Graph estimation for Gaussian data zero-inflated by double truncation

Anne Gégout-Petit, Aurélie Muller-Gueudin, Clémence Karmann

Graph estimation for Gaussian data zero-inflated by double truncation.

Anne Gégout-Petit[a], Aurélie Gueudin-Muller[a] & Clémence Karmann[a]

[a] *Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France, Inria BIGS Team ; anne.gegout-petit@univ-lorraine.fr, aurelie.gueudin@univ-lorraine.fr, clemence.karmann@inria.fr*

Abstract. We consider the problem of graph estimation in a zero-inflated Gaussian model. In this model, zero-inflation is obtained by double truncation (right and left) of a Gaussian vector. The goal is to recover the latent graph structure of the Gaussian vector with observations of the zero-inflated truncated vector. We propose a two step estimation procedure. The first step consists in estimating each term of the covariance matrix by maximising the corresponding bivariate marginal log-likelihood of the truncated vector. The second one uses the graphical lasso procedure to estimate the precision matrix sparsity, which encodes the graph structure. We then state some theoretical convergence results about the convergence rate of the covariance matrix and precision matrix estimators. These results allow us to establish consistency of our procedure with respect to graph structure recovery. We also present some simulation studies to corroborate the efficiency of our procedure.

Keywords. doubly truncated Gaussian; zero-inflation; graph estimation; precision matrix; Gaussian graphical model; graphical lasso; sparsity.

1. Introduction

Multivariate data analysis often involves describing and explaining the relationships among a set of variables. Undirected graphical models offer a way to address this demand by using a graph to represent a model. A graph is a set of nodes and edges which can be represented as a graphic in order to make it easier to study, visually or computationally. Undirected graphical models are based on the conditional independence: a relation between two variables, represented by an edge in the graph, means that the corresponding variables are conditionally dependent given all the remaining variables. Among undirected graphical models, Gaussian graphical model provides a particularly convenient framework. This model assumes that the observations have a multivariate Gaussian distribution with mean μ and covariance matrix Σ. In this Gaussian setting, a direct relation between two variables corresponds to a non-zero entry in the precision matrix Σ^{-1}. In other words, if Σ^{-1}_{ij} is zero, then variables i and j are conditionally independent given the other variables. Thus, graph estimation involves finding the zero pattern in the inverse covariance matrix. The theoretical approaches developed solve a

maximum likelihood problem with an added L_1 penalty on the precision matrix to increase sparsity of the resulting graph. Many authors like Yuan, M. and Lin, Y. (2007), Dahl, J. and Vandenberghe, L. and Roychowdhury, V. (2008) or Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008) used interior point methods to solve the exact maximisation of the L_1 penalised log-likelihood. One of the most powerful approach is the graphical lasso of Friedman, J., Hastie, T. and Tibshirani, R. (2008), who used a blockwise coordinate descent approach.

Furthermore, truncated Gaussian distributions received much attention in the second half of the last century. Cohen (Cohen, Jr., A. C. (1949), Cohen, Jr., A. C. (1950), Cohen, Jr., A. C. (1957)) studied extensively mean and standard deviation estimation with likelihood maximisation for univariate doubly truncated Gaussian. These data correspond to Gaussian distributions which fall between two points of truncation a and b, $a < b$. He distinguished cases when the number of "unmeasured" observations, that is observations which fall into the tail(s), is known or not and whether we know the number in each tail. Shah, S. M. and Jaiswal, M. C. (1966) also studied this case by estimating parameters from first four sample moments.
Bivariate case was then naturally studied. Raj D. (1953) and Cohen, Jr., A. C. (1955) studied mean, variance and covariance estimation when only one of the variables is truncated whereas Nath, G. B. (1966), Dyer, D. D. (1973) then Muthén, B. (1990) analysed it when both variables are truncated. In all these papers, as soon as one of the variables falls outside its points of truncation, none of the variables of the bivector is observed. In others words, data samples are only constituted of the Gaussian data for which the two variables are observed.
About multivariate case, Cohen, Jr., A. C. (1957) estimate model parameters where only one variable is truncated by likelihood maximisation. Singh, N. (1960) considers means and variances estimation in the case where only some variables are truncated. Later, Gupta, A. K. and Tracy, D. S. (1976), Lee, L. (1983) and Manjunath, BG and Wilhelm, S. (2009) studied moments when all the variables of the Gaussian vector are doubly truncated, that is, when all the variables of the Gaussian vector fall inside their points of truncation. Graph estimation and matrix covariance estimation do not seem to have yet been discussed in the literature.

In this paper, we address the problem of graph estimation in a zero-inflated Gaussian model. In this model, zero-inflation is obtained by double truncation (right and left) of a Gaussian vector. This means that each of the Gaussian variables are normally observed inside its points of truncation, but is null otherwise. If a variable is truncated, we then observe a zero instead of its value, but we still observe the other variables of the vector, contrary to the literature. Our goal is to recover the latent graph structure of the Gaussian vector, encoded in the precision matrix, with observations of the zero-inflated truncated vector. To retrieve this theoretical graph structure, we use the graphical lasso procedure which involves the empirical covariance matrix. Unlike the classic Gaussian setting, the Gaussian vector is not directly observed in our setting and its empirical covariance matrix is therefore inaccessible. We then propose another estimator for the

covariance matrix whose theoretical guarantees, including the control of the convergence rate in infinite norm, required for the graphical lasso procedure, are studied.

The rest of the paper is organized as follows. In Section 2, we explicit the model and present the two step estimation procedure. The first step consists in estimating the covariance matrix, by estimating each term by maximising the corresponding bivariate marginal log-likelihood of the truncated vector, which is a non-convex optimisation problem. The second one relies on the graphical lasso to estimate the precision matrix. Section 3 contains two theoretical results about the convergence rate of the covariance and the precision matrix estimators. We first use recent results of Mei, S., Bai, Y. and Montanari, A. (2017) to set out that our covariance matrix estimator concentrates well in infinite norm around the theoretical covariance matrix. These results concern properties of the stationary points of non-convex empirical risk minimisation problems. Next, we use this first result and the consistency properties of graphical lasso, studied by Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011) in a more general framework, to show consistency and sparsistency of our final estimator of the precision matrix. In Section 3, we also state the resulting theorem which establishes consistency of our procedure with respect to graph structure recovery. In Section 4, we present some simulations studies to corroborate its theoretical efficiency.

2. Model and estimation procedure

2.1. Model

Let X be a Gaussian p-vector $X \sim \mathcal{N}_p(\mu, \Sigma^*)$ where $\mu = (\mu_j)_{j=1,\dots,p} \in \mathbb{R}^p$ is the mean vector and $\Sigma^* = (\Sigma^*_{jk})_{1\leq j,k\leq p} \in \mathcal{M}_p(\mathbb{R})$ the covariance matrix. Let us consider the p-vector Y defined as:

$$Y_j = \mathbb{1}_{a_j \leq X_j \leq b_j} X_j \text{ for all } j \in \{1, \dots, p\},$$

where the points of truncation $a_j, b_j \in \mathbb{R}$, $a_j < b_j$ are known and depend on j. The Gaussian vector X is not directly observed, but it is observed through the zero-inflated truncated vector Y. Unlike what exists in the (multivariate) truncated Gaussian literature (for example, in Gupta, A. K. and Tracy, D. S. (1976), Lee, L. (1983) and Manjunath, BG and Wilhelm, S. (2009)), when one of the initial Gaussian variables falls outside its points of truncation, we observe a zero instead and we observe the rest of the vector according the same rule. In other words, our truncation does not consist in restricting to the observations of X which falls into $[a_1, b_1] \times \cdots \times [a_p, b_p] \subset \mathbb{R}^p$.

Assume that $\mu_j = 0$ and $\Sigma^*_{jj} = 1$ for all $j \in \{1, \dots, p\}$. In practice, we can estimate them with existing techniques for doubly truncated univariate Gaussian vector (for example, (Cohen, Jr., A. C., 1957)) if variables are not centered and scaled.

Gaussian graphical model is particularly appropriate for conditional dependency graph inference. Indeed, the precision matrix $\Theta^* := (\Sigma^*)^{-1}$ specifies the conditional dependency

structure (see Hastie T., Tibshirani R., and Friedman J. (2001)). More precisely, the graph contains an edge between the variables X_j and X_k iff:

$$\left.\begin{aligned} X_j \longleftrightarrow X_k &\iff X_j \not\perp\!\!\!\perp X_k \mid (X_l)_{l\neq j,k} \\ &\iff \mathrm{cor}(X_j, X_k \mid (X_l)_{l\neq j,k}) \neq 0 \\ &\iff \Theta^*_{jk} \neq 0. \end{aligned}\right\} \tag{2.1}$$

Given a symmetric positive definite matrix M, let us denote:

$$E(M) = \{(j,k) \in \{1,\dots,p\}^2,\ j \neq k,\ M_{jk} \neq 0\}. \tag{2.2}$$

In particular, $E(\Theta^*)$ denotes the set of the edges of the theoretical graph.

The goal of this paper is to recover the latent graph structure of the variables of the Gaussian vector X from observations of the zero-inflated truncated vector Y.

2.2. Some theoretical tools

To explicit the model and exhibit some complexities, we give some theoretical tools and will focus here on bivariate marginal likelihood from the truncated vector Y.

Let $(j,k) \in \{1,\dots,p\}^2,\ j < k$ and let $f_{jk}(x,y) = f(x,y,\Sigma^*_{jk})$ denotes the bivariate marginal log-likelihood function of the Gaussian couple $(X_j, X_k) \sim \mathcal{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \Sigma^*_{jk}\\ \Sigma^*_{jk} & 1\end{pmatrix}\right)$.

With these notations, the likelihood of (Y_j, Y_k) is then $\mathcal{L}_{jk}(\Sigma^*_{jk}, y)$ where y is an observation of the vector Y and:

$$\mathcal{L}_{jk}(\sigma, y) = \sum_{a,b=0}^{1} \phi_{ab,jk}(\sigma, y_j, y_k) n_{ab}(y_j, y_k), \tag{2.3}$$

with :

- $n_{ab}(y_j, y_k) = \mathbb{1}_{\zeta_j=a, \zeta_k=b}$ where $\zeta_l = \begin{cases} 1 \text{ if } y_l \in [a_l, b_l] \setminus \{0\}, \\ 0 \text{ if } y_l = 0. \end{cases}$
- $\sum_{a,b=0}^{1} n_{ab}(y_j, y_k) = 1$
- $\phi_{11,jk}(\sigma, y_j, y_k) = f(y_j, y_k, \sigma)$
- $\phi_{01,jk}(\sigma, y_j, y_k) = \phi_{01,jk}(\sigma, y_k) = \int_{[a_j,b_j]^c} f(x, y_k, \sigma)\mathrm{d}x$
- $\phi_{10,jk}(\sigma, y_j, y_k) = \phi_{10,jk}(\sigma, y_j) = \int_{[a_k,b_k]^c} f(y_j, y, \sigma)\mathrm{d}y$
- $\phi_{00,jk}(\sigma, y_j, y_k) = \phi_{00,jk}(\sigma) = \iint_{[a_j,b_j]^c \times [a_k,b_k]^c} f(x, y, \sigma)\mathrm{d}x\mathrm{d}y.$

The likelihood (and log-likelihood) of a couple of variables of Y involves four terms according to the nullity of each of the components of the couple. In the same way, the likelihood of Y would involve 2^p terms by distinguishing all possible cases: the density of the Gaussian vector X (no component of Y is null), p simple integrals (only one null component), $\binom{p}{2}$ double integrals (two null components), $\binom{p}{3}$ triple integrals, ..., one p-multiple integral (all the components are null). Writing the likelihood of the vector Y becomes than quite complicated and tedious. This is why we choose to restrict to the study of the likelihoods of couples for the estimation.

In practice, we have a n-sample $\mathbf{Y} := (Y^{(1)}, \dots, Y^{(n)})$ of the vector Y. The likelihood of the n-sample $\left((Y_j^{(i)}, Y_k^{(i)})\right)_{i=1,\dots,n}$ is then $\mathcal{L}_{jk}^{(n)}(\Sigma_{jk}^*, \mathbf{y})$ defined by:

$$\begin{aligned}\mathcal{L}_{jk}^{(n)}(\sigma, \mathbf{y}) &= \prod_{i=1}^{n} \mathcal{L}_{jk}(\sigma, y^{(i)}),\\ &= \prod_{i=1}^{n} \sum_{a,b=0}^{1} \phi_{ab,jk}(\sigma, y_j^{(i)}, y_k^{(i)}) n_{ab}(y_j^{(i)}, y_k^{(i)}),\end{aligned}$$

where $\mathbf{y} := (y^{(1)}, \dots, y^{(n)})$ is the realisation (value) of the n-sample $\mathbf{Y}$. The log-likelihood is then $L_{jk}^{(n)}(\Sigma_{jk}^*, \mathbf{y})$ where:

$$\begin{aligned}L_{jk}^{(n)}(\sigma, \mathbf{y}) &= \sum_{i=1}^{n} \sum_{a,b=0}^{1} n_{ab}(y_j^{(i)}, y_k^{(i)}) \log\left(\phi_{ab,jk}(\sigma, y_j^{(i)}, y_k^{(i)})\right)\\ &= \sum_{\substack{i=1\\ i:y_j^{(i)}=y_k^{(i)}=0}}^{n} \log\left(\phi_{00,jk}(\sigma)\right) + \sum_{\substack{i=1\\ i:y_j^{(i)}=0, y_k^{(i)}\neq 0}}^{n} \log\left(\phi_{01,jk}(\sigma, y_k^{(i)})\right)\\ &+ \sum_{\substack{i=1\\ i:y_j^{(i)}\neq 0, y_k^{(i)}=0}}^{n} \log\left(\phi_{10,jk}(\sigma, y_j^{(i)})\right) + \sum_{\substack{i=1\\ i:y_j^{(i)}\neq 0, y_k^{(i)}\neq 0}}^{n} \log\left(\phi_{11,jk}(\sigma, y_j^{(i)}, y_k^{(i)})\right).\end{aligned}$$

2.3. Estimation procedure

Our goal is to recover the latent graph structure of the Gaussian vector X, encoded in the precision matrix Θ^*, from observations of the truncated vector Y. Our estimation procedure is a two step procedure. In the first instance, we estimate the covariance matrix of the Gaussian vector X. Then, we estimate the precision matrix by using the graphical lasso procedure (Friedman, J., Hastie, T. and Tibshirani, R., 2008) to recover the underlying graph structure.

2.3.1. Step 1: covariance matrix estimation

Estimating the covariance matrix Σ^* of X as the empirical covariance matrix of the n-sample $\mathbf{Y}$ would lead to poor results because of the zero-inflation.

Another idea could be to maximise the likelihood of the vector Y. But we have seen that this likelihood involves 2^p terms and is too tedious.

Because of these difficulties, we estimate the covariance matrix by estimating each of its entries separately using the likelihood of the couples (Y_j, Y_k), $j < k$ defined in (2.3). More precisely, we estimate Σ^* by $\widetilde{\Sigma}^{(n)}$ by estimating each of its entries Σ^*_{jk} by maximisation of the log-likelihood of the n-sample $\left((Y_j^{(i)}, Y_k^{(i)})\right)_{i=1,\dots,n}$ of the couple (Y_j, Y_k), which is not convex.

***Definition* 2.1** (Estimator $\widetilde{\Sigma}^{(n)}$ of Σ^*)**.** The estimator $\widetilde{\Sigma}^{(n)} = (\widetilde{\Sigma}^{(n)}_{jk})_{1\leq j,k\leq p}$ of the covariance matrix Σ^* is defined by:

$$\begin{aligned}\widetilde{\Sigma}^{(n)}_{jk} &= \underset{|\sigma|\leq 1}{\operatorname{argmax}}\ L^{(n)}_{jk}(\sigma, \mathbf{y}) \\ &= \underset{|\sigma|\leq 1}{\operatorname{argmax}}\ \frac{1}{n} L^{(n)}_{jk}(\sigma, \mathbf{y}),\end{aligned} \tag{2.4}$$

for all $j < k$, where $\mathbf{y} := (y^{(1)}, \dots, y^{(n)})$ is the realisation of the n-sample $\mathbf{Y}$.

2.3.2. Step 2: precision matrix estimation

As our goal is to recover the conditional dependency graph, it is natural to use the estimator of the precision matrix Θ^* given by the graphical lasso (Friedman, J., Hastie, T. and Tibshirani, R., 2008). The graphical lasso is a procedure used in the Gaussian graphical model. It consists in estimating the precision matrix by maximising the penalised log-likelihood of the Gaussian model over the set $p \times p$ non-negative definite symmetric matrices:

$$\underset{\Theta \succ 0}{\operatorname{argmax}}\ \log\det(\Theta) - \operatorname{trace}(\Theta S) - \lambda_n ||\Theta||_{1,\text{off}},$$

where $||\Theta||_{1,\text{off}} = \sum_{\substack{j,k=1 \\ j\neq k}}^{p} |\Theta_{jk}|$, S is the empirical covariance matrix of X and $\lambda_n > 0$ the regularisation parameter. This optimisation problem is convex and has an unique solution (Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B., 2011).

In our case, the empirical covariance matrix of X is not directly available. Instead of obtaining this matrix as the empirical covariance matrix of Y, we replace the empirical covariance matrix S by the estimator $\widetilde{\Sigma}^{(n)}$ of Σ^* obtained at the step 1 (2.4).

***Definition* 2.2** (Estimator $\widetilde{\Theta}^{(n)}$ of Θ^*)**.** The estimator $\widetilde{\Theta}^{(n)}$ of the precision matrix Θ^* is defined as the unique solution of the following convex optimisation problem:

$$\widetilde{\Theta}^{(n)} = \underset{\Theta \succ 0}{\operatorname{argmax}} \ \log\det(\Theta) - \operatorname{trace}(\Theta\widetilde{\Sigma}^{(n)}) - \lambda_n ||\Theta||_{1,\text{off}}. \tag{2.5}$$

Theoretical results of Section 3 relate the estimators $\widetilde{\Sigma}^{(n)}$ and $\widetilde{\Theta}^{(n)}$ respectively defined in (2.4) and (2.5) when the points of truncation $(a_j)_{1\leq j\leq p}$ and $(b_j)_{1\leq j\leq p}$ are known.

3. Convergence results

The goal of this Section is to show that the estimation procedure proposed in Subsection 2.3 has strong theoretical guarantees. For that, we study theoretical properties of the estimator $\widetilde{\Theta}^{(n)}$ with regard to the recovery of the graph structure. Assume that the points of truncation $(a_j)_{1\leq j\leq p}$ and $(b_j)_{1\leq j\leq p}$ are known.

3.1. Covariance matrix estimator

3.1.1. Convergence rates in elementwise infinite norm

In a first place, we give a result about the estimator $\widetilde{\Sigma}^{(n)}$ of the covariance matrix Σ^* given by (2.4). Let us first set out two assumptions:

(H1) For all $j < k$, $|\Sigma^*_{jk}| \neq 1$. Thus, there exists $\delta > 0$ such that for all $j < k$, $|\Sigma^*_{jk}| < 1 - \delta$.

(H2) Let $j < k$ and consider the application $g : \sigma \in [-1+\delta, 1-\delta] \mapsto \mathbb{E}\Big(L^{(n)}_{jk}(\sigma, \mathbf{y})\Big)$. Then, we assume that:

- $-1+\delta$ and $1-\delta$ are not critical points of g,
- g has a finite number of critical points,
- every critical points of g, different from Σ^*_{jk}, are non-degenerate, i.e.:

$$\text{for all } \sigma \neq \Sigma^*_{jk},\ g'(\sigma) = 0 \Rightarrow g''(\sigma) \neq 0.$$

Note that Σ^*_{jk} is a non-degenerate critical point of g. This will be proved in the proof of Proposition 3.1 (see equations (3.7)).

Proposition 3.1 states rate convergence results about the estimator $\widetilde{\Sigma}^{(n)}$ of the covariance matrix Σ^* by bounding $\left|\left|\widetilde{\Sigma}^{(n)} - \Sigma^*\right|\right|_\infty$ with high probability.

***Proposition* 3.1.** Assume **(H1)** and **(H2)** and let $0 < \rho < 1$. There exist some known constants B, C and D such that letting $\frac{n}{\log n} \geq C \log\left(\frac{B}{\rho}\right)$, then the estimator of the covariance matrix $\widetilde{\Sigma}^{(n)}$ defined by (2.4) satisfies:

$$\mathbb{P}\left(\left|\left|\widetilde{\Sigma}^{(n)} - \Sigma^*\right|\right|_\infty \geq D\sqrt{\frac{\log n}{n}\log\left(\frac{B}{\rho}\right)}\right) \leq \frac{p(p-1)}{2}\rho,$$

where $||A||_\infty = \max_{j,k\in\{1,\dots,p\}} |A_{jk}|$ is the elementwise infinite norm of the matrix A.

3.1.2. Proof of Proposition 3.1

Proof relies on Theorem 2 of Mei, S., Bai, Y. and Montanari, A. (2017), who study the properties of the stationary points of non-convex empirical risk minimisation problems. We begin with three auxiliary Lemmas, proved in Appendix, which all state properties about the bivariate marginal likelihood defined in (2.3) or components of it:

***Lemma* 3.1.** There exists $\gamma > 0$ such that, for all $j < k$, if $(y_j, y_k) \in [a_j, b_j] \times [a_k, b_k]$ and $\sigma \in [-1+\delta, 1-\delta]$, then for all $a, b \in \{0, 1\}$, $\phi_{ab,jk}(\sigma, y_j, y_k) \geq \frac{1}{\gamma}$.

***Lemma* 3.2.** There exist L_1, L_2 and $L_3 > 0$ such that for all $j < k$, if $(y_j, y_k) \in [a_j, b_j] \times [a_k, b_k]$ and $\sigma \in [-1+\delta, 1-\delta]$, then for all $a, b \in \{0, 1\}$,

$$\left|\partial_\sigma^m \phi_{ab,jk}(\sigma, y_j, y_k)\right| \leq L_m, \text{ for } m \in \{1, 2, 3\}.$$

***Lemma* 3.3.** Let $j < k$.

1. For all $\sigma \in [-1+\delta, 1-\delta]$ and for all $l \in \mathbb{N}^*$,

$$\int_{\mathbb{R}^2} \partial_\sigma^l \mathcal{L}_{jk}(\sigma, y)\mathrm{d}\mu(y) = \partial_\sigma^l \int_{\mathbb{R}^2} \mathcal{L}_{jk}(\sigma, y)\mathrm{d}\mu(y) = 0,$$

   where μ is the measure on $\mathbb{R}^2$ defined by:

$$\mu := \delta_0 \otimes \delta_0 + \delta_0 \otimes \lambda + \lambda \otimes \delta_0 + \lambda \otimes \lambda, \tag{3.1}$$

   where δ_a denotes the Dirac measure in $a \in \mathbb{R}$ and λ the Lebesgue measure on $\mathbb{R}$.
2. For all $\sigma \in [-1+\delta, 1-\delta]$ and for all $l \in \mathbb{N}^*$,

$$\partial_\sigma^l \mathbb{E}_{\Sigma^*_{jk}}\Big(\log \mathcal{L}_{jk}(\sigma, Y)\Big) = \mathbb{E}_{\Sigma^*_{jk}}\Big(\partial_\sigma^l \log \mathcal{L}_{jk}(\sigma, Y)\Big),$$

   i.e.,

$$\partial_\sigma^l \int_{\mathbb{R}^2} \log \mathcal{L}_{jk}(\sigma, y)\mathcal{L}_{jk}(\Sigma^*_{jk}, y)\mathrm{d}\mu(y) = \int_{\mathbb{R}^2} \partial_\sigma^l\Big(\log \mathcal{L}_{jk}(\sigma, y)\Big)\mathcal{L}_{jk}(\Sigma^*_{jk}, y)\mathrm{d}\mu(y).$$

Fix $j < k$. With notations of Mei, S., Bai, Y. and Montanari, A. (2017), let us set:

$$\ell_{jk}(\sigma, \mathbf{y}) = \log \mathcal{L}_{jk}(\sigma, \mathbf{y})$$
$$= \sum_{a=0}^{1}\sum_{b=0}^{1} n_{ab}(y_j, y_k) \log\Big(\phi_{ab,jk}(\sigma, y_j, y_k)\Big) \tag{3.2}$$

$$\hat{R}_n(\sigma, \mathbf{Y}) = \frac{1}{n} L_{jk}^{(n)}(\sigma, \mathbf{Y}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\sigma, Y^{(i)}) \tag{3.3}$$

$$R(\sigma) = \mathbb{E}_{\Sigma_{jk}^*}\Big(\hat{R}_n(\sigma, \mathbf{Y})\Big) = \mathbb{E}_{\Sigma_{jk}^*}\Big(\ell(\sigma, Y)\Big). \tag{3.4}$$

***Remarks* 3.1.** • To lighten notations, we drop the underscripts jk and simply write ℓ, $\hat{R}_n$ and R instead of ℓ_{jk}, $\hat{R}_{n,jk}$ and R_{jk}.
• Point 2 of Lemma 3.3 can be rewritten as:
For all $\sigma \in [-1+\delta, 1-\delta]$ and for $l \in \mathbb{N}^*$:

$$R^{(l)}(\sigma) = \partial_\sigma^l \mathbb{E}_{\Sigma_{jk}^*}\Big(\ell(\sigma, Y))\Big) = \mathbb{E}_{\Sigma_{jk}^*}\Big(\partial_\sigma^l \ell(\sigma, Y))\Big). \tag{3.5}$$

Theorem 2 of Mei, S., Bai, Y. and Montanari, A. (2017) requires four assumptions. Let us check these assumptions.

(i) Gradient statistical noise. *The gradient of ℓ w.r.t. σ is τ^2-sub-Gaussian for some $\tau > 0$, i.e.:*

$$\forall \sigma \in [-1+\delta, 1-\delta],\ \forall \lambda \in \mathbb{R},\ \mathbb{E}\left[\exp\left(\lambda\Big(\partial_\sigma \ell(\sigma, Y) - \mathbb{E}\big(\partial_\sigma \ell(\sigma, Y)\big)\Big)\right)\right] \leq \exp\Big(\frac{\tau^2\lambda^2}{2}\Big).$$

Indeed, for all $y \in \prod_{j=1}^{p}[a_j, b_j]$ and $\sigma \in [-1+\delta, 1-\delta]$,

$$\partial_\sigma \ell(\sigma, y) = \sum_{a=0}^{1}\sum_{b=0}^{1} n_{ab}(y_j, y_k) \frac{\partial_\sigma \phi_{ab,jk}(\sigma, y_j, y_k)}{\phi_{ab,jk}(\sigma, y_j, y_k)}$$
$$\text{Thus: } \Big|\partial_\sigma \ell(\sigma, y)\Big| \leq \sum_{a=0}^{1}\sum_{b=0}^{1} n_{ab}(y_j, y_k) \frac{\big|\partial_\sigma \phi_{ab,jk}(\sigma, y_j, y_k)\big|}{\big|\phi_{ab,jk}(\sigma, y_j, y_k)\big|}$$
$$\leq \sum_{a=0}^{1}\sum_{b=0}^{1} n_{ab}(y_j, y_k)\gamma L_1 = \gamma L_1 \text{ by Lemmas 3.1 and 3.2.}$$

This way, $\partial_\sigma \ell(\sigma, Y) - \mathbb{E}(\partial_\sigma \ell(\sigma, Y))$ is zero-mean and bounded by $2\gamma L_1$. By Theorem 9.9 of Stromberg K. (1994), $\partial_\sigma \ell(\sigma, Y) - \mathbb{E}(\partial_\sigma \ell(\sigma, Y))$ is then τ^2-sub-Gaussian for $\tau = 2\gamma L_1$. Assumption "Gradient statistical noise" is satisfied.

(ii) Hessian statistical noise. *The second derivative of ℓ w.r.t. σ is τ^2-sub-exponential ($\tau = 2\gamma L_1$), that is:*

$$||\partial_\sigma^2 \ell(\sigma, Y)||_{\psi_1} \leq \tau^2,$$

where $||.||_{\psi_1}$ is the Orlicz ψ_1-norm defined by $||X||_{\psi_1} := \sup_{k\geq 1} \frac{1}{k}\mathbb{E}\Big(\big|X - \mathbb{E}(X)\big|^k\Big)^{\frac{1}{k}}$.

For all $y \in \prod_{j=1}^{p}[a_j, b_j]$ and $\sigma \in [-1+\delta, 1-\delta]$,

$$\partial_\sigma^2 \ell(\sigma, y) = \sum_{a=0}^{1}\sum_{b=0}^{1} n_{ab}(y_j, y_k)\Big(\frac{\partial_\sigma^2 \phi_{ab,jk}(\sigma, y_j, y_k)}{\phi_{ab,jk}(\sigma, y_j, y_k)} - \Big(\frac{\partial_\sigma \phi_{ab,jk}(\sigma, y_j, y_k)}{\phi_{ab,jk}(\sigma, y_j, y_k)}\Big)^2\Big)$$

$$\begin{aligned}
\text{Thus: } \Big|\partial_\sigma^2 \ell(\sigma, y)\Big| &\leq \sum_{a=0}^{1}\sum_{b=0}^{1} n_{ab}(y_j, y_k)\Big(\frac{\big|\partial_\sigma^2 \phi_{ab,jk}(\sigma, y_j, y_k)\big|}{\big|\phi_{ab,jk}(\sigma, y_j, y_k)\big|} + \Big(\frac{\big|\partial_\sigma \phi_{ab,jk}(\sigma, y_j, y_k)\big|}{\big|\phi_{ab,jk}(\sigma, y_j, y_k)\big|}\Big)^2\Big) \\
&\leq \sum_{a=0}^{1}\sum_{b=0}^{1} n_{ab}(y_j, y_k)(\gamma L_2 + \gamma^2 L_1^2) \text{ by Lemmas 3.1 and 3.2} \\
&= \gamma L_2 + \gamma^2 L_1^2. \qquad (3.6)
\end{aligned}$$

Therefore, $\partial_\sigma^2 \ell(\sigma, Y) - \mathbb{E}(\partial_\sigma^2 \ell(\sigma, Y))$ is bounded by $2(\gamma L_2 + \gamma^2 L_1^2)$ and for all $k \geq 1$,

$$\frac{1}{k}\mathbb{E}\Big(\big|\partial_\sigma^2 \ell(\sigma, Y) - \mathbb{E}(\partial_\sigma^2 \ell(\sigma, Y))\big|^k\Big)^{\frac{1}{k}} \leq \frac{2}{k}(\gamma L_2 + \gamma^2 L_1^2).$$

Hence, $||\partial_\sigma^2 \ell(\sigma, Y)||_{\psi_1} \leq 2(\gamma L_2 + \gamma^2 L_1^2) \leq \tau^2 = 4\gamma^2 L_1^2$ (we can possibly choose L_1 and γ larger). So, $\partial_\sigma^2 \ell(\sigma, Y)$ is τ^2-sub-exponential. Assumption "Hessian statistical noise" is satisfied.

(iii) Hessian regularity.

1. *The second derivative of R (defined in (3.4)) is bounded at one point:*

$$\text{there exists } |\sigma^*| \leq 1 - \delta \text{ and } H > 0 \text{ such that } \Big|R''(\sigma^*)\Big| \leq H.$$

2. *The second derivative of ℓ w.r.t. σ is Lipschitz continuous with integrable Lipschitz constant (w.r.t. y), i.e.:*

$$\text{there exists } J^* > 0 \text{ such that } \mathbb{E}[J(Y)] \leq J^*,$$

where $J(y) = \sup_{\substack{|\sigma_1|,|\sigma_2|\leq 1-\delta \\ \sigma_1 \neq \sigma_2}} \frac{|\partial_\sigma^2 \ell(\sigma_1, y) - \partial_\sigma^2 \ell(\sigma_2, y)|}{|\sigma_1 - \sigma_2|}$.

3. *Constants H and J^* satisfy: $H \leq \tau^2$ and $J^* \leq \tau^3$.*

First, $R''(\sigma) = \mathbb{E}_{\Sigma^*_{jk}}\Big(\partial^2_\sigma \ell(\sigma, Y)\Big)$ by the point 2 of Lemma 3.3 and (3.5). By (3.6), $\Big|\partial^2_\sigma \ell(\sigma, Y)\Big| \leq \gamma L_2 + \gamma^2 L_1^2$ for all $\sigma \in [-1+\delta, 1-\delta]$, thus any $|\sigma^*| \leq 1-\delta$ and $H = \gamma L_2 + \gamma^2 L_1^2$ are appropriate. Moreover, we have $H \leq \tau^2 = 4\gamma^2 L_1^2$ (with L_1 and γ possibly chosen larger).

For all $y \in \prod_{j=1}^{p} [a_j, b_j]$ and $\sigma \in [-1+\delta, 1-\delta]$, we have (with a slight lightening of notations):

$$\partial^3_\sigma \ell(\sigma, y) = \sum_{a=0}^{1}\sum_{b=0}^{1} n_{ab}(y_j, y_k)\Big(\frac{\partial^3_\sigma \phi_{ab,jk}}{\phi_{ab,jk}} - 3\frac{\partial_\sigma \phi_{ab,jk}\partial^2_\sigma \phi_{ab,jk}}{\phi^2_{ab,jk}} + 2\Big(\frac{\partial_\sigma \phi_{ab,jk}}{\phi_{ab,jk}}\Big)^3\Big)$$

$$\begin{aligned}\text{Thus: } \Big|\partial^3_\sigma \ell(\sigma, y)\Big| &\leq \sum_{a=0}^{1}\sum_{b=0}^{1} n_{ab}(y_j, y_k)(\gamma L_3 + 3\gamma^2 L_1 L_2 + 2\gamma^3 L_1^3) \text{ (Lemmas 3.1 and 3.2)} \\ &= \gamma L_3 + 3\gamma^2 L_1 L_2^2 + 2\gamma^3 L_1^3.\end{aligned}$$

Therefore, for all $y \in \prod_{j=1}^{p} [a_j, b_j]$, $J(y) \leq \gamma L_3 + 3\gamma^2 L_1 L_2^2 + 2\gamma^3 L_1^3$ by the mean value theorem. Taking $J^* = \gamma L_3 + 3\gamma^2 L_1 L_2^2 + 2\gamma^3 L_1^3$ with L_1 and γ possibly chosen larger, we have $J^* \leq \tau^3 = 8\gamma^3 L_1^3$. Assumption "Hessian regularity" is satisfied.

(iv) Morse. *There exists $\epsilon > 0$ and $\eta > 0$ such that R is (ϵ, η) strongly Morse, i.e.:*

1. *$|R'(\sigma)| > \epsilon$ for all σ such that $|\sigma| = 1-\delta$ and,*
2. *for all σ such that $|\sigma| < 1-\delta$:*

$$|R'(\sigma)| \leq \epsilon \Rightarrow |R''(\sigma)| \geq \eta.$$

In other words, R satisfies this assumption if $-1+\delta$ and $1-\delta$ are not critical points of R and if R has a finite number of critical points, which are moreover non-degenerate:

$$R'(\sigma) = 0 \Rightarrow R''(\sigma) \neq 0.$$

Assumption **(H2)** implies point 1. and point 2. for $\sigma \neq \Sigma^*_{jk}$. Let us prove that Σ^*_{jk} is a non-degenerate critical point by showing that Σ^*_{jk} is a global maximum of R. Indeed, for all σ such that $|\sigma| < 1$:

$$\left.\begin{aligned} R(\sigma) \leq R(\Sigma^*_{jk}) &\iff \mathbb{E}_{\Sigma^*_{jk}}\Big(\ell(\sigma, Y)\Big) \leq \mathbb{E}_{\Sigma^*_{jk}}\Big(\ell(\Sigma^*_{jk}, Y)\Big) \\ &\iff \mathbb{E}_{\Sigma^*_{jk}}\Big(\log \mathcal{L}_{jk}(\sigma, Y)\Big) \leq \mathbb{E}_{\Sigma^*_{jk}}\Big(\log \mathcal{L}_{jk}(\Sigma^*_{jk}, Y)\Big) \\ &\text{since } \ell(\sigma, y) = \log \mathcal{L}_{jk}(\sigma, y) \text{ (defined in (3.2))}, \\ &\iff \mathbb{E}_{\Sigma^*_{jk}}\Big(\log \frac{\mathcal{L}_{jk}(\sigma, Y)}{\mathcal{L}_{jk}(\Sigma^*_{jk}, Y)}\Big) \leq 0. \end{aligned}\right\} \tag{3.7}$$

By Jensen inequality,

$$\mathbb{E}_{\Sigma^*_{jk}}\left(\log\frac{\mathcal{L}_{jk}(\sigma,Y)}{\mathcal{L}_{jk}(\Sigma^*_{jk},Y)}\right) \leq \log\mathbb{E}_{\Sigma^*_{jk}}\left(\frac{\mathcal{L}_{jk}(\sigma,Y)}{\mathcal{L}_{jk}(\Sigma^*_{jk},Y)}\right)$$
$$= \log\int_{\mathbb{R}^2}\mathcal{L}_{jk}(\sigma,y)\mathrm{d}\mu(y) = 0, \tag{3.8}$$

since $y \mapsto \mathcal{L}_{jk}(\sigma,y)$ is a probability density function (see (C.1)) w.r.t. the measure μ on $\mathbb{R}^2$ defined in (3.1).

Equation (3.8) implies (3.7). Σ^*_{jk} is thus a global maximum of R and $R'(\Sigma^*_{jk}) = 0$. Let us prove that $R''(\Sigma^*_{jk}) \neq 0$:

$$\begin{aligned}
R''(\Sigma^*_{jk}) &= \mathbb{E}_{\Sigma^*_{jk}}\left(\partial^2_\sigma \ell(\Sigma^*_{jk},Y)\right) \text{ by point 2 of Lemma 3.3}\\
&= \mathbb{E}_{\Sigma^*_{jk}}\left(\frac{\partial^2_\sigma\mathcal{L}_{jk}(\Sigma^*_{jk},Y)}{\mathcal{L}_{jk}(\Sigma^*_{jk},Y)} - \left(\frac{\partial_\sigma\mathcal{L}_{jk}(\Sigma^*_{jk},Y)}{\mathcal{L}_{jk}(\Sigma^*_{jk},Y)}\right)^2\right)\\
&= -\mathbb{E}_{\Sigma^*_{jk}}\left(\left(\frac{\partial_\sigma\mathcal{L}_{jk}(\Sigma^*_{jk},Y)}{\mathcal{L}_{jk}(\Sigma^*_{jk},Y)}\right)^2\right),
\end{aligned}$$

since $\mathbb{E}_{\Sigma^*_{jk}}\left(\frac{\partial^2_\sigma\mathcal{L}_{jk}(\Sigma^*_{jk},Y)}{\mathcal{L}_{jk}(\Sigma^*_{jk},Y)}\right) = \int_{\mathbb{R}^2}\partial^2_\sigma\mathcal{L}_{jk}(\Sigma^*_{jk},y)\mathrm{d}\mu(y) = 0$ by point 1 of Lemma 3.3. If we assume that $R''(\Sigma^*_{jk}) = 0$, then $\partial_\sigma\mathcal{L}_{jk}(\Sigma^*_{jk},Y) = 0$ a.s., which contradicts the definition of $\mathcal{L}_{jk}(\Sigma^*_{jk},Y)$ given in (2.3).

Accordingly, there exists $\epsilon > 0$ and $\eta > 0$ such that R is (ϵ,η) strongly Morse. Assumption "Morse" is satisfied.

For each couple (j,k) such that $j < k$, Theorem 2 of Mei, S., Bai, Y. and Montanari, A. (2017) applied to the estimator $\widetilde{\Sigma}^{(n)}_{jk}$ yields:

Let $0 < \rho < 1$. There exists an universal constant C_0 such that letting $\frac{n}{\log n} \geq 4C_0\Big[\log\Big(\frac{\tau(1-\delta)}{\rho}\Big)\vee 1\Big]\Big(\frac{\tau^2}{\epsilon^2}\vee\frac{\tau^4}{\eta^2}\vee\frac{\tau^2L^2}{\eta^4}\Big)$ with $\tau = 2\gamma L_1$ and $L = \sup_{\sigma:|\sigma|\leq 1-\delta}\big|R^{(3)}(\sigma)\big|$,

$$\mathbb{P}\left(\big|\widetilde{\Sigma}^{(n)}_{jk} - \Sigma^*_{jk}\big| \leq \frac{2\tau}{\eta}\sqrt{C_0\frac{\log n}{n}\ \Big[\log\Big(\frac{\tau(1-\delta)}{\rho}\Big)\vee 1\Big]}\right) \geq 1-\rho.$$

It follows that, for $0 < \rho < 1$ and n such that $\frac{n}{\log n} \geq 4C_0\Big[\log\Big(\frac{\tau(1-\delta)}{\rho}\Big)\vee 1\Big]\Big(\frac{\tau^2}{\epsilon^2}\vee\frac{\tau^4}{\eta^2}\vee\frac{\tau^2L^2}{\eta^4}\Big)$, then:

$$\mathbb{P}\left(\big|\big|\widetilde{\Sigma}^{(n)} - \Sigma^*\big|\big|_\infty \leq \frac{2\tau}{\eta}\sqrt{C_0\frac{\log n}{n}\ \Big[\log\Big(\frac{\tau(1-\delta)}{\rho}\Big)\vee 1\Big]}\right) \geq 1-\rho\frac{p(p-1)}{2},$$

where $||A||_\infty = \max_{j,k\in\{1,\dots,p\}} |A_{jk}|$ denotes the elementwise infinite norm of the matrix A. This finishes the proof of Proposition 3.1.

3.2. Precision matrix estimator and graph recovery

Before giving a result about the estimator of the precision matrix Θ^*, let us state a third and last assumption:

(H3) There exists some $\alpha \in]0,1]$ such that:

$$\max_{e\in S^c} \left|\left|\Gamma^*_{eS}(\Gamma^*_{SS})^{-1}\right|\right|_1 = \left|\left|\left|\Gamma^*_{S^cS}(\Gamma^*_{SS})^{-1}\right|\right|\right|_\infty \leq 1-\alpha,$$

where:

- if $M \in \mathcal{M}_{r,m}(\mathbb{R})$, $A \subset [\![1,r]\!]$ and $B \subset [\![1,m]\!]$, M_{AB} denotes the matrix $(m_{ij})_{i\in A, j\in B}$,
- $S = S(\Theta^*) := E(\Theta^*) \cup \{(1,1),\dots,(p,p)\}$ where $\Theta^* = (\Sigma^*)^{-1}$ and $E(\Theta^*)$ is the set of the edges of the theoretical graph (see (2.2)). Let $s := |E(\Theta^*)|$, hence $|S(\Theta^*)| = |E(\Theta^*)| + p = s + p$,
- $S^c = S^c(\Theta^*) = [\![1,p]\!]^2 \setminus S(\Theta^*)$,
- $\Gamma^* = \Sigma^* \otimes \Sigma^*$ where $\otimes$ denotes the Kronecker matrix product. We have: $\Gamma^*_{(j,k),(l,m)} = \mathrm{cov}(X_jX_k, X_lX_m)$ and thus $\Gamma^*_{SS} \in \mathcal{M}_{s+p,s+p}(\mathbb{R})$,
- $||u||_1 = \sum_{j=1}^d |u_j|$ for all $u \in \mathbb{R}^d$ is the ℓ_1-norm,
- $|||U|||_\infty = \max_{j=1,\dots,d} \sum_{k=1}^m |U_{jk}|$ for all $U \in \mathcal{M}_{d,m}(\mathbb{R})$.

The underlying intuition is that this assumption **(H3)** limits the influence that the non-edge terms, indexed by S^c, can have on the edge-based terms, indexed by S (Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B., 2011).

Proposition 3.2, set out below, provides an upper-bound for the elementwise maximum norm of the precision matrix estimator $\widetilde{\Theta}^{(n)}$ obtained by the graphical lasso procedure (2.5). It also shows its sparsistency with respect to graphical model structure recovery. Here are some preliminary notations:

- d is the maximum degree:

$$d = \max_{j=1,\dots,p} \left|\{k \in [\![1,p]\!] : \Theta^*_{jk} \neq 0\}\right|. \tag{3.9}$$

- κ_{Σ^*} and κ_{Γ^*} are defined by:

$$\kappa_{\Sigma^*} := |||\Sigma^*|||_\infty = \max_{j=1,\dots,p} \sum_{k=1}^p |\Sigma^*_{jk}|, \tag{3.10}$$

$$\kappa_{\Gamma^*} := \left|\left|\left|\left(\Gamma^*_{SS}\right)^{-1}\right|\right|\right|_\infty. \tag{3.11}$$

***Proposition* 3.2.** Assume **(H3)** and assume that there exist some strictly positive constants B, C, D and $c > 2$ such that letting $\frac{n}{\log n} \geq C \log\left(Bp^c\right)$, we have:

$$\mathbb{P}\left(\left|\left|\widetilde{\Sigma}^{(n)} - \Sigma^*\right|\right|_\infty \geq D\sqrt{\frac{\log n}{n}\log\left(Bp^c\right)}\right) \leq \frac{p(p-1)}{2p^c}, \tag{3.12}$$

where $||A||_\infty = \max_{j,k\in\{1,\ldots,p\}} |A_{jk}|$ denotes the elementwise infinite norm of the matrix A. Assume that the sample size n is lower bounded as $\frac{n}{\log n} > D^2 \log\left(Bp^c\right) \max\left\{\frac{\sqrt{C}}{D}, 6(1+8\alpha^{-1})d\max\{\kappa_{\Sigma^*}\kappa_{\Gamma^*}, \kappa^3_{\Sigma^*}\kappa^2_{\Gamma^*}\}\right\}^2$, and denote $\widetilde{\Theta}^{(n)}$ the unique solution of (2.5) and $\lambda_n = \frac{8D}{\alpha}\sqrt{\frac{\log n}{n}\log\left(Bp^c\right)}$ the regularisation parameter involved in (2.5). Then, with probability greater than $1 - \frac{1}{p^{c-2}}$, we have:

(a) The estimator $\widetilde{\Theta}^{(n)}$ of Θ^* satisfies:

$$||\widetilde{\Theta}^{(n)} - \Theta^*||_\infty \leq 2D(1+8\alpha^{-1})\kappa_{\Gamma^*}\sqrt{\frac{\log n}{n}\log\left(Bp^c\right)}.$$

(b) The estimated edges set is a subset of the true edges set: $E(\widetilde{\Theta}^{(n)}) \subset E(\Theta^*)$ and $E(\widetilde{\Theta}^{(n)})$ includes all edges (j,k) with:

$$\left|\Theta^*_{jk}\right| > 2D(1+8\alpha^{-1})\kappa_{\Gamma^*}\sqrt{\frac{\log n}{n}\log\left(Bp^c\right)}.$$

Proof relies on results of Theorem 1 of Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011), in which they study the precision matrix estimation problem in the multivariate Gaussian setting.

Proof. (Proposition 3.2) Let us check the two assumptions of Theorem 1 of Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011).

- **Incoherence assumption.** This assumption is exactly our assumption **(H3)**.

- **Control of sampling noise.** A careful reading of Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011) reveals that the *tail conditions* of their Theorem 1 is not necessary. The required assumption, stated below, is in fact weaker, and is given in Lemma 8 of Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011):
There exists $v^* > 0$ such that for all $c > 2$ and n such that $\bar{\beta}_f(n,p^c) \leq \frac{1}{v_*}$, we have:

$$\mathbb{P}\left[||\widetilde{\Sigma}^{(n)} - \Sigma^*||_\infty \geq \bar{\beta}_f(n,p^c)\right] \leq \frac{1}{p^{c-2}},$$

where $\bar{\beta}_f(n,r) := \text{argmax}\{\beta : f(n,\beta) \leq r\}$ *for some function* $f(n,\beta)$.

Setting $f(n,\beta) = \frac{1}{B}\exp\left(\frac{n}{\log n}\left(\frac{\beta}{D}\right)^2\right)$ and $v_* = \frac{\sqrt{C}}{D}$ and noticing that $\frac{p(p-1)}{2} \leq p^2$, Assumption (3.12) gives this result. Indeed:

- $\bar{\beta}_f(n,r) = \text{argmax}\{\beta : f(n,\beta) \leq r\} = D\sqrt{\frac{\log n}{n}\log\left(Br\right)}$
- $\frac{n}{\log n} \geq C\log\left(Bp^c\right) \iff \bar{\beta}_f(n,p^c) \leq \frac{D}{\sqrt{C}}$

Assumption "Control of sampling noise" is satisfied.

At last, let us set $\bar{n}_f(\beta,r) := \text{argmax}\{n : f(n,\beta) \leq r\}$. Then, the condition $n > \bar{n}_f(\beta,r)$ is equivalent to $\frac{n}{\log n} > \log\left(Br\right)\frac{D^2}{\beta^2}$ since $f(n,\beta) \leq r \iff \frac{n}{\log n} \leq \log\left(Br\right)\frac{D^2}{\beta^2}$.

We complete the proof by applying Theorem 1 of Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). □

Finally, Propositions 3.1 and 3.2 provide the following theorem, which establishes consistency of the estimator $\widetilde{\Theta}^{(n)}$ in the elementwise maximum-norm:

Theorem 3.1. *Assume* ***(H1)***, ***(H2)*** *and* ***(H3)***. *Let* $c > 2$, $\widetilde{\Theta}^{(n)}$ *the unique solution of* (2.5) *and* α, d, κ_{Σ^*} *and* κ_{Γ^*} *respectively defined in* ***(H3)***, *in* (3.9), *in* (3.10) *and in* (3.11). *There exists some known constants* B, C *and* D *such that letting* n *lower bounded as* $\frac{n}{\log n} > D^2\log\left(Bp^c\right)\max\left\{\frac{\sqrt{C}}{D}, 6(1+8\alpha^{-1})d\max\{\kappa_{\Sigma^*}\kappa_{\Gamma^*}, \kappa_{\Sigma^*}^3\kappa_{\Gamma^*}^2\}\right\}^2$ *and* $\lambda_n = \frac{8D}{\alpha}\sqrt{\frac{\log n}{n}\log\left(Bp^c\right)}$ *the penalisation parameter of the equation* (2.5), *we have, with probability greater than* $1 - \frac{1}{p^{c-2}}$:

(a) The estimator $\widetilde{\Theta}^{(n)}$ *of* Θ^* *satisfies:*

$$||\widetilde{\Theta}^{(n)} - \Theta^*||_\infty \leq 2D(1+8\alpha^{-1})\kappa_{\Gamma^*}\sqrt{\frac{\log n}{n}\log\left(Bp^c\right)}.$$

(b) $E(\widetilde{\Theta}^{(n)}) \subset E(\Theta^*)$ *and* $E(\widetilde{\Theta}^{(n)})$ *includes all edges* (j,k) *with:*

$$\left|\Theta^*_{jk}\right| > 2D(1+8\alpha^{-1})\kappa_{\Gamma^*}\sqrt{\frac{\log n}{n}\log\left(Bp^c\right)}.$$

In other words, the graph structure of latent Gaussian encoded in Θ^* *is consistently recovered as long as:* $\left|\Theta^*_{jk}\right| > 2D(1+8\alpha^{-1})\kappa_{\Gamma^*}\sqrt{\frac{\log n}{n}\log\left(Bp^c\right)}$.

The parameter c of Theorem 3.1 is a user-defined parameter. The larger c is, the larger the probability for which results of Theorem 3.1 hold is. However, large values of this parameter lead to more stringent requirements on the sample size n.

4. Simulation studies

4.1. Simulation settings

In almost all of this Section (unless otherwise stated), we use the following simulation settings. We simulate $n = 500$ observations of a $p = 100$-Gaussian vector X centered and scaled. Graph structure is a chain, that is $X_1 \longleftrightarrow X_2 \longleftrightarrow \cdots \longleftrightarrow X_{100}$. The graph contains then 99 edges. Data have been simulated with the R function `huge.generator`, option `graph = "band"` of the package `huge`.

Two different settings of the points of truncation are presented:

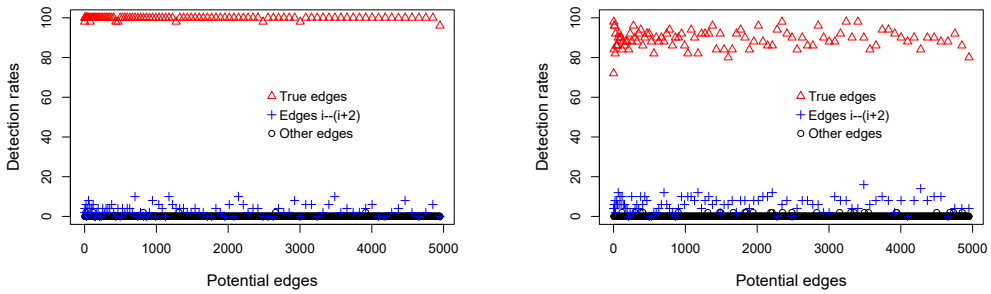
- identical points of truncation: $a = -0.5$ and $b = 2$,
- decreasing points of truncation: $a = -1$, $b =$ `seq(2,0.5, length = p)`.

We then apply the estimation procedure described in Subsection 2.3. We assume that the points of truncation are known and we use the estimators $\widetilde{\Sigma}^{(n)}$ and $\widetilde{\Theta}^{(n)}$ of the covariance and precision matrices, respectively defined at Step 1 (2.4) and at Step 2 (2.5).

For the estimation of the precision matrix with the graphical lasso procedure, we use the function `huge` of the same package, option `method = "glasso"`. Unfortunately, theoretical results of Section 3 do not give an explicit choice of the penalty parameter. We thus choose the penalty parameter with "stars" and "ebic" methods, implemented in the package `huge`.

Simulations address several problems. Let first explicit the two different procedures used for these simulations, respectively called "our procedure" and "graphical lasso directly on truncated data" (shortened in "Glasso"). The first one is our procedure, which consists in replacing the empirical covariance matrix of X in the graphical lasso by our estimator $\widetilde{\Sigma}^{(n)}$. The second one is the graphical lasso directly applied to the truncated data, which consists in replacing the empirical covariance matrix of X in the graphical lasso by the empirical covariance matrix of the truncated vector Y. Here are the problems addressed in the following subsections:

- Efficiency of our procedure. Does using our estimator for the covariance matrix really improve graph estimation?
- Impact of the points of truncation, that is how the values of the truncation points impact edges detection.
- Is our procedure as efficient with other graph structures?



(a) Our procedure, identical points of truncation: $a = -0.5$ and $b = 2$.

(b) Glasso, identical points of truncation: $a = -0.5$ and $b = 2$.

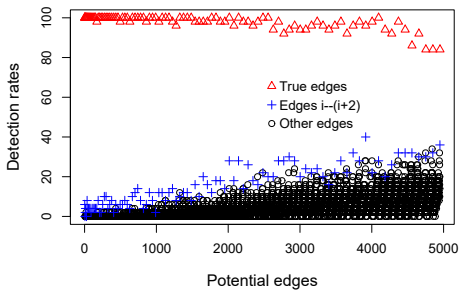
Figure 1. Comparison of detection rates obtained by our procedure and by graphical lasso directly on truncated data. Identical points of truncation setting. Detection rates are obtained on 50 i.i.d. repetitions for $n = 500$ observations of $p = 100$ variables. True edges are represented with red triangles, (false) edges of type $X_i - X_{i+2}$ with blue crosses and other false edges with black circles.

To study efficiency of these procedures on graph estimation, we make 50 i.i.d. repetitions of the procedure and we represent the detection rates of each of the $\binom{100}{2} = 4950$ potential edges.

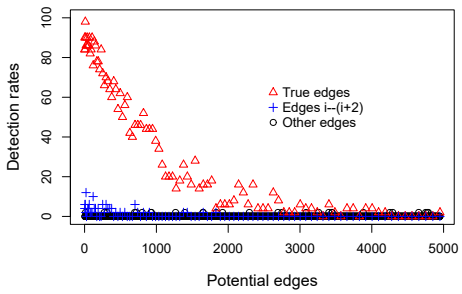
4.2. Efficiency

In this subsection, we aim at illustrating the efficiency of our procedure. For that, we compare detection rates of each potential edges obtained with our procedure and with graphical lasso directly on truncated data.

Results and comments. Figures 1 and 2 illustrate detection rates for our procedure and for graphical lasso directly on truncated data Y. We represent detection rates of each potential edge by unfolding the matrix with `upper.tri`, which implies that the first potential edge whose detection rate is displayed is $X_2 \longleftrightarrow X_1$, then $X_3 \longleftrightarrow X_1$, $X_3 \longleftrightarrow X_2$, $X_4 \longleftrightarrow X_1$, ..., $X_4 \longleftrightarrow X_3$ and so on. The 99 true theoretical edges, that is edges $X_i \longleftrightarrow X_{i+1}$, are displayed with red triangles and in the following order $X_1 \longleftrightarrow X_2$, $X_2 \longleftrightarrow X_3$, ..., $X_{99} \longleftrightarrow X_{100}$. Detection rates of edges of type $X_i \longleftrightarrow X_{i+2}$, which are not true edges, are displayed with blue crosses. We distinguish these edges because these interactions can be relatively strong because of the indirect link through X_{i+1}.
Figure 1 shows identical points of truncation setting. We can observe that our method gives better results: true edges are better detected and other edges are less detected than with Glasso. For example, true edges are all detected more than 96% whereas Glasso



(a) Our procedure, decreasing points of truncation: $a = -1$, $b =$ `seq(2,0.5, length = p)`.



(b) Glasso, decreasing points of truncation: $a = -1$, $b =$ `seq(2,0.5, length = p)`.

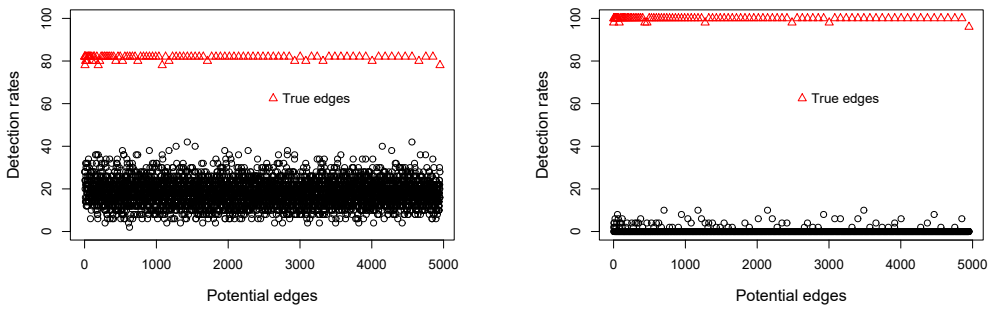
Figure 2. Comparison of detection rates obtained by our procedure and by graphical lasso directly on truncated data. Decreasing points of truncation setting. Detection rates are obtained on 50 i.i.d. repetitions for $n = 500$ observations of $p = 100$ variables. True edges are represented with red triangles, (false) edges of type $X_i - X_{i+2}$ with blue crosses and other false edges with black circles.

detects them between 80% and 100%: 66 of them are detected at most 90%. Edges of type $X_i \longleftrightarrow X_{i+2}$ tend to be more detected with Glasso.
Figure 2 shows decreasing points of truncation setting. Efficiency of our procedure is even more convincing in this setting. Indeed, true edges are much better detected by our procedure (more thant 80% whereas only 60 true edges are detected at most 80% by Glasso). These differences are probably due to the zero rate in the truncated data Y which grows from 20% to 50% according to the variables. Consequently, edges between variables whose zeros rate is high (that is edges $X_i \longleftrightarrow X_{i+1}$ for i close to 100) tend to be less detected. The phenomenon is even stronger with Glasso. Besides, the other edges (the false ones) tend to be slightly less detected with Glasso. For our procedure, the false edges are more detected when the truncation points induce a high rate of zero for the involved variables.

4.3. Impact of points of truncation

To expand this section, we illustrate the impact of the points of truncation values. For that, we briefly compare results obtained for both "identical" and "decreasing" settings. Observations of the underlying Gaussian vector are the same and we only change values of the points of truncation according to the chosen setting.

Results and comments. Figures 1(a) and 2(a) respectively show detection rates for "identical" and "decreasing" points of truncation settings.



(a) Points of truncation setting: $a = -1$ and $b = 1$.

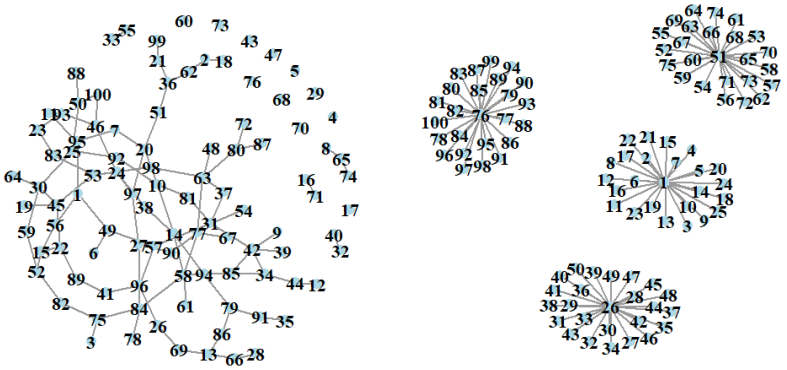
(b) Identical points of truncation: $a = -0.5$ and $b = 2$.

Figure 3. Comparison of detection rates for two points of truncation settings for which the zero inflation is similar (around 32%). Detection rates are obtained on 50 i.i.d. repetitions for $n = 500$ observations of $p = 100$ variables. True edges are represented with red triangles.

The zero inflation of truncated data Y of the first setting is around 33%. Zero inflation of truncated data of the "decreasing" setting decreases from 20% (for Y_1) to 50% (for Y_{100}). Detection rates of potential edges are represented by unfolding the matrix with `upper.tri`, which implies that the first potential edge whose detection rate is displayed is $X_2 \longleftrightarrow X_1$, then $X_3 \longleftrightarrow X_1$, $X_3 \longleftrightarrow X_2$, $X_4 \longleftrightarrow X_1$, ..., $X_4 \longleftrightarrow X_3$ etc. The 99 true theoretical edges, that is the edges $X_i \longleftrightarrow X_{i+1}$, are displayed with red triangles and in the following order $X_1 \longleftrightarrow X_2$, $X_2 \longleftrightarrow X_3$, ..., $X_{99} \longleftrightarrow X_{100}$. Thus, we can observe that edges involving variables whose zero inflation is close to 50% have worse detection rates: true edges are less detected whereas false edges have higher detection rates.

In short and as expected, zero inflation impacts detection rates: the more the zero inflation is, worse is the detection rate.

An other phenomenon, not noticeable in Figures 1(a) and 2(a) occurs. This phenomenon is also linked to zero inflation and is noticeable in Figure 3. It points out that detection rates does not only depend on zero inflation but also on the observation window. Indeed, Figure 3 exhibits detection rates for "identical" points of truncation setting, that is $a = -0.5$ and $b = 2$, and for an other setting $a = -1$ and $b = 1$. For both settings, zero inflation is about 32%. However, results obtained for "identical" points of truncation are much better. The underlying idea is that the observation of the Gaussian variable gives more informations between -0.5 and 2 than between -1 and 1, especially for covariances estimation.



(a) "Random" graph structure. (b) "Hub" graph structure.

Figure 4. Graphical representation of the two graphs used in this subsection: "random" and "hub".

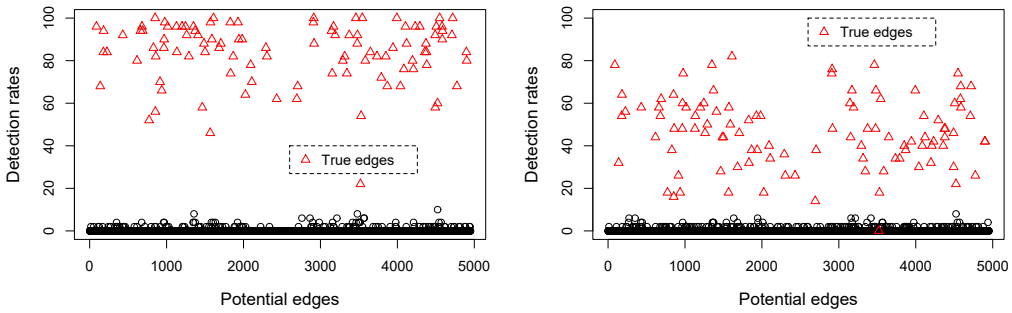
4.4. Other graph structures

Previously, we restrict to only one graph structure, the chain structure which tends to give satisfactory results in general. To fulfil these simulation studies, we present some results with other graph structures:

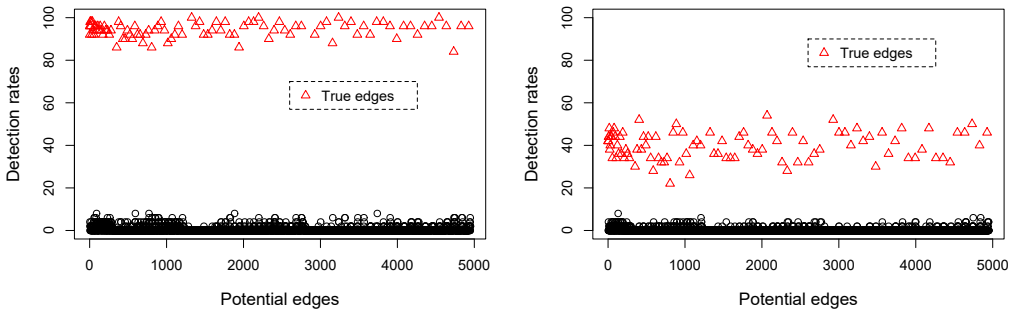
- The "random" structure. There exists an edge between two variables with probability 1/50. Data have been simulated with the R function `huge.generator`, options `graph = 'random', prob = 1/50`. Resulting graph has 103 edges.
- The "hub" structure. Variables are split into 4 groups of 25. Inside each group, one of the variable is a "hub" and is connected to all the variables of its group. Data have been simulated with the R function `huge.generator`, options `graph = 'hub', g = 4`. Resulting graph has 96 edges.

These graphs are displayed in Figure 4. Points of truncation are set to identical ($a = -0.5$ and $b = 2$). We simulate $n = 500$ observations of $p = 100$ variables and we compare detection rates of each potential edge obtained with our procedure and with graphical lasso directly on truncated data like in Subsection 4.2.

Results and comments. Figure 5 displays these comparisons for "random" and "hub" graph structures.
Results obtained with our procedure are indeed slightly less satisfying than with "chain" structure. Detection rates with "hub" graph structure are a little better than with "random". For "random", all the false edges are detected less than 10% and true edges are all detected more than 46% except for the edge $X_{34} \longleftrightarrow X_{85}$ (22%). This edge is also slightly less detected (92% whereas at least 98% for the other true edges) when we directly apply graphical lasso on untruncated Gaussian data X (see Figure 6(a)). Thus, this low detection rate is not only due to our procedure but probably also to the data

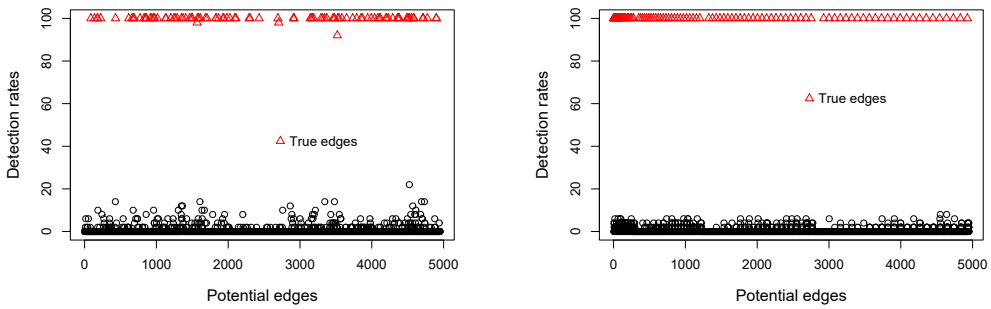


(a) "Random" graph structure.



(b) "Hub" graph structure.

Figure 5. Comparison of detection rates obtained by our procedure (left) and by graphical lasso (right) directly on truncated data. Identical points of truncation setting. "random" and "hub" graph structures. Detection rates are obtained on 50 i.i.d. repetitions for $n = 500$ observations of $p = 100$ variables. True edges are represented with red triangles.



(a) "Random" graph structure.

(b) "Hub" graph structure.

Figure 6. Detection rates obtained with graphical lasso on Gaussian data X. "random" and "hub" graph structures. Detection rates are obtained on 50 i.i.d. repetitions for $n = 500$ observations of $p = 100$ variables. True edges are represented with red triangles.

itself. "Hub" structure gives good results: false edges are detected at most 8% and true edges at most 84%, which is also the case when we apply graphical lasso on untruncated data X (see Figure 6(b)).
In comparison, Glasso (on truncated data Y) always gives less good results. For "random", the edge $X_{34} \longleftrightarrow X_{85}$ is for example never detected and the other true edges are detected between 14% and 82%. "Hub" structure exhibits the most striking difference: true edges are detected at most 54% (but at least 22%).

5. Discussion

In this paper, we proposed a procedure for graph estimation in a zero-inflated Gaussian model. In this model, zero-inflation is obtained by double truncation (left and right) of Gaussian data. More precisely, the goal is to retrieve the underlying graph structure given by the precision matrix of the Gaussian data with the doubly truncated data. Our procedure includes two steps: the first one consists in estimating the covariance matrix terms to terms by maximising the corresponding bivariate marginal log-likelihood of the truncated vector. The second one relies on the graphical lasso procedure to obtain a lasso estimation of the precision matrix. We then proved some theoretical convergence guarantees with regard to graph estimation. The first result states rate convergence about the covariance matrix estimator. The second one provides sparsistency of the precision matrix estimator with respect to graph structure recovery.
Practically, simulations studies also corroborate efficiency of our procedure. They also show that our procedure is more appropriate than using graphical lasso directly on truncated data (without a preliminary estimation of the covariance matrix).

However, results depend on graph structure and simulations exhibit well that some graph structures are more favorable to graph recovery.

This work only deals with a double truncation. Yet, our procedure seems to be practically efficient with a single truncation (right or left) but proof of our theoretical results requires the both right and left points of truncation and they do not hold in the unilateral setting. It would be interessant to address this case later, perhaps using different tools.

Appendix A: Proof of Lemma 3.1

Proof. (Lemma 3.1) It is sufficient to show the existence of such a constant $\gamma_{jk} > 0$ for $j < k$ fixed. Let $j < k$, $\sigma \in [-1+\delta, 1-\delta]$ and $(y_j, y_k) \in [a_j, b_j] \times [a_k, b_k]$. The proof naturally falls into four parts:

- $a = b = 1$:

$$\phi_{11,jk}(\sigma, y_j, y_k) = \frac{1}{2\pi\sqrt{1-\sigma^2}} \exp\left[-\frac{y_j^2 - 2\sigma y_j y_k + y_k^2}{2(1-\sigma^2)} \right]. \tag{A.1}$$

Since $(y_j, y_k) \in [a_j, b_j] \times [a_k, b_k]$ and $\delta^2 \leq 1 - \sigma^2 \leq 1$, $\phi_{11,jk}$ is continuous and strictly positive on a compact of $\mathbb{R}^3$.

- $a = 0, b = 1$: An easy computation yields that

$$\phi_{01,jk}(\sigma, y_k) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y_k^2}{2} \right) \left[1 - F\Big(\frac{b_j - \sigma y_k}{\sqrt{1-\sigma^2}} \Big) + F\Big(\frac{a_j - \sigma y_k}{\sqrt{1-\sigma^2}} \Big) \right], \tag{A.2}$$

where F denotes the c.d.f. of $\mathcal{N}(0,1)$. Since $y_k \in [a_k, b_k]$, $-\infty < a_k < b_k < \infty$ and $\delta^2 \leq 1 - \sigma^2 \leq 1$, $\phi_{01,jk}$ is hence continuous and strictly positive on a compact of $\mathbb{R}^2$.

- $a = 1, b = 0$: Analogous to $a = 0, b = 1$.

- $a = b = 0$:

$$\phi_{00,jk}(\sigma) = \int_{[a_k, b_k]^c} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2} \right) \left[1 - F\Big(\frac{b_j - \sigma y}{\sqrt{1-\sigma^2}} \Big) + F\Big(\frac{a_j - \sigma y}{\sqrt{1-\sigma^2}} \Big) \right] \mathrm{d}y.$$

Since $\delta^2 \leq 1 - \sigma^2 \leq 1$, $\phi_{00,jk}$ is hence continuous and strictly positive on a compact of $\mathbb{R}$. □

Appendix B: Proof of Lemma 3.2

Proof. (Lemma 3.2) In the same manner as for the proof of Lemma 3.1, it is sufficient to show the result for $j < k$ fixed. Fix $j < k$ and show that for all $a, b \in \{0, 1\}$, the function $\phi_{ab,jk}$ is C^3 on the compact $[-1+\delta, 1-\delta] \times [a_j, b_j] \times [a_k, b_k]$ (it is even C^∞).

Hence, for all $m \in \{1,2,3\}$, $\partial_\sigma^m \phi_{ab,jk}$ is continuous on a compact of $\mathbb{R}^3$, which establishes the result. We naturally distinguish the four cases (see (A.2) for expressions of $\phi_{ab,jk}$):

- $a = b = 1$: $\phi_{11,jk}$ is C^3 on $]-1,1[\times\mathbb{R}^2$ hence on $[-1+\delta, 1-\delta] \times [a_j, b_j] \times [a_k, b_k]$.
- $a = 0, b = 1$: Since F is C^∞ hence C^3 on $\mathbb{R}$, $\phi_{01,jk}$ is C^3 on $[-1+\delta, 1-\delta] \times [a_k, b_k]$.
- $a = 1, b = 0$: Analogous to $a = 0, b = 1$.
- $a = b = 0$: Let $m \in \{1,2,3\}$. We apply Lebesgue theorem for continuity and differentiability of integrals with parameters:
  - $\sigma \mapsto \phi_{01,jk}(\sigma, y)$ is C^3 on $[-1+\delta, 1-\delta]$.
  - Straigthforward calculations of derivatives of $\phi_{01,jk}(\sigma, y)$ w.r.t. σ show that, for $\sigma \in [-1+\delta, 1-\delta]$:

$$\left|\partial_\sigma^m \phi_{01,jk}(\sigma, y)\right| \leq C(a_j, b_j, a_k, b_k, \delta, m) \exp\left(-\frac{y^2}{2}\right),$$

  where $C(a_j, b_j, a_k, b_k, \delta, m)$ is a positive constant depending on $a_j, b_j, a_k, b_k, \delta$ and m and $y \mapsto C(a_j, b_j, a_k, b_k, \delta, m) \exp\left(-\frac{y^2}{2}\right)$ is integrable on $[a_k, b_k]^c$.

It follows that $\phi_{00,jk}$ is C^3 on $[-1+\delta, 1-\delta]$. □

Appendix C: Proof of Lemma 3.3

Proof. (Lemma 3.3)

1. Let $l \in \mathbb{N}^*$. First, for all $\sigma \in [-1+\delta, 1-\delta]$, we have:

$$\int_{\mathbb{R}^2} \mathcal{L}_{jk}(\sigma, y) \mathrm{d}\mu(y) = \iint_{\mathbb{R}^2} f(x, y, \sigma) \mathrm{d}x \mathrm{d}y = 1. \tag{C.1}$$

It remains to prove that:

$$\partial_\sigma^l \left(\int_{\mathbb{R}^2} \mathcal{L}_{jk}(\sigma, y) \mathrm{d}\mu(y) \right) = \partial_\sigma^l \phi_{00,jk}(\sigma) + \int_{a_k}^{b_k} \partial_\sigma^l \phi_{01,jk}(\sigma, y) \mathrm{d}y + \int_{a_j}^{b_j} \partial_\sigma^l \phi_{10,jk}(\sigma, x) \mathrm{d}x + \iint_{[a_j,b_j]\times[a_k,b_k]} \partial_\sigma^l \phi_{11,jk}(\sigma, x, y) \mathrm{d}x \mathrm{d}y. \tag{C.2}$$

Let us deal with each of these terms:

- For $a = 0, b = 0$: it is obvious.

For the following terms, we use Lebesgue theorem for differentiability of integrals with parameters.

• For $a=0, b=1$ (and $a=1, b=0$): According to (A.2), it is clear that $\phi_{01,jk}$ is C^∞ on the compact $[1-\delta, 1+\delta] \times [a_k, b_k]$, which establishes the formula.

• For $a=1, b=1$: Analogously, (A.1) shows that $\phi_{11,jk}$ is C^∞ on the compact $[1-\delta, 1+\delta] \times [a_j, b_j] \times [a_k, b_k]$.

2. Let $l \in \mathbb{N}^*$. Let us first clarify some notations:

$$\begin{aligned}
\mathbb{E}_{\Sigma^*_{jk}}\Big(\log \mathcal{L}_{jk}(\sigma, Y)\Big) = \mathbb{E}_{\Sigma^*_{jk}}\Big(\ell(\sigma, Y)\Big) &= R(\sigma) \\
&= \int_{\mathbb{R}^2} \log \mathcal{L}_{jk}(\sigma, y) \mathcal{L}_{jk}(\Sigma^*_{jk}, y) \mathrm{d}\mu(y) \\
&= \phi_{00,jk}(\Sigma^*_{jk}) \log \phi_{00,jk}(\sigma) + \int_{a_k}^{b_k} \phi_{01,jk}(\Sigma^*_{jk}, y) \log \phi_{01,jk}(\sigma, y) \mathrm{d}y \\
&\quad + \int_{a_j}^{b_j} \phi_{10,jk}(\Sigma^*_{jk}, x) \log \phi_{10,jk}(\sigma, x) \mathrm{d}x \\
&\quad + \iint_{[a_j, b_j] \times [a_k, b_k]} \phi_{11,jk}(\Sigma^*_{jk}, x, y) \log \phi_{11,jk}(\sigma, x, y) \mathrm{d}x \mathrm{d}y.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\mathbb{E}_{\Sigma^*_{jk}}\Big(\partial^l_\sigma \log \mathcal{L}_{jk}(\sigma, Y)\Big) &= \int_{\mathbb{R}^2} \partial^l_\sigma \Big(\log \mathcal{L}_{jk}(\sigma, y)\Big) \mathcal{L}_{jk}(\Sigma^*_{jk}, y) \mathrm{d}\mu(y) \\
&= \phi_{00,jk}(\Sigma^*_{jk}) \partial^l_\sigma \log \phi_{00,jk}(\sigma) \\
&\quad + \int_{a_k}^{b_k} \phi_{01,jk}(\Sigma^*_{jk}, y) \partial^l_\sigma \log \phi_{01,jk}(\sigma, y) \mathrm{d}y \\
&\quad + \int_{a_j}^{b_j} \phi_{10,jk}(\Sigma^*_{jk}, x) \partial^l_\sigma \log \phi_{10,jk}(\sigma, x) \mathrm{d}x \\
&\quad + \iint_{[a_j, b_j] \times [a_k, b_k]} \phi_{11,jk}(\Sigma^*_{jk}, x, y) \partial^l_\sigma \log \phi_{11,jk}(\sigma, x, y) \mathrm{d}x \mathrm{d}y.
\end{aligned}$$

To show the equality $\partial^l_\sigma \mathbb{E}_{\Sigma^*_{jk}}\Big(\log \mathcal{L}_{jk}(\sigma, Y)\Big) = \mathbb{E}_{\Sigma^*_{jk}}\Big(\partial^l_\sigma \log \mathcal{L}_{jk}(\sigma, Y)\Big)$, we show the equality for each of the four terms

• For $a=0, b=0$: it is obvious.

For the three remaining terms, we use Lebesgue theorem for differentiability of integrals with parameters.

• For $a=0, b=1$ (and $a=1, b=0$):

$$\log \phi_{01,jk}(\sigma, y) = -\frac{y^2}{2} - \log \sqrt{2\pi} + \log \left[1 - F\Big(\frac{b_j - \sigma y}{\sqrt{1-\sigma^2}}\Big) + F\Big(\frac{a_j - \sigma y}{\sqrt{1-\sigma^2}}\Big)\right].$$

The function $\log \phi_{01,jk}$ is C^∞ on the compact $[-1+\delta, 1-\delta] \times [a_k, b_k]$. Therefore, for all $\sigma \in [-1+\delta, 1-\delta]$ and $y \in [a_k, b_k]$, $\Big|\partial^l_\sigma \log \phi_{01,jk}(\sigma, y)\Big|$ is upper bounded by

a constant which is integrable on the compact $[a_k, b_k]$ (with regard to the density $y \mapsto \phi_{01,jk}(\Sigma^*_{jk}, y)$).

- For $a = 1, b = 1$:

$$\log \phi_{11,jk}(\sigma, x, y) = -\log(2\pi) - \frac{1}{2}\log(1-\sigma^2) - \frac{x^2 - 2\sigma xy + y^2}{2(1-\sigma^2)}.$$

Analogously, the function $\log \phi_{11,jk}$ is C^∞ on the compact $[-1+\delta, 1-\delta] \times [a_j, b_j] \times [a_k, b_k]$. Thus, for all $\sigma \in [-1+\delta, 1-\delta]$, $x \in [a_j, b_j]$ and $y \in [a_k, b_k]$, $\left|\partial^l_\sigma \log \phi_{11,jk}(\sigma, x, y)\right|$ is upper bounded by a constant which is integrable on the compact $[a_j, b_j] \times [a_k, b_k]$ (with regard to the density $(x, y) \mapsto \phi_{11,jk}(\Sigma^*_{jk}, x, y)$).

□

Acknowledgements

We would like to thank Stéphane Robin and Stéphane Chrétien for their expert advice and comments.

References

Banerjee, Onureena and El Ghaoui, Laurent and d'Aspremont, Alexandre (2008) Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, **9**, 485–516. MR2417243

Cohen, Jr., A. Clifford (1949) On estimating the mean and standard deviation of truncated normal distributions. *J. Amer. Statist. Assoc.*, **44**, 518–525. MR0032138

Cohen, Jr., A. Clifford (1950) Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples. *Ann. Math. Statistics*, **21**, 557–569. MR0038041

Cohen, Jr., A. Clifford (1955) Restriction and selection in samples from bivariate normal distributions. *J. Amer. Statist. Assoc.*, **50**, 884–893. MR0074740

Cohen, Jr., A. Clifford (1957) On the solution of estimating equations for truncated and censored samples from normal populations. *Biometrika*, **44**, 225–236. MR0087271

Cohen, Jr., A. Clifford (1957) Restriction and selection in multinormal distributions. *Ann. Math. Statist.*, **28**, 731–741. MR0090945

Dahl, Joachim and Vandenberghe, Lieven and Roychowdhury, Vwani (2008) Covariance selection for nonchordal graphs via chordal embedding. *Optimization Methods & Software*, **23**, 501–520.

Dyer, Danny D. (1973) On moments estimation of the parameters of a truncated bivariate normal distribution. *J. Roy. Statist. Soc. Ser. C Appl. Statist.*, **22**, 287–291. MR0331597

Friedman, Jerome and Hastie, Trevor and Tibshirani, Robert (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.

Gupta, A. K. and Tracy, D. S. (1976) Recurrence relations for the moments of truncated multinormal distribution. *Comm. Statist.–Theory Methods*, **A5**, 855–865. MR0431516

Hastie Trevor, Tibshirani Robert, and Friedman Jerome (2001) *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer-Verlag. MR1851606

Lee, Lung-fei (1983) The determination of moments of the doubly truncated multivariate normal tobit model. *Economics Letters*, **11**, 245–250.

Manjunath, BG and Wilhelm, Stefan (1966) Moments calculation for the double truncated multivariate normal density. *Ann. Inst. Statist. Math.*, **18**, 107–111.

Mei, Song, Bai, Yu and Montanari, Andrea (2017) The landscape of empirical risk for non-convex losses. arXiv preprint arXiv:1607.06534.

Muthén, Bengt (1990) Moments of the censored and truncated bivariate normal distribution. *British J. Math. Statist. Psych.*, **43**, 131–143. MR1065201

Nath, G. Baikunth (1971) Estimation in truncated bivariate normal distributions. *Applied Statistics*, 313–319.

Raj Des (1953) On estimating the parameters of bivariate normal populations from doubly and singly linearly truncated samples. *Sankhy*, **12**, 277–290. MR0057517

Ravikumar, Pradeep and Wainwright, Martin J. and Raskutti, Garvesh and Yu, Bin (2011) High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.*, **5**, 935–980. MR2836766

Shah, S. M. and Jaiswal, M. C. (1966) Estimation of parameters of doubly truncated normal distribution from first four sample moments. *Ann. Inst. Statist. Math.*, **18**, 107–111. MR0196848

Singh, Naunihal (1960) Estimation of parameters of a multivariate normal population from truncated and censored samples. *J. Roy. Statist. Soc. Ser. B*, **22**, 307–311. MR0115237

Stromberg, Karl (1994) Probability for analysts. *CRC Press*, , .

Yuan, Ming and Lin, Yi (2007) Model selection and estimation in the Gaussian graphical model. *Available at SSRN 1472153* MR2367824