\rangle \geqslant \lambda_{\min}(\Omega_{yy}^{-1})\, \mathrm{tr}(\delta\theta_{yy}\, \Omega_{yy}^{-1}\, \delta\theta_{yy}) \geqslant \frac{\|\delta\theta_{yy}\|_F^2}{4\, \lambda_{\max}^2(\Omega_{yy}^*)}$$

where the inequality in the denominator comes from $\lambda_{\max}(\Omega_{yy}) \leqslant 2\, \lambda_{\max}(\Omega_{yy}^*)$ already established in the proof of Lemma 5.8. Similarly,

$$\langle\!\langle \Omega_{yy}^{-1}, \delta\theta_{yx}\, L\, \delta\theta_{yx}^t \rangle\!\rangle \geqslant \lambda_{\min}(\Omega_{yy}^{-1})\, \mathrm{tr}(\delta\theta_{yx}\, L\, \delta\theta_{yx}^t) \geqslant \frac{\lambda_{\min}(L)\, \|\delta\theta_{yx}\|_F^2}{2\, \lambda_{\max}(\Omega_{yy}^*)}.$$

Lemma 5.7 directly enables to bound the last term,

$$\langle\!\langle \Omega_{yy}^{-1}, \delta\theta_{yy}\, S_{xx}^{(n)}\, \delta\theta_{yy} \rangle\!\rangle \geqslant \lambda_{\min}(\Omega_{yy}^{-1})\, \mathrm{tr}(\delta\theta_{yx}\, S_{xx}^{(n)}\, \delta\theta_{yx}^t) \geqslant \frac{\lambda_{\min}(\Sigma_{xx}^*)\, \|\delta\theta_{yx}\|_F^2}{20\, \lambda_{\max}(\Omega_{yy}^*)}.$$

In conclusion, combining (5.9), (5.13) and the upper bounds above,

$$\begin{aligned} e_1(\theta, \theta^*) &\geqslant \frac{a_1\, \|\delta\theta_{yy}\|_F^2}{8\, \lambda_{\max}^2(\Omega_{yy}^*)} + \frac{a_2\, \lambda_{\min}(L)\, \|\delta\theta_{yx}\|_F^2}{4\, \lambda_{\max}(\Omega_{yy}^*)} + \frac{a_3\, \lambda_{\min}(\Sigma_{xx}^*)\, \|\delta\theta_{yx}\|_F^2}{40\, \lambda_{\max}(\Omega_{yy}^*)} \\ &\geqslant \min\left\{ \frac{a_1}{8\, \lambda_{\max}^2(\Omega_{yy}^*)}, \frac{a_2\, \lambda_{\min}(L)}{4\, \lambda_{\max}(\Omega_{yy}^*)} + \frac{a_3\, \lambda_{\min}(\Sigma_{xx}^*)}{40\, \lambda_{\max}(\Omega_{yy}^*)} \right\} \|\delta\theta\|_F^2 \end{aligned}$$

and we clearly identify $\gamma_{r,\eta,\beta,p} > 0$. □

Lemma 5.11. *Assume that λ, μ and η are chosen according to the configuration of the theorem. Suppose also that h_a in (2.3) and h_b in (2.4) satisfy*

$$\max\{h_a, h_b\} < \frac{r^*\, \gamma_{r,\eta,\beta,p}}{c_{\lambda,\mu} \sqrt{|S|}}$$

where r^* is given in (A.6), $\gamma_{r,\eta,\beta,p}$ in (A.7) and $c_{\lambda,\mu}$ in (A.8). Then, under (H_1) and (H_2), the estimation error satisfies $\|\delta\vartheta\|_F \leqslant r^*$.

Proof. By convexity of the objective and optimality of $\widehat{\theta}$, each move from θ^* in the direction $t\, \delta\vartheta$ for $t \in [0,1]$ must lead to a decrease of the objective, *i.e.*

$$R_n(\theta^* + t\, \delta\vartheta) + \lambda\, |\Omega_{yy}^* + t\, \delta\vartheta_{yy}|_1^- + \mu\, |\Omega_{yx}^* + t\, \delta\vartheta_{yx}|_1 - R_n(\theta^*) - \lambda\, |\Omega_{yy}^*|_1^- - \mu\, |\Omega_{yx}^*|_1 \leqslant 0.$$

Taking the notation of (5.8), this can be rewritten as $\Delta_n(\theta^* + t\, \delta\vartheta, \theta^*) \leqslant 0$. If $\|\delta\vartheta\|_F \leqslant r^*$ then choose $t = 1$, otherwise calibrate $0 < t < 1$ such that $\|t\, \delta\vartheta\|_F = r^*$. Then, from Lemma 5.9, it clearly follows that $\theta^* + t\, \delta\vartheta \in N_{r,\alpha}(\theta^*)$. Hence, the reasoning preceding (5.8) still holds and, together with Lemma 5.10, we obtain

$$\begin{aligned} 0 &\geqslant \underline{c}\, \big(|[t\, \delta\vartheta_{yy}]_{\bar{S}}|_1 + |[t\, \delta\vartheta_{yx}]_{\bar{S}}|_1 \big) - \overline{c}\, \big(|[t\, \delta\vartheta_{yy}]_S|_1 + |[t\, \delta\vartheta_{yx}]_S|_1 \big) + \gamma_{r,\eta,\beta,p}\, \|t\, \delta\vartheta\|_F^2 \\ &\geqslant -\overline{c}\, |[t\, \delta\vartheta]_S|_1 + \gamma_{r,\eta,\beta,p}\, \|t\, \delta\vartheta\|_F^2 \\ &\geqslant -c_{\lambda,\mu} \max\{h_a, h_b\} \sqrt{|S|}\, \|t\, \delta\vartheta\|_F + \gamma_{r,\eta,\beta,p}\, \|t\, \delta\vartheta\|_F^2 \end{aligned}$$

where we used $\underline{c} > 0$ and Cauchy-Schwarz inequality to get $|[\cdot]_S|_1^2 \leqslant |S|\, \|[\cdot]_S\|_F^2$. The constant $c_{\lambda,\mu}$ may be explicitely computed from the configuration of (λ, μ, η) and is given in (A.8). Note that in the proof of Lemma 5.9,

it was sufficient to see that $R_n(\theta) - R_n(\theta^*) \geqslant \phi'(0)$ whereas here, we must consider $R_n(\theta) - R_n(\theta^*) = \phi'(0) + e_1(\theta, \theta^*)$ to meet our purposes. That explains the presence of $\gamma_{r,\eta,\beta,p} \, \|t\, \delta\vartheta\|_F^2$ in the inequality. We deduce that the error must satisfy

$$\|t\, \delta\vartheta\|_F \leqslant \frac{c_{\lambda,\mu}\sqrt{|S|}\max\{h_a, h_b\}}{\gamma_{r,\eta,\beta,p}}.$$

As a corollary, it holds that $\|\delta\vartheta\|_F > r^* \Rightarrow c_{\lambda,\mu}\sqrt{|S|}\max\{h_a, h_b\} \geqslant r^*\, \gamma_{r,\eta,\beta,p}$ or, conversely written, $c_{\lambda,\mu}\sqrt{|S|}\max\{h_a, h_b\} < r^*\, \gamma_{r,\eta,\beta,p} \Rightarrow \|\delta\vartheta\|_F \leqslant r^*$. □

Lemma 5.12. *Assume that λ, μ and η are chosen according to the configuration of the theorem. Then, under* (H_1)*, there exists absolute constants $b_1 > 0$ and $b_2 > 0$ such that, for any $b_3 \in]0, 1[$ and as soon as*

$$n \geqslant \max\left\{b_1\,(q + \lceil s_\alpha \rceil \ln(p+q)),\, \ln(10(p+q)^2) - \ln(b_3)\right\},$$

with probability no less that $1 - \mathrm{e}^{-b_2 n} - b_3$ both the random hypothesis (H_2) *is satisfied and the upper bound*

$$\max\{h_a, h_b\} \leqslant 16\, m^* \sqrt{\frac{\ln(10(p+q)^2) - \ln(b_3)}{n}}$$

holds, where h_a and h_b are given in (2.3) *and* (2.4)*, s_α is defined in* (A.5) *and m^* in* (A.9)*. Hence, one can find a minimal number of observations n_0 such that the theorem holds with high probability as soon as $n > n_0$.*

Proof. All the ingredients of the proof are established in [29]. The authors start by recalling that there exists absolute constants $b_1 > 0$ and $b_2 > 0$ such that hypothesis (H_2) is satisfied with probability no less than $1 - \mathrm{e}^{-b_2 n}$ as soon as $n \geqslant b_1\,(q + \lceil s_\alpha \rceil \ln(p+q))$. We also refer the reader to Lemma 5.1 and Theorem 5.2 of [3], or to Lemma 7.4 of [10] for the random bounds of the restricted isometry constants. Afterwards, they prove (see Prop. 4) that, as soon as $n \geqslant \ln(10(p+q)^2) - \ln(b_3)$ for some $b_3 > 0$, with probability $1 - b_3$,

$$\max\{h_a, h_b\} \leqslant 16\, m^* \sqrt{\frac{\ln(10(p+q)^2) - \ln(b_3)}{n}}.$$

To find the minimal number of observations, we just need to make sure that the above bound is itself smaller than the one of Lemma 5.11. It is then not hard to see that we may retain the minimal size n_0 given in (2.6). □

Appendix A. Some constants

This appendix is entirely dedicated to the constants appearing in the theoretical guarantees. Indeed, a centralization seemed necessary to clarify the rest of the paper, especially the understanding of the main theorem. First, we need to define some constants related to L and to the true values of the model. The bounds

$$\underline{\omega}_L = \frac{\lambda_{\min}(\Omega^*_{yx}\, L\, \Omega^{*\,t}_{yx})}{4\,\lambda_{\max}(\Omega^*_{yy})}, \qquad \overline{\omega}_L = \frac{4\,\lambda_{\max}(\Omega^*_{yx}\, L\, \Omega^{*\,t}_{yx})}{\lambda_{\min}(\Omega^*_{yy})}, \qquad \overline{\omega}_S = \frac{4\,\lambda_{\max}(\Omega^*_{yx}\, \Sigma^*_{xx}\, \Omega^{*\,t}_{yx})}{\lambda_{\min}(\Omega^*_{yy})} \tag{A.1}$$

are useful to control the eigenvalues of some recurrent expressions (Lems. 5.7 and 5.8), uniformly in a neighborhood of $\theta^* = (\Omega^*_{yy}, \Omega^*_{yx})$. The true value of the term at the heart of the structural regularization is

$$s_L = \langle\!\langle L, \Omega^{*\,t}_{yx}\, \Omega^{*\,-1}_{yy}\, \Omega^*_{yx} \rangle\!\rangle. \tag{A.2}$$

It plays a role in the proof of Lemma 5.9 and, as a consequence, in the definition of the area of validity Λ. This important lemma also requires to define

$$\ell_a = |\Omega_{yy}^{*\,-1}\,\Omega_{yx}^{*}\,L\,\Omega_{yx}^{*\,t}\,\Omega_{yy}^{*\,-1}|_\infty \quad \text{and} \quad \ell_b = 2\,|\Omega_{yy}^{*\,-1}\,\Omega_{yx}^{*}\,L|_\infty \tag{A.3}$$

and, in the context of the theorem,

$$\alpha = \frac{\max\left\{\frac{(c_\lambda+1)\lambda}{c_\lambda} + \eta\beta\, s_L^{\beta-1}\ell_a,\, \frac{(c_\mu+1)\mu}{c_\mu} + \eta\beta\, s_L^{\beta-1}\ell_b\right\}}{\min\left\{\frac{(c_\lambda-1)\lambda}{c_\lambda} - \eta\beta\, s_L^{\beta-1}\ell_a,\, \frac{(c_\mu-1)\mu}{c_\mu} - \eta\beta\, s_L^{\beta-1}\ell_b\right\}}. \tag{A.4}$$

From α and the cardinality of the true active set $|S|$, let

$$s_\alpha = |S|\left[1 + \frac{12\,\alpha^2\,\lambda_{\max}(\Sigma_{xx}^*)}{\lambda_{\min}(\Sigma_{xx}^*)}\right] \tag{A.5}$$

which serves as an upper bound in the random hypothesis (H_2). Similarly, let

$$r^* = \min\{r_1^*, r_2^*, r_3^*, r_4^*\} \tag{A.6}$$

where

$$r_1^* = \frac{\lambda_{\min}(\Omega_{yy}^*)}{2}, \qquad r_2^* = \frac{\frac{\sqrt{10}-\sqrt{7}}{\sqrt{5}}\sqrt{\lambda_{\max}(\Omega_{yx}^*\,\Sigma_{xx}^*\,\Omega_{yx}^{*\,t})}}{\frac{3\sqrt{3}}{2\sqrt{2}}\sqrt{\lambda_{\max}(\Sigma_{xx}^*)}}, \qquad r_3^* = \frac{\lambda_{\min}(\Omega_{yx}^*\,L\,\Omega_{yx}^{*\,t})}{4\,\|L\,\Omega_{yx}^{*\,t}\|_2}$$

and

$$r_4^* = \frac{(\sqrt{2}-1)\sqrt{\lambda_{\max}(\Omega_{yx}^*\,L\,\Omega_{yx}^{*\,t})}}{\sqrt{\lambda_{\max}(L)}}.$$

Together with α given above, r^* is necessary to build the so-called neighborhood $N_{r,\alpha}(\theta^*)$ defined in (5.7), which plays a fundamental role in all our reasonings. It is important to note that, under the configuration of the theorem and hypothesis (H_1), $\alpha > 0$ and $r^* > 0$. Then, Lemma 5.10 highlights a new constant, characterizing a strong local convexity of the smooth part of the objective in the neighborhood $N_{r,\alpha}(\theta^*)$,

$$\gamma_{r,\eta,\beta,p} = \min\left\{\frac{a_1}{8\,\lambda_{\max}^2(\Omega_{yy}^*)},\, \frac{a_2\,\lambda_{\min}(L)}{4\,\lambda_{\max}(\Omega_{yy}^*)} + \frac{a_3\,\lambda_{\min}(\Sigma_{xx}^*)}{40\,\lambda_{\max}(\Omega_{yy}^*)}\right\} \tag{A.7}$$

where, as it is detailed in the proof of the lemma in question,

$$a_1 = 1 - \epsilon_S\,\overline{\omega}_S - \eta\beta\,p^{\beta-1}\,\overline{\omega}_L^{\beta}\,\epsilon_L, \qquad a_2 = \frac{2\,\epsilon_S}{2+\epsilon_S} \quad \text{and} \quad a_3 = \eta\beta\,(p\,\underline{\omega}_L)^{\beta-1}\,\frac{2\,\epsilon_L}{2+\epsilon_L}$$

for some well-chosen $\epsilon_S > 0$ and $\epsilon_L > 0$. Here again, we make sure that $\gamma_{r,\eta,\beta,p} > 0$. In the same way, in the context of the theorem,

$$c_{\lambda,\mu} = \max\left\{\frac{(c_\lambda+1)\,d_\lambda}{c_\lambda} + e_\lambda,\, \frac{(c_\mu+1)\,d_\mu}{c_\mu} + e_\mu\right\} \tag{A.8}$$

is needed through Lemma 5.11. Finally, independently of the structure matrix L,

$$m^* = |\mathrm{diag}(\Sigma^*_{xx})|_\infty + |\mathrm{diag}(\Omega^{*\,-1}_{yy}\,\Omega^*_{yx}\,\Sigma^*_{xx}\,\Omega^{*\,t}_{yx}\,\Omega^{*\,-1}_{yy})|_\infty \tag{A.9}$$

is going to play a significative role in the upper bound of the theorem.

Acknowledgements and Fundings. The authors warmly thank the two anonymous reviewers for the careful reading and for making numerous useful corrections to improve the paper. We thank ALM (Angers Loire Métropole) and the ICO (Institut de Cancérologie de l'Ouest) for the financial support. This work is partially financed through the ALM grant and the "Programme opérationnel régional FEDER-FSE Pays de la Loire 2014-2020" noPL0015129 (EPICURE). The authors also thank Mario Campone (project leader and director of the ICO), Mathilde Colombié (scientific coordinator of EPICURE clinical trial) and Fadwa Ben Azzouz, biomathematician in Bioinfomics, for the initiation, the coordination and the smooth running of the project.

References

[1] G. Andrew and J. Gao, Scalable training of L_1-regularized log-linear models. *Proc. 24th Inte. Conf. Mach. Learning* (2007) 33–40.

[2] O. Banerjee, L. El Ghaoui and A. D'Aspremont, Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** (2008) 485–516.

[3] R. Baraniuk, M. Davenport, R. DeVore and M. Wakin, A simple proof of the restricted isometry property for random matrices. *Constr. Approx.* **28** (2008) 253–263.

[4] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge University Press (2004).

[5] T. Cai, H. Li, W. Liu and J. Xie, Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika* **100** (2013) 139–156.

[6] T. Cai, W. Liu and X. Luo, A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Am. Stat. Assoc.* **106** (2011) 594–607.

[7] J. Chiquet, T. Mary-Huard and S. Robin, Structured regularization for conditional Gaussian graphical models. *Stat. Comput.* **27** (2017) 789–804.

[8] J. Fan, Y. Feng and Y. Wu, Network exploration via the adaptive Lasso and SCAD penalties. *Ann. Appl. Stat.* **3** (2009) 521–541.

[9] J. Friedman, T. Hastie and R. Tibshirani, Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* **9** (2008) 432–441.

[10] C. Giraud, Introduction to High-Dimensional Statistics. *Chapman & Hall/CRC Monographs on Statistics & Applied Probability.* Taylor & Francis (2014).

[11] T. Hastie, R. Tibshirani and M. Wainwright, Statistical Learning with Sparsity: The Lasso and Generalizations. *Chapman & Hall/CRC Monographs on Statistics and Applied Probability.* CRC Press (2015).

[12] R.A. Horn and C.R. Johnson, Matrix Analysis (Second Edition). Cambridge University Press, Cambridge, New York (2012).

[13] C. Johnson, A. Jalali and P. Ravikumar, High-dimensional sparse inverse covariance estimation using greedy methods. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research.* PMLR (2012) 574–582.

[14] W. Lee and Y. Liu, Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood. *J. Multivariate. Anal.* **111** (2012) 241–255.

[15] Z. Lu, Smooth optimization approach for sparse covariance selection. *Siam. J. Optimiz.* **19** (2009) 1807–1827.

[16] M. Maathuis, M. Drton, S.L. Lauritzen and M. Wainwright, Handbook of Graphical Models. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods.* CRC Press (2018).

[17] N. Meinshausen and P. Bühlmann, High-dimensional graphs and variable selection with the Lasso. *Ann. Stat.* **34** (2006) 1436–1462.

[18] F. Pascal, L. Bombrun, J.Y. Tourneret and Y. Berthoumieu, Parameter estimation for multivariate generalized Gaussian distributions. *IEEE Trans. Signal. Process.* **61** (2013) 5960–5971.

[19] J. Peng, P. Wang, N. Zhou and J. Zhu, Partial correlation estimation by joint sparse regression models. *J. Am. Stat. Assoc.* **104** (2009) 735–746.

[20] J. Ramsay and B. Silverman, Functional Data Analysis, 2nd ed. Springer, New York (2006).

[21] P. Ravikumar, M. Wainwright, G. Raskutti and B. Yu, High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electr. J. Stat.* **5** (2011) 935–980.

[22] P. Rossi, G. Allenby and R. McCulloch, Bayesian Statistics and Marketing. *Wiley Series in Probability and Statistics.* Wiley (2012).

[23] A. Rothman, E. Levina and J. Zhu, Sparse multivariate regression with covariance estimation. *J. Comput. Graph. Stat.* **19** (2010) 947–962.

[24] M. Slawski, The structured elastic net for quantile regression and support vector classification. *Stat. Comput.* **22** (2012) 153–168.

[25] M. Slawski, W. Zu Castell and G. Tutz, Feature selection guided by structural information. *Ann. Appl. Stat.* **4** (2010) 1056–1080.

[26] K.A. Sohn and S. Kim, Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics.* Vol. 22 of *Proceedings of Machine Learning Research.* PMLR (2012) 1081–1089.

[27] J. Yin and H. Li, A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Stat.* **5** (2011) 2630–2650.

[28] M. Yuan and Y. Lin, Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** (2007) 19–35.

[29] X.T. Yuan and T. Zhang, Partial Gaussian graphical model estimation. *IEEE Trans. Inf. Theory* **60** (2014) 1673–1687.