Série des Documents de Travail

n° 2019-01

The finite sample properties of Sparse M-estimators with Pseudo-Observations

B.POIGNARD[1]
J-D.FERMANIAN[2]

[1] Osaka University, Graduate School of Engineering Science. E-mail: poignard@sigmath.es.osaka-u.ac.jp
[2] CREST; ENSAE. E-mail : jean-david.fermanian@ensae.fr

The finite sample properties of Sparse M-estimators with Pseudo-Observations

Benjamin Poignard[*], Jean-David Fermanian[†]

January 4, 2019

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 that hold with high probability. 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.

Key words: Non-convex regularizer; copulas; pseudo-observations; statistical consistency; exponential bounds.

Running title: Non-convex M-estimation with pseudo-observations.

1 Introduction

The need for a joint modeling for high-dimensional random vectors has fostered a flourishing research in sparse models. The application domains of sparse modeling has 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

[*]Osaka University, Graduate School of Engineering Science. E-mail address: poignard@sigmath.es.osaka-u.ac.jp

[†]Ensae-Crest, 5 avenue Henry le Chatelier, 91129 Palaiseau, France. E-mail address: jean-david.fermanian@ensae.fr (corresponding author)

both statistically precise and provides intuitive insights among asset relationships. This gave rise to sparse matrix precision estimation or sparse factor modeling, e.g.

Nowadays, copulas constitute a standard way of modeling the joint distribution of a random vector. They are flexible in that they allow a separate modeling 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 a standard. Beside, nonparametric estimation of copulas (since the seminal paper of Deheuvels, 1979) 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 copula may also involve numerous parameters. The use of conditional copulas has opened the door towards rich regression-type dependence models. Hopefully, regularizing a copula-based model through a penalization procedure offers an interesting strategy to tackle the potential 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), Zhang and Zou (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 allow 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 a consistency result for very general penalization functions, in which we explicitly are able to manage pseudo-observations. This framework encompasses both parametric and semi-parametric models. It is then applied to some copula-based models: Gaussian and Student copulas, mixtures, etc. Moreover, we evaluate the probabilities of satisfying the conditions so that such consistency results apply. To the best of our knowledge, this paper is the first attempt to build bridges between general penalized (non-convex) M-estimators and the semi-parametric inference of copula models.

The remainder of the paper is organized as follows. In Section 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 and we state some upper bounds for the probability of satisfying our assumptions. Section 3 is dedicated to the application of these results to some semi-parametric copula models. Section 4 illustrates these theoretical results through a short simulated experiment.

2 Nonconvex penalized criteria based on pseudo-observations

2.1 Copula models

Let us start with an i.i.d. sample of n realizations of a random vector $\boldsymbol{X} \in \mathbb{R}^q$, denoted as $\mathcal{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$. 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 would imply we should 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 $\hat{\boldsymbol{U}}_i = (\hat{U}_{i,1}, \cdots, \hat{U}_{i,q})$, $i = 1, \ldots, n$, obtained from the initial sample $\mathcal{X}$. For instance and as usual, 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 $F_{n,k}(s) := n^{-1}\sum_{i=1}^{n}\mathbf{1}_{X_{i,k}\leq s}$. 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 $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d)$, i.e. given $\boldsymbol{U}_i = \boldsymbol{u}_i$ for every $i = 1, \ldots, n$ and under $\mathbb{P}_\theta$. Typically, $\mathbb{G}_n(\theta; \boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ is the empirical loss associated to a continuous function $\ell : \Theta \times [0,1]^q \to \mathcal{R}_+$, i.e.

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

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

$$\hat{\theta} = \underset{\theta: g(\theta)\leq R}{\arg\min}\, \{\mathbb{G}_n(\theta,\hat{\mathcal{U}}) + \boldsymbol{p}(\lambda_n, \theta)\}, \tag{2.1}$$

where $\boldsymbol{p}(\lambda_n, .) : \mathbb{R}^d \to \mathbb{R}$ is a regularizer and λ_n is the regularization parameter, which depends on the sample size and enforce 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). Due to the potential non-convexity of this penalty, we include the side condition $g(\theta) \geq \|\theta\|_1$ for every θ. The function $\theta \to \mathbb{E}[\ell(\theta, \boldsymbol{U})]$ is supposed to be uniquely minimized at $\theta = \theta_0$ so that $\mathbb{E}[\nabla_\theta \mathbb{G}_n(\theta_0, \mathcal{U})] = 0$.

The true parameter θ_0 is supposed to be sparse, so that $k_0 = \text{card}(\mathcal{A})$, with $\mathcal{A} = \{i :$

$\theta_{0,i} \neq 0\}$. Note that θ_0 is independent of the sample size n. We impose that $g(\theta_0) \leq R$, so that θ_0 is a feasible point.

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 Lo and Wainwright (2015, 2017).

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

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

(i) $\rho \mapsto \boldsymbol{p}(\lambda_n, \rho)$ *is symmetric around zero and* $\boldsymbol{p}(\lambda_n, 0) = 0$.

(ii) $\rho \mapsto \boldsymbol{p}(\lambda_n, \rho)$ *is non-decreasing on* $\mathbb{R}^+$.

(iii) $\rho \mapsto \boldsymbol{p}(\lambda_n, \rho)/\rho$ *is non-increasing on* $\mathbb{R}_\star^+$.

(iv) $\rho \mapsto \boldsymbol{p}(\lambda_n, \rho)$ *is differentiable for any* $\rho \neq 0$.

(v) $\lim_{\rho \to 0^+} \boldsymbol{p}'(\lambda_n, \rho) = \lambda_n$.

(vi) $\rho \mapsto \boldsymbol{p}(\lambda_n, \rho) + \mu\rho^2/2$ *is convex for some* $\mu \geq 0$.
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* $\boldsymbol{p}'(\lambda_n, \rho) = 0$ *for* $\rho \geq \lambda_n \gamma$.

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. 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

$$
\begin{aligned}
\textbf{Lasso}: \ \boldsymbol{p}(\lambda_n, \rho) &= \lambda_n |\rho|, \\
\textbf{MCP}: \ \boldsymbol{p}(\lambda_n, \rho) &= sign(\rho)\lambda_n \textstyle\int_0^{|\rho|} (1 - z/(\lambda_n b_{mcp}))_+ dz, \\
\textbf{SCAD}: \ \boldsymbol{p}(\lambda_n, \rho) &= \begin{cases} \lambda_n |\rho|, & \text{for } |\rho| \leq \lambda_n, \\ -(\rho^2 - 2b_{scad}\lambda_n|\rho| + \lambda_n^2)/(2(b_{scad}-1)), & \text{for } \lambda_n \leq |\rho| \leq b_{scad}\lambda_n, \\ (b_{scad}+1)\lambda_n^2/2, & \text{for } |\rho| > b_{scad}\lambda_n, \end{cases}
\end{aligned}
$$

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).

Obviously, numerous copula log-densities are not concave functions of their parameters. 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 τ_1, τ_2 of (θ, n, d) such that, for any $\Delta \in \mathbb{R}^d$,

$$
\langle \nabla_\theta \mathbb{L}_n(\theta + \Delta) - \nabla_\theta \mathbb{L}_n(\theta), \Delta \rangle \geq \alpha_1 \|\Delta\|_2^2 - \tau_1 \frac{\ln d}{n} \|\Delta\|_1^2, \text{ if } \|\Delta\|_2 \leq 1, \tag{2.2}
$$

$$
\langle \nabla_\theta \mathbb{L}_n(\theta + \Delta) - \nabla_\theta \mathbb{L}_n(\theta), \Delta \rangle \geq \alpha_2 \|\Delta\|_2 - \tau_2 \sqrt{\frac{\ln d}{n}} \|\Delta\|_1, \text{ if } \|\Delta\|_2 \geq 1. \tag{2.3}
$$

Note that the (RSC) property is fundamentally local and that α_k, τ_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 (see Theorem 2.1 below) and we will need such extensions for the copula models of Section 3. Indeed, we will apply the (RSC) condition with $\mathbb{L}(\theta) = \mathbb{G}(\theta, \mathcal{U})$ ($\mathcal{U}$ containing unfeasible observations, most of the time) and/or $\mathbb{L}(\theta) = \mathbb{G}(\theta, \hat{\mathcal{U}})$ (with the so-called pseudo-observations). Moreover, to weaken notations, we simply write α_k and τ_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*

$$\begin{aligned}
\langle \nabla_{\bar{\theta}} \overline{\mathbb{L}}_n(\bar{\theta} + \bar{\Delta}) - \nabla_{\bar{\theta}} \overline{\mathbb{L}}_n(\bar{\theta}), \bar{\Delta} \rangle &\geq \bar{\alpha}_1 \|\bar{\Delta}\|_2^2 - \bar{\tau}_1 \frac{\ln d}{n} \|\bar{\Delta}\|_1^2, \ \|\bar{\Delta}\|_2 \leq \zeta, \\
\langle \nabla_{\bar{\theta}} \overline{\mathbb{L}}_n(\bar{\theta} + \bar{\Delta}) - \nabla_{\bar{\theta}} \overline{\mathbb{L}}_n(\bar{\theta}), \bar{\Delta} \rangle &\geq \bar{\alpha}_2 \|\bar{\Delta}\|_2 - \bar{\tau}_2 \sqrt{\frac{\ln d}{n}} \|\bar{\Delta}\|_1, \ \|\bar{\Delta}\|_2 \geq \zeta,
\end{aligned}$$

with the new constants $(\bar{\alpha}_1, \bar{\tau}_1, \bar{\alpha}_2, \bar{\tau}_2) := (\alpha_1/\zeta^2, \tau_1/\zeta^2, \alpha_2/\zeta, \tau_2/\zeta)$.

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 $\mathbb{G}_n(., \hat{\mathcal{U}})$.

Theorem 2.1. *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*

$$4 \max \left\{ \|\nabla_\theta \mathbb{G}_n(\theta_0, \hat{\mathcal{U}})\|_\infty, \alpha_2 \sqrt{\frac{\ln d}{n}} \right\} \leq \lambda_n \leq \frac{\alpha_2}{6R}, \tag{2.4}$$

and $n \geq 16R^2 \max\{\tau_1^2, \tau_2^2\} \ln d / \alpha_2^2$. Then, a stationary point $\hat{\theta}$ of (2.1), satisfies

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

The proof is provided in the appendix.

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.*

Remark 3. *Our proof of Theorem 2.1 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*

$$0 \leq 3\boldsymbol{p}(\lambda_n,\theta_0) - \boldsymbol{p}(\lambda_n,\hat{\theta}) \leq \lambda_n(3\|\Delta_{\mathcal{M}}\|_1 - \|\Delta_{\mathcal{M}^c}\|_1), \tag{2.5}$$

where $\mathcal{M} = \max_{i\in\mathcal{A}}\{\hat{\theta}_i - \theta_{0,i}\}$ *(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 inequality (2.5) is* $2|\beta^*| - |\beta| \leq 2|\nu_A| - |\nu_{A^c}|$*, or even* $4 - a - b \leq 2|a-2| - b$*. This is clearly false in general: set* $a = 3/2$*, for instance.*

2.5 Sufficient conditions based on exponential bounds

Now, we aim obtaining an exponential-type inequality to evaluate the probability of satisfying the condition (2.4) of Theorem 2.1. In general, this implies to evaluate non-linear functions of pseudo-observations. To derive such bound, we rely on the Bernstein inequality and suitable regularity conditions on the loss function. Since the interesting event involves pseudo-observations, its probability of occurrence is unclear, but it is possible to bound such a probability from above. Assumptions 3 to 7 are technical but are not demanding. From now on, we restrict ourselves to criteria $\mathbb{G}_n(\theta,\boldsymbol{u})$ that are written as sums over all the $\boldsymbol{u}-$components (as for log-likelihood criteria, but not only).

Assumption 3.

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

for some trimming function $\pi : [0,1]^q \to [0,1]$ *whose support is included in* $\prod_{k=1}^{q}[a_k, 1-b_k]$*, for some nonnegative numbers* a_k, b_k*,* $k = 1,\ldots,q$*,* $a_k + b_k < 1$*. The function* $\theta \to \mathbb{E}[\pi(\boldsymbol{U})\ell(\theta,\boldsymbol{U})]$ *is supposed to be uniquely minimized at* $\theta = \theta_0$*.*

As in (2.1), the optimizer of $\mathbb{G}_n(\theta,\vec{\boldsymbol{u}})$ is denoted by $\hat{\theta}$. The trimming function will avoid that some of the pseudo-observations are too close to the boundaries of the unit

hypercube $[0,1]^q$. Indeed, a lot of usual copula log-densities (the Gaussian copula density, e.g.) tend to the infinity when some of their arguments tend to 0 or 1.

Remark 4. *For the sake of simplicity, we have considered a fixed trimming functions π. It would be possible to consider instead a sequence of such functions (π_n) whose support tend to $(0,1)^q$ when n tends to the infinity, as in Fermanian and Lopez (2018). Since this would induce significant additional complexities, we have preferred to leave such extensions to some interested readers.*

Assumption 4. *For every $\boldsymbol{u}$, the function $\theta \mapsto \ell(\boldsymbol{u},\theta)$ is continuously differentiable on Θ. Its derivative at θ is denoted by $\dot{\ell}(\boldsymbol{u},\theta)$. The function $\boldsymbol{u} \mapsto \dot{\ell}(\boldsymbol{u},\theta_0)$ is two times continuously differentiable on $(0,1)^q$. For any $k=1,\ldots,d$, denote*

$$\sigma_k^2 := \mathbb{E}\Big[\big(\pi(\boldsymbol{U}_i)\frac{\partial \ell}{\partial \theta_k}(\theta,\boldsymbol{U}_i)_{|\theta=\theta_0}\big)^2\Big],$$

and assume that, for any $s \geq 3$,

$$\mathbb{E}\Big[\Big(\pi(\boldsymbol{U}_i)\frac{\partial \ell}{\partial \theta_k}(\theta,\boldsymbol{U}_i)_{|\theta=\theta_0}\Big)_+^s\Big] \leq \frac{s!}{2}\sigma_k^2 c_k^{s-2}. \tag{2.6}$$

Assumption 5. *Denote*

$$I_{kl} := \int |\pi(\boldsymbol{u})\frac{\partial^2 \ell}{\partial u_l \partial \theta_k}(\theta,\boldsymbol{u})_{|\theta=\theta_0}| c(\boldsymbol{u},\theta_0)\, d\boldsymbol{u}, \text{ and } \sigma_{kl}^2 := \mathbb{E}\Big[\Big(\pi(\boldsymbol{U}_i)\frac{\partial^2 \ell}{\partial u_l \partial \theta_k}(\theta,\boldsymbol{U}_i)_{|\theta=\theta_0}\Big)^2\Big].$$

Assume that

$$\mathbb{E}\Big[\Big(\pi(\boldsymbol{U}_i)\frac{\partial^2 \ell}{\partial u_l \partial \theta_k}(\theta,\boldsymbol{U}_i)_{|\theta=\theta_0}\Big)_+^s\Big] \leq \frac{s!}{2}\sigma_{kl}^2 c_{kl}^{s-2}, \ \forall s \geq 3. \tag{2.7}$$

Assumption 6. *There exist some measurable functions h_{klj} s.t.*

$$\sup_{\boldsymbol{u}} |\mathbf{1}(u_s \in [v_s - a_s/2; v_s + b_s/2], \forall s)\frac{\partial^3 \ell}{\partial^2 u_l u_j \partial \theta_k}(\theta_0,\boldsymbol{u})| \leq h_{klj}(\boldsymbol{v}), \tag{2.8}$$

for every $k=1,\ldots,d$, $l,j=1,\ldots,q$, every $\boldsymbol{v} \in supp(\pi)$, and $H_{klj} := \mathbb{E}\big[\pi(\boldsymbol{U})h_{klj}(\boldsymbol{U})\big] < \infty$.

Assumption 7. *Denote $\mathbb{E}\big[\pi(\boldsymbol{U})h_{klj}(\boldsymbol{U})^2\big] = \tau_{klj}^2$, and assume, for every $s \geq 3$,*

$$\mathbb{E}\Big[\pi(\boldsymbol{U})h_{klj}(\boldsymbol{U})^s\Big] \leq \frac{s!}{2}\tau_{klj}^2 d_{klj}^{s-2}. \tag{2.9}$$

Assumption 8. *The pseudo-observations are given by the usual empirical counterparts, i.e.*

$$\hat{U}_{i,k} = F_{n,k}(X_{i,k}) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}(X_{j,k} \leq X_{i,k}), \quad k = 1, \ldots, q.$$

Now, let us state our exponential bounds, whose proof has been postponed into an appendix.

Theorem 2.2. *Under Assumptions 3 to 8 and for every $\epsilon > 0$,*

$$\mathbb{P}\left(\|\nabla_\theta \mathbb{G}_n(\theta_0, \mathcal{U})\|_\infty > \epsilon\right) \leq 2\sum_{k=1}^{d} \exp\left(-\frac{n\epsilon^2}{18(\sigma_k^2 + c_k\epsilon/3)}\right), \text{ and}$$

$$\begin{aligned}
\mathbb{P}\left(\|\nabla_\theta \mathbb{G}_n(\theta_0, \hat{\mathcal{U}})\|_\infty > \epsilon\right) &\leq 2\sum_{k=1}^{d} \exp\left(-\frac{n\epsilon^2}{18(\sigma_k^2 + c_k\epsilon/3)}\right) \\
&+ \sum_{k=1}^{d}\sum_{l=1}^{q}\left\{\exp\Big(-\frac{2n\epsilon^2}{9q^2 I_{kl}^2(1 + 2c_{kl}\epsilon/(3qI_{kl}\sigma_{kl}))}\Big) + 2\exp\Big(-\frac{2n\epsilon^2}{9q^2 I_{kl}^2(1 + 2\epsilon\sigma_{kl}/(3qI_{kl}^2))^2}\Big)\right\} \\
&+ \sum_{k=1}^{d}\sum_{l,j=1}^{q}\Big\{\exp\Big(-\frac{4n\epsilon}{3q^2 H_{klj}(1 + 2d_{klj}\sqrt{2\epsilon}/(q\tau_{kl}\sqrt{3H_{klj}}))}\Big) \\
&+ 2\exp\Big(-\frac{4n\epsilon}{3q^2 H_{klj}(1 + 2\tau_{klj}\sqrt{2\epsilon}/(qH_{klj}\sqrt{3H_{klj}}))}\Big)\Big\}.
\end{aligned}$$

In particular, applying the previous result with $\epsilon = \lambda_n/4$ allows the estimation of the probability that Condition (2.4) in Theorem 2.1 in fulfilled.

3 Application to some copula families

In this section, we provide some insights regarding the applicability of the finite sample results of Section 2.4 to some copula-based likelihood functions. This means that, from now on, we choose the loss function as 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 $\boldsymbol{u}_i$, $i = 1, \ldots, n$ a set of n random vectors in $[0,1]^q$. This as a generic notations for a usual iid sample $\mathcal{U}$, or for a sample of n pseudo-observations $\hat{\mathcal{U}}$ as above. 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 2.1 apply?

3.1 Gaussian copula models

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

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

for any $\boldsymbol{u} \in (0,1)^q$, where Φ and Φ_Σ 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 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 be also 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. $\theta = vec(\Sigma)$). For convenience, we will write $\mathbb{G}_n(\Sigma, \cdot)$ instead of $\mathbb{G}_n(\theta, \cdot)$ in this subsection, with a slight abuse of notation. Therefore, the regularized statistical criterion may be written as a maximization of a multivariate (in general trimmed) Gaussian log-likelihood, i.e.

$$\left\{ \begin{array}{rcl} \hat{\Sigma} & = & \underset{\Sigma \in \Theta}{\arg\min} \, \{\mathbb{G}_n(\Sigma, \boldsymbol{u}_1, \cdots, \boldsymbol{u}_n) + \boldsymbol{p}(\lambda_n, \Sigma)\}, \ \text{with} \\ \mathbb{G}_n(\Sigma, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) & = & s_\pi\big(n \ln(2\pi)/2 + \ln|\Sigma|/2\big) + \sum_{i=1}^n \pi_i \boldsymbol{x}_i' \Sigma^{-1} \boldsymbol{x}_i/(2n), \\ \boldsymbol{x}_i & := & (\Phi^{-1}(u_{i,1}), \ldots, \Phi^{-1}(u_{i,q})), \ \ i = 1, \ldots, n, \\ \pi_i & := & \pi(\boldsymbol{u}_i), \ \text{and } s_\pi := \sum_{i=1}^n \pi_i/n. \end{array} \right.$$

Above, Θ denotes a $q \times q$-correlation matrices subset such as

$$\Theta = \{\Sigma : \Sigma = \Sigma', Diag(\Sigma) = Id, \ \ \lambda_{\min}(2\Sigma_n - s_\pi \Sigma) > a, g(\Sigma) \leq R\}, \tag{3.1}$$

for some positive constants a and R, and introducing the "empirical" covariance matrix $\Sigma_n := \sum_{i=1}^n \pi_i \boldsymbol{x}_i \boldsymbol{x}_i'/n$. Obviously and as in Subsection 2.2, the function g assumed to be convex and it satisfies $\|vech(\Sigma)\|_1 \leq g(\Sigma)$ for every correlation matrix Σ. Note that, in practice, most of the weights π_i, $i = 1, \ldots, n$ are one, and then s_π is close to one. Moreover, note that Θ is convex: if $\lambda_{\min}(2\Sigma_n - s\pi\Sigma_k) > a$, $k = 1, 2$, then

$$\lambda_{\min}(2\Sigma_n - s_\pi(t\Sigma_1 + (1-t)\Sigma_2)) \geq t\lambda_{\min}(2\Sigma_n - s_\pi\Sigma_1) + (1-t)\lambda_{\min}(2\Sigma_n - s_\pi\Sigma_2) > a.$$

Moreover, the function $\Sigma \mapsto \mathbb{G}_n(\Sigma, \boldsymbol{u}_1, \ldots \boldsymbol{u}_n)$ is convex on Θ for any values of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ (apply Boyd and Vandenberghe 2004, exercise 7.4).

Note that, if we observe realizations of $\boldsymbol{U}$, Σ_n can be the defined either as $\sum_{i=1}^n \pi(\boldsymbol{U}_i)\boldsymbol{X}_i\boldsymbol{X}_i'/n$, with $\boldsymbol{X}_i := (\Phi^{-1}(\boldsymbol{U}_{i,1}), \ldots, \Phi^{-1}(\boldsymbol{U}_{i,q}))$. Alternatively, if we only observe pseudo-realizations realizations of $\boldsymbol{U}$, Σ_n is $\sum_{i=1}^n \pi(\hat{\boldsymbol{U}}_i)\hat{\boldsymbol{X}}_i\hat{\boldsymbol{X}}_i'/n$, with $\hat{\boldsymbol{X}}_i := (\Phi^{-1}(\hat{\boldsymbol{U}}_{i,1}), \ldots, \Phi^{-1}(\hat{\boldsymbol{U}}_{i,q}))$. When dealing with true observations (resp. pseudo-observations), the function $\Sigma \mapsto \mathbb{G}_n(\Sigma, \mathcal{U})$ (resp. $\Sigma \mapsto \mathbb{G}_n(\Sigma, \hat{\mathcal{U}})$) is convex; it can be optimized with usual optimization procedures as long as the matrix $2\Sigma_n - \Sigma$ is positive. The latter condition is satisfied when n is large and when Σ is not "too far" away from the "true" matrix Σ_0. In particular, this is the case when the spectral norm of $\Sigma - \Sigma_n$ is smaller than the spectral norm of Σ_n.

The total number of nonzero entries is denoted as $k_0 = |\mathcal{A}|$ with

$$\mathcal{A} = \{(i,j) : i > j \text{ and } \Sigma_{0,(i,j)} \neq 0\}.$$

Corollary 3.1. *Suppose that λ_n satisfies*

$$2\max\left\{\|\Sigma_0^{-1} - \Sigma_0^{-1}\Sigma_n\Sigma_0^{-1}\|_\infty, a\frac{\sqrt{\ln(q(q-1)/2)}}{2q^3 n^{1/2}}\right\} \leq \lambda_n \leq \frac{a}{24q^3R}. \tag{3.2}$$

Suppose that Σ_0 belongs to the convex parameter set Θ given in (3.1) and that $2a/q^3 - 3\mu > 0$. Then, for every n, any stationary point $\hat{\Sigma}$ of (3.1) satisfies

$$\|vech(\hat{\Sigma}) - vech(\Sigma_0)\|_2 \leq \frac{6\lambda_n\sqrt{k_0}}{2a/q^3 - 3\mu}, \text{ and } \|vech(\hat{\Sigma}) - vech(\Sigma_0)\|_1 \leq \frac{6(4a/q^3 - 9\mu)\lambda_n k_0}{(2a/q^3 - 3\mu)^2}.$$

Note that the upper bounds we obtain above depend on the dimension matrix q, i.e. on the number of free parameters. The latter could depend the sample size n too.

Proof. To establish the (RSC) condition, use the differential operator applied w.r.t. Σ. Then, usual calculations provide $2\nabla_\Sigma \mathbb{G}_n(\Sigma, \vec{\boldsymbol{u}}) = \Sigma^{-1}(s_\pi \Sigma - \Sigma_n)\Sigma^{-1}$. Hence in vector form, the derivative becomes

$$2\nabla_{vec(\Sigma)} \mathbb{G}_n(\Sigma, \vec{\boldsymbol{u}}) = vec(\Sigma^{-1}(s_\pi \Sigma - \Sigma_n)\Sigma^{-1}).$$

To check the (RSC) condition, we now focus on the Hessian matrix of $\mathbb{G}_n$. The formulas in Subsection 10.6.1. in Lütkepohl (1996) yield

$$2\nabla^2_{vec(\Sigma),vec(\Sigma)'} \mathbb{G}_n(\Sigma, \vec{\boldsymbol{u}}) = \Sigma^{-1}\Sigma_n\Sigma^{-1} \otimes \Sigma^{-1} + \Sigma^{-1} \otimes \Sigma^{-1}\Sigma_n\Sigma^{-1} - s_\pi \Sigma^{-1} \otimes \Sigma^{-1}.$$

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

$$\begin{aligned} e_n(\Sigma) &:= vec(\Delta)' \nabla^2_{vec(\Sigma),vec(\Sigma)'} \mathbb{G}_n(\Sigma, \boldsymbol{u}) vec(\Delta) \\ &\geq \frac{1}{2} vec(\Delta)' \Big(\Sigma^{-1}(\Sigma_n - s_\pi \Sigma/2)\Sigma^{-1} \otimes \Sigma^{-1} + \Sigma^{-1} \otimes \Sigma^{-1}(\Sigma_n - s_\pi \Sigma/2)\Sigma^{-1} \Big) vec(\Delta) \\ &\geq \|\Delta\|_F^2 \lambda_{\min}(\Sigma_n - s_\pi \Sigma/2) \lambda_{\min}(\Sigma^{-1})^3, \end{aligned}$$

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

$$\begin{aligned} e_n(\Sigma) &\geq ||\Delta||_F^2 \lambda_{\min}(2\Sigma_n - s_\pi \Sigma) \lambda_{\max}(\Sigma)^{-3}/2 \\ &\geq ||\Delta||_F^2 \lambda_{\min}(2\Sigma_n - s_\pi \Sigma)/(2q^3) \geq ||\Delta||_F^2 a/(2q^3). \end{aligned} \tag{3.3}$$

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 a located strictly below the main diagonal of Σ. But, with obvious notations, note that $\|\Delta\|_F^2 = \|\Sigma - \Sigma_0\|_2^2 = 2\|\theta - \theta_0\|_2^2$ 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,

$$(\theta-\theta_0)'\nabla^2_{\theta,\theta'}\mathbb{G}_n(\theta^*,\vec{\boldsymbol{u}})(\theta-\theta_0)\geq||\theta-\theta_0||_2^2 a/(4q^3),$$

for any θ^* that lies between θ and θ_0. Thus, at Σ_0, the (RSC) condition is satisfied with $\alpha_1=a/(4q^3)$ and $\alpha_2=\alpha_1$, $\tau_1=\tau_2=0$. And the result follows from Theorem 2.1. □

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, through our functions $\boldsymbol{p}$. In this case, the regularized statistical criterion would become

$$\left\{\begin{array}{rcl}\hat{S} & = & \underset{S\in\bar{\Theta}}{\arg\min}\,\{\mathbb{G}_n(S,\boldsymbol{u}_1,\cdots,\boldsymbol{u}_n)+\boldsymbol{p}(\lambda_n,S)\},\ \text{ with}\\ \mathbb{G}_n(S,\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n) & = & s_\pi\big(n\ln(2\pi)/2-\ln|S|/2\big)+\sum_{i=1}^n\pi_i\boldsymbol{x}_i'S\boldsymbol{x}_i/(2n),\end{array}\right.$$

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 latter case (derivations wrt Σ): see Corollary 3 in Lo and Wainwright (2015), for instance. Unfortunately, we have to restrict ourselves on the inverse of *correlation* matrices, and then the parameter subset is no longer convex. This explains why we have preferred to parameterize 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 $\int_0^\infty r^{q-1}\psi(r^2)\,dr<\infty$. 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

$$c_g(\boldsymbol{u})=\frac{\psi\left(\vec{F}_\psi^{-1}(\boldsymbol{u})'\Sigma^{-1}\vec{F}_\psi^{-1}(\boldsymbol{u})\right)}{|\Sigma|^{1/2}\prod_{k=1}^q f_\psi(F_\psi^{-1}(u_k))},\quad \vec{F}_\psi^{-1}(\boldsymbol{u}):=[F_\psi^{-1}(u_1),\ldots,F_\psi^{-1}(u_q)]',\tag{3.4}$$

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.

$$F_\psi(x) = \int_{-\infty}^{x} \psi_1(t^2)\, dt, \quad \psi_1(u) = \frac{\pi^{(q-1)/2}}{\Gamma((q-1)/2)} \int_0^\infty \psi(u+s) s^{(q-3)/2}\, ds. \tag{3.5}$$

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 parametrize the model by Σ instead of by 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.

$$\psi(t) = \int_0^\infty (2\pi r^2)^{-q/2} \exp(-t/2r^2)\, F_\infty(dr), \tag{3.6}$$

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, (3.6) 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 Subsection 3.1, we define the statistical criterion as

$$\left\{ \begin{array}{rcl} \hat{\Sigma} & = & \arg\min\limits_{\Sigma\in\Theta} \{\mathbb{G}_n(\Sigma, \boldsymbol{y}) + \boldsymbol{p}(\lambda_n, \Sigma)\}, \text{ where} \\ \mathbb{G}_n(\Sigma, \vec{\boldsymbol{u}}) & = & s_\pi \ln|\Sigma|/2 - \sum_{i=1}^n \pi_i \ln\psi(\boldsymbol{y}_i'\Sigma^{-1}\boldsymbol{y}_i)/n, \\ \boldsymbol{y}_i & := & \big(F_\psi^{-1}(u_{i,1}), \ldots, F_\psi^{-1}(u_{i,q})\big),\ i = 1, \ldots, n, \\ \pi_i & := & \pi(\boldsymbol{u}_i), \text{ and } s_\pi := \sum_{i=1}^n \pi_i/n. \end{array} \right. \tag{3.7}$$

Denote by $\|A\|_s$ the usual spectral norm of any matrix. Θ will be a set of $q \times q$-correlation matrices such as

$$\Theta = \{\Sigma : \Sigma = \Sigma', Diag(\Sigma) = Id,\ \|\Sigma - \Sigma_0\|_s < \epsilon,\ \lambda_{\min}(2\mathcal{S}_n(\Sigma_0) - s_\pi\Sigma) > b,\ g(\Sigma) \le R\}, \tag{3.8}$$

for some positive constants $\epsilon < 1$ and b. For an arbitrary correlation matrix, we have denoted

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

Note that $\mathcal{S}_n(\Sigma)$ is nonnegative because ψ is decreasing under (3.6). It is not difficult to check that Θ is convex. Moreover, Σ_0, the true correlation matrix, is assumed to belong to Θ and satisfies $\mathbb{E}[\nabla_{vec(\Sigma)}\mathbb{G}_n(\Sigma_0, \mathcal{U})] = 0$ by assumption. The true subset model $\mathcal{A}$ admits the same cardinality k_0 as in the Gaussian copula case.

Under (3.6), note that $(\psi')^2 \leq \psi''\psi$ by the Cauchy-Schwarz inequality. Then, we set, for every $i = 1, \ldots, n$,

$$\sup_{\Sigma|\, \|\Sigma-\Sigma_0|_s<\epsilon} \Big(\frac{\psi'}{\psi}\Big)'(\boldsymbol{y}_i\Sigma^{-1}\boldsymbol{y}_i) := \theta_i^2 \text{ and } V_n := \frac{2}{n}\sum_{i=1}^{n} \pi_i\theta_i^2\|\boldsymbol{y}_i\|_2^4.$$

Corollary 3.2. *Let*

$$C_\epsilon := \frac{\epsilon\|\Sigma_0^{-1}\|_s^2}{(1-\epsilon)^2}, \text{ and } \alpha = \frac{(b/q^3 - (1+C_\epsilon)V_n)}{4}.$$

Assume (3.6), $4\alpha > 3\mu$, *and that* (λ_n, R) *satisfies*

$$2\max\{\|vec(\Sigma_0^{-1}\mathcal{S}_n(\Sigma_0)\Sigma_0^{-1} - s_\pi\Sigma_0^{-1})\|_\infty, 2\alpha\sqrt{\frac{\ln d}{n}}\} \leq \lambda_n \leq \frac{\alpha}{6R}.$$

Then, any stationary point $\hat{\Sigma}$ *of (3.7) satisfies*

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

Note that the case of elliptical copulas is more complex than the case of Gaussian copulas because the set of matrices s.t. $2\mathcal{S}_n(\Sigma) - s_\pi\Sigma$ is non convex in general. Therefore, we had to restrict the set of possible matrices Σ by adding the condition $\|\Sigma - \Sigma_0\|_s < \epsilon$.

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 arbitrary matrix* $\bar{\Sigma}$ *that is not "too far" from* Σ_0 ($\|\Sigma_0 - \bar{\Sigma}\|_s < 1$, *to be specific). In particular,* $\bar{\Sigma}$ *may be chosen as a preliminary crude estimator of* Σ_0.

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

$$\nabla_{vec(\Sigma)}\mathbb{G}_n(\Sigma,\vec{\boldsymbol{u}}) = -\frac{1}{n}\sum_{i=1}^{n}\pi_i\Big(\frac{\psi'}{\psi}\Big)(\boldsymbol{y}_i'\Sigma^{-1}\boldsymbol{y}_i)\frac{\partial \boldsymbol{y}_i'\Sigma^{-1}\boldsymbol{y}_i}{\partial vec(\Sigma)} + \frac{s_\pi}{2}\frac{\partial \ln|\Sigma|}{\partial vec(\Sigma)}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\pi_i\Big(\frac{\psi'}{\psi}\Big)(\boldsymbol{y}_i'\Sigma^{-1}\boldsymbol{y}_i)\big(\Sigma^{-1}\boldsymbol{y}_i\otimes\Sigma^{-1}\boldsymbol{y}_i\big) + \frac{s_\pi}{2}vec(\Sigma^{-1}). \tag{3.9}$$

Equivalently (Lütkepohl 1996, p.177, Eq. (10)), note that this could be rewritten as

$$2\nabla_{vec(\Sigma)}\mathbb{G}_n(\Sigma,\vec{\boldsymbol{u}}) = vec\big(-\Sigma^{-1}\mathcal{S}_n(\Sigma)\Sigma^{-1} + s_\pi\Sigma^{-1}\big).$$

By deriving (3.9), we obtain the Hessian matrix of $\mathbb{G}_n$

$$2\nabla^2_{vec(\Sigma),vec(\Sigma)'}\mathbb{G}_n(\Sigma,\vec{\boldsymbol{u}}) = -\frac{2}{n}\sum_{i=1}^{n}\pi_i\Big(\frac{\psi''}{\psi} - \frac{(\psi')^2}{\psi^2}\Big)(\boldsymbol{y}_i'\Sigma^{-1}\boldsymbol{y}_i)(\Sigma^{-1}\boldsymbol{y}_i\otimes\Sigma^{-1}\boldsymbol{y}_i)(\Sigma^{-1}\boldsymbol{y}_i\otimes\Sigma^{-1}\boldsymbol{y}_i)'$$
$$+ \quad \Sigma^{-1}\otimes\Sigma^{-1}\mathcal{S}_n(\Sigma)\Sigma^{-1} + \Sigma^{-1}\mathcal{S}_n(\Sigma)\Sigma^{-1}\otimes\Sigma^{-1} - s_\pi\Sigma^{-1}\otimes\Sigma^{-1}.$$

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,

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

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 Corollary 3.1 (see (3.3)), we obtain

$$\boldsymbol{v}'T_1\boldsymbol{v} \geq ||\boldsymbol{v}||_2^2\lambda_{\min}(2\mathcal{S}_n(\Sigma_0) - s_\pi\Sigma)/q^3 \geq ||\boldsymbol{v}||_2^2 b/q^3. \tag{3.10}$$

Note that, for any multiplicative matrix norm $\|\cdot\|$ (in particular the spectral norm $\|\cdot\|_s$), we have (Lütkepohl 1996, p.1076),

$$\|\Sigma^{-1} - \Sigma_0^{-1}\| \leq \|\Sigma_0^{-1}\|\|\Sigma^{-1}\|\|\Sigma - \Sigma_0\| \leq \frac{\|\Sigma_0^{-1}\|}{1 - \|\Sigma - \Sigma_0\|}\|\Sigma - \Sigma_0\|\cdot$$

Under our assumptions, for any vector $\boldsymbol{v} \in \mathbb{R}^q$,

$$|\boldsymbol{v}'\big(\mathcal{S}_n(\Sigma) - \mathcal{S}_n(\Sigma_0)\big)\boldsymbol{v}| \leq \frac{2}{n}\sum_{i=1}^n \pi_i\theta_i^2 \big|\boldsymbol{y}_i'\big(\Sigma^{-1} - \Sigma_0^{-1}\big)\boldsymbol{y}_i\big|(\boldsymbol{v}'\boldsymbol{y}_i)^2$$

$$\leq \quad \frac{2\|\Sigma_0^{-1}\|_s\|\Sigma - \Sigma_0\|_s}{n(1-\epsilon)}\sum_{i=1}^n \pi_i\theta_i^2\|\boldsymbol{y}_i\|_2^2(\boldsymbol{v}'\boldsymbol{y}_i)^2 \leq \frac{2\epsilon\|\Sigma_0^{-1}\|_s}{n(1-\epsilon)}\sum_{i=1}^n \pi_i\theta_i^2\|\boldsymbol{y}_i\|_2^4\|\boldsymbol{v}\|_2^2.$$

We deduce the upper bound

$$\|\mathcal{S}_n(\Sigma) - \mathcal{S}_n(\Sigma_0)\|_s \leq \frac{2\epsilon\|\Sigma_0^{-1}\|_s}{n(1-\epsilon)}\sum_{i=1}^n \pi_i\theta_i^2\|\boldsymbol{y}_i\|_2^4 = \frac{\epsilon\|\Sigma_0^{-1}\|_s V_n}{(1-\epsilon)}.$$

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 those of $\mathcal{S}_n(\Sigma) - \mathcal{S}_n(\Sigma_0)$, we obtain

$$\|T_2\|_s \leq 2\|\Sigma^{-1}\|_s\|\mathcal{S}_n(\Sigma) - \mathcal{S}_n(\Sigma_0)\|_s \leq \frac{\epsilon\|\Sigma_0^{-1}\|_s^2 V_n}{(1-\epsilon)^2} = C_\epsilon V_n,$$

and then $\boldsymbol{v}'T_2\boldsymbol{v}| \leq \|\boldsymbol{v}\|_2^2\|T_2\|_s \leq C_\epsilon V_n\|\boldsymbol{v}\|_2^2$.

Concerning the "remainder" term T_3,

$$|\boldsymbol{v}'T_3\boldsymbol{v}| \leq \frac{2}{n}\sum_{i=1}^n \pi_i\left|\Big(\frac{\psi'}{\psi}\Big)'(\boldsymbol{y}_i\Sigma^{-1}\boldsymbol{y}_i)\right|\boldsymbol{v}'\Sigma^{-1}\boldsymbol{y}_i\boldsymbol{y}_i'\Sigma^{-1}\otimes\Sigma^{-1}\boldsymbol{y}_i\boldsymbol{y}_i'\Sigma^{-1}\boldsymbol{v}$$

$$\leq \quad \frac{2\|\boldsymbol{v}\|_2^2}{n}\sum_{i=1}^n \pi_i\theta_i^2\|\Sigma^{-1}\boldsymbol{y}_i\boldsymbol{y}_i'\Sigma^{-1}\|_s^2 \leq \frac{2\|\boldsymbol{v}\|_2^2}{n}\sum_{i=1}^n \pi_i\theta_i^2\|\boldsymbol{y}_i\|_2^4 = V_n\|\boldsymbol{v}\|_2^2.$$

Finally, this yields $2\boldsymbol{v}\nabla^2_{vec(\Sigma),vec(\Sigma)'}\mathbb{G}_n(\Sigma,\vec{\boldsymbol{u}})\boldsymbol{v} \geq \|\boldsymbol{v}\|_2^2\big(b/q^3 - (1+C_\epsilon)V_n\big)$. Therefore, with the same reasoning as for the Gaussian copula case, the (RSC) condition is satisfied with $\alpha_1 = (b/q^3 - (1+C_\epsilon)V_n)/4$ and $\alpha_2 = \alpha_1$, $\tau_1 = \tau_2 = 0$. □

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, e.g.), there is a one-to-one mapping between the components of $\Sigma = [\sigma_{kl}]_{1\leq k,l\leq q}$ and all the bivariate Kendall's tau $\tau_{k,l}$ associated to 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', the statistical criterion

is based on a moment-based penalized method to estimate Σ and is given by

$$\left\{\begin{array}{rcl} \forall (k,l), \hat{\sigma}_{k,l} & = & \underset{\sigma_{k,l}: g(\Sigma)\leq R}{\arg\min} \{\mathbb{G}_n(\sigma_{k,l}, \vec{\boldsymbol{u}}) + \boldsymbol{p}(\lambda_n, \sigma_{k,l})\}, \text{ where} \\ \mathbb{G}_n(\sigma_{k,l}, \vec{\boldsymbol{u}}) & = & \big(\sigma_{k,l} - \sin(\pi\hat{\tau}_{k,l}/2)\big)^{\alpha}, \ \alpha \geq 1, \text{ with} \\ \hat{\tau}_{k,l} & := & \frac{2}{n(n-1)} \sum_{i<j} \Big(\mathbf{1}(X_{i,k} \leq X_{i,l}, X_{j,k} \leq X_{j,l}) - \mathbf{1}(X_{i,k} \geq X_{i,l}, X_{j,k} \leq X_{j,l}) \Big). \end{array}\right. \tag{3.11}$$

Note that this way of working allows 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 could 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. Even if it is always possible to project $\hat{\Sigma}$ on the subset of correlation matrices, the associated theoretical properties of the 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 though mixtures. Indeed, consider a family of fixed q-dimensional copulas $(C_k, k = 1, \ldots, m)$. 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 $\theta := (\omega_1, \ldots, \omega_m)'$, with $\omega_k \in [0,1]$ for every $k = 1, \ldots, m$ and $\sum_{k=1}^m \omega_k = 1$. The associated loss function is (minus) the corresponding log-likelihood. Denoting by c_k is the copula density associated with C_k, $k = 1, \ldots, m$, the statistical criterion is thus given by

$$\left\{\begin{array}{rcl} \hat{\theta} & = & \underset{\theta\in\Theta}{\arg\min} \{\mathbb{G}_n(\theta, \boldsymbol{u}) + \boldsymbol{p}(\lambda_n, \theta)\}, \text{ where} \\ \mathbb{G}_n(\theta, \vec{\boldsymbol{u}}) & = & -\sum_{i=1}^n \pi_i \ln\left(\sum_{k=1}^m \omega_k c_k(\boldsymbol{u}_i)\right)/n, \end{array}\right. \tag{3.12}$$

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

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

Corollary 3.3. *For any $\theta \neq \mathbf{0}$, let $\alpha = \lambda_{\min}\big(n^{-1}\sum_{i=1}^n \pi_i\mu_i^2\vec{c}(\boldsymbol{u}_i)\vec{c}(\boldsymbol{u}_i)'\big)$, and assume $\alpha > 3\mu/4$. Suppose that (λ_n, R) satisfy*

$$4\max\left\{\|n^{-1}\sum_{i=1}^n \pi_i\mu_i\vec{c}(\boldsymbol{u}_i)\|_\infty, \alpha\sqrt{\frac{\ln m}{n}}\right\} \leq \lambda_n \leq \frac{\alpha}{6R}.$$

Then any stationary point $\hat{\theta}$ of (3.13) satisfies

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

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

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

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

$$\theta'\vec{c}(\boldsymbol{u}_i) \leq \theta_0'\vec{c}(\boldsymbol{u}_i) + \|\theta_0-\theta_1\|_2\|\vec{c}(\boldsymbol{u}_i)\|_2 \leq \mu_i^{-1}.$$

Therefore, this yields

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

and the (RSC) applies with $\alpha_1=\alpha_2=\lambda_{\min}\big(n^{-1}\sum_{i=1}^n \pi_i\mu_i^2\vec{c}(\boldsymbol{u}_i)\vec{c}(\boldsymbol{u}_i)'\big)$, $\tau_1=\tau_2=0$. □

Remark 6. *As for 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 applied, replacing θ_0 (in Θ and μ_i) by any feasible parameter $\bar{\theta}$ s.t. $\|\theta_0-\bar{\theta}\|_2<1$.*

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 $\mathcal{C}_k := \{C_{k,\theta_k} \text{ copula on } [0,1]^q; \theta_k \in \Theta_k \subset \mathbb{R}^{d_k}\}$, 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

$$\left\{\begin{array}{rcl}\hat{\theta} & = & \arg\min\limits_{\theta\in\Theta}\{\mathbb{G}_n(\theta,\boldsymbol{u}) + \boldsymbol{p}(\lambda_n,\theta)\}, \text{ where} \\ \mathbb{G}_n(\theta,\vec{\boldsymbol{u}}) & = & -\sum_{i=1}^n \pi_i \ln\left(\sum_{k=1}^m \omega_k c_{k,\theta_k}(\boldsymbol{u}_i)\right)/n,\end{array}\right. \tag{3.13}$$

$$\Theta := \{\theta \in \mathbb{R}_+^m \times \bigtimes_{k=1}^m \Theta_k,\ \sum_{k=1}^m \omega_k = 1,\ \omega_k \in [\underline{\omega}_k, \overline{\omega}_k] \text{ for all } k,\ \|\theta - \theta_0\|_2 \leq 1,\ g(\theta) \leq R\},$$

for some positive constants $\underline{\omega}$ and $\overline{\omega}$, $k = 1, \ldots, n$. The dimension of θ is then $d := m + d_1 + \ldots, d_m$.

We will assume that a (RSC)-type condition applies on every model "marginal" parameterized by θ_k, $k = 1, \ldots, m$: for every $k = 1, \ldots, m$, there exist some constants $\alpha_{1,k} > 0$ and $\tau_{1,k} \geq 0$ s.t.

$$\boldsymbol{v}_k' \nabla_{\theta_k, \theta_k'} \mathbb{G}_n(\theta, \vec{\boldsymbol{u}}) \boldsymbol{v}_k \geq \alpha_{1,k} \|\boldsymbol{v}_k\|_2^2 - \tau_{1,k} \frac{\ln d_k}{n} \|\boldsymbol{v}_k\|_1^2,$$

for every $\theta \in \Theta$ and $\boldsymbol{v}_k \in \mathbb{R}^{d_k}$, $\|\boldsymbol{v}_k\|_2 \leq 1$.

Set $\omega = [\omega_1, \ldots, \omega_m]'$ and $\vec{c}_\theta(\boldsymbol{u}) := [c_{1,\theta_1}(\boldsymbol{u}), \ldots, c_{m,\theta_m}(\boldsymbol{u})]'$. For every $i = 1, \ldots, n$, let $\psi_i := \sup_{\theta \in \Theta} \sup_k \|\partial_{\theta_k} \ln c_{k,\theta_k}(\boldsymbol{u}_i)\|_\infty$, and $\mu_i(\theta) := (\omega' \vec{c}_\theta(\boldsymbol{u}_i))^{-1}$ for every $\theta \in \Theta$. We introduce the constants

$$\alpha_1 := \min\left(\inf_{\theta \in \Theta} \lambda_{\min}\Big(\frac{1}{n} \sum_{i=1}^n \mu_i^2(\theta) \vec{c}_\theta(\boldsymbol{u}_i) \vec{c}_\theta'(\boldsymbol{u}_i)\Big); \inf_k \big(\underline{\omega_k}^2 \alpha_{1k}\big) \|\boldsymbol{v}\|_2^2 \right),$$

$$\tau := \sup_k \big(\overline{\omega_k}^2 \tau_{1k}\big) + \frac{1}{n} \sum_{i=1}^n \pi_i \psi_i^2 + \frac{2}{n} \sum_{i=1}^n \pi_i \psi_i \sup_{\theta \in \Theta} \Big\{ \frac{\max_k c_{k,\theta_k}(\boldsymbol{u}_i)}{\sum_{l=1}^m \underline{\omega}_l c_{l,\theta_l}(\boldsymbol{u}_i)} \Big\} + \frac{\lambda_n}{2 \min_k \underline{\omega}_k}.$$

Corollary 3.4. *Assume that $4\alpha_1 > 3\mu$, and that (λ_n, R) satisfies*

$$4 \max\left\{ \|\nabla_\theta \mathbb{G}_n(\theta, \vec{\boldsymbol{u}})\|_\infty, \alpha_2 \sqrt{\frac{\ln d}{n}} \right\} \leq \lambda_n \leq \frac{\alpha_2}{6R}.$$

for some positive constant α_2. Then, $n > 16R^2\tau^2 \ln d / \alpha_2^2$, any stationary point $\hat{\theta}$ of (3.13) satisfies

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

The proof is given in the appendix.

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 Subsection 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. At the opposite, Hierarchical Archimedean copulas (HAC) are nice generalizations. They allow asymmetries and different dependencies for couples of variables, by combining a hierarchy of Archimedean copulas C_j, $j = 1, \ldots, m$, with different parameters θ_j. Obviously, the whole model is known once we have known/estimated $\theta := (\theta_1, \ldots, \theta_m)$. 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. W.l.o.g., we keep these assumptions, but our ideas appliy 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 every case, under our penalized point-of-view, there is not 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 to 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 Subsection 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: $\phi_j(\tau_j) = \theta_j$, $j = 1, \ldots, m$. 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 $\theta_1, \ldots, \theta_{j-1}$ 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 empirical Kendall's tau, as in (3.11), denoted by $\hat{\tau}_j$, and to estimate θ_j as

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

And the process goes on, allowing the estimation of all parameters θ_k successively. Nonetheless, we will not try to detail technical conditions to apply Theorem 2.1 for such models. Actually, this task is unfeasible in generality, and analytic calculations have to be done for every particular parametric model.

3.5 Conditional copula models

At first glance, there do not exist so many highly dimensional parametric copula models, in the literature, beside elliptical copulas and mixtures of copulas. In particular, most popular archimedean copulas depend on only one or two parameters. Nonetheless, a natural source of highly parameterized specifications is given by conditional copula models. In such a case, the observed random vector $\boldsymbol{X} \in \mathbb{R}^q$ is split into two parts, as $\boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{Z})$, where $\boldsymbol{Y} \in \mathbb{R}^p$ is the explained random vector and $\boldsymbol{Z} \in \mathbb{R}^m$ are covariates (fixed or random). Therefore, the previous results can be adapted, by considering pseudo-observations $\hat{U}_{i,k} := \hat{F}_k(Y_{i,k}|\boldsymbol{Z}_i)$, $i = 1, \ldots, n$, $k = 1, \ldots, q$, where $\hat{F}_k(y|\boldsymbol{z})$ denotes a consistent estimator of $F_k(y|\boldsymbol{z}) := \mathbb{P}(Y_k \leq y|\boldsymbol{Z} = \boldsymbol{z})$. Typically, set

$$\hat{F}_k(y|\boldsymbol{z}) := \frac{\sum_{i=1}^n \mathbf{1}(Y_{i,k} \leq y) K((\boldsymbol{z} - \boldsymbol{Z}_i)/h)}{\sum_{i=1}^n K((\boldsymbol{z} - \boldsymbol{Z}_i)/h)},$$

for some kernel $K : \mathbb{R}^m \mapsto \mathbb{R}$ and a bandwidth sequence $h = h(n)$ that tends to zero with n. Alternatively, it is always possible to specify some parametric models for the conditional of some Y_k given $\boldsymbol{Z}$ instead.

Our Theorem 2.1 directly applies with such new pseudo-observations, because it is not based on a probabilistic reasoning. In such a case, the "constants" (α_k, τ_k), $k = 1, 2$, in the (RSC) condition possibly depend on the sample of explanatory variables $(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ and on the way we have defined our pseudo-observations (conditional parametric/semi-or nonparametric models). Moreover, Theorem 2.2 has to be modified: we have to replace the DKW inequality by an exponential inequality related to the new pseudo-observations.

Such inequalities are available in the literature: see Proposition 11 in Fermanian and Lopez (2018), that builds on Einmahl and Mason (2005). Details are left to the reader.

For such conditional copula models, the model parameters are those given by the conditional law of $\boldsymbol{Y}$ given $\boldsymbol{Z}$, once the effect of the conditional margins Y_k given $\boldsymbol{Z}$, $k = 1, \ldots, q$, have been removed by Sklar's theorem. The corresponding criteria $\mathbb{G}_n$ are similar to the previous ones, but with conditional distributions instead. This induces more parameters than previously, due to the model specification of the effect of covariates. For instance, in the case of a conditional Gaussian copula, it could be assumed that every coefficient of Σ (or Σ^{-1}) is a function $\Sigma(\boldsymbol{z})$, given $\boldsymbol{Z} = \boldsymbol{z}$. A usual difficulty would be find the right functional to insure its positiveness for every $\boldsymbol{z}$. Several solutions have been proposed in the literature, as spectral or Cholevski decompositions, hypersherical coordinates (Jaeckel and Rebonato 2001, e.g.) or vines (Poignard and Fermanian, 2018), for instance.

4 Empirical study

In this section, we carry out a short simulation study to illustrate the theoretical results on the regularized M-estimator in the presence of pseudo-observations. To do so, we consider the Gaussian copula family described in Subsection 3.1: the data generating process is induced by a Gaussian copula with parameter Σ_0 (a correlation matrix), which is supposed to be sparse, so that the number of non-zero components is arbitrarily set depending on the problem size.

We consider two cases, whether the margins of $\boldsymbol{U}_i = \boldsymbol{u}_i$ are known or unknown. In the first case, we simulate a random vector $\boldsymbol{Z}$ according to a Gaussian distribution $\mathcal{N}(0, \Sigma_0)$ in $\mathbb{R}^q$, and we apply the parametric transform $U_{i,k} = \Phi(Z_{i,k})$, $k = 1, \ldots, q$, so that we obtain our sample $\mathcal{U}$. In the case of unknown margins, we compute pseudo-observations $\hat{\boldsymbol{U}}$ by applying non-parametric transforms $\hat{U}_{i,k} = \hat{F}_k(Z_{i,k})$ for each observation i and component $k = 1, \cdots, q$. The transform $\hat{F}_k(\cdot)$ is specified as the usual empirical cdf.

To recover the sparse support $\mathcal{A}$, we consider the regularized problem as detailed in Section 3.1. Denoting $\theta = vech(\Sigma)$, we set $g(\theta) = \|\theta\|_1$. To solve this optimization problem, we follow the composite gradient descent procedure of Loh and Wainwright (2015, section 4), which consists in a three step updating procedure of the optimized parameter value. Importantly, due to the constraints on the (RSC) coefficients, we considered $a = 0.1$, where a is the lower bound of $\lambda_{\min}(2\Sigma_n - s_\pi \Sigma)$ specified in Θ, and thus set

$b_{scad} = b_{mcp} = 15003$ (resp. 12000), the values from which $\alpha_1 > \frac{3}{4}\mu$ is satisfied when $p = 10$ (resp. $p = 20$). Hence for $p = 10$, $\alpha = 5 \times 10^{-5}$, $\mu = 6.6664 \times 10^{-5}$ for both SCAD and MCP. For $p = 20$, $\alpha = 6.25 \times 10{-}6$, $\mu = 8.333 \times 10^{-6}$ for both SCAD and MCP.

As for the regularization parameters, following Loh and Wainwright (2015,2017), we select $R = \boldsymbol{p}(\lambda_n, \theta_0)/\lambda_n$. Furthermore, we set $\lambda_n = c\sqrt{\log d}/\sqrt{n}$, where d is the problem size, that is $d = q(q-1)/2$. The constant c is chosen among a grid of four values within $[1, 2]$ so that we perform a cross-validation procedure to choose the optimal λ_n.

We consider two problem sizes: $q = 10$ and $q = 20$, so that the total number of parameters is 45 and 190, respectively. We set $k_0 := |\mathcal{A}| = 22$ (resp. $k_0 = 95$) for $q = 10$ (resp. $q = 20$) and arbitrarily fix this true sparse correlation matrix for each sample size. For $q = 10$ (resp. $q = 20$), $\|vech(\Sigma_0)\|_1 = 4.34$ (resp. $\|vech(\Sigma_0)\|_1 = 13.08$) and $\|vech(\Sigma_0)\|_2 = 1.21$ (resp. $\|vech(\Sigma_0)\|_2 = 1.66$). Then for each sample size, we simulate 200 times the random vector $\boldsymbol{Z}$, and thus we obtain 200 sparsity-based estimates $\hat{\Sigma}$. Figures 1a and 1b show their $\|.\|_1$ consistency with respect to the sample size for both dimensions q. Each point represents the average error over the 200 simulations. As predicted in Corollary 3.1, the three curves for the MCP, SCAD and Lasso converge toward zero as the number of samples increases. The same remark holds for the $\|.\|_2$ consistency displayed in Figures 1c and 1d. Interestingly, each plot displays the sparsity-based estimation under $\hat{\boldsymbol{U}}$ and $\boldsymbol{U}$. Although the statistical error decreases, the estimation is less precise under the $\hat{\boldsymbol{U}}$ case due to the non-parametric transform $\hat{F}_k(\cdot)$ to each margin and its amount of additional noise.

References

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[2] I.S. Borisov. Approximation of distributions of von Mises statistics with multidimensional kernels. *Siberian Math. J.*, 32:554-566, 1991.

[3] S. Boucheron, G. Lugosi, G. and P. Massart. *Concentration inequalities. A non asymptotic theory of independence.* Oxford UP, 2013.

[4] S. Cambanis, S. Huang and G. Simons. On the theory of elliptically contoured distrbutions. *Journal of Multivariate Analysis*, 11:368-385, 1981.

[5] P. Deheuvels. La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. *Académie Royale de Belgique, Bulletin de la Classe des Sciences*, 65:274292, 1979.

[6] U. Einmahl and D. Mason. Uniform in bandwidth consistency of kernel-type function estimators. *Annals of Statistics*, 33:1380-1403, 2005.

[7] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:13481360, 2001.

[8] J.-D. Fermanian and O. Lopez. Single-index copulas. *Journal of Mulivariate Analysis*, 165:27-55, 2018.

[9] C. Genest, K. Ghoudi and L.-P. Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82:543552, 1995.

[10] E.M. Gómez, A. Gómez-Villegas and J.M. Marín. A survey on continuous elliptical vector distributions. *Revista matemática complutense*, 16:345-361, 2003.

[11] J. Górecki, M. Hofert and M. Holeňa. On structure, family and parameter estimation of hierarchical archimedean copulas. arXiv preprint arXiv:1611.09225, 2016.

[12] M. Hofert and D. Pham. Densities of nested archimedean copulas. *Journal of Multivariate Analysis*, 118:37-52, 2013.

[13] P. Jaeckel and R. Rebonato. The Most General Methodology for Creating a Valid Correlation Matrix for Risk Management and Option Pricing Purposes. *Journal of Risk*, 2:1728, 2000.

[14] K. Knight and W. Fu. Asymptotics for Lasso-Type Estimators. *Annals of statistics*, 28:1356-1378, 2000.

[15] P.L. Loh. Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Annals of Statistics*, 45:866-896, 2017.

[16] P.L. Loh and M.J. Wainwright. Regularized M-estimators with non-convexity: statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16:559-616, 2015.

[17] P.L. Loh and M.J. Wainwright. Support recovery without incoherence: a case for non-convex regularization. *Annals of Statistics*, 45:2455-2482, 2017.

[18] P.L. Loh and M.J. Wainwright. Supplement to Support recovery without incoherence: a case for nonconvex regularization. DOI:10.1214/16-AOS1530SUPP, 2017.

[19] H. Lütkepohl. *Handbook of matrices*. Wiley, 1996.

[20] A.J. McNeil. Sampling nested Archimedean copulas. *Journal of Statistical Computation and Simulation*, 78:567-581, 2008.

[21] S.N. Negahban, P. Ravikumar, M.J. Wainwright and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27:538-557, 2012.

[22] O. Okhrin, Y. Okhrin and W. Schmid. On the Structure and Estimation of Hierarchical Archimedean Copulas. *Journal of Econometrics*, 173:189-204, 2013a.

[23] O. Okhrin, Y. Okhrin and W. Schmid. Properties of Hierarchical Archimedean Copulas. *Statistics & Risk Modeling*, 30:21-54, 2013b.

[24] B. Poignard and J.-D. Fermanian. Dynamic asset correlations based on vines. *Econometric Theory*. Online. doi:10.1017/S026646661800004X, 2018.

[25] P.S. Ruzankin. On exponential inequalities for V-statistics with unbounded kernels. (Russian. English summary) *Sib. Elektron. Mat. Izv.* 11:200-206, electronic only, 2014.

[26] J. Segers and N. Uyttendaele. Nonparametric estimation of the tree structure of a nested archimedean copula. *Computational Statistics & Data Analysis*, 72:190-204, 2014.

[27] J. Shi and T. Louis. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 51:1384-1399, 1995.

[28] S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:13601392, 2009.

[29] M. Wegkamp and Y. Zhao. Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli*, 22:1184-1226, 2016.

[30] T. Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine. Learning Research*, 11:1081-1107, 2010.

[31] H. Zou and H.H. Zhang. On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, 37:1733-1751, 2009.

5 Appendix

5.1 Proof of Theorem 2.1

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

$$\langle \nabla_\theta \mathbb{G}_n(\hat{\theta}; \hat{\mathcal{U}}) - \nabla_\theta \mathbb{G}_n(\theta_0; \hat{\mathcal{U}}), \Delta \rangle \geq \alpha_2 \|\Delta\|_2 - \tau_2 \sqrt{\frac{\ln d}{n}} \|\Delta\|_1. \tag{5.1}$$

Moreover, we have

$$\langle \nabla_\theta \mathbb{G}_n(\hat{\theta}; \hat{\mathcal{U}}) + \nabla_\theta \boldsymbol{p}(\lambda_n, \hat{\theta}), \theta_0 - \hat{\theta} \rangle \geq 0. \tag{5.2}$$

The true parameter θ_0 is feasible, so that we can chose $\theta = \theta_0$ in (5.2) and using (5.1), we have

$$\langle -\nabla_\theta \boldsymbol{p}(\lambda_n, \hat{\theta}) - \nabla_\theta \mathbb{G}_n(\theta_0; \hat{\mathcal{U}}), \Delta \rangle \geq \alpha_2 \|\Delta\|_2 - \tau_2 \sqrt{\frac{\ln d}{n}} \|\Delta\|_1. \tag{5.3}$$

Then, by Hölder's inequality, we have

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

where the last inequality follows from the bound in (2.4) with $\|\nabla_\theta \mathbb{G}_n(\theta_0; \hat{\mathcal{U}})\|_\infty \leq \lambda_n/4$ and Lemma 4 of Loh and Wainwright (2015) implies $\|\nabla_\theta \boldsymbol{p}(\lambda_n, \hat{\theta})\|_\infty \leq \lambda_n$. Hence, inequality (5.3) becomes

$$\|\Delta\|_2 \leq \frac{\|\Delta\|_1}{\alpha_2} \Big(\frac{5\lambda_n}{4} + \tau_2 \sqrt{\frac{\ln d}{n}} \Big) \leq \frac{2R}{\alpha_2} \Big(\frac{5\lambda_n}{4} + \tau_2 \sqrt{\frac{\ln d}{n}} \Big).$$

Using the bounds (2.4) and the lower bound on n, the right hand-side is upper bounded by 1, which implies $\|\Delta\|_2 \leq 1$. We may then apply the (RSC) condition for the case $\|\Delta\|_2 \leq 1$, that is

$$\langle \nabla_\theta \mathbb{G}_n(\hat{\theta}; \hat{\mathcal{U}}) - \nabla_\theta \mathbb{G}_n(\theta_0; \hat{\mathcal{U}}), \Delta \rangle \geq \alpha_1 \|\Delta\|_2^2 - \tau_1 \frac{\ln d}{n} \|\Delta\|_1^2. \tag{5.4}$$

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

$$\boldsymbol{p}(\lambda_n,\theta_0)+\frac{\mu}{2}\|\theta_0\|_2^2-\boldsymbol{p}(\lambda_n,\hat{\theta})-\frac{\mu}{2}\|\hat{\theta}\|_2^2 \geq \langle\nabla_\theta\{\boldsymbol{p}(\lambda_n,\hat{\theta})+\frac{\mu}{2}\|\hat{\theta}\|_2^2\},\theta_0-\hat{\theta}\rangle = \langle\nabla_\theta\boldsymbol{p}(\lambda_n,\hat{\theta})+\mu\hat{\theta},\theta_0-\hat{\theta}\rangle,$$

which yields

$$\langle\nabla_\theta\boldsymbol{p}(\lambda_n,\hat{\theta}),\theta_0-\hat{\theta}\rangle \leq \boldsymbol{p}(\lambda_n,\theta_0)-\boldsymbol{p}(\lambda_n,\hat{\theta})+\frac{\mu}{2}\|\Delta\|_2^2. \tag{5.5}$$

Hence, using (5.4), (5.2) and (5.5), we obtain

$$\alpha_1\|\Delta\|_2^2-\tau_1\frac{\ln d}{n}\|\Delta\|_1^2 \leq -\langle\nabla_\theta\mathbb{G}_n(\theta_0,\hat{\mathcal{U}}),\Delta\rangle+\boldsymbol{p}(\lambda_n,\theta_0)-\boldsymbol{p}(\lambda_n,\tilde{\theta})+\frac{\mu}{2}\|\Delta\|_2^2.$$

By Hölder's inequality, we have

$$\begin{aligned}(\alpha_1-\frac{\mu}{2})\|\Delta\|_2^2 &\leq \boldsymbol{p}(\lambda_n,\theta_0)-\boldsymbol{p}(\lambda_n,\hat{\theta})+\|\nabla_\theta\mathbb{G}_n(\theta_0;\hat{\mathcal{U}})\|_\infty\|\Delta\|_1+\tau_1\frac{\ln d}{n}\|\Delta\|_1^2\\ &\leq \ \boldsymbol{p}(\lambda_n,\theta_0)-\boldsymbol{p}(\lambda_n,\hat{\theta})+\big(\|\nabla_\theta\mathbb{G}_n(\theta_0;\hat{\mathcal{U}})\|_\infty+4R\tau_1\frac{\ln d}{n}\big)\|\Delta\|_1. \end{aligned}\tag{5.6}$$

Moreover, by assumption, we have

$$\|\nabla_\theta\mathbb{G}_n(\theta_0;\hat{\mathcal{U}})\|_\infty+4R\tau_1\frac{\ln d}{n} \leq \frac{\lambda_n}{4}+\alpha_2\sqrt{\frac{\ln d}{n}} \leq \frac{\lambda_n}{2}.$$

Using (5.6) and Lemma 4 of Loh and Wainwright (2015), we obtain

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

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

$$0 \leq (\alpha_1-\frac{3\mu}{4})\|\Delta\|_2^2 \leq \boldsymbol{p}(\lambda_n,\theta_0)-\boldsymbol{p}(\lambda_n,\hat{\theta})+\frac{1}{2}\boldsymbol{p}(\lambda_n,\Delta). \tag{5.7}$$

Therefore, this provides

$$0 \leq (\alpha_1 - \frac{3\mu}{4})\|\Delta\|_2^2 \leq \sum_{k\in\mathcal{A}} \Big\{ p(\lambda_n, |\theta_{0,k}|) - p(\lambda_n, |\hat{\theta}_k|) \Big\} - \sum_{k\notin\mathcal{A}} p(\lambda_n, |\hat{\theta}_k|) + \frac{1}{2}\sum_k p(\lambda_n, \Delta)$$

$$\leq \quad \lambda_n \sum_{k\in\mathcal{A}} |(|\theta_{0,k}| - |\hat{\theta}_k|)| + \frac{1}{2}\Big(\sum_{k\in\mathcal{A}} p(\lambda_n, \Delta) - \sum_{k\notin\mathcal{A}} p(\lambda_n, \Delta) \Big) \tag{5.8}$$

$$\leq \quad \lambda_n \|\Delta_{\mathcal{A}}\|_1 + \frac{\lambda_n}{2}\|\Delta_{\mathcal{A}}\|_1 - 0 \leq \frac{3\lambda_n}{2}\|\Delta_{\mathcal{A}}\|_1 \leq \frac{3\lambda_n\sqrt{k_0}}{2}\|\Delta\|_2.$$

Consequently, we obtain the upper bound

$$\|\hat{\theta} - \theta_0\|_2 \leq \frac{6\lambda_n\sqrt{k_0}}{4\alpha_1 - 3\mu} \tag{5.9}$$

Concerning the upper bound of $\|\hat{\theta} - \theta_0\|_1$, note that (5.8) implies

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

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

$$\lambda_n \sum_{k\notin\mathcal{A}} |\Delta_k| \leq 3\lambda_n \|\Delta_{\mathcal{A}}\|_1 + \frac{\mu\|\Delta_{\mathcal{A}^c}\|_2^2}{2}. \tag{5.10}$$

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

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

We deduce from (5.10),

$$\|\Delta_{\mathcal{A}^c}\|_1 \leq 3\|\Delta_{\mathcal{A}}\|_1 + \frac{3\mu\sqrt{k_0}}{(4\alpha_1 - 3\mu)}\|\Delta_{\mathcal{A}}\|_2.$$

Invoking (5.9), this yields

$$\|\Delta\|_1 \leq \|\Delta_{\mathcal{A}}\|_1 + \|\Delta_{\mathcal{A}^c}\|_1 \leq 4\|\Delta_{\mathcal{A}}\|_1 + \frac{3\mu\sqrt{k_0}}{(4\alpha_1 - 3\mu)}\|\Delta\|_2$$

$$\leq \quad \Big(4 + \frac{3\mu}{(4\alpha_1 - 3\mu)}\Big)\sqrt{k_0}\|\Delta\|_2 \leq \frac{6(16\alpha_1 - 9\mu)}{(4\alpha_1 - 3\mu)^2}\lambda_n k_0,$$

proving the result. □

5.2 Proof of Theorem 2.2

Proof. By a usual Taylor expansion, we can write

$$\dot{\ell}(\theta_0, \hat{\boldsymbol{U}}_i) = \dot{\ell}(\theta_0, \boldsymbol{U}_i) + \partial_{\boldsymbol{u}}\dot{\ell}(\theta_0, \boldsymbol{U}_i).(\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i) + \frac{1}{2}\partial^2_{\boldsymbol{u}}\dot{\ell}(\theta_0, \boldsymbol{U}_i^*).(\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i)^{[2]},$$

for some random vector $\boldsymbol{U}_i^*$ s.t. $\|\boldsymbol{U}_i^* - \boldsymbol{U}_i\| < \|\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i\|$.

For every $\epsilon > 0$, we have

$$\mathbb{P}\left(\|\nabla_\theta \mathbb{G}_n(\theta_0, \hat{\mathcal{U}})\|_\infty > \epsilon\right) \leq \mathbb{P}\left(\|\nabla_\theta \mathbb{G}_n(\theta_0, \mathcal{U})\|_\infty > \epsilon/3\right)$$
$$+ \quad \mathbb{P}\left(\|\frac{1}{n}\sum_{i=1}^n \pi(\boldsymbol{U}_i)\partial_{\boldsymbol{u}}\dot{\ell}(\theta_0, \boldsymbol{U}_i).(\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i)\|_\infty > \epsilon/3\right)$$
$$+ \quad \mathbb{P}\left(\|\frac{1}{2n}\sum_{i=1}^n \pi(\boldsymbol{U}_i)\partial^2_{\boldsymbol{u}}\dot{\ell}(\theta_0, \boldsymbol{U}_i^*).(\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i)^{[2]}\|_\infty > \epsilon/3\right) =: T_1 + T_2 + T_3.$$

The first term T_1 can be bounded by invoking Bernstein's inequality (see Corollary 2.11 in Boucheron et al., 2013). Indeed, $\mathbb{E}[\pi(\boldsymbol{U}_i)\dot{\ell}(\theta_0, \boldsymbol{U}_i)] = 0$ by assumption. Under (2.6), we get

$$T_1 \leq \sum_{k=1}^d \mathbb{P}\left(\frac{1}{n}|\sum_{i=1}^n \pi(\boldsymbol{U}_i)\frac{\partial \ell}{\partial \theta_k}(\theta, \boldsymbol{U}_i)_{|\theta=\theta_0}| > \epsilon/3\right) \leq 2\sum_{k=1}^d \exp\left(-\frac{n\epsilon^2}{18(\sigma_k^2 + c_k\epsilon/3)}\right). \tag{5.11}$$

Moreover, we have

$$
\begin{aligned}
T_2 &\leq \sum_{k=1}^{d} \mathbb{P}\left(\frac{1}{n} | \sum_{i=1}^{n} \pi(\boldsymbol{U}_i) \frac{\partial^2 \ell}{\partial \boldsymbol{u} \partial \theta_k}(\theta, \boldsymbol{U}_i)_{|\theta=\theta_0} . (\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i)| > \epsilon/3 \right) \\
&\leq \sum_{k=1}^{d} \sum_{l=1}^{q} \mathbb{P}\left(\frac{1}{n} | \sum_{i=1}^{n} \pi(\boldsymbol{U}_i) \frac{\partial^2 \ell}{\partial u_l \partial \theta_k}(\theta, \boldsymbol{U}_i)_{|\theta=\theta_0} . (\hat{U}_{i,l} - U_{i,l})| > \epsilon/(3q) \right) \\
&\leq \sum_{k=1}^{d} \sum_{l=1}^{q} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} |\pi(\boldsymbol{U}_i) \frac{\partial^2 \ell}{\partial u_l \partial \theta_k}(\theta_0, \boldsymbol{U}_i)| \times \|F_{n,l} - F_l\|_\infty > \epsilon/(3q) \right) \\
&\leq \sum_{k=1}^{d} \sum_{l=1}^{q} \mathbb{P}\Big(\frac{1}{n} \sum_{i=1}^{n} |\pi(\boldsymbol{U}_i) \frac{\partial^2 \ell}{\partial u_l \partial \theta_k}(\theta_0, \boldsymbol{U}_i)| - I_{kl} > a \Big) \\
&+ \mathbb{P}\big(\|F_{n,l} - F_l\|_\infty (I_{kl} + a) > \epsilon/(3q) \big),
\end{aligned}
$$

for every constant $a > 0$. By the Dvoretsky-Kiefer-Wolfovitz (DKW) inequality, we know that, for every l and every $x > 0$, $\mathbb{P}\Big(\|F_{n,l} - F_l\|_\infty > x \Big) \leq 2\exp(-2nx^2)$. Moreover, by (2.7) and Bernstein's inequality,

$$
\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} |\pi(\boldsymbol{U}_i) \frac{\partial^2 \ell}{\partial u_l \partial \theta_k}(\theta_0, \boldsymbol{U}_i)| - I_{kl} > a \right) \leq \exp\Big(- \frac{na^2}{2(\sigma_{kl}^2 + c_{kl} a)} \Big).
$$

This yields

$$
T_2 \leq \sum_{k=1}^{d} \sum_{l=1}^{q} \left\{ \exp\Big(- \frac{na^2}{2(\sigma_{kl}^2 + c_{kl} a)} \Big) + 2\exp\big(- \frac{2n\epsilon^2}{9q^2 (I_{kl} + a)^2} \big) \right\}. \tag{5.12}
$$

Even if a formal minimization of the r.h.s. of the latter equation with respect to a would be better, we propose a rougher but more explicit upper bound, by imposing the setting $a := 2\epsilon\sigma_{kl}/(3qI_{kl})$. This is reasonable when ϵ is significantly smaller than the other constants indexed by (k, l) that we consider. Therefore, this provides

$$
T_2 \leq \sum_{k=1}^{d} \sum_{l=1}^{q} \left\{ \exp\Big(- \frac{2n\epsilon^2}{9q^2 I_{kl}^2 (1 + 2c_{kl}\epsilon/(3qI_{kl}\sigma_{kl}))} \Big) + 2\exp\big(- \frac{2n\epsilon^2}{9q^2 I_{kl}^2 (1 + 2\epsilon\sigma_{kl}/(3qI_{kl}^2))^2} \big) \right\}. \tag{5.13}
$$

Remark 7. *Sharper bounds for T_2 could be obtained by invoking some exponential inequalities for U-statistics; In our case, the kernel is unbounded in general but some results are available in the literature, notably in Theorem 1 in Borisov (1991) or Theorem B in*

Ruzankin (2014). Unfortunately, in the former paper, explicit constants do not appear, and it is not guaranteed that any positive ϵ value can be considered in the latter case. When ℓ and its derivatives are bounded on the hypercube, it is possible to invoke more standard results: see Boucheron et al. (2013) and the references therein. Therefore, for the sake of simplicity, we have preferred to invoke the usual DKW inequality. It will provide some bounds that are of the same order as those obtained for T_1.

To manage the third term T_3, we need to control the distance between the pseudo-observations and the boundaries of $[0,1]^q$, through the trimming function. As for T_2 and due to Assumption 6, we get

$$\begin{aligned}
T_3 &\leq \sum_{k=1}^{d}\sum_{l,j=1}^{q} \mathbb{P}\left(\frac{1}{n}|\sum_{i=1}^{n}\pi(\boldsymbol{U}_i)\frac{\partial^3\ell}{\partial u_l\partial u_j\partial\theta_k}(\theta_0,\boldsymbol{U}_i^*).(\hat{U}_{i,l}-U_{i,l}).(\hat{U}_{i,j}-U_{i,j})| > 2\epsilon/(3q^2)\right)\\
&\leq \sum_{k=1}^{d}\sum_{l,j=1}^{q} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\pi(\boldsymbol{U}_i)h_{klj}(\boldsymbol{U}_i)\times\|F_{n,l}-F_l\|_\infty\|F_{n,j}-F_j\|_\infty > 2\epsilon/(3q^2)\right)\\
&\leq \sum_{k=1}^{d}\sum_{l,j=1}^{q} \Big\{\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\pi(\boldsymbol{U}_i)h_{klj}(\boldsymbol{U}_i)-H_{klj}>b\right)\\
&+ \mathbb{P}\Big(\|F_{n,l}-F_l\|_\infty(H_{klj}+b)^{1/2} > \big(2\epsilon/(3q^2)\big)^{1/2}\Big)\Big\}\\
&+ \mathbb{P}\Big(\|F_{n,j}-F_j\|_\infty(H_{klj}+b)^{1/2} > \big(2\epsilon/(3q^2)\big)^{1/2}\Big)\Big\},
\end{aligned}$$

for any $b>0$. Then, by (2.9) and Bernstein's inequality,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\pi(\boldsymbol{U}_i)h_{klj}(\boldsymbol{U}_i)-H_{klj}>b\right) \leq \exp\Big(-\frac{nb^2}{2(\tau_{klj}^2+d_{klj}b)}\Big), \text{ and}$$

$$T_3 \leq \sum_{k=1}^{d}\sum_{l,j=1}^{q}\Big\{\exp\big(-\frac{nb^2}{2(\tau_{klj}^2+d_{klj}b)}\big)+4\exp\big(-\frac{4n\epsilon}{3q^2(H_{klj}+b)}\big)\Big\}. \tag{5.14}$$

By setting $b^2 := 8\tau_{klj}^2\epsilon/(3q^2H_{klj})$, we obtain

$$\begin{aligned}
T_3 \leq \sum_{k=1}^{d}\sum_{l,j=1}^{q}\Big\{&\exp\Big(-\frac{4n\epsilon}{3q^2H_{klj}(1+2d_{klj}\sqrt{2\epsilon}/(q\tau_{kl}\sqrt{3H_{klj}}))}\Big)\\
&+ 2\exp\Big(-\frac{4n\epsilon}{3q^2H_{klj}(1+2\tau_{klj}\sqrt{2\epsilon}/(qH_{klj}\sqrt{3H_{klj}}))}\Big)\Big\}.
\end{aligned} \tag{5.15}$$

Finally, the inequalities (5.11),(5.13) and (5.15) provide the result. □

5.3 Proof of Corollary 3.4.

Proof. that is a d-dimensional column vector. By obvious calculations, we obtain

$$\nabla_\theta \mathbb{G}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{(-\pi_i)}{\omega' \vec{c}_\theta(\boldsymbol{u}_i)} V_\theta(\boldsymbol{u}_i),$$

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

that is a d-dimensional column vector. To lighten notations, we will set $\mu_i(\theta) := (\omega' \vec{c}_\theta(\boldsymbol{u}_i))^{-1}$, that 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 $\boldsymbol{v} = \theta - \theta_0$, and note that, by the definition of Θ, $\|\boldsymbol{v}\| \leq 1$. Then, simple calculations provide

$$\boldsymbol{v}' \nabla^2_{\theta,\theta'} \mathbb{G}_n(\theta) \boldsymbol{v} = \frac{1}{n}\sum_{i=1}^{n} \pi_i \big(\mu_i^2 V_\theta V_\theta' - \mu_i W_\theta\big)(\boldsymbol{u}_i),$$

and the "Hessian" matrix $W_\theta(\boldsymbol{u}) = \partial_{\theta'} V_\theta(\boldsymbol{u}_i)$ 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 $[\boldsymbol{v}_0', \boldsymbol{v}_1', \ldots, \boldsymbol{v}_m']'$, so that it is conformable with the gradient vectors $V_\theta(\boldsymbol{u})$. To lighten notations, set, for every $k = 0, \ldots, m$ and every $i = 1, \ldots, n$, $\zeta_{k,i} := \boldsymbol{v}_k' \partial_{\theta_k} c_{k,\theta_k}(\boldsymbol{u}_i)$; and $\nu_{k,i} := \boldsymbol{v}_k' \partial^2_{\theta_k,\theta_k'} c_{k,\theta_k} \boldsymbol{v}_k$. Therefore, simple

calculations yield

$$\boldsymbol{v}'\nabla^2_{\theta,\theta'}\mathbb{G}_n(\theta)\boldsymbol{v} = \frac{1}{n}\sum_{i=1}^n \pi_i\mu_i^2\big(\boldsymbol{v}_0'\vec{c}_\theta(\boldsymbol{u}_i)\big)^2 + \frac{1}{n}\sum_{k,l=1}^m\sum_{i=1}^n \pi_i\mu_i^2\omega_k\omega_l\Big\{\zeta_{k,i}\zeta_{l,i} - \nu_{k,i}c_{l,\theta_l}(\boldsymbol{u}_i)\Big\}$$

$$+ \quad \frac{2}{n}\sum_{k,l=1}^m\sum_{i=1}^n \pi_i\mu_i^2\Big\{v_{0,l}\omega_k - v_{0,k}\omega_l\Big\}\zeta_{k,i}c_{l,\theta_l}(\boldsymbol{u}_i) =: T_0 + T_1 + T_2.$$

By assumption, for every $k = 1, \ldots, m$ and every $\theta \in \Theta$

$$\frac{1}{n}\sum_{i=1}^n \pi_i\mu_i^2(\theta)\Big\{\zeta_{k,i}^2 - \nu_{k,i}c_{k,\theta_k}(\boldsymbol{u}_i)\Big\} \geq \alpha_{1,k}\|\boldsymbol{v}_k\|_2^2 - \tau_{1,k}\frac{\ln d_k}{n}\|\boldsymbol{v}_k\|_1^2,$$

because $\|\boldsymbol{v}_k\|_2 \leq \|\boldsymbol{v}\|_2 \leq 1$. Therefore,

$$T_1 = \frac{1}{n}\sum_{k=1}^m\sum_{i=1}^n \pi_i\mu_i^2(\theta)\omega_k^2\Big\{\zeta_{k,i}^2 - \nu_{k,i}c_{k,\theta_k}(\boldsymbol{u}_i)\Big\} + \frac{1}{n}\sum_{k,l=1,k\neq l}^m\sum_{i=1}^n \pi_i\mu_i^2(\theta)\omega_k\omega_l\zeta_{k,i}\zeta_{l,i}$$

$$=: \quad T_1' + T_1'', \text{ where}$$

$$T_1' \geq \sum_{k=1}^m \omega_k^2\alpha_{1k}\|\boldsymbol{v}_k\|_2^2 - \sum_{k=1}^m \omega_k^2\tau_{1k}\frac{\ln d_k}{n}\|\boldsymbol{v}_k\|_1^2 \geq \inf_k\big(\omega_k^2\alpha_{1k}\sum_{k=1}^m\big)\|\boldsymbol{v}_k\|_2^2 - \sup_k\big(\omega_k^2\tau_{1k}\frac{\ln d_k}{n}\big)\|\boldsymbol{v}\|_1^2,$$

because $\sum_k \|\boldsymbol{v}_k\|_1^2 \leq \|\boldsymbol{v}\|_1^2$. Moreover, for every i, we have $\mu_i|\sum_k \omega_k\zeta_{k,i}| \leq \psi_i\|\boldsymbol{v}\|_1$ and

$$T_1'' \leq \frac{1}{n}\sum_{i=1}^n \pi_i\mu_i^2(\theta)\Big(\sum_{k=1}^n \omega_k\zeta_{k,i}\Big)^2 \leq \frac{1}{n}\sum_{i=1}^n \pi_i\psi_i^2\|\boldsymbol{v}\|_1^2, \quad \text{implying}$$

$$T_1 \geq \inf_k\big(\underline{\omega}^2\alpha_{1k}\big)\|\sum_{k=1}^m \boldsymbol{v}_k\|_2^2 - \sup_k\big(\overline{\omega_k}^2\tau_{1k}\frac{\ln d_k}{n}\big)\|\boldsymbol{v}\|_1^2 - \frac{1}{n}\sum_{i=1}^n \pi_i\psi_i^2\|\boldsymbol{v}\|_1^2.$$

Concerning T_2, we have

$$|T_2| \leq \frac{2}{n}\sum_{i=1}^n \pi_i|\sum_{k=1}^m \mu_i\omega_k\zeta_{k,i}| \times \big|\mu_i(\theta)\boldsymbol{v}_0'\vec{c}_\theta(\boldsymbol{u}_i)\big| + \frac{2}{n}\sum_{k=1}^m \frac{|v_{0,k}|}{\omega_k} \times \Big|\omega_k\sum_{i=1}^n \pi_i\mu_i(\theta)\zeta_{k,i}\Big|$$

$$\leq \quad \frac{2\|\boldsymbol{v}\|_1}{n}\sum_{i=1}^n \pi_i\psi_i\|\boldsymbol{v}_0\|_1 \max_k|\mu_i(\theta)c_{k,\theta_k}(\boldsymbol{u}_i)| + \frac{\lambda_n}{2}\sum_{k=1}^m \|\boldsymbol{v}_k\|_1\frac{\|\boldsymbol{v}_0\|_\infty}{\underline{\omega}_k}$$

$$\leq \quad \|\boldsymbol{v}\|_1^2\bigg(\frac{2}{n}\sum_{i=1}^n \pi_i\psi_i\sup_{\theta\in\Theta}\Big\{\frac{\max_k c_{k,\theta_k}(\boldsymbol{u}_i)}{\sum_{l=1}^m \underline{\omega}_l c_{l,\theta_l}(\boldsymbol{u}_i)}\Big\} + \frac{\lambda_n}{2\min_k\underline{\omega}_k}\bigg).$$

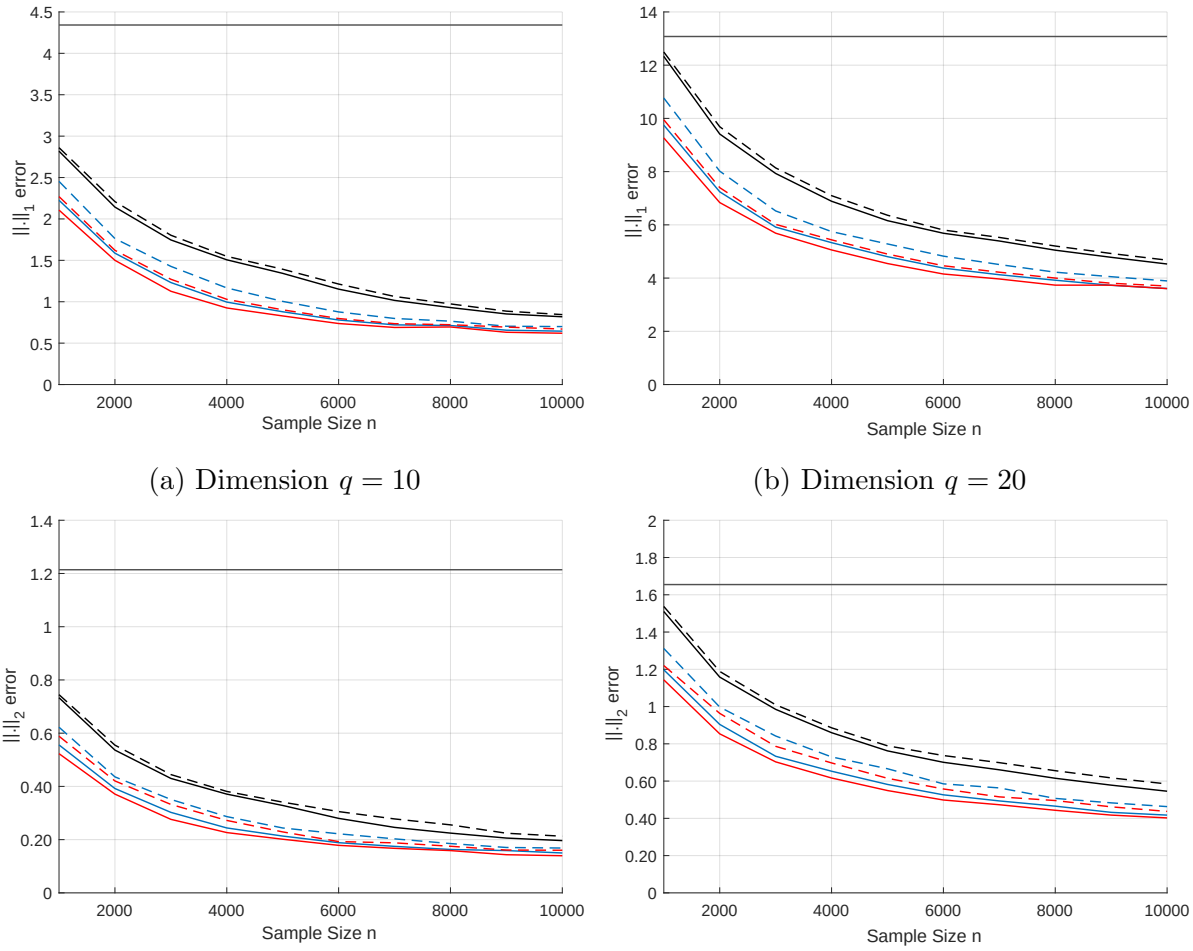
Finally, we manage T_0 as in the proof of Corollary 3.3:

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

proving the result. □

Remark 8. *Note that, in the definition of Θ, we have imposed (mainly for convenience) that $\|\theta-\theta_0\|_2 \leq 1$ for every $\theta \in \Theta$. Therefore, $\|\boldsymbol{v}\|_2 \leq 1$ and the constants α_2 and τ_2 can be arbitrarily chosen. The constraint $\|\theta-\theta_0\|_2 \leq 1$ could be removed at the price of painful additional technicalities. Indeed, it would then be necessary to distinguish the cases $\|\boldsymbol{v}_k\|_2$ and $\|\boldsymbol{v}_k\|_1$ are smaller/larger than one, when $\|\boldsymbol{v}\|_2 > 1$. Such extensions are left for the interested reader.*

5.4 Figures



(a) Dimension $q = 10$

(b) Dimension $q = 20$

(c) Dimension $q = 10$

(d) Dimension $q = 20$

Figure 1: Statistical consistency in the $\|.\|_1$ (panels (a) and (b)) and $\|.\|_2$ (panels (c) and (d)) sense of the sparse Gaussian copula correlation estimator. SCAD, MCP and Lasso results are represented in blue, red and black respectively. The case $\boldsymbol{U}$ (resp. $\hat{\boldsymbol{U}}$) is represented in solid (resp. dashed) line. Each point represents an average of 200 trials and the x-axis represents the sample size n. For each dimension, $\|\theta_0\|_1$ and $\|\theta_0\|_2$ is represented by the gray solid line.