Annals of the Institute of Statistical Mathematics (2022) 74:1–31
https://doi.org/10.1007/s10463-021-00785-4

The finite sample properties of sparse M-estimators with pseudo-observations

Benjamin Poignard[1,2,3] · **Jean-David Fermanian**[4]

Received: 14 April 2020 / Revised: 26 October 2020 / Accepted: 25 December 2020 /
Published online: 8 April 2021
© The Institute of Statistical Mathematics, Tokyo 2021

Abstract
We provide finite sample properties of general regularized statistical criteria in the presence of pseudo-observations. Under the restricted strong convexity assumption of the unpenalized loss function and regularity conditions on the penalty, we derive non-asymptotic error bounds on the regularized M-estimator. This penalized framework with pseudo-observations is then applied to the M-estimation of some usual copula-based models. These theoretical results are supported by an empirical study.

Keywords Copulas · Non-convex regularizer · Pseudo-observations · Statistical consistency

1 Introduction

The need for a joint modelling for high-dimensional random vectors has fostered a flourishing research in sparse models. The application domains of sparse modelling have been substantially widened by the availability of massive data. For instance, when dealing with significantly large financial portfolio sizes, it is arduous to build a realistic model that is both statistically precise and provides intuitive insights among

✉ Benjamin Poignard
bpoignard@econ.osaka-u.ac.jp

Jean-David Fermanian
jean-david.fermanian@ensae.fr

[1] Graduate School of Economics, Osaka University, 1-7, Machikaneyama, Toyonaka-Shi, Osaka-Fu 560-0043, Japan

[2] Jointly Affiliated at High-Dimensional Statistical Modeling Team, RIKEN Center for Advanced Intelligence Project (AIP), 2-1 Hirosawa, Wako-Shi, Saitama-Ken 351-0198, Japan

[3] CREST-LFA, 5 avenue le Chatelier, 91120 Palaiseau, France

[4] Ensae-Crest, 5 avenue Henry le Chatelier, 91129 Palaiseau, France

asset relationships. This gave rise to sparse matrix precision estimation or sparse factor modelling, for example.

Nowadays, copulas constitute a standard way of modelling the joint distribution of a random vector. They are flexible in that they allow a separate modelling between the dependence structure and the marginal distributions of the vector components. Fully parametric copula-based models can be estimated by assuming parametric models for both the copula and the marginals and then performing maximum likelihood estimation. As an alternative, the empirical cumulative distribution of each margin can be plugged at the maximization step of the likelihood function. This semi-parametric (CML, or canonical maximum likelihood) approach has been introduced first in Genest et al. (1995) or Shi and Louis (1995), and it has become standard. Besides, nonparametric estimation of copulas treats both the copula and the marginals parameter-free and thus offers the greatest generality.

In this paper, we consider the semi-parametric approach for copula estimation. A typical problem that often arises is the model complexity in that the parameterization requires the estimation of a significantly large number of parameters. For instance, the variance–covariance matrix of a Gaussian copula involves the estimation of $q(q-1)/2$ components of the correlation matrix of a q-dimensional random vector. Mixtures of copulas may also involve numerous parameters. Hopefully, regularizing a copula-based model through a penalization procedure offers an interesting way to tackle the over-fitting issue.

Most of the theoretical analysis of sparsity-based estimators has been developed for i.i.d. variables and convex loss functions: see Knight and Fu (2000), Fan and Li (2001), Zou and Zhang (2009), concerning their asymptotic properties; see also van de Geer and Bühlmann (2009), for finite-sample properties, for instance. Recent studies proposed theoretical results for sparse estimators that explicitly manage potentially non-convex statistical criteria. For instance, Loh and Wainwright (2015) derive finite-sample error bounds on penalized M-estimators, where the non-convexity potentially comes from the objective function or from the regularizer. Using the same setting, Loh and Wainwright (2017) provide the support recovery property for a broad range of penalized models such as the Gaussian graphical model, or the corrected LASSO for error-in-variables linear models. In both studies, the restricted strong convexity (Negahban et al. 2012) of the unpenalized loss function and suitable regularity conditions on the penalty function enable us to prove that any local minimum of the penalized function lies within statistical precision of the true sparse parameter, and to provide conditions for variable selection consistency. In our study, we propose to extend their framework to pseudo-observation-based models for some loss functions that satisfy the restricted strong convexity condition. To do so, we state consistency results for very general penalization functions, in which we explicitly are able to manage pseudo-observations. It is widely recognized that replacing "true" observations by estimated ones (typically after transformations by marginal empirical distributions) significantly complicates inference theory: see Ghoudi and Rémillard (1998, 2004), van der Vaart and Wellner (2007), for instance. Our framework encompasses both parametric and semi-parametric models. It is then applied to some usual copula models: Gaussian and Student copulas, mixtures, etc. To the best of our knowledge, this paper is the

first attempt to build bridges between general penalized (non-convex) M-estimators, pseudo-observations and the semi-parametric inference of copula models.

The remainder of the paper is organized as follows. In Sect. 2, we start with a description of the copula-based model framework and of our penalized statistical criterion. Then, we provide some finite sample error bounds on the regularized estimators for pseudo-observation-based models. Incidentally, we correct a mistake in the initial result (Theorem 1 in Loh and Wainwright 2015) that was stated in the usual case of "true" observations. Section 3 is dedicated to the application of these results to some usual semi-parametric copula models. Section 4 illustrates these theoretical results through a short simulated experiment. The main proofs and additional elements are postponed into the "Appendix".

2 Nonconvex penalized criteria based on pseudo-observations

2.1 Copula models

Let us start with a n sample of n realizations of a random vector $\boldsymbol{X} \in \mathbb{R}^q$, $\boldsymbol{X} := (X_1, \ldots, X_q)'$. This sample is denoted as $\mathcal{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$. Note that the observations may be dependent or not. As usual in the copula world (or elsewhere), we are more interested in the "reduced" random variables $U_k = F_k(X_k)$, $k = 1, \ldots, q$, where F_k denotes the cdf of X_k. When the underlying laws are continuous, the variables U_k are uniformly distributed on $[0, 1]$ and the joint law of $\boldsymbol{U} := (U_1, \ldots, U_q)$ is the uniquely defined copula of $\boldsymbol{X}$. This should imply we could work with the sample $\mathcal{U} = (\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n)$ instead of $\mathcal{X}$. Nonetheless, since the marginal cdfs' X_k are unknown, they have to be replaced by consistent estimates. Therefore, we rather build a sample of pseudo-observations $\hat{\boldsymbol{U}}_i = (\hat{U}_{i,1}, \ldots, \hat{U}_{i,q})$, $i = 1, \ldots, n$, obtained from the initial sample $\mathcal{X}$. For instance, it is a usual practice to set $\hat{U}_{i,k} = \hat{F}_k(X_{i,k})$ for every $i = 1, \ldots, n$ and every $k = 1, \ldots, m$, where $\hat{F}_k$ denotes a consistent estimate of F_k. Obviously, the most straightforward estimate of F_k is given by the usual empirical cdf $F_{n,k}(s) := n^{-1} \sum_{i=1}^{n} \mathbf{1}_{X_{i,k} \leqslant s}$ or one of its rescaled versions, but such choices are not mandatory hereafter. Indeed, we will state our results under some assumptions on the pseudo-observations themselves. Since we consider parametric copula models, the law of $\boldsymbol{U}$ belongs to a family $\mathcal{P} := \{\mathbb{P}_\theta, \theta \in \Theta\}$, where Θ denotes a convex subset of $\mathbb{R}^d$. The "true" value of the parameter is denoted by θ_0.

2.2 The optimization program

We are interested in the finite-sample properties of regularized M-estimators for both parametric and semi-parametric models. The non-convexity in the statistical criterion can potentially come from the unpenalized loss function, from the regularizer, or even from both of them.

More precisely, consider a loss function $\mathbb{G}_n$ from $\Theta \times [0, 1]^{qn}$ to $\mathbb{R}$. The value $\mathbb{G}_n(\theta; \boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ evaluates the quality of the "fit" when the sample $\mathcal{U}$ is given by $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$, i.e. given $\boldsymbol{U}_i = \boldsymbol{u}_i$ for every $i = 1, \ldots, n$ and under $\mathbb{P}_\theta$. Typically, with

i.i.d. data, $\mathbb{G}_n(\theta;\boldsymbol{u}_1,\dots,\boldsymbol{u}_n)$ is the empirical loss associated with a continuous function $\ell : \Theta \times [0,1]^q \to \mathcal{R}_+$, i.e.

$$\mathbb{G}_n(\theta;\boldsymbol{u}_1,\dots,\boldsymbol{u}_n) := \frac{1}{n}\sum_{i=1}^{n}\ell(\theta,\boldsymbol{u}_i).$$

Typically, the function ℓ is defined as a least square error, or minus a log-likelihood function, but our framework is more general for the moment.

The quantity $\mathbb{G}_n(\theta,\mathcal{U})$ cannot be calculated since we do not observe realizations of $\boldsymbol{U}$ in practice. Therefore, denoting $\hat{\mathcal{U}} := (\hat{\boldsymbol{U}}_1,\cdots,\hat{\boldsymbol{U}}_n)$, the loss function $\mathbb{G}_n(\theta,\mathcal{U})$ will be approximated by $\mathbb{G}_n(\theta,\hat{\mathcal{U}})$, a quantity called "pseudo-empirical" loss function. Then, the problem of interest becomes

$$\hat{\theta} = \underset{\theta : g(\theta)\leqslant R}{\arg\min}\ \{\mathbb{G}_n(\theta,\hat{\mathcal{U}}) + \boldsymbol{p}(\lambda_n,\theta)\}, \tag{1}$$

where $\boldsymbol{p}(\lambda_n,.) : \mathbb{R}^d \to \mathbb{R}$ is a regularizer and $\lambda_n \geqslant 0$ is the regularization parameter, which depends on the sample size and enforces a particular type of sparse structure in the solution. Moreover, $g : \mathbb{R}^d \to \mathbb{R}$, a convex function, and a supplementary regularization parameter $R > 0$ ensure the existence of local/global optima (see Loh and Wainwright 2015). For technical reasons, we include the side condition $g(\theta) \geqslant \|\theta\|_1$ for every θ. The function $\theta \to \mathrm{E}[\ell(\theta,\boldsymbol{U})]$ is supposed to be uniquely minimized at $\theta = \theta_0$ so that $\mathrm{E}[\nabla_\theta \mathbb{G}_n(\theta_0,\mathcal{U})] = 0$.

We impose that $g(\theta_0) \leqslant R$, so that θ_0 is a feasible point.

Hereafter, we will consider general losses and penalties, both being non-convex possibly. As a consequence, due to the potential optimal duality gap between primal and dual optimization programs, it would not be possible to remove the constraint $g(\theta) \leqslant R$ by considering penalized losses (or the opposite).

2.3 Potentially non-convex losses and regularization functions

This section provides the assumptions required for our theoretical setting. They mostly come from the framework of Loh and Wainwright (2015, 2017).

Assumption 1 Sparsity assumption: $\mathrm{card}(\mathcal{A}) = k_0 < d, \mathcal{A} = \{i : \theta_{0,i} \neq 0\}$.

Assumption 2 We consider coordinate-separable penalty (or regularizer) functions $\boldsymbol{p}(.,.) : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$, i.e. $\boldsymbol{p}(\lambda_n,\theta) = \sum_{k=1}^{d} p(\lambda_n,\theta_k)$. More over, for some $\mu \geqslant 0$, the regularizer $\boldsymbol{p}(\lambda_n,.)$ is assumed to be μ-amenable, in the sense that

(i) $\rho \mapsto p(\lambda_n,\rho)$ is symmetric around zero and $\boldsymbol{p}(\lambda_n,0) = 0$.
(ii) $\rho \mapsto p(\lambda_n,\rho)$ is non-decreasing on $\mathbb{R}_+$.
(iii) $\rho \mapsto p(\lambda_n,\rho)/\rho$ is non-increasing on $\mathbb{R}_+\backslash\{0\}$.
(iv) $\rho \mapsto p(\lambda_n,\rho)$ is differentiable for any $\rho \neq 0$.
(v) $\lim_{\rho\to 0^+} p'(\lambda_n,\rho) = \lambda_n$.
(vi) $\rho \mapsto p(\lambda_n,\rho) + \mu\rho^2/2$ is convex for some $\mu \geqslant 0$.

The regularizer $\boldsymbol{p}(\lambda_n, .)$ is said to be (μ, γ)-amenable if, in addition,

(vii) there exists $\gamma \in (0, \infty)$ such that $p'(\lambda_n, \rho) = 0$ for $\rho \geqslant \lambda_n \gamma$.

We denote by $\boldsymbol{q} : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ the function $\boldsymbol{q}(\lambda_n, \theta) = \lambda_n \|\theta\|_1 - \boldsymbol{p}(\lambda_n, \theta)$ so that the function $\mu\|\theta\|_2^2/2 - \boldsymbol{q}(\lambda_n, \theta)$ is convex.

Assumption 1 implies that the true (unknown) support is sparse, that is the vector θ_0 contains some zero components. Note that θ_0 is independent of the sample size n. To derive our theoretical properties, Assumption 2 provides regularity conditions that potentially encompass non-convex functions. These regularity conditions are the same as in Loh and Wainwright (2015, 2017) or Loh (2017). In this paper, we focus on the LASSO, the SCAD due to Fan and Li (2001) and the MCP due to Zhang (2010), given by

$$
\begin{aligned}
\textbf{LASSO}: \; p(\lambda_n, \rho) &= \lambda_n |\rho|, \\
\textbf{MCP}: \; p(\lambda_n, \rho) &= \operatorname{sign}(\rho) \lambda_n \textstyle\int_0^{|\rho|} (1 - z/(\lambda_n b_{mcp}))_+ \mathrm{d}z, \\
\textbf{SCAD}: \; p(\lambda_n, \rho) &= \begin{cases} \lambda_n |\rho|, & \text{for } |\rho| \leqslant \lambda_n, \\ -(\rho^2 - 2 b_{\text{scad}} \lambda_n |\rho| + \lambda_n^2)/(2(b_{\text{scad}} - 1)), & \text{for } \lambda_n \leqslant |\rho| \leqslant b_{\text{scad}} \lambda_n, \\ (b_{\text{scad}} + 1)\lambda_n^2/2, & \text{for } |\rho| > b_{\text{scad}} \lambda_n, \end{cases}
\end{aligned}
$$

where $b_{scad} > 2$ and $b_{mcp} > 0$ are fixed parameters for the SCAD and MCP, respectively. The LASSO is a μ-amenable regularizer, whereas the SCAD and the MCP regularizers are (μ, γ)-amenable. More precisely, $\mu = 0$ (resp. $\mu = 1/(b_{scad} - 1)$, resp. $\mu = 1/b_{mcp}$) for the LASSO (resp. SCAD, resp. MCP).

As for many parametric models, numerous empirical log-likelihoods associated with copulas are not concave functions in their parameters, at finite distance and globally on $\boldsymbol{\Theta}$. Moreover, this is still the case for some popular regularizers, as SCAD. Therefore, we would like to weaken such convexity/concavity assumption so that $\hat{\theta}$ would be a consistent estimate of θ_0, for which we could evaluate its accuracy. To this goal, the restricted strong convexity is a key ingredient that allows the management of non-convex loss functions. Intuitively, we would like to handle a loss function that locally admits some curvature. To do so, we will weaken the most often assumed local strong convexity property of the loss function. Remind that the strong convexity of a differentiable loss function corresponds to a strictly positive lower bound on the eigenvalues of the Hessian matrix uniformly valid over a local region around the true parameter. The notion of restricted strong convexity weakens the (local) strong convexity by adding a tolerance term. A detailed explanation is provided in Negahban et al. (2012).

Being more specific and slightly extending the definition of Loh and Wainwright (2017), we say that an empirical loss function $\mathbb{L}_n$ satisfies the restricted strong convexity condition (RSC) at θ if there exist two positive functions α_1, α_2 and two nonnegative functions ν_1, ν_2 of (θ, n, d) such that, for any $\Delta \in \mathbb{R}^d$,

$$
\langle \nabla_\theta \mathbb{L}_n(\theta + \Delta) - \nabla_\theta \mathbb{L}_n(\theta), \Delta \rangle \geqslant \alpha_1 \|\Delta\|_2^2 - \nu_1 \|\Delta\|_1^2, \text{ if } \|\Delta\|_2 \leqslant 1, \tag{2}
$$

$$
\langle \nabla_\theta \mathbb{L}_n(\theta + \Delta) - \nabla_\theta \mathbb{L}_n(\theta), \Delta \rangle \geqslant \alpha_2 \|\Delta\|_2 - \nu_2 \|\Delta\|_1, \text{ if } \|\Delta\|_2 > 1. \tag{3}
$$

Note that the (RSC) property is fundamentally local and that α_k, ν_k, $k = 1, 2$, depend on the chosen θ. In Loh and Wainwright (2017), their so-called (RSC) condition is similar, but the latter coefficients do not depend on (n, d). This is not necessary in general, but we will need such extensions for some copula models of Sect. 3. In the latter section, we will apply the (RSC) condition with $\mathbb{L}_n(\theta) = \mathbb{G}_n(\theta, \mathcal{U})$ ($\mathcal{U}$ containing unfeasible observations, most of the time) and/or $\mathbb{L}_n(\theta) = \mathbb{G}_n(\theta, \hat{\mathcal{U}})$ (with the so-called pseudo-observations). Moreover, to weaken notations, we simply write α_k and ν_k, $k = 1, 2$, by skipping their implicit arguments (θ, n, d).

Remark 1 In the latter (RSC) condition, the threshold "one" for $\|\Delta\|_2$ has been chosen for convenience. Actually, it is always possible to reparameterize the model with $\bar{\theta} := \zeta\theta$ for some $\zeta > 0$. Therefore, the criterion becomes $\overline{\mathbb{L}}_n(\bar{\theta}) := \mathbb{L}_n(\zeta\theta)$. Since $\nabla_\theta \overline{\mathbb{L}}_n(\theta) = \zeta \nabla_{\bar{\theta}} \overline{\mathbb{L}}_n(\bar{\theta})$, the (RSC) is rewritten as

$$\langle \nabla_{\bar{\theta}} \overline{\mathbb{L}}_n(\bar{\theta} + \bar{\Delta}) - \nabla_{\bar{\theta}} \overline{\mathbb{L}}_n(\bar{\theta}), \bar{\Delta} \rangle \geqslant \bar{\alpha}_1 \|\bar{\Delta}\|_2^2 - \bar{\nu}_1 \|\bar{\Delta}\|_1^2, \ \|\bar{\Delta}\|_2 \leqslant \zeta,$$
$$\langle \nabla_{\bar{\theta}} \overline{\mathbb{L}}_n(\bar{\theta} + \bar{\Delta}) - \nabla_{\bar{\theta}} \overline{\mathbb{L}}_n(\bar{\theta}), \bar{\Delta} \rangle \geqslant \bar{\alpha}_2 \|\bar{\Delta}\|_2 - \bar{\nu}_2 \|\bar{\Delta}\|_1, \ \|\bar{\Delta}\|_2 \geqslant \zeta,$$

with the new constants $(\bar{\alpha}_1, \bar{\nu}_1, \bar{\alpha}_2, \bar{\nu}_2) := (\alpha_1/\zeta^2, \nu_1/\zeta^2, \alpha_2/\zeta, \nu_2/\zeta)$.

2.4 Finite sample consistency results

Now, following Loh and Wainwright (2015), we provide some error bounds over the penalized parameters, assuming that the loss function satisfies the (RSC) condition and the penalty is μ-amenable. This is the purpose of the next theorem, which is stated in a deterministic manner. The bounds can actually hold with a high probability, depending on the upper bound over the loss function $\mathbb{G}_n(., \hat{\mathcal{U}})$ and the choice of λ_n: see the discussion in Sect. 2.5.

Theorem 1 *Suppose the objective function $\mathbb{G}_n(\cdot, \hat{\mathcal{U}}) : \mathbb{R}^d \mapsto \mathbb{R}$ satisfies the (RSC) condition at θ_0. Moreover, $\boldsymbol{p}(\lambda_n, .)$ is assumed to be μ-amenable, with $3\mu < 4\alpha_1$ and $4R\nu_2 \leqslant \alpha_2$. Assume*

$$4 \max \left\{ \|\nabla_\theta \mathbb{G}_n(\theta_0, \hat{\mathcal{U}})\|_\infty, 2R\nu_1 \right\} \leqslant \lambda_n \leqslant \frac{\alpha_2}{6R}. \tag{4}$$

Then, for every n, any stationary point $\hat{\theta}$ of (1) satisfies

$$\|\hat{\theta} - \theta_0\|_2 \leqslant \frac{6\lambda_n \sqrt{k_0}}{4\alpha_1 - 3\mu}, \text{ and } \|\hat{\theta} - \theta_0\|_1 \leqslant \frac{6(16\alpha_1 - 9\mu)}{(4\alpha_1 - 3\mu)^2} \lambda_n k_0.$$

The proof is provided in Appendix A.1. The latter result is an improvement in Th. 1 in Loh (2017), since our bounds are sharper under similar assumptions.

In Theorem 1, Condition (4) is satisfied even for large n because, under such circumstances, λ_n and ν_1 both tend to zero. Indeed, the "constant" ν_1 is typically a function of (n, q, d, θ_0) and of $O(n^{-1/2})$ in many models.

Remark 2 The result above is based on an optimization reasoning only, and not on probabilistic arguments. Then, the previous theorem could be rewritten exactly similarly, replacing $\mathbb{G}_n(\theta, \hat{\mathcal{U}})$ by $\mathbb{G}_n(\theta, \mathcal{U})$ or even by any empirical loss function $\mathbb{L}_n(\theta)$ that satisfies the (RSC) condition. In particular, it is not necessary to deal with pseudo-observations.

Our proof of Theorem 1 follows the proof of Theorem 1 in Loh and Wainwright (2015) but is not identical. Indeed, a key argument of the latter authors comes from their Lemma 5, that would imply in our proof

$$0 \leqslant 3\boldsymbol{p}(\lambda_n, \theta_0) - \boldsymbol{p}(\lambda_n, \hat{\theta}) \leqslant \lambda_n(3\|\Delta_{\mathcal{M}}\|_1 - \|\Delta_{\mathcal{M}^c}\|_1), \tag{5}$$

where $\mathcal{M}$ denotes the index set of the k_0 largest elements $|\hat{\theta}_i - \theta_{0,i}|$ and $\Delta := \hat{\theta} - \theta_0$ (see their Equation (25)). Unfortunately, this lemma is wrong. Indeed, with its notations, choose $\beta^* = (2, 0)$, $\beta = (a, b)$, for some positive constants a and $b < 1$. Moreover, set $\rho_\lambda(\beta) = \lambda|\beta|$ (with $L = 1$). Set $\xi = 2$. Then, $\nu = (a - 2, b)$, $\nu_A = (a - 2, 0)$ and $\nu_{A^c} = (0, b)$. The asserted result of this Lemma 5 is here $2|\beta^*| - |\beta| \leqslant 2|\nu_A| - |\nu_{A^c}|$, or even $4 - a - b \leqslant 2|a - 2| - b$. This is clearly false in general: set $a = 3/2$, for instance.

For the sake of completeness, let us state a corrected version of Th. 1 in Loh and Wainwright (2015), using their own assumptions. They have reparameterized the (RSC) assumption by setting $\nu_1 = \tau_1 \ln d/n$ and $\nu_2 = \tau_2(\ln d/n)^{1/2}$.

Corollary 1 *Suppose the objective function $\mathbb{G}_n(\cdot, \hat{\mathcal{U}}) : \mathbb{R}^d \mapsto \mathbb{R}$ satisfies the (RSC) condition at θ_0. Moreover, $\boldsymbol{p}(\lambda_n, .)$ is assumed to be μ-amenable, with $3\mu < 4\alpha_1$. Assume*

$$4 \max\left\{ \|\nabla_\theta \mathbb{G}_n(\theta_0, \hat{\mathcal{U}})\|_\infty, \frac{\alpha_2}{2}\sqrt{\frac{\ln d}{n}} \right\} \leqslant \lambda_n \leqslant \frac{\alpha_2}{6R}, \tag{6}$$

and $n \geqslant 16R^2 \max\{\tau_1^2, \tau_2^2\} \ln d/\alpha_2^2$. Then, any stationary point $\hat{\theta}$ of (1) satisfies

$$\|\hat{\theta} - \theta_0\|_2 \leqslant \frac{6\lambda_n\sqrt{k_0}}{4\alpha_1 - 3\mu}, \text{ and } \|\hat{\theta} - \theta_0\|_1 \leqslant \frac{6(16\alpha_1 - 9\mu)}{(4\alpha_1 - 3\mu)^2}\lambda_n k_0.$$

The proof can be deduced from a few straightforward modifications of Theorem 1's proof. Check that the assumptions of Corollary 1 are stronger than those of Theorem 1: indeed, $n \geqslant 16R^2\tau_2^2 \ln d/\alpha_2^2$ is equivalent to our assumption $\alpha_2 \geqslant 4R\nu_2$ and $n \geqslant 16R^2\tau_1^2 \ln d/\alpha_2^2$ means $\alpha_2(\ln d/n)^{1/2} \geqslant 4R\nu_1$. The latter inequality and the assumed condition $2\alpha_n(\ln d/n)^{1/2} \leqslant \lambda_n$, that appears in (6), imply $8R\nu_1 \leqslant \lambda_n$, as in (4).

2.5 Discussion

To evaluate the scope of our result, consider again a general penalized M-estimator based on pseudo-observations: $\hat{\theta} = \underset{\theta : g(\theta) \leqslant R}{\arg\min} \{\mathbb{G}_n(\theta, \hat{\mathcal{U}}) + \boldsymbol{p}(\lambda_n, \theta)\}$. Therefore, by usual differentiation, we have

$$\langle \nabla_\theta \mathbb{G}_n(\theta+\Delta,\hat{\mathcal{U}}) - \nabla_\theta \mathbb{G}_n(\theta,\hat{\mathcal{U}}),\Delta\rangle = \langle \nabla_\theta \mathbb{G}_n(\theta+\Delta,\mathcal{U}) - \nabla_\theta \mathbb{G}_n(\theta,\mathcal{U}),\Delta\rangle$$
$$+ \Delta'\Big(\nabla_{\theta',\theta,\mathcal{U}}\mathbb{G}_n(\theta^*,\mathcal{U}^*)\cdot\big(\hat{\mathcal{U}}-\mathcal{U}\big)\Big)\Delta := W_{1,n}+W_{2,n},$$

with obvious notations. Assume that the (RSC) assumption applies with $\mathcal{U}$, an unobservable i.i.d. sample (the "usual" situation): when $\|\Delta\|_2 \leqslant 1$,

$$\langle \nabla_\theta \mathbb{G}_n(\theta+\Delta,\mathcal{U}) - \nabla_\theta \mathbb{G}_n(\theta,\mathcal{U}),\Delta\rangle \geqslant \alpha_1\|\Delta\|_2^2 - \nu_1\|\Delta\|_1^2.$$

Working with pseudo-observations induces an additional amount of noise, summarized through $W_{2,n}$. But this noisy term, due to the discrepancy between the pseudo-observations $\hat{\boldsymbol{U}}_i$ and their unobservable targets $\boldsymbol{U}_i$, can be controlled. Indeed, if we evaluate our pseudo-observations with usual empirical cdfs', the DKW inequality for i.i.d. observations yields

$$\mathbb{P}\Big(\sup_{k=1,\dots,q}\sup_{i=1,\dots,n}|\hat{U}_{i,k}-U_{i,k}|^2>\epsilon\Big) \leqslant 2q\exp\big(-2n\epsilon\big).$$

Therefore, for any positive constant C_0, $\sup_{i,k}|\hat{U}_{i,k}-U_{i,k}| \leqslant C_0/\sqrt{n}$ with a probability larger than $1-2q\exp(-2C_0^2)$, yielding

$$|W_{2,n}| \leqslant \sum_{k,l=1}^{d}|\Delta_k\Delta_l|\|\partial^2_{\theta_k,\theta_l}\nabla_{\mathcal{U}}\mathbb{G}_n(\theta^*,\mathcal{U}^*)\|_1\|\hat{\mathcal{U}}-\mathcal{U}\|_\infty$$
$$\leqslant \frac{C_0}{\sqrt{n}}\sup_{k,l}\|\partial^2_{\theta_k,\theta_l}\nabla_{\mathcal{U}}\mathbb{G}_n(\theta^*,\mathcal{U}^*)\|_1\|\Delta\|_1^2.$$

We deduce $\langle \nabla_\theta \mathbb{G}_n(\theta+\Delta,\hat{\mathcal{U}}) - \nabla_\theta \mathbb{G}_n(\theta,\hat{\mathcal{U}}),\Delta\rangle \geqslant \alpha_1\|\Delta\|_2^2 - \nu_1'\|\Delta\|_1^2$, with

$$\nu_1' := \nu_1 + \frac{C_0}{\sqrt{n}}\sup_{k,l}\|\partial^2_{\theta_k,\theta_l}\nabla_{\mathcal{U}}\mathbb{G}_n(\theta^*,\mathcal{U}^*)\|_1, \tag{7}$$

with the same probability as above. In Loh and Wainwright's papers and usual samples, the constant ν_1 is rather of magnitude $\ln d/n$ (when it is nonzero). Equation (7) means that ν_1' is larger than ν_1, and the gap is $O_P(d^2p/\sqrt{n})$ when $\mathbb{G}_n$ is a sample average, as usual. Thus, the (RSC) condition is satisfied more easily with ν_1' than with ν_1, but there is a price to be paid: this larger constant tends to increase λ_n, and then the upper bound of $\|\hat{\theta}-\theta_0\|_k$, $k\in\{1,2\}$. In particular, λ_n has to be at least of order $d^2p/\sqrt{n}$ "in general" (i.e. without taking into account some particular model features). Coming back to (4), note that $\|\nabla_\theta\mathbb{G}_n(\theta_0,\hat{\mathcal{U}})\|_\infty$ is typically of order $O_P(d^2pn^{-1/2})$ too. Indeed, as above, a limited expansion yields

$$\|\nabla_\theta\mathbb{G}_n(\theta_0,\hat{\mathcal{U}})\|_\infty \leqslant \|\nabla_\theta\mathbb{G}_n(\theta_0,\mathcal{U})\|_\infty + \sum_{k=1}^{d}\|\partial_{\theta_k}\nabla_{\mathcal{U}}\mathbb{G}_n(\theta_0,\mathcal{U}^*)\|_1\|\hat{\mathcal{U}}-\mathcal{U}\|_\infty$$
$$\leqslant \|\nabla_\theta\mathbb{G}_n(\theta_0,\mathcal{U})\|_\infty + \frac{C_0 d}{\sqrt{n}}\sup_k\|\partial_{\theta_k}\nabla_{\mathcal{U}}\mathbb{G}_n(\theta_0,\mathcal{U}^*)\|_1,$$

with a probability larger than $1 - 2q\exp(-2C_0^2)$. The first term on the r.h.s. is of order $(\ln d/n)^{1/2}$ most often (Loh, 2017, p.876), and the second one is $O_P(d^2p/\sqrt{n})$, the same rate as for ν_1'. Therefore, by working with pseudo-observations, the usual rate $\lambda_n \asymp (\ln d/n)^{1/2}$ will be replaced by the more demanding rate $d^2p/\sqrt{n}$, a reasonable cost when d is not "too large".

Remark 3 Alternatively, if we are able to bound from below the Hessian matrix $\nabla_{\theta,\theta'}\mathbb{G}_n(\theta,\hat{\mathcal{U}})$ by a positive definite matrix Ω_0 uniformly w.r.t. $\theta \in \Theta$, we obviously satisfy the (RSC) condition with $\alpha_k = \lambda_{\min}(\Omega_0)$ and $\nu_k = 0$, $k = 1, 2$. The price to be paid is to restrain the domain Θ, sometimes excessively. We will see an example of such a situation in Sect. 3.1.

Among the numerous potential loss functions, a particularly interesting case is obtained through Bregman divergences (Bregman 1967; Censor and Zenios 1998): let $\phi(\cdot)$ be a differentiable and strictly convex function defined on a convex subset Θ of $\mathbb{R}^d$. The Bregman divergence between two vectors θ_1 and θ_2 in Θ is $D(\theta_1,\theta_2) := \phi(\theta_1) - \phi(\theta_2) - \nabla\phi(\theta_2)'\big(\theta_1 - \theta_2\big)$. The latter quantity is nonnegative, is zero when $\theta_1 = \theta_2$, can be easily symmetrized, induces nice optimization algorithms and satisfies many interesting properties (see "Appendix A" in Ravikumar et al. 2011). Bregman divergence has been used in many areas, including clustering (Banerjee et al. 2005), graphical models (Cai and Zhou 2012), speech processing (Gray et al. 1980), etc. The squared Euclidian norm, the Kullback–Leibler divergence or the Mahalanobis distance, among many others, are particular cases of Bregman divergences. In our case, we are interested in situations, where there exists some empirical counterpart of the model parameter θ, called θ_n, and the empirical loss is given by

$$\begin{aligned}\mathbb{G}_{n,\omega}(\theta) &:= \omega D(\theta,\theta_n) + (1-\omega)D(\theta_n,\theta)\\ &= (2\omega-1)\big\{\phi(\theta) - \phi(\theta_n)\big\} - \big\{\omega\nabla\phi(\theta_n) - (1-\omega)\nabla\phi(\theta)\big\}'\big(\theta - \theta_n\big),\end{aligned}$$

for some known constant $\omega \in [0,1]$. This is an extension of the standard Bregman divergence framework as the loss is now a balance between $D(\theta,\theta_n)$ and its switched argument version. We skip the dependence of $\mathbb{G}_{n,\omega}$ w.r.t. the sample or the pseudo-sample, because both cases are similar here. For instance, in many copula models, there exists a one-to-one mapping between the parameter θ and some dependence measures ρ (as Kendall's tau, Spearman's rho, typically). In other words, the model can be re-parameterized by ρ. Since there exist natural and rather simple empirical estimators for many dependence measures, we can estimate ρ by minimizing a divergence between this vector and ρ_n. Another example with matrices of parameters is given in Sect. 3.1.

Simple calculations yield

$$\nabla_\theta\mathbb{G}_{n,\omega}(\theta) = \omega\big\{\nabla\phi(\theta) - \nabla\phi(\theta_n)\big\} + (1-\omega)\nabla^2\phi(\theta)\big(\theta - \theta_n\big),$$

and this implies

$$\nabla_\theta \mathbb{G}_{n,\omega}(\theta) - \nabla_\theta \mathbb{G}_{n,\omega}(\theta_0) = \omega\big\{\nabla\phi(\theta) - \nabla\phi(\theta_0)\big\} + (1-\omega)\nabla^2\phi(\theta_0)(\theta-\theta_0) + (1-\omega)\big\{\nabla^2\phi(\theta) - \nabla^2\phi(\theta_0)\big\}(\theta-\theta_n).$$

If ϕ is strongly convex, then there exists a nonnegative constant α_1 s.t.

$$\langle \nabla\phi(\theta) - \nabla\phi(\theta_0), \theta - \theta_0\rangle \geqslant \alpha_1 \|\theta-\theta_0\|_2^2, \tag{8}$$

and this constant α_1 is smaller than the minimum eigenvalue of the Hessian matrix $\nabla^2\phi(\theta_0)$. If ϕ is three times differentiable on the interior of Θ and its derivatives are bounded, then, for every $\Delta := \theta - \theta_0$ s.t. $\theta_0 + \Delta \in \Theta$,

$$\begin{aligned}\langle \nabla_\theta \mathbb{G}_{n,\omega}(\theta_0+\Delta) - \nabla_\theta \mathbb{G}_{n,\omega}(\theta_0), \Delta\rangle &\geqslant \omega\alpha_1\|\Delta\|_2^2 + (1-\omega)\Delta'\nabla^2\phi(\theta_0)\Delta \\ &- (1-\omega)|\Delta'\big\{\nabla^2\phi(\theta_0+\Delta) - \nabla^2\phi(\theta_0)\big\}(\theta_0+\Delta-\theta_n)| \\ &\geqslant \alpha_1\|\Delta\|_2^2 - \nu_1\|\Delta\|_1^2,\end{aligned} \tag{9}$$

by setting $\nu_1 := 2(1-\omega)\sup_{\theta\in\Theta}\|\theta\|_1 \sup_{i,j,k}\sup_{\theta\in\Theta}|\partial^3_{\theta_i,\theta_j,\theta_k}\phi(\theta)|$. This provides another justification of the (RSC) condition (in particular the need for the constant ν_1), in the case of usual samples or pseudo-observations. Interestingly, $\nu_1 = 0$ by choosing $\omega = 1$ and then $\mathbb{G}_n(\theta) = D(\theta, \theta_n)$.

Remark 4 As in Remark 3, it may be interesting to impose the (RSC) condition with $\nu_1 = 0$, by noticing that

$$\nabla^2\phi(\theta)(\theta-\theta_n) - \nabla^2\phi(\theta_0)(\theta_0-\theta_n) = \big\{\nabla^3\phi(\theta^*)(\theta^*-\theta_n) + \nabla^2\phi(\theta^*)\big\}(\theta-\theta_0),$$

for some vector θ^* between θ and θ_0, and under some conditions of regularity on ϕ. Then, by constraining Θ, it is sometimes possible to impose $\nabla^3\phi(\theta^*)(\theta^*-\theta_n) + \nabla^2\phi(\theta^*) \succeq \Omega$, for some definite positive matrix Ω. This would imply

$$\langle \nabla_\theta \mathbb{G}_{n,\omega}(\theta_0+\Delta) - \nabla_\theta \mathbb{G}_{n,\omega}(\theta_0), \Delta\rangle \geqslant \omega\alpha_1\|\Delta\|_2^2 + \Delta'\Omega\Delta \geqslant \big(\omega\alpha_1 + \lambda_{\min}(\Omega)\big)\|\Delta\|_2^2.$$

3 Application to some copula families

In this section, we provide some insights regarding the applicability of the finite sample results of Sects. 2.4–2.5 to some copula-based models. We restrict ourselves to i.i.d. samples hereafter. This means we choose a loss function $\mathbb{G}_n$ that will be typically but not exclusively given by (minus) the log-likelihood: $\ell(\theta, \boldsymbol{u}) = -\ln c(\boldsymbol{u}, \theta)$, where $c(\cdot, \theta)$ denotes the copula density of $\boldsymbol{X}$ (or $\boldsymbol{U}$, equivalently), given the parameter value θ. In particular, we will check when the (RSC) condition applies. Hereafter, we will denote by $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ a set of n random vectors in $[0,1]^q$. This is a generic notation for a usual i.i.d. sample $\mathcal{U}$ or for a sample of n pseudo-observations $\hat{\mathcal{U}}$ (as defined above), unless explicitly stated otherwise. Therefore, we will simultaneously cover the two

cases of known and/or unknown margins. In other words, setting $\vec{\boldsymbol{u}} := (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ as the second argument of $\mathbb{G}_n$, this means that $\vec{\boldsymbol{u}}$ can represent $\mathcal{U}$ or $\hat{\mathcal{U}}$.

Now, for every copula family, we will try to answer the following questions: what is the associated criterion $\mathbb{G}_n$? Is the optimization program a concave function of θ? Is the (RSC) satisfied? And, finally, can Theorem 1 be applied?

3.1 Gaussian copula models

If the underlying copula of the random vectors $\boldsymbol{X}$ is Gaussian, this means this copula is

$$C(\boldsymbol{u}, \theta) = \boldsymbol{\Phi}_{\Sigma}(\boldsymbol{\Phi}^{-1}(u_1), \ldots, \boldsymbol{\Phi}^{-1}(u_q)),$$

for any $\boldsymbol{u} \in (0,1)^q$, where $\boldsymbol{\Phi}$ and $\boldsymbol{\Phi}_{\Sigma}$, respectively, denote the cdf of a standard univariate Gaussian r.v. and of a centered Gaussian vector whose covariance matrix is Σ. Actually, there are ones in the diagonal of Σ, meaning that this is a correlation matrix. Note that Σ is a $q \times q$ matrix, and the number of free parameters is $d = q(q-1)/2$.

The parameter θ will be defined as the column vector of the Σ-components that are located below the main diagonal, excluding the diagonal. It could also be possible to include the ones of the diagonal into θ (i.e. $\theta = vech(\Sigma)$), or even to consider all the stacked coefficients of Σ itself (i.e. $\theta = vec(\Sigma)$). We denote the total number of nonzero entries of Σ as $k_0 = |\mathcal{A}|$, with $\mathcal{A} = \{(i,j) : i > j \text{ and } \Sigma_{0,(i,j)} \neq 0\}$. For convenience and with a slight abuse of notation, we write the loss $\mathbb{G}_n(\Sigma, \cdot)$ or $\mathbb{G}_n(\theta, \cdot)$ alternatively in this subsection. The objective is the estimation of a sparse correlation matrix Σ from the sample covariance matrix (of the pseudo-observations) Σ_n. The latter is defined as $\Sigma_n = n^{-1} \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i'$ with $\boldsymbol{x}_i = (\boldsymbol{\Phi}^{-1}(u_{i,1}), \ldots, \boldsymbol{\Phi}^{-1}(u_{i,q}))'$. Both Σ and Σ_n can be linked through some Bregman divergences for matrices, defined for two nonnegative conformable matrices Σ_1 and Σ_2 as

$$D(\Sigma_1, \Sigma_2) := \phi(\Sigma_1) - \phi(\Sigma_2) - \mathrm{tr}\big(\nabla\phi(\Sigma_2)'(\Sigma_1 - \Sigma_2)\big),$$

where $\phi(\cdot)$ is a differentiable and strictly convex function over the space of real and symmetric nonnegative matrices. As in Sect. 2.5 and for some fixed known $\omega \in [0,1]$, our loss function will be

$$\mathbb{G}_{n,\omega}(\Sigma) := \omega D(\Sigma, \Sigma_n) + (1-\omega) D(\Sigma_n, \Sigma).$$

In this section, we propose three options for $\phi(\cdot)$ and hence $D(\cdot,\cdot)$:

(i) $\phi(\Sigma) = -\log(|\Sigma|)$, called the Burg divergence, yielding the loss function

$$\mathbb{G}_{n,\omega}(\Sigma) = (1-\omega)\mathrm{tr}(\Sigma_n \Sigma^{-1}) + \omega \mathrm{tr}(\Sigma \Sigma_n^{-1}) + (2\omega - 1)\log(|\Sigma_n \Sigma^{-1}|) - q.$$

When $\omega = 0$, this is equivalent to the Canonical ML criterion.

(ii) $\phi(\Sigma) = \mathrm{tr}(\Sigma^2)$, whose symmetrized version is called Jeffreys divergence. Since the matrix derivative of $\phi(\Sigma)$ is 2Σ and after some simple calculations, the associated loss is the least squares-based criterion

$$\mathbb{G}_n(\Sigma) = \|\Sigma_n - \Sigma\|_2^2 = \mathrm{tr}\big((\Sigma_n - \Sigma)^2\big),$$

that does not depend on ω. We have introduced the Frobenius matrix norm: $\|A\|_2^2 = Tr(AA') = \sum_{i,j=1}^q a_{i,j}^2$ for any squared matrix $A := [a_{i,j}]_{1 \leqslant i,j \leqslant q}$.

(iii) $\phi(\Sigma) = \mathrm{tr}(\Sigma \log(\Sigma) - \Sigma)$, known as the von Neumann divergence. Since the matrix derivative of $\phi(\Sigma)$ (see Magnus and Neudecker 2019, Chapter 9) is $\log(\Sigma)$, the associated loss function is

$$\begin{aligned}\mathbb{G}_{n,\omega}(\Sigma) = \omega\,\mathrm{tr}(\Sigma \log \Sigma) - (\omega - 1)\mathrm{tr}(\Sigma_n \log \Sigma_n) \\ + (2\omega - 1)\mathrm{tr}(\Sigma_n - \Sigma) - \omega\mathrm{tr}(\log \Sigma_n \Sigma) - (1 - \omega)\mathrm{tr}(\Sigma_n \log \Sigma).\end{aligned}$$

In particular, when $\omega = 1$, this loss function nicely becomes

$$\mathbb{G}_{n,1}(\Sigma) = \mathrm{tr}(\Sigma \log \Sigma - \Sigma + \Sigma_n - \Sigma \log \Sigma_n).$$

We are now in a position to check the conditions of applicability of Theorem 1 for several estimators of Σ. For the three loss functions we will consider, the function $g(\cdot)$ may simply be chosen as the L^1-norm of the underlying parameter, meaning $g(\Sigma) = \|vech(\Sigma)\|_1$ for every correlation matrix Σ.

In case (i), $\phi(\Sigma) = -\log(|\Sigma|)$ provides a Gaussian-type estimator

$$\hat{\Sigma}^g := \underset{\Sigma \in \Theta_1}{\arg\min} \{\mathbb{G}_{n,\omega}(\Sigma, \boldsymbol{u}_1, \cdots, \boldsymbol{u}_n) + \boldsymbol{p}(\lambda_n, \Sigma)\}. \tag{10}$$

Above, Θ_1 denotes the convex subset of $q \times q$-correlation matrices such as

$$\Theta_1 = \{\Sigma : \Sigma = \Sigma', Diag(\Sigma) = Id, \quad \lambda_{\min}(2(1-\omega)\Sigma_n - (1-2\omega)\Sigma) \geqslant a_\omega, g(\Sigma) \leqslant R\},$$

for some positive constants a_ω and R. Note that the function $\Sigma \mapsto \mathbb{G}_{n,\omega}(\Sigma, \boldsymbol{u}_1, \ldots \boldsymbol{u}_n)$ is convex on Θ_1 for any values of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ (apply Boyd and Vandenberghe 2004, exercise 7.4).

Proposition 1 *Suppose that Σ_n is invertible and λ_n satisfies*

$$4\|(1-2\omega)\Sigma_0^{-1} + \omega\Sigma_n^{-1} - (1-\omega)\Sigma_0^{-1}\Sigma_n\Sigma_0^{-1}\|_\infty \leqslant \lambda_n \leqslant \frac{a_\omega}{12q^3R}. \tag{11}$$

Suppose that Σ_0 belongs to the convex parameter set Θ_1 and that $2a_\omega/q^3 - 3\mu > 0$. Then, for every n, any stationary point $\hat{\Sigma}$ of (10) satisfies

$$\|vech(\hat{\Sigma}^g) - vech(\Sigma_0)\|_2 \leqslant \frac{6\lambda_n\sqrt{k_0}}{2a_\omega/q^3 - 3\mu}, \ \|vech(\hat{\Sigma}^g) - vech(\Sigma_0)\|_1 \leqslant \frac{6(8a_\omega/q^3 - 9\mu)\lambda_n k_0}{(2a_\omega/q^3 - 3\mu)^2}.$$

The latter upper bounds actually depend on the dimension q, i.e. on the number of free parameters. The latter could depend on the sample size n too.

Proof To establish the (RSC) condition, use the differential operator w.r.t. Σ. Then, usual calculations provide

$$\nabla_{\Sigma}\mathbb{G}_{n,\omega}(\Sigma,\vec{\boldsymbol{u}}) = (1-2\omega)\Sigma^{-1} + \omega\Sigma_n^{-1} - (1-\omega)\Sigma^{-1}\Sigma_n\Sigma^{-1}.$$

To check the (RSC) condition, we focus on the Hessian matrix of $\mathbb{G}_n$. Using the formulas of Section 10.6.1. in Lütkepohl (1996), the Hessian is

$$\begin{aligned}&\nabla^2_{vec(\Sigma),vec(\Sigma)'}\mathbb{G}_{n,\omega}(\Sigma,\vec{\boldsymbol{u}})\\ &\quad= (1-\omega)\left\{\Sigma^{-1}\Sigma_n\Sigma^{-1}\otimes\Sigma^{-1} + \Sigma^{-1}\otimes\Sigma^{-1}\Sigma_n\Sigma^{-1}\right\} + (2\omega-1)\Sigma^{-1}\otimes\Sigma^{-1}.\end{aligned}$$

For some $\Sigma_1 \in \Theta$ and some $t \in [0,1]$, let $\Sigma := \Sigma_0 + t\Delta$, $\Delta := \Sigma_1 - \Sigma_0$. Then,

$$\begin{aligned}e_n(\Sigma) &:= vec(\Delta)'\nabla^2_{vec(\Sigma),vec(\Sigma)'}\mathbb{G}_n(\Sigma,\boldsymbol{u})vec(\Delta)\\ &\geqslant vec(\Delta)'\Big(\Sigma^{-1}\big((1-\omega)\Sigma_n - (1/2-\omega)\Sigma\big)\Sigma^{-1}\otimes\Sigma^{-1}\\ &\quad+\Sigma^{-1}\otimes\Sigma^{-1}\big((1-\omega)\Sigma_n - (1/2-\omega)\Sigma\big)\Sigma^{-1}\Big)vec(\Delta)\\ &\geqslant 2\|\Delta\|_2^2\lambda_{\min}((1-\omega)\Sigma_n - (1/2-\omega)\Sigma)\lambda_{\min}(\Sigma^{-1})^3,\end{aligned}$$

because the spectrum of $A\otimes B$ is the cross-product of the spectrums of A and B (Lütkepohl 1996, Section 5.2.1), and $\lambda_{\min}(\Sigma) = \inf_{\boldsymbol{x}}\boldsymbol{x}'\Sigma\boldsymbol{x}/\|\boldsymbol{x}\|_2^2$. Therefore, since $\lambda_{\max}(\Sigma)\leqslant Tr(\Sigma) = q$, we get

$$\begin{aligned}e_n(\Sigma) &\geqslant ||\Delta||_2^2\lambda_{\min}(2(1-\omega)\Sigma_n - (1-2\omega)\Sigma)\lambda_{\max}(\Sigma)^{-3}\\ &\geqslant ||\Delta||_2^2\lambda_{\min}(2(1-\omega)\Sigma_n - (1-2\omega)\Sigma)/q^3 \geqslant ||\Delta||_F^2 a_\omega/q^3.\end{aligned} \tag{12}$$

Here, the constraint $\lambda_{\min}(2(1-\omega)\Sigma_n - (1-2\omega)\Sigma)\geqslant a_\omega$ is key to ensure that the minimum eigenvalue is strictly positive. Now recall that the true vector of parameters is not Σ nor $vec(\Sigma)$ but rather the so-called vector θ, that stacks all coefficients of Σ that are located strictly below the main diagonal of Σ. With obvious notations, note that $\|\Delta\|_2^2 = \|\Sigma-\Sigma_0\|_2^2 = 2\|\theta-\theta_0\|_2^2$ for any correlation matrix Σ. Moreover, note that $e_n(\Sigma) = 4(\theta-\theta_0)'\nabla^2_{\theta,\theta'}\mathbb{G}_n(\Sigma,\vec{\boldsymbol{u}})(\theta-\theta_0)$. We deduce $(\theta-\theta_0)'\nabla^2_{\theta,\theta'}\mathbb{G}_n(\theta^*,\vec{\boldsymbol{u}})(\theta-\theta_0)\geqslant||\theta-\theta_0||_2^2 a_\omega/(2q^3)$, for any θ^* that lies between θ and θ_0. Thus, at Σ_0, the (RSC) condition is satisfied with $\alpha_1 = a_\omega/(2q^3)$ and $\alpha_2=\alpha_1$, $\nu_1=\nu_2=0$, and the result follows from Theorem 1. □

The constraint $\lambda_{\min}(2(1-\omega)\Sigma_n - (1-2\omega)\Sigma) > 0$ is easily satisfied when $\omega > 1/2$. In particular, when $\omega = 1$, check that Proposition 1 applies, replacing a_ω/q^3 by $1/q^2$. Otherwise, the definition of Θ_1 may appear somewhat restrictive. Indeed, when Σ_n has zero eigenvalues (the case of high-dimensional models $q >> 1$, most often) and $\omega\leqslant 1/2$, the parameter set Θ_1 is empty.

The choice of λ_n depends on the distance (in sup-norm) between the true correlation (resp. precision) matrix Σ_0 (resp. Σ_0^{-1}) and its empirical counterparts Σ_n (resp. Σ_n^{-1}), once obtained with pseudo-observations. Indeed, deduce from (11) that a convenient choice would be

$$\lambda_n \geqslant 4 \max \left\{ \| \Sigma_0^{-1} \left(\Sigma_0 - \Sigma_n \right) \Sigma_0^{-1} \|_\infty, \| \Sigma_0^{-1} - \Sigma_n^{-1} \|_\infty \right\}.$$

From Liu et al. (2009), Corollary 5 & 6, and using their truncated empirical marginal cdfs', $\| \Sigma_0 - \Sigma_n \|_\infty = O_P\big(\ln n \sqrt{\ln q}/n^{1/4} \big)$, and

$$\| \Sigma_0^{-1} - \Sigma_n^{-1} \|_\infty \leqslant \| \Sigma_0^{-1} - \Sigma_n^{-1} \|_{Frob} = O_P\Big(\ln n \sqrt{(k_0 + q) \ln q}/n^{1/4} \Big).$$

We deduce $\lambda_n \asymp q^2 \ln n \sqrt{\ln q}/n^{1/4}$ is acceptable to apply Proposition 1. Nonetheless, for small q and large n, this choice is not probably the best one (c.f. our discussion in Sect. 2.5). The same conclusion applies for the two next loss estimators of Σ (Propositions 14 and 16).

Alternatively and to weaken the latter problem, we can check the (RSC) condition for another estimator

$$\tilde{\Sigma}^g := \underset{\Sigma \in \tilde{\Theta}_1}{\arg \min} \{ \mathbb{G}_n(\Sigma, \boldsymbol{u}_1, \cdots, \boldsymbol{u}_n) + \boldsymbol{p}(\lambda_n, \Sigma) \}, \tag{13}$$

where $\tilde{\Theta}_1 := \{ \Sigma : \Sigma = \Sigma', Diag(\Sigma) = Id, \ \lambda_{\min}(2(1-\omega)\Sigma_0 - (1-2\omega)\Sigma) > \tilde{a}_\omega, \ g(\Sigma) \leqslant R \}$, for some positive constants $\tilde{a}_\omega$. Indeed, invoke inequalities (8) and (9). Clearly, $\tilde{\Theta}_1$ is generally larger than Θ_1 when Σ_0 is positive definite and when the considered matrices Σ are "not too far" from Σ_0. The price to be paid is coming from nonzero coefficients ν_1 and ν_2. Moreover, Σ_0 is unknown in $\tilde{\Theta}_1$, contrary to Σ_n in Θ_1. This is clearly a drawback.

Now, let us consider the case (ii), i.e. the Bregman divergence $\phi(\Sigma) = \mathrm{tr}(\Sigma^2)$. This provides a least squares-based estimator

$$\hat{\Sigma}^{ls} := \underset{\Sigma \in \Theta_2}{\arg \min} \{ \mathbb{G}_{n,\omega}(\Sigma, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) + \boldsymbol{p}(\lambda_n, \Sigma) \}, \tag{14}$$

with $\mathbb{G}_{n,\omega}(\Sigma, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) := \| \Sigma_n - \Sigma \|_2^2$ and the convex parameter space is

$$\Theta_2 := \{ \Sigma : \Sigma = \Sigma', Diag(\Sigma) = Id, \ \ g(\Sigma) \leqslant R \}.$$

Proposition 2 *Suppose that λ_n satisfies*

$$8 \| \Sigma_n - \Sigma_0 \|_\infty \leqslant \lambda_n \leqslant 1/(6R) \cdot \tag{15}$$

Suppose that Σ_0 belongs to the convex parameter set Θ_2 and that $\mu < 4/3$. Then, for every n, any stationary point $\hat{\Sigma}^{ls}$ of (14) satisfies

$$\| vech(\hat{\Sigma}^{ls}) - vech(\Sigma_0) \|_2 \leqslant \frac{6\lambda_n \sqrt{k_0}}{4 - 3\mu}, \ \| vech(\hat{\Sigma}^{ls}) - vech(\Sigma_0) \|_1 \leqslant \frac{6(16 - 9\mu)\lambda_n k_0}{(4 - 3\mu)^2} \cdot$$

Proof Using the differential operator with respect to Σ, we easily obtain $\nabla_{vec(\Sigma)} \mathbb{G}_n(\Sigma, \vec{\boldsymbol{u}}) = 2\mathrm{vec}(\Sigma - \Sigma_n)$. This directly implies

$$\langle \nabla_{vec(\Sigma)} \mathbb{G}_n(\Sigma, \vec{\boldsymbol{u}}) - \nabla_{vec(\Sigma)} \mathbb{G}_n(\Sigma_0, \vec{\boldsymbol{u}}), \Sigma - \Sigma_0 \rangle \geqslant 2 \| \Sigma - \Sigma_0 \|_2^2.$$

Since $e_n(\Sigma) = 4(\theta-\theta_0)'\nabla^2_{\theta,\theta'}\mathbb{G}_n(\Sigma,\vec{\boldsymbol{u}})(\theta-\theta_0)$ and $\|\Sigma-\Sigma_0\|_2^2 = 2\|\theta-\theta_0\|_2^2$, we get $(\theta-\theta_0)'\nabla^2_{\theta\theta'}\mathbb{G}_n(\theta^*,\vec{\boldsymbol{u}})(\theta-\theta_0) \geqslant \|\theta-\theta_0\|_2^2$, for any θ^* between θ_0 and θ. The (RSC) condition is satisfied for the parameters $\alpha_1=\alpha_2=1$ and $\tau_1=\tau_2=0$. Hence, Theorem 1 yields the desired upper bounds. □

Finally, we turn to case (iii), when $\phi(\Sigma) = \mathrm{tr}(\Sigma\log\Sigma-\Sigma)$. Here, we define the logarithm of any squared matrix S by $\log S = -\sum_{j\geqslant 1}(Id-S)^j/j$. Such series are conveniently defined when the largest eigenvalue of S in absolute value is less than one. This can be satisfied replacing our "usual" correlation matrices Σ by Σ/q. We prefer to directly control such a constraint in the parameter space: there will exist a constant $b>0$ s.t. all the eigenvalues of our matrices Σ and Σ_n will be less than b. Moreover, we restrict ourselves to the case $\omega=1$ to simplify. The corresponding estimator of Σ is then

$$\begin{aligned}\hat{\Sigma}^n &:= \underset{\Sigma\in\Theta_3}{\arg\min}\ \{\mathbb{G}_{n,1}(\Sigma,\boldsymbol{u}_1,\cdots,\boldsymbol{u}_n)+\boldsymbol{p}(\lambda_n,\Sigma)\},\\ \mathbb{G}_{n,1}(\Sigma,\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n) &= \mathrm{tr}(\log(\Sigma/b)\Sigma-\Sigma+\Sigma_n-\log(\Sigma_n/b)\Sigma)/b,\end{aligned} \tag{16}$$

where $\Theta_3 := \{\Sigma : \Sigma=\Sigma', Diag(\Sigma)=Id, Sp(\Sigma)\in]0,b[,\ \ g(\Sigma)\leqslant R\}$, a subset of matrices in Θ_2 with positive eigenvalues. Note that it is always possible to set $b=q$ and the parameter Θ_3 is then reduced to Θ_2.

Proposition 3 *Suppose that Σ_0 and Σ_n belong to Θ_3, $2/(bq)>3\mu$ and that λ_n satisfies*

$$4\|\log(\Sigma_0/b)-\log(\Sigma_n/b)\|_\infty \leqslant \lambda_n \leqslant 1/(12bqR).$$

Then, for every n, any stationary point $\hat{\Sigma}^n$ of (16) satisfies

$$\|vech(\hat{\Sigma}^n)-\mathrm{vech}(\Sigma_0)\|_2 \leqslant \frac{6\lambda_n\sqrt{k_0}}{2/(bq)-3\mu},\ \ \|vech(\hat{\Sigma}^n)-\mathrm{vech}(\Sigma_0)\|_1 \leqslant \frac{6(8/(bq)-9\mu)\lambda_n k_0}{2/(bq)-3\mu)^2}.$$

Proof Let us compute the gradient and Hessian of $\mathbb{G}_{n,1}(.,\vec{\boldsymbol{u}})$. Since $\nabla_\Sigma\phi(\Sigma)=\log\Sigma$, we get $\nabla_{vec(\Sigma)}\mathbb{G}_{n,1}(\Sigma,\vec{\boldsymbol{u}}) = vec\big(\log(\Sigma/b)-\log(\Sigma_n/b)\big)/b$. Since $\log(S) = -\sum_{j=1}^{+\infty}(I-S)^j/j$, the Hessian matrix is given by

$$\nabla^2_{vec(\Sigma)vec(\Sigma)'}\mathbb{G}_{n,\omega}(\Sigma,\vec{\boldsymbol{u}}) = \sum_{j=1}^{+\infty}\frac{1}{b^2 j}\sum_{k=0}^{j-1}(I-\Sigma/b)^{j-1-k}\otimes(I-\Sigma/b)^k,$$

applying Lütkepohl (1996), Section 10.5.1, Eq. (14). For some $\Sigma_1\in\Theta_3$ and $u\in[0,1]$, define $\Sigma=\Sigma_0+u\Delta$ with $\Delta=\Sigma_1-\Sigma_0$. Then,

$$\begin{aligned}e_n(\Sigma) &:= \mathrm{vec}(\Delta)'\nabla^2_{vec(\Sigma)vec(\Sigma)'}\mathbb{G}_{n,\omega}(\Sigma,\vec{\boldsymbol{u}})\mathrm{vec}(\Delta)\\ &= \sum_{j=1}^{+\infty}\frac{1}{b^2 j}\sum_{k=0}^{j-1}\mathrm{tr}\big((I-\Sigma/b)^{j-1-k}\Delta(I-\Sigma/b)^k\Delta\big),\end{aligned}$$

from Lütkepohl (1996), Section 2.4, Eq. (15). Since the two nonnegative matrices Σ_1 and Σ_0 can be diagonalized in the same basis, this is the case for Σ too. In the latter

basis, Σ_0 and Σ_1 are denoted as $\mathrm{Diag}(\lambda_{0,1}, \dots, \lambda_{0,q})$ and $\mathrm{Diag}(\lambda_{1,1}, \dots, \lambda_{1,q})$, respectively, with positive spectrums. As a consequence,

$$\mathrm{tr}\big((I - \Sigma/b)^{j-1-k}\Delta(I - \Sigma/b)^k\Delta\big) = \mathrm{tr}\big((I - \Sigma/b)^{j-1}\Delta^2\big), \text{ and}$$

$$e_n(\Sigma) \geqslant \frac{1}{b^2}\sum_{j=1}^{+\infty}\sum_{i=1}^{q}(1 - \lambda_i/b)^{j-1}(\lambda_{1,i} - \lambda_{0,i})^2 = \sum_{i=1}^{q}(b\lambda_i)^{-1}(\lambda_{1,i} - \lambda_{0,i})^2$$

$$\geqslant \sum_{i=1}^{q}(\lambda_{1,i} - \lambda_{0,i})^2 / \max\left(b\lambda_{\max}(\Sigma_0), b\lambda_{\max}(\Sigma_1)\right) \geqslant \|\Delta\|_2^2/(bq).$$

This yields $(\theta - \theta_0)'\nabla^2_{\theta\theta'}\mathbb{G}_{n,1}(\theta^*, \vec{\boldsymbol{u}})(\theta - \theta_0) \geqslant \|\theta - \theta_0\|_2^2/(2bq)$, for any θ^* between θ_0 and θ. The (RSC) condition would thus be satisfied for the parameters $\alpha_1 = \alpha_2 = 1/(2bq)$, $\tau_1 = \tau_2 = 0$. □

The latter results dedicated to the sparse estimation of Σ deserve a few comments. The error bounds are significantly altered by the choice of the loss function. Should we consider the Burg divergence (case (i)), the lack of smoothness of $\mathbb{G}_n(\cdot, \vec{\boldsymbol{u}})$, reflected by small α_k's, especially once the dimension becomes larger—the denominator of the α_1 is of order q^3, which rapidly shrinks to zero with q -, enforces a small μ, meaning less non-convexity of the penalty function. As a consequence, the theoretical upper bounds are less precise since their denominators are sensitive to the trade-off (α_1, μ). For the von Neumann case (iii), the α_i coefficients are improved compared to the previous case by two aspects: no constraint on the spectrum of $2\Sigma_n - \Sigma$, and the denominator of the upper bound is of order bq at most instead of q^3. Thus, we expect more informative theoretical upper bounds. For the least squares loss function (case (ii)), the RSC coefficients are dimension/sample free. This significantly improves the precision of the error bounds. Indeed, the trade-off (α_1, μ) is not altered by the dimension/sample and can accommodate sufficiently large μ. Therefore, we promote the use of $\hat{\Sigma}^{ls}$ defined in (14). For the latter estimator, it is tempting to evaluate the likelihood of satisfying Condition (15). In the case of usual empirical covariance matrices Σ_n, that would be evaluated from the unobservable sample $\mathcal{U}$, this can be done using the minimax bounds obtained by Bickel and Levina (2008), Cai et al. (2010), Cai and Zhou (2012), among others. Nonetheless, in the case of empirical covariance/correlation matrices calculated with pseudo-observations, we are not aware of similar results.

Alternatively, it would be tempting to parameterize this Gaussian copula model with the precision matrix $S := \Sigma^{-1}$ (or its lower diagonal components) instead of the correlation matrix Σ. Indeed, the coefficients of the precision matrix are partial correlations, that are of interest by themselves. Therefore, this would make sense to penalize partial-correlations instead of correlations. In the Gaussian loss case, the regularized statistical criterion would become

$$\begin{cases} \hat{S} & = \underset{S\in\bar{\mathcal{O}}}{\arg\min}\ \{\mathbb{G}_n(S, \boldsymbol{u}_1, \dots, \boldsymbol{u}_n) + \boldsymbol{p}(\lambda_n, S)\}, \quad \text{with} \\ \mathbb{G}_n(S, \boldsymbol{u}_1, \dots, \boldsymbol{u}_n) & = n\ln(2\pi)/2 - \ln|S|/2 + \sum_{i=1}^n \boldsymbol{x}_i' S \boldsymbol{x}_i/(2n), \end{cases}$$

where $\bar{\Theta}$ is a convenient subset of $q \times q$ nonnegative matrices. Moreover, the derivatives of such criteria wrt S are simpler than in the case of derivations wrt Σ (Corollary 3 in Loh and Wainwright 2015). Unfortunately, we have to restrict ourselves to the inverse of *correlation* matrices, and then the corresponding parameter subset would not be convex. This explains why we have parameterized the Gaussian copula model with Σ instead of Σ^{-1}.

3.2 Elliptical copula models

Elliptical copulas are generalizations of Gaussian copulas. They are defined by the density generator ψ of a centered elliptical distribution $\boldsymbol{Y}$ in $\mathbb{R}^q$ and a correlation matrix Σ. We recall that the density of such a q-random vector $\boldsymbol{Y}$ is given by $f_{\boldsymbol{Y}}(\boldsymbol{y}) = |\Sigma|^{-1/2}\psi(\boldsymbol{y}'\Sigma^{-1}\boldsymbol{y})$, for some function ψ that must satisfy $\int_0^\infty r^{q-1}\psi(r^2)\,dr < \infty$. In particular, we recover Gaussian distributions by setting $\psi(t) = \exp(-t/2)$. See Section 4 in Cambanis et al. (1981) for a reminder about elliptical distributions. We deduce that the elliptical copula density w.r.t. the Lebesgue measure in $\mathbb{R}^q$ is

$$c_g(\boldsymbol{u}) = \frac{\psi\left(\vec{F}_\psi^{-1}(\boldsymbol{u})'\,\Sigma^{-1}\vec{F}_\psi^{-1}(\boldsymbol{u})\right)}{|\Sigma|^{1/2}\prod_{k=1}^q f_\psi(F_\psi^{-1}(u_k))}, \quad \vec{F}_\psi^{-1}(\boldsymbol{u}) := [F_\psi^{-1}(u_1), \ldots, F_\psi^{-1}(u_q)]',$$

where F_ψ (resp. f_ψ) denotes the cdf (resp. density) of any margin of a q-dimensional centered and reduced elliptical random vector whose density, generator is ψ, i.e.

$$F_\psi(x) = \int_{-\infty}^x \psi_1(t^2)\,\mathrm{d}t, \quad \psi_1(u) = \frac{\pi^{(q-1)/2}}{\Gamma((q-1)/2)}\int_0^\infty \psi(u+s)s^{(q-3)/2}\,\mathrm{d}s.$$

See Cambanis et al. (1981) or Gómez et al. (2003).

We assume this generator ψ is known and that the single unknown parameter of the elliptical copula is the correlation matrix Σ. As for the case of Gaussian copulas and for the same reason, we parameterize the model by Σ instead of Σ^{-1}. Note that ψ is most often convex. Indeed, for most density generators, there exists a distribution F_∞ on the positive real line s.t.

$$\psi(t) = \int_0^\infty (2\pi r^2)^{-q/2}\exp(-t/2r^2)\,F_\infty(\mathrm{d}r), \tag{17}$$

for any positive t. This is the case for elliptical distributions that have been obtained with "universal" (independent of the dimension q) characteristic generators: see Equation (24) in Cambanis et al. (1981). Nonetheless, (17) does not imply that $\Sigma \mapsto \mathbb{G}_n(\Sigma, \boldsymbol{y})$ is a convex function in general.

Therefore, with the same notations as in Sect. 3.1, we define the statistical criterion as

$$\left\{\begin{array}{rcl} \hat{\Sigma} & = & \underset{\Sigma\in\Theta}{\arg\min}\,\{\mathbb{G}_n(\Sigma,\boldsymbol{y})+\boldsymbol{p}(\lambda_n,\Sigma)\},\ \text{where} \\ \mathbb{G}_n(\Sigma,\vec{\boldsymbol{u}}) & = & \ln|\Sigma|/2-\sum_{i=1}^n \ln\psi(\boldsymbol{y}_i'\Sigma^{-1}\boldsymbol{y}_i)/n, \\ \boldsymbol{y}_i & := & \left(F_\psi^{-1}(u_{i,1}),\ldots,F_\psi^{-1}(u_{i,q})\right),\ i=1,\ldots,n. \end{array}\right. \tag{18}$$

This is the "usual" penalized canonical maximum likelihood criterion. Denote by $\|A\|_s$ the usual spectral norm of any matrix. Θ will be the convex set of $q\times q$-correlation matrices such as

$$\Theta=\{\Sigma:\Sigma=\Sigma',\mathrm{Diag}(\Sigma)=Id,\ \lambda_{\min}(\Sigma)\geqslant a,\ \lambda_{\min}(2\mathcal{S}_n(\Sigma_0)-\Sigma)>b,\ g(\Sigma)\leqslant R\}, \tag{19}$$

for some positive constants a and b. For an arbitrary correlation matrix, we have denoted

$$\mathcal{S}_n(\Sigma):=\frac{(-2)}{n}\sum_{i=1}^n\left(\frac{\psi'}{\psi}\right)(\boldsymbol{y}_i'\Sigma^{-1}\boldsymbol{y}_i)\boldsymbol{y}_i\boldsymbol{y}_i'.$$

Note that $\mathcal{S}_n(\Sigma)$ is nonnegative because ψ is decreasing under (17). Moreover, Σ_0, the true correlation matrix, is assumed to belong to Θ and satisfies $\mathrm{E}[\nabla_{vec(\Sigma)}\mathbb{G}_n(\Sigma_0,\mathcal{U})]=0$. The true subset model $\mathcal{A}$ admits the same cardinality k_0 as in the Gaussian copula case.

Under (17), note that $(\psi')^2\leqslant\psi''\psi$ by the Cauchy–Schwarz inequality. Then, we can set, for every $i=1,\ldots,n$,

$$\sup_{\Sigma\in\Theta|}\left(\frac{\psi'}{\psi}\right)'(\boldsymbol{y}_i\Sigma^{-1}\boldsymbol{y}_i)=:\theta_i^2\ \text{and}\ V_n:=\frac{2}{n}\sum_{i=1}^n\theta_i^2\|\boldsymbol{y}_i\|_2^4.$$

Proposition 4 *Let* $\alpha=(b/q^3-V_n)/4$. *Assume* (17), $4\alpha>3\mu$, $R\geqslant 1/6$ *and that* (λ_n,R) *satisfies*

$$2\max\left\{\|vec(\Sigma_0^{-1}\mathcal{S}_n(\Sigma_0)\Sigma_0^{-1}-\Sigma_0^{-1})\|_\infty,\frac{8R}{a^3}\|\mathcal{S}_n(\Sigma)-\mathcal{S}_n(\Sigma_0)\|_s\right\}\leqslant\lambda_n\leqslant\frac{\alpha}{6R}.$$

Then, any stationary point $\hat{\Sigma}$ *of* (18) *satisfies*

$$\|vech(\hat{\Sigma}-\Sigma_0)\|_2\leqslant\frac{6\lambda_n\sqrt{k_0}}{4\alpha-3\mu},\ \|vech(\hat{\Sigma}-\Sigma_0)\|_1\leqslant\frac{6(16\alpha-9\mu)\lambda_n k_0}{(4\alpha-3\mu)^2}.$$

The proof has been postponed in Appendix A.2. Elliptical copulas provide an interesting case where the constants ν_k, $k=1,2$ of the (RSC) condition are nonzero. When $\mathcal{S}_n(\Sigma)$ does not depend on Σ, as for the Gaussian copula, we recover Proposition 1.

Remark 5 The set Θ depends on the unknown matrix Σ_0. Then, it may appear as only theoretical. Actually, in the definition of Θ, the true matrix Σ_0 can be replaced by any preliminary crude estimator $\bar{\Sigma}$ that is not "too far" from Σ_0 ($\|\Sigma_0-\bar{\Sigma}\|_s<1$, to be specific).

Alternatively, there is another way of estimating Σ without calculating the marginal distribution F_ψ, its derivative and the elliptical copula. Indeed, this is often a boring task in analytical terms, and the evaluation of F_ψ usually requires numerical analysis routines. As it is well known (see Wegkamp and Zhao 2016, for example), there is a one-to-one mapping between the components of $\Sigma = [\sigma_{kl}]_{1\leqslant k,l\leqslant q}$ and all the bivariate Kendall's tau $\tau_{k,l}$ associated with the underlying random vector $\boldsymbol{X}$: for every couple of indices (k, l), $k \neq l$, $\sigma_{k,l} = \sin(\pi\tau_{k,l}/2)$. Therefore, invoking empirical Kendall's taus', a statistical criterion may be based on a moment-based penalized method to estimate Σ. It is given by

$$\left\{\begin{array}{rcl} \forall (k,l), \hat{\sigma}_{k,l} & = & \underset{\sigma_{k,l}:\, g(\Sigma)\leqslant R}{\arg\min}\ \{\mathbb{G}_n(\sigma_{k,l},\vec{\boldsymbol{u}}) + \boldsymbol{p}(\lambda_n,\sigma_{k,l})\}, \quad \text{where} \\ \mathbb{G}_n(\sigma_{k,l},\vec{\boldsymbol{u}}) & = & \big(\sigma_{k,l} - \sin(\pi\hat{\tau}_{k,l}/2)\big)^{\alpha}, \ \alpha \geqslant 1, \quad \text{with} \\ \hat{\tau}_{k,l} & := & 2\sum_{i<j}\big(\mathbf{1}(X_{i,k}\leqslant X_{i,l}, X_{j,k}\leqslant X_{j,l}) - \mathbf{1}(X_{i,k} > X_{i,l}, X_{j,k}\leqslant X_{j,l})\big)/(n^2-n). \end{array}\right. \tag{20}$$

Note that this way of working enables one to split the global criteria $\mathbb{G}_n(\Sigma,\vec{\boldsymbol{u}}) + \boldsymbol{p}(\lambda_T,\Sigma)$ as a sum of univariate functions. Therefore, we would replace a global optimization in $\mathbb{R}^{q(q-1)/2}$ by $q(q-1)/2$ univariate optimization programs, what is clearly a nice feature. Obviously, the (RSC) condition would apply in this case. Unfortunately, the obtained matrix $\hat{\Sigma} := [\hat{\sigma}_{k,l}]$ has no reasons to be nonnegative definite. Even if it is always possible to project $\hat{\Sigma}$ on the subset of correlation matrices, the associated theoretical properties of this final output are far from clear and we prefer not to develop more this idea here.

3.3 Mixtures of copula models

An easy way of building highly parameterized copula models is through mixtures. Indeed, consider a family of fixed q-dimensional copulas $\{C_k, k = 1,\dots,m\}$. We can assume the true copula C is a linear combination of all the latter ones, i.e. $C(\boldsymbol{u}) = \sum_{k=1}^m \omega_k C_k(\boldsymbol{u})$, for every $\boldsymbol{u}\in[0,1]^q$. Obviously, the parameter is $\theta := (\omega_1,\dots,\omega_m)'$, with $\omega_k \in [0,1]$ for every $k = 1,\dots,m$ and $\sum_{k=1}^m \omega_k = 1$. The associated loss function is (minus) the corresponding log-likelihood. Denoting by c_k the copula density associated with C_k, $k = 1,\dots,m$, the statistical criterion is thus given by

$$\left\{\begin{array}{rcl} \hat{\theta} & = & \underset{\theta\in\Theta}{\arg\min}\ \{\mathbb{G}_n(\theta,\vec{\boldsymbol{u}}) + \boldsymbol{p}(\lambda_n,\theta)\}, \quad \text{where} \\ \mathbb{G}_n(\theta,\vec{\boldsymbol{u}}) & = & -\sum_{i=1}^n \ln\big(\sum_{k=1}^m \omega_k c_k(\boldsymbol{u}_i)\big)/n, \quad \text{with} \end{array}\right. \tag{21}$$

$\Theta = \{(\omega_1,\dots,\omega_m)\in\mathbb{R}_+^m,\ \sum_{k=1}^m\omega_k = 1,\ \|\theta-\theta_0\|_2 < \epsilon,\ g(\theta)\leqslant R\}$, for $\epsilon > 0$.

For convenience, introduce the column vector $\vec{c}(\boldsymbol{u}_i) := \big(c_1(\boldsymbol{u}_i),\dots,c_m(\boldsymbol{u}_i)\big)'$ for every i, and set $\mu_{i,0} := \big(\theta_0'\vec{c}(\boldsymbol{u}_i) + \epsilon\|\vec{c}(\boldsymbol{u}_i)\|_2\big)^{-1}$.

Proposition 5 *For any $\theta \neq \mathbf{0}$, let $\alpha = \lambda_{\min}\big(n^{-1}\sum_{i=1}^n \mu_{i,0}^2\vec{c}(\boldsymbol{u}_i)\vec{c}(\boldsymbol{u}_i)'\big)$, and assume $\alpha > 3\mu/4$. Suppose that (λ_n, R) satisfy*

$$4\|n^{-1}\sum_{i=1}^{n}\mu_i\vec{c}(\boldsymbol{u}_i)\|_\infty \leqslant \lambda_n \leqslant \alpha/(6R)\cdot$$

Then, any stationary point $\hat{\theta}$ *of* (21) *satisfies*

$$\|\hat{\theta}-\theta_0\|_2 \leqslant \frac{6\lambda_n\sqrt{k_0}}{4\alpha-3\mu}, \quad \|\hat{\theta}-\theta_0\|_1 \leqslant \frac{6(16\alpha-9\mu)\lambda_n k_0}{(4\alpha-3\mu)^2}\cdot$$

Proof Since $\mathbb{G}_n(\theta,\vec{\boldsymbol{u}}) = -\sum_{i=1}^{n}\ln\left(\theta'\vec{c}(\boldsymbol{u}_i)\right)/n$, simple calculations provide

$$\nabla_\theta\mathbb{G}_n(\theta,\vec{\boldsymbol{u}}) = -\sum_{i=1}^{n}\frac{\vec{c}(\boldsymbol{u}_i)}{n\theta'\vec{c}(\boldsymbol{u}_i)}, \text{ and } \nabla^2_{\theta,\theta'}\mathbb{G}_n(\theta,\vec{\boldsymbol{u}}) = \sum_{i=1}^{n}\frac{\vec{c}(\boldsymbol{u}_i)\vec{c}(\boldsymbol{u}_i)'}{n(\theta'\vec{c}(\boldsymbol{u}_i))^2}\cdot$$

Consider the parameter $\theta_1 \in \Theta$, and $\theta = t\theta_0 + (1-t)\theta_1$ for some $t \in [0,1]$. Since $\theta'\vec{c}(\boldsymbol{u}_i)$ is nonnegative for every $t \in [0,1]$, we have

$$\theta'\vec{c}(\boldsymbol{u}_i) \leqslant \theta_0'\vec{c}(\boldsymbol{u}_i) + \|\theta_0-\theta_1\|_2\|\vec{c}(\boldsymbol{u}_i)\|_2 \leqslant \mu_{i,0}^{-1}.$$

Therefore, this yields

$$\begin{aligned}(\theta_1-\theta_0)'\nabla^2_{\theta,\theta'}\mathbb{G}_n(\theta,\vec{\boldsymbol{u}})(\theta_1-\theta_0) &\geqslant \sum_{i=1}^{n}\mu_{i,0}^2(\theta_1-\theta_0)'\vec{c}(\boldsymbol{u}_i)\vec{c}(\boldsymbol{u}_i)'(\theta_1-\theta_0)/n\\ &\geqslant \|\theta_1-\theta_0\|_2^2\lambda_{\min}\Big(n^{-1}\sum_{i=1}^{n}\mu_{i,0}^2\vec{c}(\boldsymbol{u}_i)\vec{c}(\boldsymbol{u}_i)'\Big),\end{aligned}$$

so that $\alpha_1 = \alpha_2 = \lambda_{\min}\left(n^{-1}\sum_{i=1}^{n}\mu_{i,0}^2\vec{c}(\boldsymbol{u}_i)\vec{c}(\boldsymbol{u}_i)'\right)$, $\nu_1 = \nu_2 = 0$. □

Remark 6 As in the case of elliptical copulas, the set Θ and the constants μ_i depend on the unknown parameter θ_0. Nonetheless, it can be easily checked that the previous result holds, replacing θ_0 (in Θ and μ_i) by any feasible and consistent parameter $\bar{\theta}$.

It is possible to extend the latter analysis towards mixtures of parametric copulas with *unknown* parameters. In this case, $C(\boldsymbol{u}) = \sum_{k=1}^{m}\omega_k C_{k,\theta_k}(\boldsymbol{u})$ for every $\boldsymbol{u} \in [0,1]^q$. Now, for any $k = 1,\ldots,m$, C_{k,θ_k} belongs to a given parametric copula family $\mathcal{C}_k := \{C_{k,\theta_k}$ copula on $[0,1]^q; \theta_k \in \Theta_k \subset \mathbb{R}^{d_k}\}$, and the associated copula densities are denoted by c_{k,θ_k}. Now, the unknown parameter is $\theta := (\omega_1,\ldots,\omega_m,\theta_1,\ldots,\theta_m)$, with $\omega_k \in [0,1]$ for every $k = 1,\ldots,m$ and $\sum_{k=1}^{m}\omega_k = 1$. The statistical criterion is thus given by

$$\begin{cases} \hat{\theta} & = \arg\min_{\theta\in\Theta} \{\mathbb{G}_n(\theta, \boldsymbol{u}) + \boldsymbol{p}(\lambda_n, \theta)\}, \quad \text{where} \\ \mathbb{G}_n(\theta, \boldsymbol{u}) & = -\sum_{i=1}^{n} \ln\left(\sum_{k=1}^{m} \omega_k c_{k,\theta_k}(\boldsymbol{u}_i)\right)/n, \end{cases} \tag{22}$$
$$\Theta := \{\theta \in \mathbb{R}_+^m \times \bigtimes_{k=1}^{m} \Theta_k,\ \sum_{k=1}^{m} \omega_k = 1,\ \|\theta - \theta_0\|_2 \leqslant \epsilon,\ g(\theta) \leqslant R\},$$

for some positive constant ϵ. The dimension of θ is then $d := m + d_1 + \cdots, d_m$.

It is tempting to assume a (RSC)-type condition for every "component copula" model c_{k,θ_k}, $k = 1, \ldots, m$ and to deduce such a condition for the mixture model above. Unfortunately, the latter "componentwise" (RSC) conditions are not sufficient because they do not allow to control the terms that involve some products of c_{k,θ_k} and c_{l,θ_l} and their derivatives, when $k \neq l$. Therefore, we will assume a stronger condition: a (RSC)-type condition applies on every mixture model, *for any given set of weights*. For such models, the unknown vector of parameters becomes $\bar{\theta} := (\theta_1, \ldots, \theta_m)$. Its dimension is denoted by $\bar{d}$ and its true value implicitly depends on the chosen vector of weights. Then, we now assume that, for every $\omega = [\omega_1, \ldots, \omega_m]'$, there exist some constants $\alpha_{j,\omega} > 0$ and $\nu_{j,\omega} \geqslant 0$, $j = 1, 2$, s.t.

$$\bar{\boldsymbol{v}}' \nabla_{\bar{\theta},\bar{\theta}'} \mathbb{G}_n\big((\omega, \bar{\theta}), \vec{\boldsymbol{u}}\big) \bar{\boldsymbol{v}} \geqslant \alpha_{1,\omega} \|\bar{\boldsymbol{v}}\|_2^2 - \nu_{1,\omega} \|\bar{\boldsymbol{v}}\|_1^2,\ \|\bar{\boldsymbol{v}}\|_2 \leqslant 1, \tag{23}$$

$$\bar{\boldsymbol{v}}' \nabla_{\bar{\theta},\bar{\theta}'} \mathbb{G}_n\big((\omega, \bar{\theta}), \vec{\boldsymbol{u}}\big) \bar{\boldsymbol{v}} \geqslant \alpha_{2,\omega} \|\bar{\boldsymbol{v}}\|_2 - \nu_{2,\omega} \|\bar{\boldsymbol{v}}\|_1,\ \|\bar{\boldsymbol{v}}\|_2 > 1, \tag{24}$$

for every $\bar{\theta}$ s.t. $(\omega, \bar{\theta}) \in \Theta$. We will assume that, for $j = 1, 2$,

$$\underline{\alpha}_j := \inf_{\omega} \alpha_{j,\omega} > 0 \text{ and } \overline{\nu}_j := \sup_{\omega} \nu_{j,\omega} < \infty. \tag{25}$$

Denote $\vec{c}_\theta(\boldsymbol{u}) := \big(c_{1,\theta_1}(\boldsymbol{u}), \ldots, c_{m,\theta_m}(\boldsymbol{u})\big)'$. For every $i = 1, \ldots, n$ and every $\theta \in \Theta$, set $\mu_i(\theta) := (\omega' \vec{c}_\theta(\boldsymbol{u}_i))^{-1}$. We introduce the constants

$$\alpha_1 := \min\left(\inf_{\theta\in\Theta} \lambda_{\min}\Big(\frac{1}{n}\sum_{i=1}^{n} \mu_i^2(\theta) \vec{c}_\theta(\boldsymbol{u}_i) \vec{c}_\theta'(\boldsymbol{u}_i)\Big); \underline{\alpha}_1\right),\ \alpha_2 := \min\left(\alpha_1; \underline{\alpha}_2\right),$$
$$\tau_0 := \frac{2}{n} \sup_{\theta\in\Theta} \sum_{i=1}^{n} \left(\mu_i(\theta)^2 \Big(\sum_{k=1}^{m} \omega_k \|\partial_{\theta_k} c_{k,\theta_k}(\boldsymbol{u}_i)\|_\infty\Big) \sup_l c_{l,\theta_l}(\boldsymbol{u}_i) + \mu_i(\theta) \sup_k \|\partial_{\theta_k} c_{k,\theta_k}(\boldsymbol{u}_i)\|_\infty\right),$$
$$\nu_1 := \overline{\nu}_1 + \tau_0,\ \nu_2 := \max\left(\overline{\nu}_2, \overline{\nu}_1\right) + \tau_0.$$

Proposition 6 *Assume that* $4\alpha_1 > 3\mu$, *and that* (λ_n, R) *satisfies*

$$4 \max\left\{\|\nabla_\theta \mathbb{G}_n(\theta, \vec{\boldsymbol{u}})\|_\infty, 2R\nu_1\right\} \leqslant \lambda_n \leqslant \alpha_2/(6R),$$

for some positive constant α_2. *Then, if* $4R\nu_2 \leqslant \alpha_2$, *any stationary point* $\hat{\theta}$ *of* (22) *satisfies*

$$\|\hat{\theta} - \theta_0\|_2 \leqslant \frac{6\lambda_n\sqrt{k_0}}{4\alpha_1 - 3\mu}, \quad \|\hat{\theta} - \theta_0\|_1 \leqslant \frac{6(16\alpha_1 - 9\mu)\lambda_n k_0}{(4\alpha_1 - 3\mu)^2}.$$

The proof is given in Appendix 6.

3.4 Archimedean copulas

Archimedean copulas are specified by their generator $g : [0,1] \mapsto \mathbb{R}_+ \cup \{+\infty\}$. Most often, this generator is assumed to belong to a parametric family $\mathcal{F}_{gen} := \{g_\theta, \theta \in \Theta\}$. Many popular copula families are obtained by conveniently choosing such families $\mathcal{F}_{gen}$: Clayton, Gumbel, Frank, etc. Very often, θ is a single number and the value $\theta = 0$ is related to the independence copula. Since this parameter θ is easily and explicitly mapped to the underlying Kendall's taus', nice and simple GMM-type estimation procedures are often available, as in the end of Sect. 3.2. And such criteria can be penalized, obviously.

Despite their popularity, highly flexible and highly parameterized Archimedean copulas are not available, to the best of our knowledge. This significantly decreases the interest of our penalized techniques in such particular cases, that are well suited when the dimension d is relatively large. At the opposite, hierarchical Archimedean copulas (HAC) are nice and richly parameterized generalizations. They allow asymmetries and different dependencies for couples of variables, by combining a hierarchy of Archimedean copulas C_j, $j = 1, \dots, m$, with different parameters θ_j. Obviously, the whole model is known once we have known/estimated $\theta := (\theta_1, \dots, \theta_m)$. See McNeil (2008), Okhrin et al. (2013 a,b), Segers and Uyttendaele (2014), Górecki et al. (2016), among others. As a standard situation, all invoked copulas in a HAC are bivariate and belong to the same family, and the successive parameter values are ordered so that we get a true q-dimensional copula. Let us keep the latter framework, even if our techniques apply in the case of more general HAC constructions.

The densities of (nested) HAC can be computed analytically (Hofert and Pham 2013), but calculations and coding become rapidly very tedious when the underlying dimension is "large". Therefore, a full MLE of the underlying parameters is feasible only when q is "small". In any case, under our penalized point of view, there is no guarantee that the (RSC) condition is satisfied for most Archimedean families, neither for HAC models a fortiori.

Therefore, we promote an adaptation of the recursive maximum likelihood method (RMLE), as exposed in Okhrin et al. (2013b) for instance. If every underlying copula C_j that defines a given HAC structure satisfies the (RSC) condition, the penalized RMLE is rather simple: as explained in Okhrin et al. (2013b), successively estimate the parameter(s) associated with every copula with pseudo-observations that are built with the previously estimated parameters. The novelty would come here from the penalization.

Alternatively, if the (RSC) is not fulfilled for some of the underlying copulas C_j, we propose to adapt the methodology of Sect. 3.1. To simplify, assume that every copula C_j is bivariate, that its parameter θ_j is a real number and that there is an explicit one-to-one analytic relationship between the Kendall tau of C_j and θ_j: $\phi_j(\tau_j) = \theta_j$,

$j = 1, \dots, m$. The RMLE process is based on the fact that C_j is the copula between some random variables $Z_{j,1}$ and $Z_{j,2}$ that are functions of $\theta_1, \dots, \theta_{j-1}$ and some of the components of $\boldsymbol{U}$. Therefore, using empirical counterparts and the previously estimated values θ_k, $k < j$, we can build a "pseudo-sample" of $(Z_{j,1}, Z_{j,2})$. Then, we are able to calculate the associated empirical Kendall's tau, as in (20), denoted by $\hat{\tau}_j$, and to estimate θ_j as

$$\hat{\theta}_j := \arg\min_{\theta_j} (\phi_j(\hat{\tau}_j) - \theta_j)^\alpha + \boldsymbol{p}(\lambda_n, \theta_j), \quad \alpha \geqslant 1. \tag{26}$$

And the process goes on, allowing the estimation of all parameters θ_k successively. Indeed, for most usual penalties, the latter program satisfies the (RSC) condition, as for penalized least-squares criteria ($\alpha = 2$). Note that any dependence measure can be applied here, once there exists a one-to-one mapping between such measure and the underlying parameter, that is univariate here. Nonetheless, we will not try to detail technical conditions to apply Theorem 1 for such models. This general task seems to be unfeasible, and analytic calculations have to be done for every particular parametric model.

Remark 7 Note that Okhrin et al. (2015) have proposed a recursive penalized MLE procedure for HAC models. It can be seen as the natural alternative to (26), in particular when (RSC) is always satisfied.

4 Empirical study

In this section, we carry out a simulation study to illustrate the theoretical results of our method in the presence of pseudo-observations. To do so, we consider mixtures of copula models, as described in Sect. 3.3: the data generating process is induced by a linear combination of copulas with known parameters, where the combination depends on weights ω, which are supposed to be sparse, so that the number of nonzero components is arbitrarily set. We consider the problem dimension $m = 5$ with $\omega = (0, 0.2, 0.8, 0, 0)'$ and bivariate copulas ($q = 2$). We select five Archimedean copulas for solving the problem (21) as follows: following the notations in Nelsen (2006), $c_1(\boldsymbol{u}_i)$ is Gumbel with parameter 30; $c_2(\boldsymbol{u}_i)$ is Clayton with parameter 0.5; $c_3(\boldsymbol{u}_i)$ is Gumbel with parameter 8; $c_4(\boldsymbol{u}_i)$ is Clayton with parameter 2; $c_5(\boldsymbol{u}_i)$ is Frank with parameter 15. Denote by ψ_j, $j = 1, \dots, 5$, the generators of the five latter Archimedean copulas. To generate a realization of $\boldsymbol{U}$ along our given mixture of copulas, we apply the following simulation procedure:

(i) randomly draw the identity of the copula (an index $j \in \{1, \dots, 5\}$), the randomization being determined by the weights in ω;
(ii) draw $V_j \sim L^{-1}(\psi_j)$, where $L^{-1}(\psi_j)$ is the inverse of the Laplace transform of ψ_j;
(iii) simulate i.i.d. realizations $X_k \underset{\text{i.i.d.}}{\sim} \mathcal{U}([0, 1])$, $k = 1, \dots, q$;
(iv) set $\boldsymbol{U} := (U_1, \cdots, U_q)$ where $U_k = \psi_j(-\ln(X_k)/V_j)$, $k = 1, \dots, q$.

In that case, the marginals are uniform on [0, 1] and the realizations $\boldsymbol{U}_i$ are i.i.d. When the margins of $\boldsymbol{U}_i$ are unknown, we compute pseudo-observations $\hat{\boldsymbol{U}}_i$ through the empirical ranks $\hat{\boldsymbol{U}}_{ik} = R_{ik}/(n+1)$, $i = 1, \dots, n$, $k = 1, \dots, q$, where R_{ik} is the rank of $U_{i,k}$ among the kth univariate sample $(U_{i,k})_{i=1,\dots,n}$.

To recover the sparse support $\mathcal{A}$ with $\text{card}(\mathcal{A}) = 2$, we consider the regularized problem as detailed in Sect. 3.3. Since the copula parameters are known here, denote $\theta = \omega$ the vector of weights in our problem (21), and we set $g(\theta) = \|\theta\|_1$. Our optimization procedure relies on a numerical scheme under the linear constraint $\sum_{k=1}^5 \omega_k = 1$, carried out by the function *fmincon(.)* on MATLAB and based on the second-order information matrix. Note that the latter can encompass nonlinear constraints such as $\{\Sigma : \lambda_{\min}(2\Sigma_n - \Sigma) > a\}$ when one considers problem (10). Alternatively, in their section 4, Loh and Wainwright (2015) developed a composite gradient descent procedure, which consists in a three-step updating procedure of the optimized parameter value, to solve (1). This first-order-based algorithm can be adapted to the sparse estimation of matrix parameters under positive-definiteness constraints, such as the alternating direction method of multipliers devised by Bien and Tibshirani (2011): see their appendix 3.

As for the regularization parameters, since we work under the constraint $\|\theta\|_1 = 1$, the specification of R can be left aside. Should we drop the weight condition on θ, then following Loh and Wainwright (2015, 2017), we would select $R = \boldsymbol{p}(\lambda_n, \theta_0)/\lambda_n$. Furthermore, we set $\lambda_n = 4\alpha\sqrt{\log m}/\sqrt{n}$, where $m = 5$ is the problem dimension and α is the minimum eigenvalue of the Hessian $\nabla^2_{\theta\theta'}\mathbb{G}_n(\theta_0, \vec{\boldsymbol{u}}) = \sum_{i=1}^n \mu_i^2 \vec{c}(\boldsymbol{u}_i)\vec{c}(\boldsymbol{u}_i)'/n$. To obtain an estimated value of α provided in Proposition 5, we set $\epsilon = 0.2$ in μ_i and simulated a sample $(\boldsymbol{U}_i)_{i=1,\dots,n}$, $n = 20,000$, according to the mixing procedure previously described. We numerically obtained $\alpha = 0.0815$. Importantly, due to the constraints on the (RSC) coefficients, mainly $\alpha > 3\mu/4$, where $\mu = 1/(b_{\text{scad}} - 1)$ (resp. $\mu = 1/b_{mcp}$, resp. $\mu = 0$) in the SCAD case (resp. MCP, resp. LASSO), we considered the following setting for $\alpha = 0.0815$: $b_{\text{scad}} = 22$ (resp. $b_{mcp} = 18$) so that $4\alpha - 3\mu = 0.1831$ (resp. 0.1593) for the SCAD (resp. MCP) penalty. In the LASSO case, since $\mu = 0$, we have $4\alpha = 0.3260$.

In addition to the estimated error $\|\hat{\theta} - \theta_0\|_l$, $l = 1, 2$, we reported on figures 1 and 2 the theoretical upper bounds in case of known margins, for each norm and using the previous parameter setting. We also reported with the light grey line the values $\|\theta_0\|_2 = 0.8246$ and $\|\theta_0\|_1 = 1$. For each sample size, we replicated 200 times the simulation set-up and obtained 200 sparsity-based estimates $\hat{\theta}$. Figure 1a (resp. Fig. 1b), provided in the supplementary material, illustrates their $\|.\|_2$ (resp. $\|.\|_1$) consistency with respect to the sample size. As predicted in Proposition 5, the three curves for the MCP, SCAD and LASSO converge toward zero as the number of samples increases. Interestingly, each plot displays the sparsity-based estimation with $\boldsymbol{U}$-samples or only pseudo-observations $\hat{\boldsymbol{U}}$. Although the statistical error decreases with n in both cases, the estimation is less precise in the $\hat{\boldsymbol{U}}$-case due to the nonparametric transform to each margin and its amount of additional noise. Note that the theoretical $\|.\|_2$-based upper bounds for parameter consistency are "informative" (in the sense they are not unrealistic), at least when n is larger than several thousands. This illustrates the practical add-in of such results. This is less the case with $\|.\|_1$-based upper bounds that are too wide.

In the latter of SCAD and MCP penalty cases, they do not appear on the figure: These bounds are, respectively, close to 1.29 and 1.57 for $n = 10,000$.

Proofs

Proof of Theorem 1

Proof Let $\Delta = \hat{\theta} - \theta_0$. We first show that $\|\Delta\|_2 \leqslant 1$. If this is not satisfied, then we have

$$\langle \nabla_\theta \mathbb{G}_n(\hat{\theta};\hat{\mathcal{U}}) - \nabla_\theta \mathbb{G}_n(\theta_0;\hat{\mathcal{U}}), \Delta \rangle \geqslant \alpha_2 \|\Delta\|_2 - \nu_2 \|\Delta\|_1. \tag{27}$$

Moreover, we have

$$\langle \nabla_\theta \mathbb{G}_n(\hat{\theta};\hat{\mathcal{U}}) + \nabla_\theta \boldsymbol{p}(\lambda_n, \hat{\theta}), \theta_0 - \hat{\theta} \rangle \geqslant 0. \tag{28}$$

The true parameter θ_0 is feasible, so that we can chose $\theta = \theta_0$ in (28) and using (27), we have

$$\langle -\nabla_\theta \boldsymbol{p}(\lambda_n, \hat{\theta}) - \nabla_\theta \mathbb{G}_n(\theta_0;\hat{\mathcal{U}}), \Delta \rangle \geqslant \alpha_2 \|\Delta\|_2 - \nu_2 \|\Delta\|_1. \tag{29}$$

Then, by Hölder's inequality, we have

$$\begin{aligned} \langle -\nabla_\theta \boldsymbol{p}(\lambda_n, \hat{\theta}) - \nabla_\theta \mathbb{G}_n(\theta_0;\hat{\mathcal{U}}), \Delta \rangle &\leqslant \{ \|\nabla_\theta \boldsymbol{p}(\lambda_n, \hat{\theta})\|_\infty + \|\nabla_\theta \mathbb{G}_n(\theta_0;\hat{\mathcal{U}})\|_\infty \} \|\Delta\|_1 \\ &\leqslant \{ \lambda_n + \lambda_n/4 \} \|\Delta\|_1, \end{aligned}$$

where the last inequality follows from the bound in (4) with $\|\nabla_\theta \mathbb{G}_n(\theta_0;\hat{\mathcal{U}})\|_\infty \leqslant \lambda_n/4$ and Lemma 4 of Loh and Wainwright (2015) implies $\|\nabla_\theta \boldsymbol{p}(\lambda_n, \hat{\theta})\|_\infty \leqslant \lambda_n$. Hence, inequality (29) becomes

$$\|\Delta\|_2 \leqslant \frac{\|\Delta\|_1}{\alpha_2} \Big(\frac{5\lambda_n}{4} + \nu_2 \Big) \leqslant \frac{2R}{\alpha_2} \Big(\frac{5\lambda_n}{4} + \frac{\alpha_2}{4R} \Big). \tag{30}$$

Using the bounds (4), the right-hand side is upper-bounded by 1, which means $\|\Delta\|_2 \leqslant 1$. We may then apply the (RSC) condition for the case $\|\Delta\|_2 \leqslant 1$, that is

$$\langle \nabla_\theta \mathbb{G}_n(\hat{\theta};\hat{\mathcal{U}}) - \nabla_\theta \mathbb{G}_n(\theta_0;\hat{\mathcal{U}}), \Delta \rangle \geqslant \alpha_1 \|\Delta\|_2^2 - \nu_1 \|\Delta\|_1^2. \tag{31}$$

By convexity of $\boldsymbol{p}(\lambda_n, \theta) + \mu \|\theta\|_2^2/2$, we obtain

$$\begin{aligned} &\boldsymbol{p}(\lambda_n, \theta_0) + \frac{\mu}{2} \|\theta_0\|_2^2 - \boldsymbol{p}(\lambda_n, \hat{\theta}) - \frac{\mu}{2} \|\hat{\theta}\|_2^2 \geqslant \langle \nabla_\theta \{ \boldsymbol{p}(\lambda_n, \hat{\theta}) + \frac{\mu}{2} \|\hat{\theta}\|_2^2 \}, \theta_0 - \hat{\theta} \rangle \\ &\quad = \langle \nabla_\theta \boldsymbol{p}(\lambda_n, \hat{\theta}) + \mu \hat{\theta}, \theta_0 - \hat{\theta} \rangle, \end{aligned}$$

which yields

$$\langle \nabla_\theta \boldsymbol{p}(\lambda_n, \hat{\theta}), \theta_0 - \hat{\theta} \rangle \leqslant \boldsymbol{p}(\lambda_n, \theta_0) - \boldsymbol{p}(\lambda_n, \hat{\theta}) + \frac{\mu}{2} \|\Delta\|_2^2. \tag{32}$$

Hence, using (31), (28) and (32), we obtain

$$\alpha_1\|\Delta\|_2^2 - \nu_1\|\Delta\|_1^2 \leqslant -\langle\nabla_\theta\mathbb{G}_n(\theta_0,\hat{\mathcal{U}}),\Delta\rangle + \boldsymbol{p}(\lambda_n,\theta_0) - \boldsymbol{p}(\lambda_n,\tilde{\theta}) + \frac{\mu}{2}\|\Delta\|_2^2.$$

By Hölder's inequality, we get

$$\begin{aligned}(\alpha_1 - \frac{\mu}{2})\|\Delta\|_2^2 &\leqslant \boldsymbol{p}(\lambda_n,\theta_0) - \boldsymbol{p}(\lambda_n,\hat{\theta}) + \|\nabla_\theta\mathbb{G}_n(\theta_0;\hat{\mathcal{U}})\|_\infty\|\Delta\|_1 + \nu_1\|\Delta\|_1^2\\ &\leqslant \boldsymbol{p}(\lambda_n,\theta_0) - \boldsymbol{p}(\lambda_n,\hat{\theta}) + \big(\|\nabla_\theta\mathbb{G}_n(\theta_0;\hat{\mathcal{U}})\|_\infty + 2R\nu_1\big)\|\Delta\|_1.\end{aligned} \tag{33}$$

Moreover, by assumption, we have

$$\|\nabla_\theta\mathbb{G}_n(\theta_0;\hat{\mathcal{U}})\|_\infty + 2R\nu_1 \leqslant \frac{\lambda_n}{4} + \frac{\lambda_n}{4} \leqslant \frac{\lambda_n}{2}\cdot$$

Using (33) and Lemma 4 of Loh and Wainwright (2015), we obtain

$$(\alpha_1 - \frac{\mu}{2})\|\Delta\|_2^2 \leqslant \boldsymbol{p}(\lambda_n,\theta_0) - \boldsymbol{p}(\lambda_n,\hat{\theta}) + \frac{\lambda_n}{2}\Big\{\frac{\boldsymbol{p}(\lambda_n,\Delta)}{\lambda_n} + \frac{\mu}{2\lambda_n}\|\Delta\|_2^2\Big\}.$$

Note that, for any couple (t,t') of positive numbers, $t > t'$, and any $\lambda > 0$, we have $\big(p(\lambda,t) - p(\lambda,t')\big)/(t-t') \leqslant p(\lambda,t)/t \leqslant \lambda$, because $t \mapsto p(\lambda,t)/t$ is non-increasing. By assumption, $4\alpha_1/3 \geqslant \mu$. Thus, we have

$$0 \leqslant (\alpha_1 - \frac{3\mu}{4})\|\Delta\|_2^2 \leqslant \boldsymbol{p}(\lambda_n,\theta_0) - \boldsymbol{p}(\lambda_n,\hat{\theta}) + \frac{1}{2}\boldsymbol{p}(\lambda_n,\Delta). \tag{34}$$

Therefore, this provides

$$\begin{aligned}0 \leqslant (\alpha_1 - \frac{3\mu}{4})\|\Delta\|_2^2 &\leqslant \sum_{k\in\mathcal{A}}\big\{p(\lambda_n,|\theta_{0,k}|) - p(\lambda_n,|\hat{\theta}_k|)\big\} - \sum_{k\notin\mathcal{A}} p(\lambda_n,|\hat{\theta}_k|) + \frac{1}{2}\sum_k p(\lambda_n,\Delta)\\ &\leqslant \lambda_n\sum_{k\in\mathcal{A}}|(|\theta_{0,k}| - |\hat{\theta}_k|)| + \frac{1}{2}\Big(\sum_{k\in\mathcal{A}} p(\lambda_n,\Delta) - \sum_{k\notin\mathcal{A}} p(\lambda_n,\Delta)\Big)\\ &\leqslant \lambda_n\|\Delta_{\mathcal{A}}\|_1 + \frac{\lambda_n}{2}\|\Delta_{\mathcal{A}}\|_1 - 0 \leqslant \frac{3\lambda_n}{2}\|\Delta_{\mathcal{A}}\|_1 \leqslant \frac{3\lambda_n\sqrt{k_0}}{2}\|\Delta\|_2.\end{aligned} \tag{35}$$

Consequently, we obtain the upper bound

$$\|\hat{\theta} - \theta_0\|_2 \leqslant \frac{6\lambda_n\sqrt{k_0}}{4\alpha_1 - 3\mu}\cdot \tag{36}$$

Concerning the upper bound of $\|\hat{\theta} - \theta_0\|_1$, note that (35) implies

$$\frac{1}{2}\sum_{k\notin\mathcal{A}} p(\lambda_n,\Delta) \leqslant \lambda_n\sum_{k\in\mathcal{A}}(|\theta_{0,k}| - |\hat{\theta}_k|) + \frac{1}{2}\sum_{k\in\mathcal{A}} p(\lambda_n,\Delta) \leqslant \frac{3\lambda_n}{2}\|\Delta_{\mathcal{A}}\|_1.$$

From Lemma 4 in Loh and Wainwright (2015), for every real number t, we have $\lambda_n t \leqslant p(\lambda_n,t) + \mu t^2/2$. Applying this identity for every Δ_k, $k \notin \mathcal{A}$, this implies

$$\lambda_n \sum_{k \notin \mathcal{A}} |\Delta_k| \leqslant 3\lambda_n \|\Delta_{\mathcal{A}}\|_1 + \frac{\mu \|\Delta_{\mathcal{A}^c}\|_2^2}{2}. \tag{37}$$

We had proven above that $(\alpha_1 - 3\mu/4)\|\Delta\|_2^2 \leqslant 3\lambda_n \sqrt{k_0}\|\Delta_{\mathcal{A}}\|_2/2$, implying

$$\|\Delta_{\mathcal{A}^c}\|_2^2 \leqslant \frac{6\lambda_n \sqrt{k_0}}{(4\alpha_1 - 3\mu)} \|\Delta_{\mathcal{A}}\|_2.$$

We deduce from (37), $\|\Delta_{\mathcal{A}^c}\|_1 \leqslant 3\|\Delta_{\mathcal{A}}\|_1 + \frac{3\mu\sqrt{k_0}}{(4\alpha_1 - 3\mu)}\|\Delta_{\mathcal{A}}\|_2$. Invoking (36), this yields

$$\begin{aligned}\|\Delta\|_1 &\leqslant \|\Delta_{\mathcal{A}}\|_1 + \|\Delta_{\mathcal{A}^c}\|_1 \leqslant 4\|\Delta_{\mathcal{A}}\|_1 + \frac{3\mu\sqrt{k_0}}{(4\alpha_1 - 3\mu)}\|\Delta\|_2 \\ &\leqslant \left(4 + \frac{3\mu}{(4\alpha_1 - 3\mu)}\right)\sqrt{k_0}\|\Delta\|_2 \leqslant \frac{6(16\alpha_1 - 9\mu)}{(4\alpha_1 - 3\mu)^2}\lambda_n k_0,\end{aligned}$$

proving the result. □

Proof of Proposition 4.

Proof Let us establish that $\mathbb{G}_n(., \boldsymbol{y})$ satisfies the (RSC) condition. By the chain rule and usual calculations (Lütkepohl 1996, 10.6.1, Eq. (1)), the first-order conditions are

$$\begin{aligned}\nabla_{vec(\Sigma)}\mathbb{G}_n(\Sigma, \vec{\boldsymbol{u}}) &= -\frac{1}{n}\sum_{i=1}^n \left(\frac{\psi'}{\psi}\right)(\boldsymbol{y}_i'\Sigma^{-1}\boldsymbol{y}_i)\frac{\partial \boldsymbol{y}_i'\Sigma^{-1}\boldsymbol{y}_i}{\partial vec(\Sigma)} + \frac{1}{2}\frac{\partial \ln|\Sigma|}{\partial vec(\Sigma)} \\ &= \frac{1}{n}\sum_{i=1}^n \left(\frac{\psi'}{\psi}\right)(\boldsymbol{y}_i'\Sigma^{-1}\boldsymbol{y}_i)\left(\Sigma^{-1}\boldsymbol{y}_i \otimes \Sigma^{-1}\boldsymbol{y}_i\right) + \frac{vec(\Sigma^{-1})}{2}.\end{aligned} \tag{38}$$

By deriving (38), we obtain the Hessian matrix of $\mathbb{G}_n$

$$\begin{aligned}2\nabla^2_{vec(\Sigma),vec(\Sigma)'}\mathbb{G}_n(\Sigma, \vec{\boldsymbol{u}}) &= -\frac{2}{n}\sum_{i=1}^n \left(\frac{\psi''}{\psi} - \frac{(\psi')^2}{\psi^2}\right)(\boldsymbol{y}_i'\Sigma^{-1}\boldsymbol{y}_i)(\Sigma^{-1}\boldsymbol{y}_i \otimes \Sigma^{-1}\boldsymbol{y}_i)(\Sigma^{-1}\boldsymbol{y}_i \otimes \Sigma^{-1}\boldsymbol{y}_i)' \\ &\quad + \Sigma^{-1} \otimes \Sigma^{-1}\mathcal{S}_n(\Sigma)\Sigma^{-1} + \Sigma^{-1}\mathcal{S}_n(\Sigma)\Sigma^{-1} \otimes \Sigma^{-1} - \Sigma^{-1} \otimes \Sigma^{-1}.\end{aligned}$$

Note that the matrix $(\Sigma^{-1}\boldsymbol{y}_i \otimes \Sigma^{-1}\boldsymbol{y}_i)(\Sigma^{-1}\boldsymbol{y}_i \otimes \Sigma^{-1}\boldsymbol{y}_i)' = \Sigma^{-1}\boldsymbol{y}_i\boldsymbol{y}_i'\Sigma^{-1} \otimes \Sigma^{-1}\boldsymbol{y}_i\boldsymbol{y}_i'\Sigma^{-1}$ is nonnegative. Thus, with obvious notations,

$$\begin{aligned}2\nabla^2_{vec(\Sigma),vec(\Sigma)'}\mathbb{G}_n(\Sigma, \vec{\boldsymbol{u}}) &= \Sigma^{-1} \otimes \Sigma^{-1}\big(\mathcal{S}_n(\Sigma_0) - \Sigma/2\big)\Sigma^{-1} + \Sigma^{-1}\big(\mathcal{S}_n(\Sigma_0) - \Sigma/2\big)\Sigma^{-1} \otimes \Sigma^{-1} \\ &\quad + \Sigma^{-1} \otimes \Sigma^{-1}\big(\mathcal{S}_n(\Sigma) - \mathcal{S}_n(\Sigma_0)\big)\Sigma^{-1} + \Sigma^{-1}\big(\mathcal{S}_n(\Sigma) - \mathcal{S}_n(\Sigma_0)\big)\Sigma^{-1} \otimes \Sigma^{-1} \\ &\quad - \frac{2}{n}\sum_{i=1}^n \left(\frac{\psi'}{\psi}\right)'(\boldsymbol{y}_i\Sigma^{-1}\boldsymbol{y}_i)\Sigma^{-1}\boldsymbol{y}_i\boldsymbol{y}_i'\Sigma^{-1} \otimes \Sigma^{-1}\boldsymbol{y}_i\boldsymbol{y}_i'\Sigma^{-1} =: T_1 + T_2 + T_3.\end{aligned}$$

Consider $\Delta := \Sigma_1 - \Sigma_0$, $\Sigma_1 \in \Theta$, $\Sigma = \Sigma_0 + t\Delta$ for some $t \in [0,1]$ and $\boldsymbol{v} = vec(\Delta)$. As in the proof of Proposition 1 (see (12)), we obtain

$$\boldsymbol{v}' T_1 \boldsymbol{v} \geqslant ||\boldsymbol{v}||_2^2 \lambda_{\min}(2\mathcal{S}_n(\Sigma_0) - \Sigma)/q^3 \geqslant ||\boldsymbol{v}||_2^2 b/q^3. \tag{39}$$

Since the spectrum of $\Sigma^{-1} \otimes \Sigma^{-1}(\mathcal{S}_n(\Sigma) - \mathcal{S}_n(\Sigma_0))\Sigma^{-1}$ is the product of eigenvalues of Σ^{-1} and of $\mathcal{S}_n(\Sigma) - \mathcal{S}_n(\Sigma_0)$, we obtain

$$\|T_2\|_s \leqslant 2\|\Sigma^{-1}\|_s^3 \|\mathcal{S}_n(\Sigma) - \mathcal{S}_n(\Sigma_0)\|_s \leqslant 2\|\mathcal{S}_n(\Sigma) - \mathcal{S}_n(\Sigma_0)\|_s / \lambda_{min}(\Sigma)^3,$$

and then $|\boldsymbol{v}' T_2 \boldsymbol{v}| \leqslant \|\boldsymbol{v}\|_2^2 \|T_2\|_s \leqslant 2\|\mathcal{S}_n(\Sigma) - \mathcal{S}_n(\Sigma_0)\|_s \|\boldsymbol{v}\|_1^2 / a^3$.

Concerning T_3,

$$\begin{aligned} |\boldsymbol{v}' T_3 \boldsymbol{v}| &\leqslant \frac{2}{n} \sum_{i=1}^n \left| \left(\frac{\psi'}{\psi} \right)' (\boldsymbol{y}_i \Sigma^{-1} \boldsymbol{y}_i) \right| \boldsymbol{v}' \Sigma^{-1} \boldsymbol{y}_i \boldsymbol{y}_i' \Sigma^{-1} \otimes \Sigma^{-1} \boldsymbol{y}_i \boldsymbol{y}_i' \Sigma^{-1} \boldsymbol{v} \\ &\leqslant \frac{2\|\boldsymbol{v}\|_2^2}{n} \sum_{i=1}^n \theta_i^2 \|\Sigma^{-1} \boldsymbol{y}_i \boldsymbol{y}_i' \Sigma^{-1}\|_s^2 \leqslant \frac{2\|\boldsymbol{v}\|_2^2}{n} \sum_{i=1}^n \theta_i^2 \|\boldsymbol{y}_i\|_2^4 = V_n \|\boldsymbol{v}\|_2^2. \end{aligned}$$

Finally, this yields $2\boldsymbol{v} \nabla^2_{vec(\Sigma), vec(\Sigma)'} \mathbb{G}_n(\Sigma, \vec{\boldsymbol{u}}) \boldsymbol{v} \geqslant \|\boldsymbol{v}\|_2^2 \big(b/q^3 - (1 + C_\epsilon) V_n \big)$. Therefore, with the same reasoning as for the Gaussian copula case, the (RSC) condition is satisfied with $\alpha_1 = \alpha_2 = (b/q^3 - V_n)/4$ and $\nu_1 = 2\|\mathcal{S}_n(\Sigma) - \mathcal{S}_n(\Sigma_0)\|_s / a^3$, $\nu_2 = R\nu_1$.

□

Proof of Proposition 6.

Proof By obvious calculations, we obtain $\nabla_\theta \mathbb{G}_n(\theta) = -n^{-1} \sum_{i=1}^n V_\theta(\boldsymbol{u}_i) / \big(\omega' \vec{c}_\theta(\boldsymbol{u}_i) \big)$,

$$V_\theta(\boldsymbol{u}_i) := \big[\vec{c}_\theta(\boldsymbol{u}_i)', \omega_1 \partial_{\theta_1'} c_{1,\theta_1}(\boldsymbol{u}_i), \ldots, \omega_m \partial_{\theta_m'} c_{m,\theta_m}(\boldsymbol{u}_i) \big]',$$

that is a d-dimensional column vector. To lighten notations, $\mu_i(\theta) := (\omega' \vec{c}_\theta(\boldsymbol{u}_i))^{-1}$ is simply written μ_i when there is no ambiguity. As usual, such a θ belongs to the segment between the true parameter θ_0 and an arbitrarily chosen vector $\theta_1 \in \Theta$. In other words, $\theta = \theta_0 + t(\theta_1 - \theta_0)$, for some $t \in (0,1)$. Let us set $\boldsymbol{v} = \theta - \theta_0$. Then, simple calculations provide $\nabla^2_{\theta,\theta'} \mathbb{G}_n(\theta) = n^{-1} \sum_{i=1}^n \big(\mu_i^2 V_\theta V_\theta' - \mu_i W_\theta \big)(\boldsymbol{u}_i)$, and the "Hessian" matrix $W_\theta(\boldsymbol{u}) = \partial_{\theta'} V_\theta(\boldsymbol{u})$ is

$$\begin{bmatrix} 0 & \ldots & \ldots & 0 & \partial_{\theta_1'} c_{1,\theta_1} & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 & \partial_{\theta_2'} c_{2,\theta_2} & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & \ldots & 0 & 0 & \ldots & 0 & \partial_{\theta_m'} c_{m,\theta_m} \\ \partial_{\theta_1} c_{1,\theta_1} & 0 & \ldots & 0 & \omega_1 \partial^2_{\theta_1,\theta_1'} c_{1,\theta_1} & 0 & \ldots & 0 \\ 0 & \partial_{\theta_2} c_{2,\theta_2} & \ddots & \vdots & 0 & \omega_2 \partial^2_{\theta_2,\theta_2'} c_{2,\theta_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \partial_{\theta_m} c_{m,\theta_m} & 0 & \ldots & 0 & \omega_m \partial^2_{\theta_m,\theta_m'} c_{m,\theta_m} \end{bmatrix} (\boldsymbol{u}).$$

We rewrite the column vector $\boldsymbol{v}$ as a block column $[\boldsymbol{v}_0', \boldsymbol{v}_1', \ldots, \boldsymbol{v}_m']'$ or $[\boldsymbol{v}_0', \bar{\boldsymbol{v}}']'$, so that it is conformable with the gradient vectors $V_\theta(\boldsymbol{u})$. To lighten notations, for every $k = 0, \ldots, m$ and every $i = 1, \ldots, n$, set $\zeta_{k,i} := \boldsymbol{v}_k' \partial_{\theta_k} c_{k,\theta_k}(\boldsymbol{u}_i)$ and $\nu_{k,i} := \boldsymbol{v}_k' \partial^2_{\theta_k,\theta_k'} c_{k,\theta_k}(\boldsymbol{u}_i) \boldsymbol{v}_k$. Therefore, simple calculations yield

$$\begin{aligned}\boldsymbol{v}' \nabla^2_{\theta,\theta'} \mathbb{G}_n(\theta) \boldsymbol{v} &= \frac{1}{n} \sum_{i=1}^n \mu_i^2 \big(\boldsymbol{v}_0' \vec{c}_\theta(\boldsymbol{u}_i)\big)^2 + \frac{1}{n} \sum_{i=1}^n \Big\{ \Big(\frac{\sum_{k=1}^m \omega_k \zeta_{k,i}}{\sum_{k=1}^m \omega_k c_{k,\theta_k}(\boldsymbol{u}_i)} \Big)^2 - \frac{\sum_{k=1}^m \omega_k \nu_{k,i}}{\sum_{k=1}^m \omega_k c_{k,\theta_k}(\boldsymbol{u}_i)} \Big\} \\ &\quad + \frac{2}{n} \sum_{k,l=1}^m \sum_{i=1}^n \mu_i^2 v_{0,l} \omega_k \zeta_{k,i} c_{l,\theta_l}(\boldsymbol{u}_i) - \frac{2}{n} \sum_{k=1}^m \sum_{i=1}^n \mu_i v_{0,k} \zeta_{k,i} =: T_0 + T_1 + T_2 + T_3.\end{aligned}$$

We manage T_0 as in the proof of Proposition 5:

$$T_0 \geqslant \|\boldsymbol{v}_0\|_2^2 \inf_{\theta \in \Theta} \lambda_{\min} \Big(\frac{1}{n} \sum_{i=1}^n \mu_i^2(\theta) \vec{c}_\theta(\boldsymbol{u}_i) \vec{c}_\theta'(\boldsymbol{u}_i) \Big) =: \|\boldsymbol{v}_0\|_2^2 \, C(T_0).$$

By Assumption (23) and obvious notations, we have $T_1 \geqslant \alpha_{1,\omega} \|\bar{\boldsymbol{v}}\|_2^2 - \nu_{1,\omega} \|\bar{\boldsymbol{v}}\|_1^2$, $\|\bar{\boldsymbol{v}}\|_2 \leqslant 1$ and $T_1 \geqslant \alpha_{2,\omega} \|\bar{\boldsymbol{v}}\|_2 - \nu_{2,\omega} \|\bar{\boldsymbol{v}}\|_1$, when $\|\bar{\boldsymbol{v}}\|_2 > 1$. Moreover, we get

$$\begin{aligned}|T_2| &\leqslant \frac{2}{n} \sum_{i=1}^n \sum_{k=1}^m \mu_i^2 \omega_k |\zeta_{k,i}| \Big(\sum_{l=1}^m c_{l,\theta_l}(\boldsymbol{u}_i) |v_{0,l}| \Big) \\ &\leqslant \frac{2}{n} \sum_{i=1}^n \mu_i^2 \Big(\sum_{k=1}^m \omega_k \|\partial_{\theta_k} c_{k,\theta_k}(\boldsymbol{u}_i)\|_\infty \Big) \|\boldsymbol{v}\|_1 \sup_l c_{l,\theta_l}(\boldsymbol{u}_i) \|\boldsymbol{v}_0\|_1 \\ &\leqslant \frac{2\|\boldsymbol{v}\|_1 \min(\|\boldsymbol{v}\|_1, 1)}{n} \sum_{i=1}^n \mu_i^2 \Big(\sum_{k=1}^m \omega_k \|\partial_{\theta_k} c_{k,\theta_k}(\boldsymbol{u}_i)\|_\infty \Big) \sup_l c_{l,\theta_l}(\boldsymbol{u}_i) =: \|\boldsymbol{v}\|_1 \min(\|\boldsymbol{v}\|_1, 1) C(T_2),\end{aligned}$$

because $\|\boldsymbol{v}_0\|_1 \leqslant 1$. Similarly, we obtain

$$\begin{aligned}|T_3| &\leqslant \frac{2}{n} \sum_{i=1}^n \mu_i \Big(\sum_{k=1}^m \|\partial_{\theta_k} c_{k,\theta_k}(\boldsymbol{u}_i)\|_\infty |v_{0,k}| \Big) \|\boldsymbol{v}\|_1 \\ &\leqslant \frac{2\|\boldsymbol{v}\|_1 \min(\|\boldsymbol{v}\|_1, 1)}{n} \sum_{i=1}^n \mu_i \sup_k \|\partial_{\theta_k} c_{k,\theta_k}(\boldsymbol{u}_i)\|_\infty =: \|\boldsymbol{v}\|_1 \min(\|\boldsymbol{v}\|_1, 1) C(T_3).\end{aligned}$$

To summarize, if $\|\boldsymbol{v}\|_2 \leqslant 1$, we have obtained

$$\begin{aligned}\boldsymbol{v}' \nabla^2_{\theta,\theta'} \mathbb{G}_n(\theta) \boldsymbol{v} &\geqslant \|\boldsymbol{v}_0\|_2^2 C(T_0) + \alpha_{1,\omega} \|\bar{\boldsymbol{v}}\|_2^2 - \nu_{1,\omega} \|\bar{\boldsymbol{v}}\|_1^2 - \|\boldsymbol{v}\|_1^2 \big(C(T_2) + C(T_3) \big) \\ &\geqslant \|\boldsymbol{v}\|_2^2 \min \big(C(T_0), \inf_\omega \alpha_{1,\omega} \big) - \|\boldsymbol{v}\|_1^2 \big(\sup_\omega \nu_{1,\omega} + C(T_2) + C(T_3) \big).\end{aligned}$$

Moreover, if $\|\bar{\boldsymbol{v}}\|_2 > 1$ and then $\|\boldsymbol{v}\|_2 > 1$, we have got

$$\begin{aligned}\boldsymbol{v}'\nabla^2_{\theta,\theta'}\mathbb{G}_n(\theta)\boldsymbol{v} &\geqslant \|\boldsymbol{v}_0\|_2^2 C(T_0) + \alpha_{2,\omega}\|\bar{\boldsymbol{v}}\|_2 - \nu_{2,\omega}\|\bar{\boldsymbol{v}}\|_1 - \|\boldsymbol{v}\|_1\big(C(T_2)+C(T_3)\big)\\ &\geqslant \|\boldsymbol{v}\|_2 \min\big(C(T_0), \inf_\omega \alpha_{2,\omega}\big) - \|\boldsymbol{v}\|_1\big(\sup_\omega \nu_{2,\omega} + C(T_2) + C(T_3)\big),\end{aligned}$$

since $\|\boldsymbol{v}_0\|_2^2 + \|\bar{\boldsymbol{v}}\|_2 \geqslant \|\boldsymbol{v}\|_2 = \sqrt{\|\boldsymbol{v}_0\|_2^2 + \|\bar{\boldsymbol{v}}\|_2^2}$ when $\|\bar{\boldsymbol{v}}\|_2 > 1$. Finally, if $\|\bar{\boldsymbol{v}}\|_2 \leqslant 1$ and $\|\boldsymbol{v}\|_2 > 1$, we get

$$\begin{aligned}\boldsymbol{v}'\nabla^2_{\theta,\theta'}\mathbb{G}_n(\theta)\boldsymbol{v} &\geqslant \|\boldsymbol{v}_0\|_2^2 C(T_0) + \alpha_{1,\omega}\|\bar{\boldsymbol{v}}\|_2^2 - \nu_{1,\omega}\|\bar{\boldsymbol{v}}\|_1^2 - \|\boldsymbol{v}\|_1\big(C(T_2)+C(T_3)\big)\\ &\geqslant \|\boldsymbol{v}\|_2 \min\big(C(T_0), \inf_\omega \alpha_{1,\omega}\big) - \|\boldsymbol{v}\|_1\big(\sup_\omega \nu_{1,\omega} + C(T_2) + C(T_3)\big),\end{aligned}$$

because $\|\bar{\boldsymbol{v}}\|_1^2 \leqslant \|\boldsymbol{v}\|_1$ in this case. Therefore, the (RSC) condition is satisfied with the defined constants α_1, α_2, ν_1 and ν_2, proving the result. □

Supplementary Information The online version supplementary material available at https://doi.org/10.1007/s10463-021-00785-4.

Acknowledgements The authors have been supported by the labex Ecodec (Reference Project ANR-11-LABEX-0047) and the Japanese Society for the Promotion of Science (Grant 19K23193).

References

Banerjee, A., Merugu, S., Dhillon, I. S., Ghosh, J. (2005). Clustering with Bregman divergences. *Journal of Machine Learning Research*, *6*, 1705–1749.

Bickel, P. J., Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, *36*(6), 2577–2604.

Bien, J., Tibshirani, R. (2011). Sparse estimation of a covariance matrix. *Biometrika*, *98*(4), 807–820.

Boyd, S., Vandenberghe, L. (2004). *Convex optimization*. New York, NY: Cambridge University Press.

Bregman, L. M. (1967). The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, *7*, 200–217.

Cai, T. T., Zhou, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, *40*(5), 2389–2420.

Cai, T. T., Zhang, C. H., Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, *38*(4), 2118–2144.

Cambanis, S., Huang, S., Simons, G. (1981). On the theory of elliptically contoured distrbutions. *Journal of Multivariate Analysis*, *11*, 368–385.

Censor, Y., Zenios, S. (1998). *Parallel optimization: Theory, algorithms, and applications*. New York, NY: Oxford University Press.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalised likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Genest, C., Ghoudi, K., Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, *82*, 543–552.

Ghoudi, K., Rémillard, B. (1998). Empirical processes based on pseudo-observations. In *Asymptotic Methods in Probability and Statistics*, 171–197. Amsterdam, North-Holland: Elsevier.

Ghoudi, K., Rémillard, B. (2004). Empirical processes based on pseudo-observations. II. The multivariate case. In *Asymptotic Methods in Stochastics*, 381–406, Fields Institute Communications, 44. American Mathematical Society.

Gómez, E. M., Gómez-Villegas, A., Marín, J. M. (2003). A survey on continuous elliptical vector distributions. *Revista Matemática Complutense*, *16*, 345–361.
Górecki, J., Hofert, M., Holeňa, M. (2016). On structure, family and parameter estimation of hierarchical archimedean copulas. arXiv:1611.09225.
Gray, R., Buzo, A., Gray, A., Matsuyama, Y. (1980). Distortion measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 367–376.
Hofert, M., Pham, D. (2013). Densities of nested archimedean copulas. *Journal of Multivariate Analysis*, *118*, 37–52.
Knight, K., Fu, W. (2000). Asymptotics for Lasso-Type Estimators. *The Annals of statistics*, *28*, 1356–1378.
Liu, H., Lafferty, J., Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, *10*, 2295–2328.
Loh, P. L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics*, *45*, 866–896.
Loh, P. L., Wainwright, M. J. (2015). Regularized M-estimators with non-convexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, *16*, 559–616.
Loh, P. L., Wainwright, M. J. (2017). Support recovery without incoherence: A case for non-convex regularization. *The Annals of Statistics*, *45*, 2455–2482.
Lütkepohl, H. (1996). *Handbook of matrices*. Chichester, WS: Wiley.
Magnus, J. R., Neudecker, H. (2019). *Matrix differential calculus with applications in statistics and econometrics*. Hoboken, NJ: Wiley.
McNeil, A. J. (2008). Sampling nested Archimedean copulas. *Journal of Statistical Computation and Simulation*, *78*, 567–581.
Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, *27*, 538–557.
Nelsen, R. B. (2006). *An introduction to copulas*. New York, NY: Springer.
Okhrin, O., Okhrin, Y., Schmid, W. (2013a). On the structure and estimation of hierarchical Archimedean copulas. *Journal of Econometrics*, *173*, 189–204.
Okhrin, O., Okhrin, Y., Schmid, W. (2013b). Properties of hierarchical Archimedean copulas. *Statistics & Risk Modeling*, *30*, 21–54.
Okhrin, O., Ristig, A., Sheen, J.R., Trück, S. (2015). Conditional systemic risk with penalized copula (No. 2015-038). SFB 649 Discussion Paper.
Ravikumar, P., Wainwright, M. J., Raskutti, G., Yu, B. (2011). High-dimensional covariance estimation by minimizing 1-penalized log-determinant divergence. *Electronic Journal of Statistics*, *5*, 935–980.
Segers, J., Uyttendaele, N. (2014). Nonparametric estimation of the tree structure of a nested archimedean copula. *Computational Statistics & Data Analysis*, *72*, 190–204.
Shi, J., Louis, T. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, *51*, 1384–1399.
Van de Geer, S., Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, *3*, 1360–1392.
Van der Vaart, A., Wellner, J. (2007). Empirical processes indexed by estimated functions. Asymptotics: Particles, Processes and Inverse Problems. *Institute of Mathematical Statistics Lecture Notes*, *55*, 234–252.
Wegkamp, M., Zhao, Y. (2016). Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli*, *22*, 1184–1226.
Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2), 894–942.
Zou, H., Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, *37*, 1733–1751.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.