Electronic Journal of Statistics
Vol. 14 (2020) 3315–3365
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1745

Statistical analysis of sparse approximate factor models

Benjamin Poignard[*] and **Yoshikazu Terada**[†]

Osaka University, Graduate School of Economics,
1-7, Machikaneyama, Toyonaka, 560-0043, Japan
Osaka University, Graduate School of Engineering Science,
1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan
e-mail: bpoignard@econ.osaka-u.ac.jp; terada@sigmath.es.osaka-u.ac.jp

Abstract: We consider the problem of estimating sparse approximate factor models. In a first step, we jointly estimate the factor loading parameters and the error - or idiosyncratic - covariance matrix based on the Gaussian quasi-maximum likelihood method. Conditionally on these first step estimators, using the SCAD, MCP and Lasso regularisers, we obtain a sparse error covariance matrix based on a Gaussian QML and, as an alternative criterion, a least squares loss function. Under suitable regularity conditions, we derive error bounds for the regularised idiosyncratic factor model matrix for both Gaussian QML and least squares losses. Moreover, we establish the support recovery property, including the case when the regulariser is non-convex. These theoretical results are supported by empirical studies.

MSC2020 subject classifications: Primary 62H25, 62F99.
Keywords and phrases: Approximate factor analysis, non-convex regulariser, statistical consistency, support recovery.

Received March 2019.

Contents

[*] Jointly affiliated at High-Dimensional Statistical Modeling Team, RIKEN Center for Advanced Intelligence Project (AIP), 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan.

[†] Jointly affiliated at Mathematical Statistics Team, RIKEN Center for Advanced Intelligence Project (AIP), 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan.

1. Introduction

The need of a joint modelling for high-dimensional random vectors has fostered a flourishing research in sparse models. The application domains of sparse modelling 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 both statistically precise and provides intuitive insights among asset relationships: as an example, this gave rise to sparse matrix precision methods - see the review of [13] - or variance covariance estimation based on factor models - see [10]. Factor modelling aims at summarizing the information from large data sets through a small number of variables called factors. For example, [28] developed an arbitrage pricing theory built on a multiple factor model for asset returns to gain statistical effectiveness when estimating the co-movements and common shocks from a large portfolio size.

In factor models, the quantity of interest is the variance covariance matrix Σ of the vector of observations, which is decomposed as a sum of the quadratic product of the factor loading matrix Λ and the covariance matrix of the idiosyncratic errors Ψ. In the standard factor analysis setting in which the dimension p - the number of variables composing the vector of observations - is fixed, it is commonly assumed that the idiosyncratic covariance matrix Ψ is diagonal. Under such structure assumption for Ψ, [2] derived the large sample properties of the likelihood-based factor model estimators. [4] extended these asymptotic results for a potentially diverging p.

[9] developed the notion of approximate factor models, where the matrix Ψ is not required to be diagonal, which enables the idiosyncratic errors to be cross-sectionally correlated. Under the assumption of bounded eigenvalues for Ψ, [5] studied the large sample properties of the likelihood-based factor model for non-diagonal Ψ and diverging p. However, if we consider the estimation of all elements of Ψ, the number of parameters exceeds the number of estimating equations. Thus, the estimation of the diagonal elements of Ψ only is considered in [5]. As the first study that explicitly modelled sparsity on Ψ and thus relaxed the diagonal assumption within the likelihood based setting, [6] considered a conditionally sparse factor framework, in the sense that Ψ is a sparse matrix with bounded eigenvalues. Using a Gaussian quasi-maximum likelihood approach, [6] proposed a joint estimation of Λ and Ψ while penalizing Ψ through the adaptive Lasso and SCAD methods, the so-called Penalised Maximum Likelihood (PML) method. They derived asymptotic consistency results under the rate $\log(p) =$

$o(n)$ and a convergence rate for the sparse estimator of Ψ, which is not minimax optimal.

In this study, we consider the following problem: given n observations of a p dimensional random vector X_i, conditionally on a suitable first step estimator of the loading factor matrix Λ, estimate a sparse Ψ. Our main contributions are as follows: first, we provide error bounds for the sparse estimator of Ψ in the ℓ_1, ℓ_2 and ℓ_∞ senses for specific scaling behaviours of (n, p, k_0), where k_0 is the cardinality of the true unknown sparse support; second, we provide the conditions to satisfy the support recovery property. Assuming a non-diagonal and sparse Ψ is intuitively meaningful. Indeed, for a portfolio composed with stocks, when the idiosyncratic components represent stocks' individual shocks, they are non-correlated or weakly correlated among the stocks across different industries/countries/and the like, since the industries/countries/and the like specific components are not necessarily pervasive for the whole portfolio. This feature also holds for a portfolio composed with different types of assets (stocks, bonds, commodities, and the like). Our approach is based on a two-step estimation, which allows for proper regularity conditions. In a first step, both the loading factor matrix Λ and idiosyncratic covariance matrix Ψ are obtained based on a Gaussian quasi-maximum likelihood estimator, whose theoretical properties were derived by [4, 5]. Importantly at this stage, Ψ is assumed diagonal, which is a required assumption to obtain consistent estimators: see [5]. Conditionally on these first step estimators, we then relax the diagonal assumption on Ψ so that the regularised idiosyncratic matrix corresponds to the solution of a penalised M-estimator minimized with respect to Ψ only. In this second step, we consider the Gaussian quasi-maximum likelihood (QML) and, as an alternative, the least squares loss function, which was proposed by [18]. To the best of our knowledge, this paper is the first attempt to link general penalised - potentially non-convex - M-estimators and the likelihood-based inference for conditionally sparse factor model. Our study shares a similar spirit to that of [6], who derived asymptotic rates for consistency of the penalised estimator of Ψ. But our work differs from theirs in two main respects: we provide bounds on ℓ_1-, ℓ_2- and ℓ_∞-errors and the conditions to satisfy the support recovery property, this for two different losses - the Gaussian QML and the least squares losses.

The framework we use to derive such error bounds is closely related to the studies of [21, 22], which covers a broad range of non-convex objective functions for sparse estimation. Under the assumption of restricted strong convexity (see e.g. [24]) of the unpenalised loss function and suitable regularity conditions on the penalty, they derive some error bounds for the penalised estimators and provide conditions for variable selection consistency. In our study, we extend their results to the sparse factor model analysis and check the conditions for which any local minimum lies within statistical precision of the true sparse parameter for both Gaussian QML and least squares based loss functions. Our main contribution is to quantify the statistical accuracy of the sparse approximate covariance estimator by deriving error bounds between the penalised estimator and the true parameter, and discuss the relevance of these theoretical bounds for each M-criterion. Besides, following [22, 26] we provide sufficient conditions for the sparse estimator to satisfy the support recovery property. Within such penalised M-estimator setting, the scaling behaviours with respect to (n, p, k_0) that we derive for support recovery highly depend on the regularity of the loss function and the effect of the two-step estimation.

The remainder of the paper is organized as follows. In Section 2, we describe the approximate factor model framework and the penalised statistical criteria. In Section 3, we derive the error bounds in the ℓ_1, ℓ_2 senses. In Section 4, we provide the conditions for which the support recovery property is satisfied. In Section 5, we discuss specific applications of the proposed penalised framework. Section 6 illustrates these theoretical results through simulated and real data experiments. All intermediary results and proofs are contained in Appendix A.

Notations. Throughout this paper, we denote the cardinality of a set E by $|E|$. For a vector $v \in \mathbb{R}^d$, the ℓ_p norm is $\|v\|_p = \big(\sum_{k=1}^p |v_k|^p\big)^{1/p}$ for $p > 0$, and $\|v\|_\infty = \max_i |v_i|$. Let the subset $\mathcal{A} \subseteq \{1, \cdots, d\}$, then $v_\mathcal{A} \in \mathbb{R}^{|\mathcal{A}|}$ is the vector v restricted to $\mathcal{A}$. For a matrix A, $\|A\|_s$, $\|A\|_\infty$ and $\|A\|_F$ are the spectral, infinity and Frobenius norms, respectively, and $\|A\|_{\max} = \max_{ij} |A_{i,j}|$ is the coordinate-wise maximum (in absolute value). We write A' (resp. v') to denote the transpose of the matrix A (resp. the vector v). We write $\text{vec}(A)$ to denote the vectorization operator that stacks the columns of A on top of one another into a vector. We denote by $A \succ 0$ (resp. $A \succeq 0$) the positive definiteness (resp. semi-definiteness) of A and $\text{vech}(A)$ the $p(p+1)/2$ vector that stacks the columns of the lower triangular part of A. For a function $f : \mathbb{R}^d \to \mathbb{R}$, we denote by ∇f the gradient or subgradient of f and $\nabla^2 f$ the Hessian of f. We denote by $(\nabla^2 f)_{\mathcal{A}\mathcal{A}}$ the Hessian of f restricted to the block $\mathcal{A}$. We write $\mathcal{A}^c$ to denote the complement of the set $\mathcal{A}$. The expression *with high probability* refers to event occurring with probability approaching one when (n, p, k_0) tend to infinity. The scaling results for (n, p, k_0) are expressed as $f(n) \geq M g(k_0, p)$ for $0 < M < \infty$ some universal constant and continuous functions $f(.), g(.)$.

2. Framework

We consider a sequence of n observations of a p-dimensional random vector (X_i), assumed to be independent and identically distributed and following the factor structure

$$X_i = \Lambda F_i + \epsilon_i, \tag{2.1}$$

where $\Lambda \in \mathcal{M}_{p\times m}(\mathbb{R})$ is the loading matrix, F_i is the $\mathbb{R}^m$ vector of factor variables and ϵ_i the $\mathbb{R}^p$ vector of errors - or idiosyncratic variables. In (2.1), the vector X_i is observable; none of the right-hand side variables are observable and the dimension $m > 0$ is known. We assume $\mathbb{E}[F_i] = \mathbf{0} \in \mathbb{R}^m$, $\mathbb{E}[F_i F_i'] = I_m$ the identity matrix of size m, $\mathbb{E}[F_i \epsilon_i'] = \mathbf{0} \in \mathcal{M}_{m\times p}(\mathbb{R})$ and $\mathbb{E}[\epsilon_i \epsilon_i'] = \Psi \in \mathcal{M}_{p\times p}(\mathbb{R})$ non-diagonal. The idiosyncratic components (ϵ_i) are assumed to be correlated following the setting of [6, 9]. Based on these assumptions, $\mathbb{E}[X_i] = \mathbf{0}$, $\text{Var}(X_i) = \Lambda\Lambda' + \Psi$. The object of interest is to recover the matrix quantity $\Sigma : \mathcal{M}_{p\times m}(\mathbb{R}) \times \mathcal{M}_{p\times p}(\mathbb{R})$ defined as $\Sigma(\Lambda, \Psi) = \Lambda\Lambda' + \Psi$. We assume that the factors and idiosyncratic variables are uniformly sub-Gaussian.

Assumption 1. *The $(X_{ij})_{1\leq i\leq n, 1\leq j\leq p}$ are uniformly sub-Gaussian random variables, that is, the common factors $(F_{ik})_{1\leq i\leq n, 1\leq k\leq m}$ and the unique factors $(\epsilon_{ij})_{1\leq i\leq n, 1\leq j\leq p}$ are uniformly sub-Gaussian. More formally, we assume that $\exists K \in \mathbb{R}$; $\exists \sigma_0^2 > 0$; $\forall i \in \{1,\dots,n\}$; $\forall j \in \{1,\dots,p\}$ such that*

$$K^2\left(\mathbb{E}[\exp(|X_{ij}|^2/K^2)] - 1\right) \leq \sigma_0^2.$$

Remark. The i.i.d. and sub-Gaussian assumptions can be relaxed in favour of dependent data such as time series with mixing conditions. Whether the i.i.d. setting is assumed or not, the theoretical error bounds we propose to derive in Section 3 together with the scaling behaviours would not be altered. However, it is worth noting that the exponential bound we use to evaluate the probability to satisfy the former error bounds would be different, should we consider, e.g., strongly mixing time series.

The factor model studies that considered a sparse estimator $\hat{\Psi}$ provided asymptotic probability bounds only. Based on the principal component method, [14] considered a two-step estimation, where the loading factor matrix Λ is estimated through an OLS procedure in a first step; in a second step, they obtained a sparse estimator $\hat{\Psi}$ using an adaptive threshold technique. [15] considered the POET estimator, where both $\Lambda\mathbb{E}[F_t F_t']\Lambda'$ and Ψ are decomposed based on a principal component approach and the eigenvalues of the PCA-based representation of Ψ are penalised. Using the likelihood-based inference approach, [6] considered a Gaussian QMLE for the joint estimation of (Λ, Ψ) and provided asymptotic rates for consistency, where Ψ only is regularised by the adaptive Lasso and SCAD methods.

The main contribution of our study is to provide error bounds for the regularised $\hat{\Psi}$ and the conditions to satisfy the support recovery property. We quantify the statistical accuracy of the latter based on two different unpenalised loss functions and for potentially non-convex regularisers. In the rest of the paper, we denote (Λ_0, Ψ_0) the true parameters and $\Sigma_0 := \Sigma(\Lambda_0, \Psi_0) = \Lambda_0\Lambda_0' + \Psi_0$.

Assumption 2. *The true parameter $\theta_{\Psi_0} = vech(\Psi_0)$ is assumed to be sparse so that $k_0 = card(\mathcal{A})$, where $\mathcal{A} = \{1 \leq i \leq p(p+1)/2 : \theta_{i,\Psi_0} := vech(\Psi_0)_i \neq 0\}$ with $k_0 < p(p+1)/2$ the total number of parameters.*

Under the sparsity assumption, we aim at recovering the true sparse support $\mathcal{A}$. To do so, we consider a regularisation procedure on the variance-covariance Ψ so that the problem of interest is

$$\left\{\begin{array}{ccl} \hat{\Psi}^g & = & \underset{\Psi\in\Omega}{\arg\min}\left\{\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi) + p(\gamma_n, \theta_\Psi)\right\}, \text{ where} \\ (\tilde{\Lambda}, \tilde{\Psi}) & = & \underset{(\Lambda,\Psi)\in\Theta}{\arg\min}\left\{\mathbb{G}_{n,p}(\Lambda;\Psi)\right\}, \text{ with} \\ \mathbb{G}_{n,p}(\Lambda;\Psi) & = & \frac{1}{2p}\big(\log(|\Sigma(\Lambda,\Psi)|) + \text{tr}(\hat{S}\Sigma(\Lambda,\Psi)^{-1})\big), \end{array}\right. \tag{2.2}$$

where $\hat{S}$ is the sample variance covariance estimator, $\theta_\Psi = \text{vech}(\Psi)$, and $(\tilde{\Lambda}, \tilde{\Psi})$ are first step estimators obtained by estimating without regularisation the factor model based on the Gaussian QML function $\mathbb{G}_{n,p}(.;.)$ for $\Lambda, \Psi \in \Theta$ - Θ

will be made explicit later -. In the second step, the regularisation procedure is performed by the regulariser $p(\gamma_n, .) : \mathbb{R}^{p(p+1)/2} \rightarrow \mathbb{R}$, where γ_n is the regularisation parameter, which depends on the sample size, and enforce a particular type of sparse structure in the solution $\hat{\Psi}^g$. Ω denotes a $p \times p$-variance covariance matrices subset defined as

$$\begin{aligned}\Omega \quad = \quad & \Big\{\Psi : \Sigma := \Sigma(\tilde{\Lambda}, \Psi) = \tilde{\Lambda}\tilde{\Lambda}' + \Psi, \;\; \Psi = \Psi', \;\; \Psi \succ 0, \\ & b_1 < \lambda_{\min}(\Psi) < \lambda_{\max}(\Psi) < b_2, \;\; a < \lambda_{\min}(2\hat{S} - \Sigma), \;\; g(\theta_\Psi) \leq R\Big\},\end{aligned}$$

for any fixed matrix $\tilde{\Lambda}$ and for some positive constants b_1, b_2, a and R. Due to the potential non-convexity of this penalty, we include the side condition $g(\theta_\Psi) \geq \|\theta_\Psi\|_1$ with $g : \mathbb{R}^{p(p+1)/2} \rightarrow \mathbb{R}$ a convex function, typically $g(\theta_\Psi) = \sum_{i \leq j}^p |\Psi_{ij}|$, and R a supplementary regularisation parameter to ensure the existence of local/global optima. We also impose $g(\theta_{\Psi_0}) \leq R$. Note that Ω is convex: if $\lambda_{\min}(2\hat{S} - \tilde{\Lambda}\tilde{\Lambda}' - \Psi_k) > a$, with $k = 1, 2$, then

$$\begin{aligned} & \lambda_{\min}(2\hat{S} - \tilde{\Lambda}\tilde{\Lambda}' - (t\Psi_1 + (1-t)\Psi_2) \\ \geq \quad & t\lambda_{\min}(2\hat{S} - \tilde{\Lambda}\tilde{\Lambda}' - \Psi_1) + (1-t)\lambda_{\min}(2\hat{S} - \tilde{\Lambda}\tilde{\Lambda}' - \Psi_2) > a.\end{aligned}$$

As an alternative, we consider the regularised estimator $\hat{\Psi}^{ls}$ based on the least squares type contrast $\mathbb{F}_{n,p}(.)$ defined as

$$\left\{\begin{array}{ccl} \hat{\Psi}^{ls} & = & \underset{\Psi \in \bar{\Omega}}{\arg\min} \Big\{\mathbb{F}_{n,p}(\tilde{\Lambda}; \Psi) + p(\gamma_n, \theta_\Psi)\Big\}, \text{ where} \\ \mathbb{F}_{n,p}(\tilde{\Lambda}; \Psi) & = & \frac{1}{2p}\|\hat{\Sigma} - \tilde{\Lambda}\tilde{\Lambda}' - \Psi\|_F^2, \text{ and} \\ (\tilde{\Lambda}, \tilde{\Psi}) & = & \underset{(\Lambda,\Psi) \in \Theta}{\arg\min} \Big\{\mathbb{G}_{n,p}(\Lambda; \Psi)\Big\}, \text{ with} \\ \mathbb{G}_{n,p}(\Lambda; \Psi) & = & \frac{1}{2p}\big(\log(|\Sigma(\Lambda, \Psi)|) + \text{tr}(\hat{S}\Sigma(\Lambda, \Psi)^{-1})\big), \end{array}\right. \tag{2.3}$$

where the parameter set $\bar{\Omega}$ is defined as

$$\begin{aligned} & \bar{\Omega} = \Big\{\Psi : \Sigma := \Sigma(\tilde{\Lambda}, \Psi) = \tilde{\Lambda}\tilde{\Lambda}' + \Psi, \Psi = \Psi', \\ & \Psi \succ 0, l_1 < \lambda_{\min}(\Psi) < \lambda_{\max}(\Psi) < l_2, g(\theta_\Psi) \leq R\Big\},\end{aligned}$$

for some positive constants l_1, l_2, R. The notation $\hat{\Psi}^g$ (resp. $\hat{\Psi}^{ls}$) refers to the Gaussian (resp. least squares) based two-step estimator satisfying (2.2) (resp. (2.3)). The non-penalised population level parameters correspond to $\Psi_0^g = \arg\min \mathbb{E}[\mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi)]$ and $\Psi_0^{ls} = \arg\min \mathbb{E}[\mathbb{F}_{n,p}(\tilde{\Lambda}; \Psi)]$, where both are assumed to be unique, so that $\Psi_0^g = \Psi_0^{ls} = \hat{S} - \tilde{\Lambda}\tilde{\Lambda}'$ and both asymptotically converge to the true parameter Ψ_0 (see [4]). Note that in both regularised problems, $\tilde{\Psi}$ must be estimated jointly with $\tilde{\Lambda}$ in the first step, though it does not enter the second-step objective.

There actually exist alternative presentations of the sparse approximate factor model, where only the off-diagonal entries of Ψ are penalised. Similar results for statistical consistency actually hold in this case. In our framework, we assume that all components are equally penalised to clarify our arguments. Similar settings in the context of the graphical Lasso were considered: e.g. [21] or [11].

Furthermore, in many existing studies, the focus is on the sparse estimation of the covariance matrix. As indicated in [15], the sparsity assumption directly on the covariance matrix Σ, in which many observed variables are uncorrelated, is not appropriate in empirical situations since it is natural to think that several common factors exist for the underlying structure of the observed variables. Factor analysis stands as the natural method to appropriately deal with the common factors. In standard factor analysis, the idiosyncratic components are assumed uncorrelated, that is, the idiosyncratic covariance matrix is diagonal, which corresponds to the so-called strict factor model. However, this diagonal assumption is too restrictive in practice: see [6, 15]. Instead of this restrictive assumption, we assume the sparsity of the idiosyncratic covariance, which allows for the existence of correlation among the idiosyncratic components. Indeed, factor models are often treated as approximate, where the observations (X_i) are correlated given the factors: this is the object of the approximate factor model of [9].

Finally, in the standard factor analysis, the Gaussian QML estimator is the most commonly used approach since the corresponding estimator is more preferable than the least squares estimator from the viewpoint of efficiency under the typical Gaussian assumption (e.g., see [2] and [19]). Our analysis will allow to compare both estimators from theoretical and empirical viewpoints.

The framework we use in Section 3 requires specific regularity conditions on the non-penalised loss function, namely the restricted strong convexity. Our two-step estimation allows for such property with respect to the parameter Ψ, conditionally on the first step estimators. In both (2.2) and (2.3) statistical criteria, the first step estimators $(\tilde{\Lambda}, \tilde{\Psi})$ are defined as

$$\begin{cases} (\tilde{\Lambda}, \tilde{\Psi}) & = \underset{(\Lambda,\Psi)\in\Theta}{\arg\min} \Big\{ \mathbb{G}_{n,p}(\Lambda; \Psi) \Big\}, \text{ with} \\ \mathbb{G}_{n,p}(\Lambda; \Psi) & = \frac{1}{2p}\big(\log(|\Sigma(\Lambda, \Psi)|) + \operatorname{tr}(\hat{S}\Sigma(\Lambda, \Psi)^{-1}) \big), \end{cases} \tag{2.4}$$

where Θ is the parameter space. Importantly, specific conditions are required for identifiability. Indeed, for any orthonormal matrix $A \in O(m \times m)$ (i.e., $A'A = AA' = I_m$), the rotated estimator $\bar{\Lambda} = \tilde{\Lambda}A$ is also the minimizer of the above contrast function. To avoid this rotational indeterminacy, we use the following identifiability condition IC5 in [4], assuming $\mathbb{E}[F_i F_i'] = I_m$:

$$\Lambda = (\Lambda_1', \Lambda_2')' \text{ with } \Lambda_1 = \begin{pmatrix} \lambda_{11} & 0 & \cdots & 0 \\ \lambda_{21} & \lambda_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{m1} & \lambda_{m2} & \cdots & \lambda_{mm} \end{pmatrix} \text{ and } \lambda_{ii} \neq 0 \; (i = 1, \ldots, r).$$

Our framework can also accommodate alternative restrictions for identifiability. For instance, [6] use the constraint $\mathbb{E}[F_i F_i'] = I_m$ and $\Lambda'\Psi^{-1}\Lambda$ diagonal in their joint estimation of Λ, Ψ, a restriction corresponding to IC3' in Table 1 of [4] for Ψ diagonal.

In the first step, we consider a diagonal constrained estimation for Ψ whereas we assume that Ψ is not a diagonal but a sparse matrix. Without the diagonal constraint, we cannot construct any reasonable estimator for Λ. In fact, for any $\Lambda \in \mathcal{M}_{p\times m}(\mathbb{R})$, we set $\tilde{\Psi}(\Lambda) = \hat{\Sigma} - \Lambda\Lambda'$ and then any discrepancy function including $\mathbb{G}_{n,p}(\Lambda;\Psi)$ attains the minimum value at $(\Lambda, \tilde{\Psi}(\Lambda))$. From the results in [5], we deduce that the estimator $\tilde{\Lambda}$ with the diagonal constraint provides reasonable estimators. We will describe more precisely this fact later in the paper. Moreover, the loading factor matrix Λ is assumed to satisfy the pervasiveness condition, which is a technical but standard assumption in the factor analysis literature: see for e.g. [6, 15].

Assumption 3. *$\exists\delta$, $0 < \delta < \infty$, so that $\delta^{-1} < \lambda_{\min}(\frac{1}{p}\Lambda_0'\Lambda_0) \leq \lambda_{\max}(\frac{1}{p}\Lambda_0'\Lambda_0) < \delta$.*

Thus, in view of assumption 3, the parameter space Θ for the initial estimator $(\tilde{\Lambda}, \tilde{\Psi})$ in problem (2.4) is defined as

$$\Theta = \Big\{(\Lambda,\Psi) : \Lambda \text{ satisfies IC5}, \; \delta^{-1} < p^{-1}\lambda_{\min}(\Lambda'\Lambda) < p^{-1}\lambda_{\max}(\Lambda'\Lambda) < \delta, \\ \text{and } \Psi \text{ is diagonal with positive diagonal components}\Big\}.$$

Since our framework compels the sample size n to be large with respect to the dimension p, we use the probability bounds derived by [5] to control this first step estimator of the loading factor. In Proposition 1 of [5], these first step estimators admit the probability bounds

$$\frac{1}{p}\sum_{k=1}^{p}\frac{1}{\tilde{\Psi}_{kk}}\|\tilde{\lambda}_k - \lambda_{0,k}\|_2^2 = O_p\left(\frac{1}{n}\right) + O_p\left(\frac{1}{p^2}\right),$$
$$\frac{1}{p}\sum_{k=1}^{p}(\tilde{\Psi}_{kk} - \Psi_{0,kk})^2 = O_p\left(\frac{1}{n}\right) + O_p\left(\frac{1}{p^2}\right),$$

where $\forall k, \|\lambda_{0,k}\|_2 \leq C$ and $C^{-2} \leq \tilde{\Psi}_{kk} \leq C^2$ with C a sufficiently large constant, $\lambda_k = (\lambda_{k1},\cdots,\lambda_{km})'$. Based on these results, the convergence rate of the loading factor matrix becomes

$$\|\tilde{\Lambda} - \Lambda_0\|_F = O_p\left(\sqrt{\frac{p}{n}}\right) + O_p\left(\sqrt{\frac{1}{p}}\right).$$

Hence based on the identifiability condition IC5 for Λ of [4], we assume the following convergence rate of the first step estimator $\tilde{\Lambda}$.

Assumption 4. *If $\forall k, \|\lambda_{0,k}\|_2 \leq C$ and $C^{-2} \leq \tilde{\Psi}_{kk} \leq C^2$ with C a sufficiently large constant, then the first step estimator $\tilde{\Lambda}$ satisfies the convergence rate*

$$\|\tilde{\Lambda} - \Lambda_0\|_F = O_p\left(\sqrt{\frac{p}{n}}\right) + O_p\left(\sqrt{\frac{1}{p}}\right).$$

Finally, our framework relies on the following regularity conditions on the penalty function.

Assumption 5. *We consider penalty functions that are assumed to be amenable regularisers defined as follows. We denote $p(.,.) : \mathbb{R}_+ \times \mathbb{R}^d$ the penalty function - or regulariser -, which is assumed to be coordinate-separable with respect to $\theta \in \mathbb{R}^d$, idest*

$$p(\gamma_n, \theta) = \sum_{k=1}^{d} p(\gamma_n, \theta_i).$$

Furthermore, let $\mu \geq 0$, and $p(\gamma_n, .)$ is μ-amenable if

(i) $\rho \mapsto p(\gamma_n, \rho)$ *is symmetric around zero and* $p(\gamma_n, 0) = 0$.
(ii) $\rho \mapsto p(\gamma_n, \rho)$ *is non-decreasing on* $\mathbb{R}^+$.
(iii) $\rho \mapsto \frac{p(\gamma_n,\rho)}{\rho}$ *is non-increasing on* $\mathbb{R}^+_\star$.
(iv) $\rho \mapsto p(\gamma_n, \rho)$ *is differentiable for any* $\rho \neq 0$.
(v) $\lim_{\rho \to 0^+} \partial_\rho p(\gamma_n, \rho) = \gamma_n$.
(vi) $\rho \mapsto p(\gamma_n, \rho) + \frac{\mu}{2}\rho^2$ *is convex for some* $\mu \geq 0$.
The regulariser $p(\gamma_n, .)$ *is* (μ, ζ)*-amenable if in addition*
(vii) *There exists* $\zeta \in (0, \infty)$ *such that* $\partial_\rho p(\gamma_n, \rho) = 0$ *for* $\rho \geq \gamma_n \zeta$.

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

Assumption 2 implies that the true support (unknown) is sparse, that is the matrix Ψ_0 contains zero components. The regularisation - or penalisation - procedure provides an estimator of $\mathcal{A}$ by discarding the covariances among the idiosyncratic components. Assumption 3, also assumed in [15, 21], is standard in factor analysis. It requires the factors to impact a non-vanishing proportion of the observations $X_{1,1}, \cdots, X_{n,p}$. Assumption 4 allows us to control for the first step estimator $\tilde{\Lambda}$. To derive our theoretical properties, assumption 5 provides regularity conditions that potentially encompass non-convex functions. These regularity conditions are the same than [20, 21, 22]. In this paper, we focus on the Lasso, the SCAD due to [12] and the MCP due to [31], given by

$$\textbf{Lasso}: \; p(\gamma_n, \rho) = \gamma_n |\rho|,$$

$$\textbf{MCP}: \; p(\gamma_n, \rho) = \text{sign}(\rho)\gamma_n \int_0^{|\rho|} (1 - z/(\gamma_n b_{mcp}))_+ dz,$$

$$\textbf{SCAD}: \; p(\gamma_n, \rho) = \begin{cases} \gamma_n |\rho|, \\ \quad \text{for } |\rho| \leq \gamma_n, \\ -\frac{1}{(2(b_{scad}-1))}(\rho^2 - 2b_{scad}\gamma_n|\rho| + \gamma_n^2), \\ \quad \text{for } \gamma_n \leq |\rho| \leq b_{scad}\gamma_n, \\ (b_{scad}+1)\gamma_n^2/2, \\ \quad \text{for } |\rho| > b_{scad}\gamma_n, \end{cases}$$

where $b_{scad} > 2$ and $b_{mcp} > 0$ are fixed parameters for the SCAD and MCP respectively. The Lasso is a μ-amenable regulariser, whereas the SCAD and the MCP regularisers are (μ, ζ)-amenable. More precisely, $\mu = 0$ (resp. $\mu = 1/(b_{scad} - 1)$, resp. $\mu = 1/b_{mcp}$) for the Lasso (resp. SCAD, resp. MCP). The parameter μ can be interpreted as a coefficient of non-convexity level: the larger, the more non-convex the penalty becomes.

Our proposed sparse estimation method of Ψ can be summarized as follows:

Step 1. Solve problem (2.4), where $(\Lambda, \Psi) \in \Theta$ and obtain $(\tilde{\Lambda}, \tilde{\Psi})$. The estimator $\tilde{\Lambda}$ satisfies the rate given in assumption 4.

Step 2. Conditionally on $\tilde{\Lambda}$ obtained in **Step 1.**, solve the problem

$$\hat{\Psi}^g = \arg \min_{\Psi \in \Omega} \left\{ \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi) + p(\gamma_n, \theta_\Psi) \right\},$$

for a specific penalty $p(\gamma_n, .)$ satisfying assumption 5 and where $\mathbb{G}_{n,p}(.)$ is the Gaussian QML. This two step procedure corresponds to problem (2.2).

or alternatively

Conditionally on $\tilde{\Lambda}$ obtained in **Step 1.**, solve the problem

$$\hat{\Psi}^{ls} = \arg \min_{\Psi \in \bar{\Omega}} \left\{ \mathbb{F}_{n,p}(\tilde{\Lambda}; \Psi) + p(\gamma_n, \theta_\Psi) \right\},$$

for a specific penalty $p(\gamma_n, .)$ satisfying assumption 5 and where $\mathbb{F}_{n,p}(.)$ is the least squares loss. This two step procedure corresponds to problem (2.3).

We highlight that Ω and $\bar{\Omega}$ impose similar eigenvalue conditions on Ψ and side condition through R; in the Gaussian loss function case, Ω includes the additional constraint $a < \lambda_{\min}(2\hat{S} - \Sigma)$, which will be key in view of the regularity conditions we will rely on to derive our consistency results. Intuitively, the constant a controls for the curvature of the Gaussian loss function as it will enter in the so-called restricted eigenvalue parameters of this Gaussian loss function in Corollary 3.3.

Obviously, the regularised problem is not convex with respect to the parameter when considering the SCAD or MCP penalty. Therefore, we would like to weaken the convexity assumption so that we could evaluate the accuracy of $\hat{\theta}_\Psi$. To this goal, the restricted strong convexity is a key ingredient to allow the management of non-convex loss functions. Intuitively, we would like to handle a loss function that locally admits some curvature. To ensure this property, we rely on the strong convexity (local) of the loss function. 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. This amounts to a curvature condition. More precisely, we are interested only in a particular direction, that is the difference $\Delta = \hat{\theta} - \theta_0$ between the estimator $\hat{\theta}$ and true parameter θ_0. Hence the notion of restricted strong convexity weakens the (local) strong convexity by adding a tolerance term. A detailed explanation is provided in [24].

Slightly extending the definition of [22], 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 $\Delta \in \mathbb{R}^d$,

$$\begin{aligned}
\langle \nabla_\theta \mathbb{L}_n(\theta+\Delta) - \nabla_\theta \mathbb{L}_n(\theta), \Delta \rangle &\geq \alpha_1 \|\Delta\|_2^2 - \tau_1 \frac{\log d}{n} \|\Delta\|_1^2, \text{ if } \|\Delta\|_2 \leq 1, \\
\langle \nabla_\theta \mathbb{L}_n(\theta+\Delta) - \nabla_\theta \mathbb{L}_n(\theta), \Delta \rangle &\geq \alpha_2 \|\Delta\|_2 - \tau_2 \sqrt{\frac{\log d}{n}} \|\Delta\|_1, \text{ if } \|\Delta\|_2 \geq 1.
\end{aligned}$$

Note that the (RSC) property is fundamentally local and that α_k, τ_k, $k = 1, 2$ depend on the chosen θ. In [22], their so-called (RSC) condition is similar but uniform with respect to (n, d). Moreover, to weaken notations, we simply write α_k and τ_k, $k = 1, 2$, by skipping their implicit arguments (θ, n, d).

Remark. 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 $\bar\theta := r\theta$ for some $r > 0$. Therefore, the criterion becomes $\bar{\mathbb{L}}_n(\bar\theta) := \mathbb{L}_n(r\theta)$. Since $\nabla_\theta \bar{\mathbb{L}}_n(\theta) = r\nabla_{\bar\theta} \bar{\mathbb{L}}_n(\bar\theta)$, the (RSC) is rewritten as

$$\begin{aligned}
\langle \nabla_{\bar\theta} \bar{\mathbb{L}}_n(\bar\theta+\bar\Delta) - \nabla_{\bar\theta} \bar{\mathbb{L}}_n(\bar\theta), \bar\Delta \rangle &\geq \bar\alpha_1 \|\bar\Delta\|_2^2 - \bar\tau_1 \frac{\log d}{n} \|\bar\Delta\|_1^2, \; \|\bar\Delta\|_2 \leq r, \\
\langle \nabla_{\bar\theta} \bar{\mathbb{L}}_n(\bar\theta+\bar\Delta) - \nabla_{\bar\theta} \bar{\mathbb{L}}_n(\bar\theta), \bar\Delta \rangle &\geq \bar\alpha_2 \|\bar\Delta\|_2 - \bar\tau_2 \sqrt{\frac{\log d}{n}} \|\bar\Delta\|_1, \; \|\bar\Delta\|_2 \geq r,
\end{aligned}$$

with the new constants $(\bar\alpha_1, \bar\tau_1, \bar\alpha_2, \bar\tau_2) := (\alpha_1/r^2, \tau_1/r^2, \alpha_2/r, \tau_2/r)$.

3. Error bounds

We first provide some error bounds under the assumption that the loss function, say $\mathbb{L}_n(.)$, satisfies the RSC condition and the penalty is μ-amenable. More precisely, $\mathbb{L}_n(.)$ is a generic empirical loss function so that the population risk function is defined as $\mathbb{L}(\theta) = \mathbb{E}[\mathbb{L}_n(\theta)]$ assumed to be uniquely minimized at θ_0, which is independent of the sample size n. The regularised estimator thus satisfies

$$\hat\theta = \underset{\theta:\|\theta\|_1 \leq R, \theta \in \Theta}{\arg\min} \Big\{ \mathbb{L}_n(\theta) + p(\gamma_n, \theta) \Big\}, \tag{3.1}$$

where $R > 0$ and $\|\theta_0\| \leq R$ so that θ_0 is a feasible point of the problem and Θ is a convex set. Then we have the following Theorem.

Theorem 3.1. *Suppose $\theta \in \mathbb{R}^d$ and the objective function $\mathbb{L}_n(.) : \mathbb{R}^d \mapsto \mathbb{R}$ satisfies the RSC condition and $p(\gamma_n, .)$ is μ-amenable, with $\frac{3}{4}\mu < \alpha_1$. Suppose the choice*

$$4 \max\Big\{ \|\nabla_\theta \mathbb{L}_n(\theta_0)\|_\infty, \alpha_2 \sqrt{\frac{\log d}{n}} \Big\} \leq \gamma_n \leq \frac{\alpha_2}{6R}, \tag{3.2}$$

and suppose $n \geq \dfrac{16R^2 \max\{\tau_1^2, \tau_2^2\}}{\alpha_2^2} \log d$. *Let* $\hat{\theta}$ *be a stationary point of (3.1). Then* $\hat{\theta}$ *satisfies*

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

where $k_0 = card(\mathcal{A})$ *with* $\mathcal{A} = supp(\theta_0) = \{i : \theta_{0,i} \neq 0\}$.

Remark.

(i) This result is based on an optimization reasoning only and is obtained in a deterministic way. The proof can be found in [25]. As will be clarified in the following Corollaries, to apply Theorem 3.1, we will need to check the conditions for which the second step loss functions $\mathbb{G}_{n,p}(\tilde{\Lambda}; .)$ for the Gaussian case and $\mathbb{F}_{n,p}(\tilde{\Lambda}; .)$ for the least squares case satisfy the RSC condition. Moreover, we will show that suitable choices of γ_n and R provide the probability to satisfy the conditions of Theorem 3.1 with high probability.

(ii) About (α_1, μ): the tightness of the error bounds are sensitive to the difference $4\alpha_1 - 3\mu$, assuming γ_n, k_0 fixed. Here, α_1 should be thought as the curvature of $\mathbb{L}_n$: the bigger α_1 is, the larger the curvature becomes. On the other hand, μ measures the non-convexity of the penalty function: the larger μ is, the more non-convex $p(\gamma_n, .)$ becomes. Thus, there is a trade-off between α_1 and μ when satisfying the constraint $4\alpha_1 > 3\mu$.

Now, we derive an exponential-type inequality to evaluate the probability of satisfying the condition (3.2) of Theorem 3.1. To derive such bound, we rely on the Bernstein inequality applied to the difference $\|\hat{S} - \Sigma_0\|_{\max}$, which appears in the gradient (with respect to Ψ) of $\mathbb{G}_{n,p}(\tilde{\Lambda}; .)$ and $\mathbb{F}_{n,p}(\tilde{\Lambda}; .)$. The sample variance covariance matrix is defined as $\hat{S} := \frac{1}{n}\sum_{i=1}^n X_i X_i'$.

Lemma 3.2. *Under assumption 1, we have the bound*

$$\mathbb{P}\left(\|\hat{S} - \Sigma_0\|_{\max} \geq h(t; n, p, K, \sigma_0^2)\right) \leq \exp(-nt),$$

where

$$h(t; n, p, K, \sigma_0^2) = 2K^2 t + 4K\sigma_0\sqrt{t} + 2\sqrt{2}K\sigma_0\lambda\left(K/(\sqrt{2}\sigma_0), n, \binom{p}{2}\right)$$

with $\lambda(K, n, p) := \sqrt{\frac{2\log(2p)}{n}} + \frac{K\log(2p)}{n}$.

Remark.

(i) This concentration inequality will be applied for bounding the random quantities $\|\nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi)\|_\infty$ and $\|\nabla_{\theta_\Psi}\mathbb{F}_{n,p}(\tilde{\Lambda}; \Psi)\|_\infty$, where the difference $\hat{S} - \Sigma_0$ must be bounded in these score functions. Moreover, our losses are functions of $\tilde{\Lambda}$ so that $\tilde{\Lambda}\tilde{\Lambda}' + \Psi_0$ appears in the gradient with respect to Ψ. The transition from $\tilde{\Lambda}\tilde{\Lambda}' + \Psi_0$ to $\Lambda_0\Lambda_0' + \Psi_0$ is feasible using assumption 4.

(ii) Resuming the remark that follows Assumption 1, should we consider e.g. strongly mixing time series, where $(F_t, \epsilon_t)_t$ would be a strongly mixing process with mixing coefficient $\alpha(.)$ satisfying $\alpha(\varsigma) \leq \kappa\rho^{\varsigma}$ with $\varsigma > 0$ and $0 < \rho < 1$, then our exponential bound would need to be adapted to the strongly mixing case. To do so, Theorem 2 of [23] could be used under strongly mixing and bounded random variables. Although the probability of satisfying the condition (3.2) would be different in such dependent framework, the error bounds together with the scaling (n, p) assumptions would not be altered. The strong mixing assumption is used in the Penalised Gaussian Maximum Likelihood framework of [6].

In Lemma 3.2, setting $t = \frac{\log p}{n}$ implies that $\|\hat{S} - \Sigma_0\|_{\max} \leq K\sqrt{\frac{\log p}{n}}$ with probability at least $1 - \exp(-\log p)$, for $K > 0$ sufficiently large. The choice of $t = \frac{\log p}{n}$ is motivated by condition (3.2) in Theorem 3.1, where we aim at evaluating the probability of satisfying such condition.

Indeed, Theorem 3.1 is stated in a deterministic manner. When applied to the approximate factor model, we derive corresponding probabilistic results, where we establish that for suitable parameter choices (γ_n, R), the conditions of Theorem 3.1 hold with high probability. To do so, this requires bounding the random quantity $\|\nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0)\|_\infty$ (resp. $\|\nabla_{\theta_\Psi}\mathbb{F}_{n,p}(\tilde{\Lambda};\Psi_0)\|_\infty$) in the Gaussian (resp. least squares) based M-estimation problem and verifying the RSC conditions. This motivates the use of Lemma 3.2.

Corollary 3.3. *Suppose the regulariser is μ-amenable, conditionally on the estimator $\tilde{\Lambda}$ satisfying $\|\tilde{\Lambda} - \Lambda_0\|_F = O_p(\sqrt{\frac{p}{n}}) + O_p(\sqrt{\frac{1}{p}})$, under the sample size $n \geq CR^2\alpha_2^{-2}\log(p(p+1)/2)$, with $C > 0$ a sufficiently large constant, with $\alpha_2 = \{\lambda_{\max}(\tilde{\Lambda}\tilde{\Lambda}') + \lambda_{\max}(\Psi_0) + 1\}^{-3}a/2p$, if the regularisation parameter satisfies*

$$4\max\left\{\frac{\lambda_{\max}(\Psi_0^{-1})^2}{2p}\|\tilde{\Lambda}\tilde{\Lambda}' + \Psi_0 - \hat{S}\|_s, \alpha_2\sqrt{\frac{\log p(p+1)/2}{n}}\right\} \leq \gamma_n \leq \frac{\alpha_2}{6R}, \quad (3.3)$$

where $\tilde{\Lambda}$ satisfies (2.4) and $\Psi_0 \in \Omega$, suppose $\frac{3}{4}\mu < \alpha_1$ with $\alpha_1 = \alpha_2$. Then any local optimum $\hat{\Psi}^g$ of the nonconvex program (2.2) satisfies

$$\begin{aligned} \|vech(\hat{\Psi}^g - \Psi_0)\|_2 &\leq \frac{6\gamma_n\sqrt{k_0}}{2\{\lambda_{\max}(\tilde{\Lambda}\tilde{\Lambda}') + \lambda_{\max}(\Psi_0) + 1\}^{-3}a/p - 3\mu}, \\ \|vech(\hat{\Psi}^g - \Psi_0)\|_1 &\leq \frac{6(8\{\lambda_{\max}(\tilde{\Lambda}\tilde{\Lambda}') + \lambda_{\max}(\Psi_0) + 1\}^{-3}a/p - 9\mu)\gamma_n k_0}{(2\{\lambda_{\max}(\tilde{\Lambda}\tilde{\Lambda}') + \lambda_{\max}(\Psi_0) + 1\}^{-3}a/p - 3\mu)^2}, \end{aligned} \quad (3.4)$$

with $a \in \Omega$ so that $a > 0$, $k_0 = |\mathcal{A}|$ and $\mathcal{A} = \{1 \leq i \leq p(p+1)/2 : \theta_{i,\Psi_0} \neq 0\}$.

Furthermore, under assumption 1 so that the sample variance-covariance estimator satisfies the bound in Lemma 3.2, if (γ_n, R) are chosen so that for C_1, C_2, M large constants, $C_1\sqrt{p/n} \leq \gamma_n \leq C_2/R$ and for a sample size $n \geq M\lambda_{\max}(\Psi_0^{-1})^4 p\max\left(R^2, k_0\right)$, then (3.4) hold with probability at least $1 - \exp(-\log p)$.

Remark. Corollary 3.3 and its conditions justify a few comments.

(i) The proof relies on the following two steps:

(a) We verify the RSC condition and derive the quantities $\alpha_1, \alpha_2, \tau_1, \tau_2$ by lower bounding $\lambda_{\min}(\nabla^2_{\theta_\Psi \theta'_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi))$. Here is the key role of the minimum eigenvalue constraint $a < \lambda_{\min}(2\hat{S} - \Sigma)$ included in Ω: this constraint allows us to lower bound the latter minimum eigenvalue. Applying Theorem 3.1, we obtain the error bounds (3.4) in a deterministic manner.

(b) We bound $\|\nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)\|_\infty$ using Lemma 3.2 for a fixed $t = \frac{\log p}{n}$. Indeed, the sub-Gaussian assumption enables to evaluate the probability of satisfying (3.3). In that case, assuming $\lambda_{\max}(\Psi_0^{-1})^4$ as a constant using the eigenvalue condition from the parameter set Ω, the required rate becomes $\max\big(R^2, k_0\big)p = O(n)$ due to the control of the first step estimator. This is a cost not encountered, e.g., by [21]: in their Corollary 3, which provides error bounds on the Graphical Lasso estimator, [21] assume $\|\hat{S} - \Sigma_0\|_{\max} \leq K\sqrt{\log p/n}$, which implies that γ_n should satisfy $C_1\sqrt{\log p/n} \leq \gamma_n \leq C_2/R$. Note that the bound constraint in Ω on the eigenvalues of Ψ_0 is the same as assumption 3.2-(ii) of [6].

(ii) When $p(\gamma_n, \theta_\Psi) = \gamma_n\|\theta_\Psi\|_1$ and $g(\theta_\Psi) = \|\theta_\Psi\|_1$, then setting $\gamma_n \geq L\sqrt{p/n}$ and $R = m_0\sqrt{k_0}$ with a constant $m_0 \geq \|\theta_{\Psi_0}\|_2$, we have the scale $n \geq Mk_0p$. If we consider these bounds in a deterministic manner and fix $\gamma_n \geq L\sqrt{\log p/n}$, then we would obtain a scaling $n \geq Mk_0 \log(p)$, which agrees with Theorem 3.1 of [6].

(iii) We highlight that using $\|\tilde{\Lambda}\tilde{\Lambda}'\|_s \leq \|\tilde{\Lambda}\tilde{\Lambda}' - \Lambda_0\Lambda_0'\|_s + \|\Lambda_0\Lambda_0'\|_s$, α_2 (and α_1) can be expressed with respect to Λ_0. Based on the probability bound of assumption 4 to control for $\|\tilde{\Lambda}\tilde{\Lambda}' - \Lambda_0\Lambda_0'\|_s$, with high probability and for a sufficiently large $C > 0$, the RSC parameter could be written as

$$\alpha_2 = \Big(2\{C^2(\frac{p}{n} + 2\sqrt{\frac{1}{n}} + \frac{1}{p}) + C\|\Lambda_0\|_F(\sqrt{\frac{p}{n}} + \sqrt{\frac{1}{p}})\} + \lambda_{\max}(\Lambda_0\Lambda_0') + \lambda_{\max}(\Psi_0) + 1\Big)^{-3}\frac{a}{2p}.$$

We now focus on the error bounds of $\hat{\Psi}^{ls}$ and check the conditions of applicability of Theorem 3.1.

Corollary 3.4. *Suppose the regulariser is μ-amenable, conditionally on the estimator $\tilde{\Lambda}$ satisfying $\|\tilde{\Lambda} - \Lambda_0\|_F = O_p(\sqrt{\frac{p}{n}}) + O_p(\sqrt{\frac{1}{p}})$, under the sample size $n \geq CR^2\alpha_2^{-2}\log(p(p+1)/2)$, with $C > 0$ a sufficiently large constant, with $\alpha_2 = \frac{1}{p}$, if the regularisation parameter satisfies*

$$4\max\Big\{\frac{1}{p}\|\hat{S} - \tilde{\Lambda}\tilde{\Lambda}' - \Psi_0\|_{\max}, \frac{1}{p}\sqrt{\frac{\log p(p+1)/2}{n}}\Big\} \leq \gamma_n \leq \frac{\alpha_2}{6R},$$

where $\tilde{\Lambda}$ satisfies (2.4) and $\Psi_0 \in \bar{\Omega}$, suppose $\frac{3}{4}\mu < \alpha_1$ with $\alpha_1 = \alpha_2$. Then any local optimum $\hat{\Psi}^{ls}$ of the nonconvex program (2.3) satisfies

$$\begin{aligned} \|vech(\hat{\Psi}^{ls} - \Psi_0)\|_2 &\leq \frac{6\gamma_n\sqrt{k_0}}{4/p - 3\mu}, \\ \|vech(\hat{\Psi}^{ls} - \Psi_0)\|_1 &\leq \frac{6(16/p - 9\mu)\gamma_n k_0}{(4/p - 3\mu)^2}, \end{aligned} \tag{3.5}$$

with $k_0 = |\mathcal{A}|$ and $\mathcal{A} = \{1 \leq i \leq p(p+1)/2 : \theta_{i,\Psi_0} \neq 0\}$.

Furthermore, under assumption 1 so that the sample variance-covariance estimator satisfies the bound in Lemma 3.2, if (γ_n, R) are chosen so that $C_1\sqrt{p/n} \leq \gamma_n \leq C_2/R$ and for a sample size $n \geq Mp \max\left(R^2, k_0\right)$, for C_1, C_2, M large constants, then (3.5) hold with probability at least $1 - \exp(-\log p)$.

Remark. The proof can be decomposed as in Corollary 3.3. Interestingly, the RSC parameters are much more simple ($\alpha_1 = \alpha_2 = 1/p$, $\tau_1 = \tau_2 = 0$) because lower bounding $\lambda_{\min}(\nabla^2_{\theta_\Psi \theta'_\Psi} \mathbb{F}_{n,p}(\tilde{\Lambda}; \Psi))$ is much more straightforward compared to the Gaussian case. This emphasizes that our bounds are sensitive to the curvature of the non-penalised loss function.

4. Support recovery

Based on the Karush-Kuhn-Tucker optimality conditions, [30] developed the primal dual witness (PDW) approach to derive selection consistency for convex problems. There exist similar approaches in [8, 32]. The PDW approach consists in plugging the true subset model $\mathcal{A}$ in the KKT optimality conditions, which are necessary and sufficient if the problem is convex, and checking if they can be satisfied. It means that any solution of the non restricted problem (the original problem providing $\mathcal{A}$) is also a solution to the restricted problem (the regularised one). [22] showed that this approach can be extended to a nonconvex problem and thus to any stationary point, which is their key contribution. They prove that all stationary points are consistent for variable selection via a strict dual feasibility condition and second-order conditions. To obtain the support recovery property, the RSC condition of the loss function with parameters $(\alpha_k, \tau_k)_{k=1,2}$ and the μ-amenability of the penalty function are key assumptions. More details can be found in Subsection A.1: there, in Theorem A.1, we provide the conditions of [22] to ensure the success of the PDW construction - corresponding to **Step 3.** -, that is the scaling of (γ_n, R) and the so-called strict feasibility condition, which characterize the solution of the PDW construction; then Theorem A.2 establishes the support recovery property together with consistency in the $\|.\|_\infty$-sense under the RSC condition, μ-amenable penalties and strict dual feasibility; finally, two sufficient conditions in Proposition A.3 ensure that strict dual feasibility holds for (μ, ζ)-amenable penalty functions. We discuss the use of these results after each Corollary we establish for support recovery for the sparse approximate factor model estimator. Within this setting, we provide ℓ_∞-guarantees for the regularised approximate factor estimator together with the conditions to satisfy the support recovery property. Rather than stating the support recovery property in a deterministic manner, we evaluate the probability of satisfying the latter property. More precisely, we show that any local/global optimum of (2.2) corresponds to the oracle estimator with high probability. The latter is defined as

$$\hat{\Psi}^{\mathcal{O}} = \underset{\Psi \in \Omega, \text{supp}(\Psi) \subseteq \text{supp}(\Psi_0)}{\arg\min} \Big\{ \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi) \Big\}, \tag{4.1}$$

with $\text{vec}(\hat{\Psi}^{\mathcal{O}}) = (\text{vec}(\hat{\Psi}^{\mathcal{O}}_{\mathcal{A}})', \mathbf{0}'_{\mathcal{A}^c})'$ and $k_0 = \text{card}(\text{supp}(\Psi_0))$ the total number of nonzero entries. The matrix $K_0 = \mathbb{E}[\nabla^2_{\text{vec}(\Psi)\text{vec}(\Psi)'} \mathbb{G}_{n,p}(\Lambda_0; \Psi_0)]$ denotes the Fisher information matrix. Importantly, the conditions we derive hold for all stationary points of (2.2), idest for local/global optimum. Note that for the sake of clarification, we express all quantities with respect to $\text{vec}(\Psi)$. Thus k_0 denotes the cardinality of $\mathcal{A} := \text{supp}(\Psi_0) = \{1 \leq i \leq p^2 : \text{vec}(\Psi_0) \neq 0\}$.

Corollary 4.1. *Suppose the sample size satisfies $n \geq C \|\Psi_0^{-1}\|_F^{12} k_0^4 p$ with $C > 0$ large enough, the regularisation parameters (γ_n, R) are chosen so that $\|vec(\Psi_0)\|_1 \leq \frac{R}{2}$ and for $C_1, C_2 > 0$*

$$C_1 \sqrt{\frac{p}{n}} \leq \gamma_n \leq \frac{C_2}{R},$$

suppose $\|K_0^{-1}\|_\infty \leq \beta_\infty$ and assumption 1 holds together with $\underset{1 \leq j \leq p}{\sup} |X_{i,j}| \leq M < \infty, \forall i = 1, \cdots, n$. Then

(i) Suppose $p(\gamma_n, .)$ is μ-amenable and the incoherence condition is satisfied, that is

$$\|K_{0,\mathcal{A}^c\mathcal{A}} K^{-1}_{0,\mathcal{A}\mathcal{A}}\|_\infty \leq \eta < 1. \tag{4.2}$$

Then with probability $1 - \exp(-\log p)$, the objective function (2.2) admits a unique optimum so that $\hat{\mathcal{A}} \subseteq \mathcal{A}$ and for a sufficiently large $\tilde{L} > 0$

$$\|\hat{\Psi}^g - \Psi_0\|_{\max} \leq \tilde{L} \sqrt{\frac{p}{n}} + \gamma_n \beta_\infty.$$

(ii) Suppose $p(\gamma_n, .)$ is (μ, ζ)-amenable and for a sufficiently large $\tilde{L} > 0$

$$\min_{i \in \mathcal{A}} |vech(\Psi_0)_i| \geq \gamma_n(\zeta + 2\beta_\infty) + \tilde{L} \sqrt{\frac{p}{n}},$$

then with probability $1 - \exp(-\log p)$, (2.2) admits a unique optimum $\hat{\Psi}$, which agrees with the oracle estimator $\hat{\Psi}^{\mathcal{O}}$ defined in (4.1) so that

$$\|\hat{\Psi}^g - \Psi_0\|_{\max} \leq \tilde{L} \sqrt{\frac{p}{n}}.$$

Remark. The assumptions and the proof steps of Corollary 4.1 deserve a few comments.

(i) The proof relies on the use of Theorem A.2. To do so, strict dual feasibility must be proved (since Theorem A.2 relies on the conditions of Theorem A.1 and strict dual feasibility). To establish strict dual feasibility, we use Theorem A.1 for μ-amenable penalty functions and Proposition A.3 for (μ,ζ)-amenable penalty functions. Thus the main proof steps can be summarized according to the following steps:

  (a) Establishing strict dual feasibility by upper bounding the quantities $\|\nabla_{\text{vec}(\Psi)}\mathbb{G}_{n,p}(\tilde{\Lambda},\Psi_0)\|_\infty$ and $\|\hat{K}_{\mathcal{A}^c\mathcal{A}}\hat{K}^{-1}_{\mathcal{A}\mathcal{A}}\nabla_{\text{vec}(\Psi)}\mathbb{G}_{n,p}(\tilde{\Lambda},\Psi_0)\|_\infty$ by $(1-\delta)/2\gamma_n$ for $\delta\in[0,1]$ defined in Theorem A.1 - with $\tau_1=0$ since the RSC condition for the Gaussian loss is satisfied with $\tau_1=\tau_2=0$; here $\hat{K}$ is defined as in Theorem A.2 for the Gaussian loss, conditionally on $\tilde{\Lambda}$. These bounds correspond to inequalities (A.6) and (A.7) in Proposition A.3. Note that for the μ-amenable penalty case, the additional quantity $\|\hat{K}_{\mathcal{A}^c\mathcal{A}}\hat{K}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty$ must be upper-bounded.

  (b) Once strict dual feasibility is established, we can express the upper bound of $\|\hat{\Psi}^g-\Psi_0\|_{\max}$ in point (i) of Theorem A.2.

  (c) Establishing (ii) of Corollary 4.1 uses the exact same steps as in (i), except that the (μ,ζ)-amenability allows for a simplification in the upper bound of $\|\hat{\Psi}^g-\Psi_0\|_{\max}$: the term involving $\|\hat{K}_{\mathcal{A}^c\mathcal{A}}\hat{K}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty$ can be discarded as well as the so-called incoherence condition.

(ii) The uniform boundedness assumption on $\sup_{1\leq j\leq p}|X_{i,j}|\leq M<\infty$ for any $i=1,\cdots,n$ is a drawback also encountered in Corollary 3 of [22]. This assumption is required to control for the third order derivative, a quantity that vanishes when considering the least squares loss. In the case of categorical data only, this assumption is always satisfied.

(iii) Corollary 4.1 is not expressed in a deterministic manner since we propose to evaluate the probability of satisfying the inequalities (A.3) and (A.4) in Theorem A.1. This implies controlling, among others, for the infimum norm of the score function evaluated at Ψ_0, conditionally on $\tilde{\Lambda}$. We thus obtain γ_n proportional to $\sqrt{p/n}$, which differs with the usual rate $\sqrt{\log p/n}$ obtained in linear/generalized linear or Graphical Lasso models with Sub-Gaussian variables.

(iv) Let us suppose that the parameter $\|\Psi_0^{-1}\|_F^{12}$ is viewed as a constant. Then the rate $k_0^4p=O(n)$ is necessary due to the first step estimation and the non-linearity with respect to Σ in the Gaussian QML loss. For sparse matrix precision estimation, that is estimating Σ^{-1} (without factor structure), [22] obtained the rate[1] $d_0^2\log(p)=O(n)$ to satisfy the support recovery, where the loss corresponds to a Gaussian QML and d_0 is the maximum number of non-zero coefficients in any row/column of the true matrix precision. Such scaling is obtained for the side constraint $\|\Sigma^{-1}\|_s\leq\kappa$ rather than $\|\text{vec}(\Sigma^{-1})\|_1\leq R$.

[1] Here, we note that the parameters such as $\|\Psi_0^{-1}\|_F^{12}$ are also viewed as constants in [22].

(v) Note that there is an alternative method for constructing $\text{vec}(\hat{\Psi}^g)_{\mathcal{A}}$ such that $\text{supp}(\hat{\Psi}^g_{\mathcal{A}}) \subseteq \mathcal{A}$ and $\text{vec}(\hat{\Psi}^g)_{\mathcal{A}}$ is a zero-subgradient point of the program (A.1) in **Step 1** of the Primal Dual Witness method: this method is based on the Brouwer's fixed point Theorem. Intuitively, the idea consists of proving that if there is a zero sub-gradient vector of the penalised estimator $\mathbb{G}_{n,p}(\tilde{\Lambda}, \Psi) + p(\gamma_n; \text{vec}(\Psi))$ within the set $\{\Psi \in \Omega, \text{supp}(\Psi) \subseteq \text{supp}(\Psi_0)\}$, then this vector is the unique optimum. Then Brouwer's fixed point Theorem is used to show that such optimum lies in a neighbourhood of the true value $\text{vec}(\Psi_0)$ in the $\|.\|_\infty$-sense. Such method was developed by [27] or [22] for sparse precision matrix estimation based on a Gaussian ML criterion. More details on the use the Brouwer's fixed point Theorem can be found in Chapter 13 of [29].

We now provide the conditions to satisfy the support recovery property for the least squares type loss function. To do so, we define the oracle estimator as

$$\hat{\Psi}^{\mathcal{O}} = \underset{\Psi \in \Omega, \text{supp}(\Psi) \subseteq \text{supp}(\Psi_0)}{\arg\min} \Big\{ \mathbb{F}_{n,p}(\tilde{\Lambda}; \Psi) \Big\}, \tag{4.3}$$

with $\text{vec}(\hat{\Psi}^{\mathcal{O}}) = (\text{vec}(\hat{\Psi}^{\mathcal{O}}_{\mathcal{A}})', \mathbf{0}'_{\mathcal{A}^c})'$.

Corollary 4.2. *Under assumption 1, if the sample size satisfies $n \geq Cp$ with $C > 0$ large enough, suppose the regularisation parameters (γ_n, R) are chosen so that $\|vec(\Psi_0)\|_1 \leq \frac{R}{2}$ and*

$$C\sqrt{\frac{p}{n}} \leq \gamma_n \leq \frac{\tilde{C}}{R}.$$

Suppose $p(\gamma_n, .)$ is (μ, ζ)-amenable and for a sufficiently large $L > 0$

$$\min_{i \in \mathcal{A}} |vech(\Psi_0)_i| \geq \gamma_n(\zeta + 2\beta_\infty) + L\sqrt{\frac{p}{n}},$$

then with probability $1 - \exp(-\log p)$, (2.3) admits a unique optimum $\hat{\Psi}$, which agrees with the oracle estimator $\hat{\Psi}^{\mathcal{O}}$ defined in (4.3) so that

$$\|\hat{\Psi}^{ls} - \Psi_0\|_{\max} \leq L\sqrt{\frac{p}{n}}.$$

Remark.

(i) The proof follows the same steps as in Corollary 4.1: since we consider (μ, ζ)-amenable penalty functions, we simply establish inequalities (A.6) and (A.7) in Proposition A.3 to use Theorem A.2. Note that upper bounding $\|\nabla_{\text{vec}(\Psi)}\mathbb{F}_{n,p}(\tilde{\Lambda}, \Psi_0)\|_\infty$ and $\|\hat{K}_{\mathcal{A}^c\mathcal{A}}\hat{K}^{-1}_{\mathcal{A}\mathcal{A}}\nabla_{\text{vec}(\Psi)}\mathbb{F}_{n,p}(\tilde{\Lambda}, \Psi_0)\|_\infty$ by $(1-\delta)/2\gamma_n$, for $\delta \in [0, 1]$ defined in Theorem A.1 - with $\tau_1 = 0$ since the RSC condition for the least squares loss is satisfied with $\tau_1 = \tau_2 = 0$ - and $\hat{K}$ defined as in Theorem A.2 for the least squares loss, is more straightforward due to the linearity of the least squares loss.

(ii) In the Lasso case, which is a μ-amenable regulariser, the mutual incoherence condition does not hold since $\|\hat{K}_{\mathcal{A}^c\mathcal{A}}\hat{K}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty = 1$. As a consequence, strict dual feasibility can not be established for μ-amenable penalties when the least squares type loss function is considered. Moreover, the required rate is $p = O(n)$ since the loss function is much easier to manipulate.

5. Applications of the two step sparse approximate factor model

5.1. *Risk management*

One important application of Corollaries 4.1 and 4.2 concerns risk management. Indeed, as [16] pointed out, the estimation error $\|\Sigma(\tilde{\Lambda},\hat{\Psi}) - \Sigma_0\|_{\max}$ appears in the sensitivity of the investor's utility function. If we consider the latter as the risk minimization with no short-sale constraint $\omega'\Sigma\omega$, with ω a fixed portfolio allocation vector, then the estimation error given in equation (2.4) of [16] satisfies the inequality

$$|\omega'\Sigma(\tilde{\Lambda},\hat{\Psi})\omega - \omega'\Sigma_0\omega| \leq \|\Sigma(\tilde{\Lambda},\hat{\Psi}) - \Sigma_0\|_{\max}\|\omega\|_1^2.$$

This upper bound can be decomposed as

$$\|\Sigma(\tilde{\Lambda},\hat{\Psi}) - \Sigma_0\|_{\max} \leq \|\tilde{\Lambda}\tilde{\Lambda}' - \Lambda_0\Lambda_0'\|_s + \|\hat{\Psi} - \Psi_0\|_{\max}$$

Corollaries 4.1 and 4.2 quantify $\|\hat{\Psi} - \Psi_0\|_{\max}$ for both loss function based estimator $\hat{\Psi}$, whereas $\|\tilde{\Lambda}\tilde{\Lambda}' - \Lambda_0\Lambda_0'\|_s$ is managed using assumption 4.

5.2. *Precision matrix*

Knowledge of the so-called precision matrix - also known as concentration matrix -, that is the inverse of the variance-covariance matrix, is key for a broad range of applications such as graphical models, portfolio optimization, statistical testing, and the like. If Σ_0 admits a factor based decomposition, using

$$\|\Sigma(\tilde{\Lambda},\hat{\Psi}) - \Sigma_0\|_s \leq \|\tilde{\Lambda}\tilde{\Lambda}' - \Lambda_0\Lambda_0'\|_s + \|\hat{\Psi} - \Psi_0\|_s,$$

we can obtain the rate for $\|\hat{\Sigma} - \Sigma_0\|_s$. As for $\|\hat{\Sigma}^{-1} - \Sigma_0^{-1}\|_s$, using the Woodbury identity, we have the following bound:

$$\begin{aligned}
&\|\hat{\Sigma}^{-1} - \Sigma_0^{-1}\|_s \\
&\leq \|\hat{\Psi}^{-1} - \Psi_0^{-1}\|_s + \|(\hat{\Psi}^{-1} - \Psi_0^{-1})A\hat{\Psi}^{-1}\|_s + \|(\hat{\Psi}^{-1} - \Psi_0^{-1})A\Psi^{-1}\|_s \\
&\quad + \|\Psi_0^{-1}(\tilde{\Lambda} - \Lambda_0)B\tilde{\Lambda}'\Psi_0^{-1}\|_s + \|\Psi_0^{-1}(\tilde{\Lambda} - \Lambda_0)B\Lambda_0'\Psi_0^{-1}\|_s \\
&\quad + \|\Psi_0^{-1}\Lambda_0(\tilde{\Lambda}'\hat{\Psi}^{-1}\tilde{\Lambda} - \Lambda_0'\Psi_0^{-1}\Lambda_0)\Lambda_0'\Psi_0^{-1}\|_s,
\end{aligned}$$

where $A := \tilde{\Lambda}(I_r + \tilde{\Lambda}'\hat{\Psi}^{-1}\tilde{\Lambda})\tilde{\Lambda}'$ and $B := I_r + \tilde{\Lambda}'\hat{\Psi}^{-1}\tilde{\Lambda}$. From Lemma 11 of [22] or Lemma 2 of [10], it follows that

$$\|\hat{\Psi}^{-1} - \Psi_0^{-1}\|_s \leq \frac{\|\Psi_0^{-1}\|_s^2\|\hat{\Psi} - \Psi_0\|_s}{1 - \|\Psi_0^{-1}\|_s\|\hat{\Psi} - \Psi_0\|_s}.$$

Since it is assumed that the minimum eigenvalue of Ψ_0 is bounded below by the constant $b_1 > 0$, the upper bound can be represented as $O(b_1^{-2})\|\hat{\Psi} - \Psi_0\|_s$. Thus, we would obtain a rate for $\|\hat{\Sigma}^{-1} - \Sigma_0^{-1}\|_s$ using the same approach with the proof of Theorem 3 in [10] (see, Section C.4.2 of [10]).

6. Empirical applications

6.1. Simulation setting

In all our simulation experiments, we simulate the p-dimensional random vector X_i based on the data generating process

$$\begin{cases} X_i & \sim & \mathcal{N}_{\mathbb{R}^p}(0, \Sigma_0), \\ \Sigma_0 & = & \Lambda_0 \Lambda_0' + \Psi_0, \end{cases}$$

where Λ_0 satisfies the identifiability condition IC5 provided in the parameter set Ω in Section 2. Each component of Λ_0 are simulated in the uniform distribution $\mathcal{U}([-0.3, 0.3])$ and we fix $m = 5$ for all simulated experiments performed from subsection 6.1 to subsection 6.5. The matrix Ψ_0 is assumed to be k_0-sparse (off-diagonal elements only), where k_0 depends on the size of the problem (arbitrarily set) and is fixed once only.

First, we propose to illustrate the statistical consistency for $p = 500$. In the second penalised step, the total number of parameters is 125000. We set the number of zero parameters as 106462, which represents approximately 85% of the total number of parameters. The true subset model is thus $k_0 := |\mathcal{A}| = 18538$. Regarding the non-zero components of Ψ_0, the off-diagonal elements are simulated in $\mathcal{U}([-2, 2])$ and the diagonal elements in $\mathcal{U}([6, 9])$. In that case, we obtain $\|\text{vech}(\Psi_0)\|_1 = 8991.3$, $\|\text{vech}(\Psi_0)\|_2 = 184.9$ and $\|\text{vech}(\Psi_0)\|_\infty = 8.99$. To recover the sparse support $\mathcal{A}$, we consider both regularised problems (2.2) for the Gaussian based second step objective function and (2.3) for the least squares based second step objective function. In both problems, non-convexity can potentially come from the regulariser, and the second step parameter sets Ω in (2.2) and $\bar{\Omega}$ in (2.3) are convex. In the first step, $(\Lambda, \Psi) \in \Theta$ and are jointly estimated. To solve the regularised optimization problem, we follow the composite gradient descent procedure of [21] (see their section 4), which consists in a three step updating procedure of the optimized parameter value. As an initial value for the algorithm, we start with $\Psi^{(0)} = \hat{S} - \tilde{\Lambda}\tilde{\Lambda}'$. Importantly, due to the constraints on the RSC coefficients and the trade-off between α_1 and μ for both the Gaussian based estimator and the least squares based estimator, we consider the following setting:

- *Gaussian loss:* $a = 1.6245$, where a is the lower bound of $\lambda_{\min}(2\hat{S} - \Sigma(\tilde{\Lambda}, \Psi))$ and thus $b_{scad} = 1.0038e + 07, b_{mcp} = 1.0038e + 07$, the values from which $\alpha_1 > \frac{3}{4}\mu$ is satisfied. To compute α_1, we replace the first step estimate $\tilde{\Lambda}$ by its true value Λ_0 and obtained $\alpha_1 = 7.4719e - 08$. For the SCAD, $4\alpha_1 - 3\mu = 1.2737e - 11$; for the MCP, $4\alpha_1 - 3\mu = 1.2766e - 11$; finally, for the lasso, $4\alpha_1 - 3\mu = 2.9888e - 07$.

- *Least squares loss:* $\alpha_1 = 0.002$. We choose $b_{scad} = 520, b_{mcp} = 460$. For the SCAD, $4\alpha_1 - 3\mu = 0.0022$; for the MCP, $4\alpha_1 - 3\mu = 0.0015$; finally, for the lasso, $4\alpha_1 - 3\mu = 0.008$.

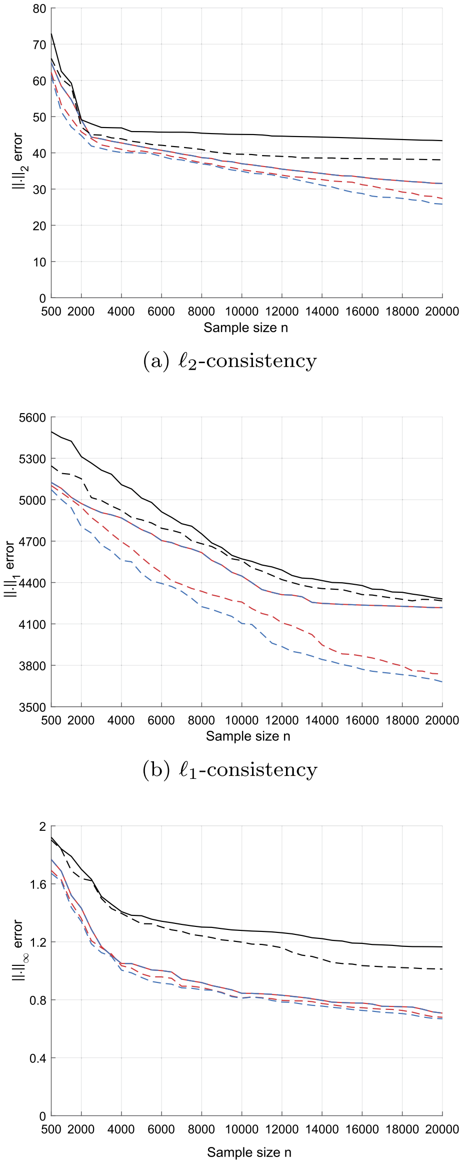
We highlight that the calibration of (α_1, μ) has a significant impact for the convergence of the gradient type algorithm. As described by [22] in their Section 5, when the condition $4\alpha_1 > 3\mu$ is violated, multiple stationary points may emerge in the case of non-convex penalties. Convergence of the gradient descent type algorithm and statistical consistency are no longer ensured. Although these authors obtained the convergence of the algorithm to a single optimum with multiple initial values when $4\alpha_1$ is slightly smaller than 3μ, we restrict our analysis and the following simulations to the $4\alpha_1 > 3\mu$ case only: this ensures statistical consistency and allows us to report the theoretical upper bounds.

As for the (γ_n, R) parameters, we select $R = \frac{4}{\gamma_n} p(\gamma_n, \text{vech}(\Psi_0))$ to ensure the feasibility of Ψ_0 following [21, 22]. Furthermore, we set $\gamma_n = c\sqrt{\log(p(p+1)/2)/n}$ with $c = 0.5$, a constant selected as optimal for both loss functions by a cross-validation procedure for $n = 20000$. For general data sets, R cannot be computed since the true underlying model is unknown, so that a data-driven method such as cross-validation is required.

We consider samples with size $500, 1000, 1500, \cdots, 20000$ and for each sample size, we simulate 200 times the random vector (X_i) and thus obtain 200 sparsity-based estimates $\hat{\Psi}^g$ and $\hat{\Psi}^{ls}$ of the theoretical matrix Ψ_0. Figure 1a show the $\|.\|_2$ consistency with respect to the sample size for both estimate $\hat{\Psi}^g$ and $\hat{\Psi}^{ls}$. Each point represents the average error of the 200 simulations. The theoretical bounds are reported for the least squares based estimate and $\gamma_n = 0.5\sqrt{\log(p(p+1)/2)/n}$. As predicted in Corollaries 3.3 and 3.4, the three curves for the MCP, SCAD and Lasso converge toward zero as the number of samples increases. The same remark holds for the $\|.\|_1$ consistency displayed in Figure 1b and the $\|.\|_\infty$ displayed in Figure 1c, where the incoherence condition is not necessary. Although the Lasso requires this condition for the support-recovery, the $\|.\|_\infty$-error for the lasso-based estimates is reported for information purposes. The figures highlight that when the problem dimension increases, the Lasso is significantly outperformed. Besides, the MCP and SCAD for the least squares case perform better than the Gaussian case for small samples.

6.2. *A sensitivity analysis*

In this section we perform a sensitivity analysis of the statistical consistency and the theoretical error bound with respect to γ_n based on two settings: Ψ_0 banded and non-banded. The banded covariance case is relevant for a time series framework and means that the entries decay based on their distance from the diagonal elements. We consider two penalisation rates for γ_n: proportional to $\sqrt{\log(p(p+1)/2)/n}$; proportional to $\sqrt{p/n}$.



(a) ℓ_2-consistency

(b) ℓ_1-consistency

(c) ℓ_∞-consistency

Fig 1. $\|.\|_2, \|.\|_1, \|.\|_\infty$ *consistencies for the setting of Subsection* 6.1. *SCAD, MCP and Lasso are represented in red, blue and black respectively. The least squares case and Gaussian case are represented in solid lines and dashed lines respectively. Each point represents an average of* 200 *trials for each sample size.*

Non-banded case.

We consider the same simulation setting as in the previous subsection for $p = 50$, where the diagonal coefficients of Ψ_0 are simulated in $\mathcal{U}([-2.5, 2.5])$ and the diagonal coefficients in $\mathcal{U}([4, 6.5])$. We propose to analyse how the convergence and theoretical bounds are altered depending on the choice of the tuning γ_n. The total number of parameters is $p(p+1)/2 = 1275$; the total number of zero parameters is set to 1084, which represents 85% of the total number of parameters; the true subset model is thus $k_0 = 191$. In such setting, Ψ_0 satisfies: $\|\text{vech}(\Psi_0)\|_2 = 39.90$ and $\|\text{vech}(\Psi_0)\|_1 = 364.82$. To recover the sparse support and estimate a sparse Ψ, we use the same two-step estimation methods (2.2) and (2.3). We consider two different regularisation parameters γ_n: after applying a cross-validation procedure for $n = 20000$, we selected $\gamma_n = 0.3\sqrt{p/n}$; alternatively, we considered $\gamma_n = 0.3\sqrt{\log(p(p+1)/2)/n}$. We set $R = \frac{4}{\gamma_n} p(\gamma_n, \text{vech}(\Psi_0))$.

Regarding the parameters related to the RSC condition and the constraint $\alpha_1 > \frac{3}{4}\mu$, we consider the following setting:

- *Gaussian loss:* $a = 1.76$ and thus $b_{scad} = 43880, b_{mcp} = 43879$. To compute α_1, we replace the first step estimate $\tilde{\Lambda}$ by its true value Λ_0 and obtained $\alpha_1 = 1.7093e - 05$. For the SCAD, $4\alpha_1 - 3\mu = 1.55e - 09$; for the MCP, $4\alpha_1 - 3\mu = 1.51e - 09$; for the Lasso, $4\alpha_1 - 3\mu = 6.84e - 05$.
- *Least squares loss:* $\alpha_1 = 0.02$. We choose $b_{scad} = 310, b_{mcp} = 230$. For the SCAD, $4\alpha_1 - 3\mu = 0.0703$; for the MCP $4\alpha_1 - 3\mu = 0.0669$; for the Lasso, $4\alpha_1 - 3\mu = 0.08$.

Banded case.

To further explore the effect of the regularisation rate, we propose an additional sensitivity analysis based on a banded Ψ_0 matrix for $p = 100$. To do so, we replicate the simulation setting of [6] provided in their Subsection 5.1, except that the coefficients $\{a_i, b_i, c_i\}_{i=1}^{p}$ are simulated in $0.9\mathcal{N}_{\mathbb{R}}(0, 1)$. In such setting, the total number of parameters is 5050; the total number of zero coefficients is set to 4656 so that $k_0 = 394$. Then, Ψ_0 satisfies: $\|\text{vech}(\Psi_0)\|_2 = 63.59$ and $\|\text{vech}(\Psi_0)\|_1 = 829.30$. We also considered two different regularisation parameters γ_n: after applying a cross-validation procedure for $n = 20000$, we selected $\gamma_n = 0.2\sqrt{p/n}$; alternatively, we considered $\gamma_n = 0.2\sqrt{\log(p(p+1)/2)/n}$. We set $R = \frac{4}{\gamma_n} p(\gamma_n, \text{vech}(\Psi_0))$.

As for the parameters related to the RSC conditions, we consider the setting:

- *Gaussian loss:* $a = 0.0015$ and thus $b_{scad} = 1.05 + 09, b_{mcp} = 1.05 + 09$. We obtained $\alpha_1 = 7.16e - 10$. For the SCAD, $4\alpha_1 - 3\mu = 9.80e - 12$; for the MCP, $4\alpha_1 - 3\mu = 9.80e - 12$; for the Lasso, $4\alpha_1 - 3\mu = 2.86e - 09$.
- *Least squares loss:* $\alpha_1 = 0.01$. We choose $b_{scad} = 350, b_{mcp} = 280$. For the SCAD, $4\alpha_1 - 3\mu = 0.0314$; for the MCP $4\alpha_1 - 3\mu = 0.0293$; for the Lasso, $4\alpha_1 - 3\mu = 0.04$.

For both Ψ_0 cases, the consistency patterns, the theoretical upper bounds and their sensitivities with respect to γ_n are reported in Panels 2a-2f for the

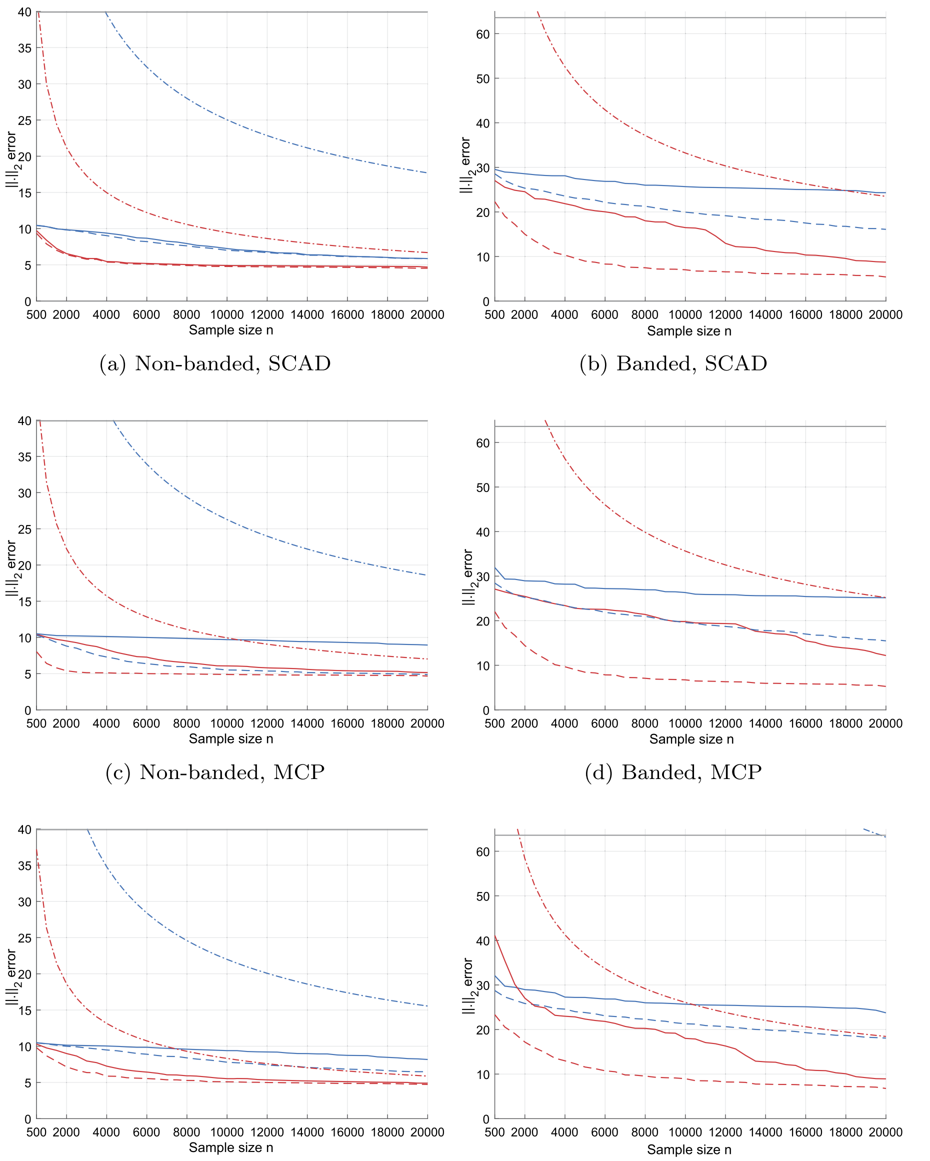
$\|.\|_2$ sense and Panels 3a-3f for the $\|.\|_1$ sense. Panels 4a-4f contains the consistency patterns in the $\|.\|_\infty$ sense only since the corresponding theoretical upper bounds are not as explicit as in the $\|.\|_2, \|.\|_1$ cases. For all cases, the consistency is more favorable when using the rate $\sqrt{\log(p(p+1)/2)/n}$. This is in line with the theoretical upper bounds, which depend on γ_n: using a tighter rate provides tighter bounds as depicted in our figures. Note that for $\|.\|_1$ consistency, we reported in the non-banded/Lasso case the line corresponding to $\|\text{vech}(\Psi_0)\|_1$ (figure 3e, gray colour) since the theoretical upper becomes informative for $n < 20000$. Finally, the choice of the SCAD/MCP values (b_{scad}, b_{mcp}), although larger than the optimal SCAD value $b_{scad} = 3.7$ identified by [12] and MCP value $b_{mcp} = 3.5$ selected by [21], they allow for informative theoretical upper bounds. However, for both sizes p, the theoretical upper bound for the $\|.\|_1$-error is not reported due to the large sample size $n \geq 30000$ required to reach from below $\|\text{vech}(\Psi_0)\|_1$ for all regularisation cases.

6.3. Relevance of the error bounds

An important issue is how “informative” these error bounds are. Their rates depend on the regularisation parameter γ_n, as highlighted in subsection 6.2, on the curvature of the loss function through the RSC parameters and the non-convexity of the penalty, where the trade-off expressed through the constraint $4\alpha_1 > \mu$ is a key element in our theoretical analysis. For the Gaussian loss function, the constraint $4\alpha_1 > \mu$ is satisfied for significantly large values of b_{scad} and b_{mcp} so that μ is small enough compared to α_1, which depends on a controlling for the minimum eigenvalues of $2\hat{S} - \Sigma$. Hence the denominator implies that the upper bounds for both the $\|.\|_1$ and $\|.\|_2$ errors are non-informative. The Gaussian-based theoretical upper bounds would thus require a sample size of significantly large order to obtain informative upper bounds. This setting changes for the least squares loss function, where the RSC parameter α_1 is large enough so that the denominator becomes larger. This requires from b_{scad} and b_{mcp} to still be large enough to obtain informative theoretical upper bounds.

6.4. A comparison to some competitors

We now propose a comparison of our penalised estimation method with some alternative factor based approaches. To do so, we consider a simulation setting for $p = 200$ and non-banded Ψ_0. The total number of parameters is 20100 in the second step for a fixed Λ. The off-diagonal coefficients of Ψ_0 are simulated in the uniform distribution $\mathcal{U}([-3,3])$ and the diagonal coefficients in $\mathcal{U}([6,9])$. The total number of zero parameters is set to 17085, which represents 85% of the total number of parameters, so that the true subset model is given by $k_0 = 3015$. To estimate Ψ_0, we consider the following penalisation approaches: our proposed two-step Gaussian and least squares based penalised losses for the Lasso, SCAD and MCP; the PML approach of [6], which consists in the joint estimation of (Λ, Ψ) based on a Gaussian QML, where Ψ only is penalised by



(a) Non-banded, SCAD

(b) Banded, SCAD

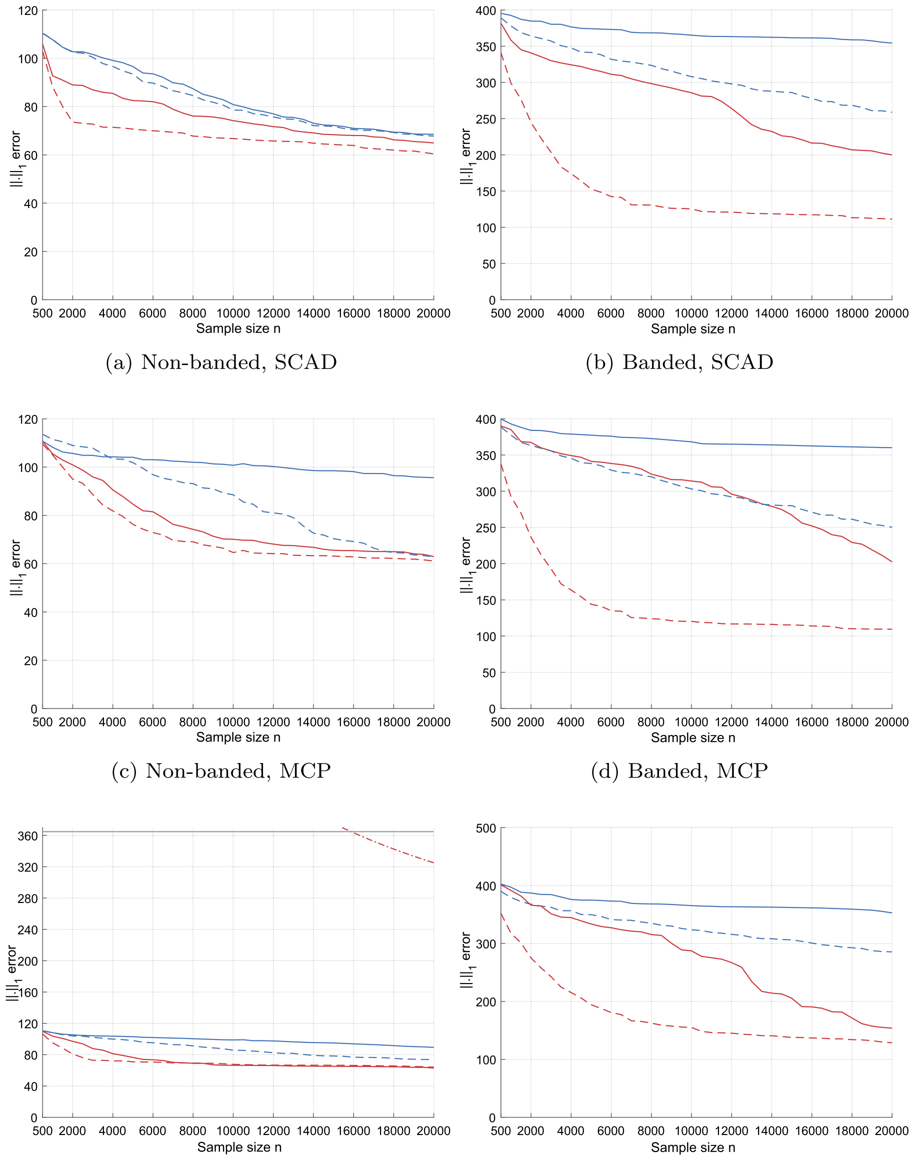
(c) Non-banded, MCP

(d) Banded, MCP

(e) Non-banded, Lasso

(f) Banded, Lasso

FIG 2. $\|.\|_2$ *consistency for the setting of Subsection 6.2. For each penalty case, the results for* $\gamma_n = 0.3\sqrt{\log(p(p+1)/2)/n}$ *(resp.* $\gamma_n = 0.3\sqrt{p/n}$*) represented in red (resp. blue). The Gaussian case and least squares case are represented in solid lines and dashed lines respectively. The horizontal gray line represents* $\|vech(\Psi_0)\|_2$*. The theoretical upper bounds are represented in dashed-dotted lines. Each point represents an average of* 200 *trials for each sample size.*



(a) Non-banded, SCAD

(b) Banded, SCAD

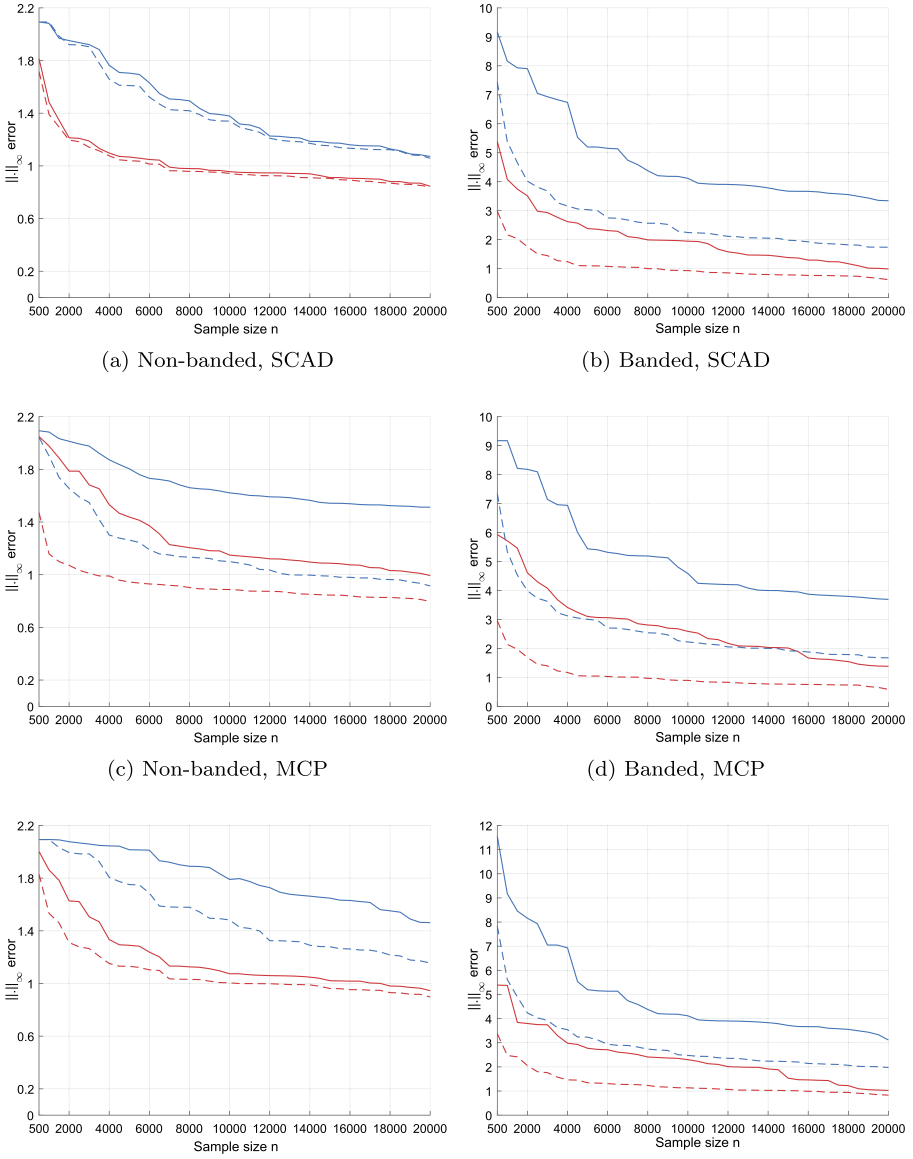
(c) Non-banded, MCP

(d) Banded, MCP

(e) Non-banded, Lasso

(f) Banded, Lasso

Fig 3. *$\|.\|_1$ consistency for the setting of Subsection 6.2. For each penalty case, the results for $\gamma_n = 0.3\sqrt{\log(p(p+1)/2)/n}$ (resp. $\gamma_n = 0.3\sqrt{p/n}$) represented in red (resp. blue). The Gaussian case and least squares case are represented in solid lines and dashed lines respectively. The horizontal gray line represents $\|vech(\Psi_0)\|_1$. The theoretical upper bounds are represented in dashed-dotted lines. Each point represents an average of 200 trials for each sample size.*



(a) Non-banded, SCAD

(b) Banded, SCAD

(c) Non-banded, MCP

(d) Banded, MCP

(e) Non-banded, Lasso

(f) Banded, Lasso

FIG 4. $\|.\|_\infty$ *consistency for the setting of Subsection* 6.2. *For each penalty case, the results for* $\gamma_n = 0.3\sqrt{\log(p(p+1)/2)/n}$ *(resp.* $\gamma_n = 0.3\sqrt{p/n}$*) represented in red (resp. blue). The Gaussian case and least squares case are represented in solid lines and dashed lines respectively. Each point represents an average of* 200 *trials for each sample size.*

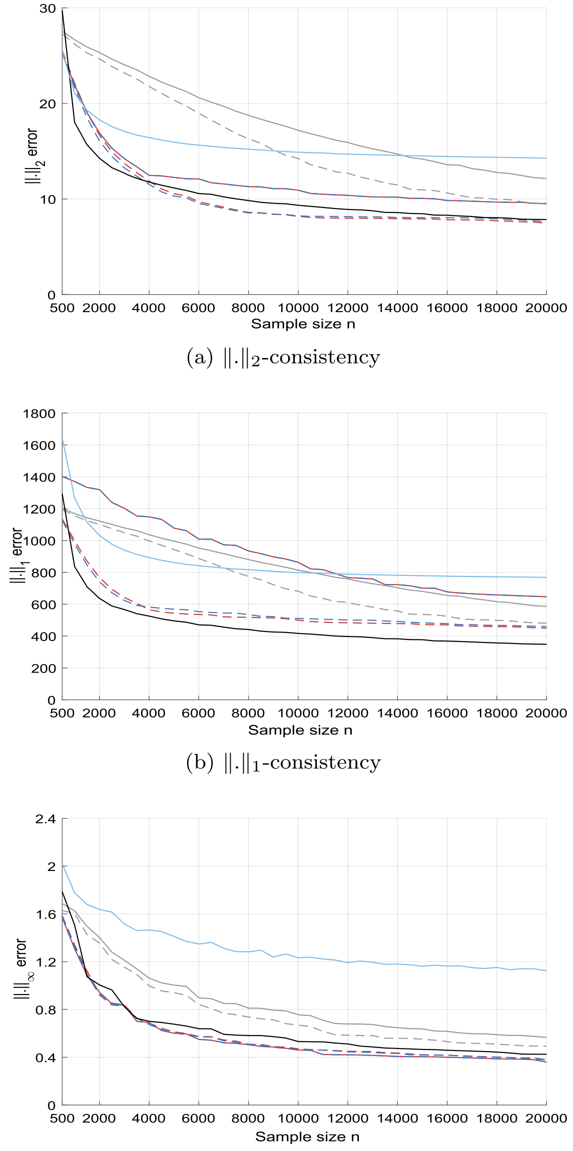
the adpative Lasso; the POET[2] of [15], where Ψ is penalised using an adaptive hard-thresholding rule. As for the choice of (γ_n, R) in the penalised two-step estimation and joint estimation (that is the method of [6]) cases, we set $\gamma_n = c\sqrt{\log(p(p+1)/2)/n}$. The parameter R is defined as in the previous section. For the POET case, we employed the soft thresholding $p_{ij}(\rho) = \text{sgn}(\rho)(|\rho| - \tau_{ij})_+$ with the adaptive threshold $\tau_{ij} = c\gamma_n\sqrt{\hat{\theta}_{ij}}$ $(\gamma_n = 1/\sqrt{p} + \sqrt{\log(p(p+1)/2)/n})$ described in equation (3.2) of [15], where $\hat{\theta}_{ij}$ depends on the first step estimation of the loading factors. For all these cases, we set the constant $c = 0.3$, a value selected by cross-validation for $n = 20000$ for the joint estimation case.

The consistency results in the $\|.\|_2, \|.\|_1, \|.\|_\infty$ senses are reported in Panels 5a, 5b and 5c, respectively. For all cases, the error decreases with the sample size, which agrees with our Corollaries and the consistency results established in [6] (see their Theorems 3.1 and Theorem 3.2) and [15] (see their Theorem 1). The two-step SCAD and MCP penalised least squares based cases and the PML provide similar patterns. Note that the two-step SCAD and MCP penalised Gaussian based case provide the same pattern (in these Panels, their corresponding errors are overlapping) since their component b_{scad}, b_{mcp} are significantly large to satisfy the condition $4\alpha_1 > 3\mu$. Interestingly, when n is large, the POET estimator is less desirable than alternative estimators.

6.5. Illustration of the support recovery property

For both banded and non-banded Ψ_0 cases, given the sparse approximate factor model structure $\Sigma_0 = \Lambda_0\Lambda_0' + \Psi_0$ and sample size n, we drew hundred batches of n independent samples from the associate Gaussian distribution $X_i \sim \mathcal{N}_{\mathbb{R}^p}(0, \Sigma_0)$, where we considered two cases, $p = 100, 300$. Λ_0 is defined as in the previous simulation settings. For the banded case, we considered the same setting as in subsection 6.2 such that the number of zero coefficients is 4656 and $k_0 = 394$ for $p = 100$. When $p = 300$, the number of zero coefficients is set as 43956 and the non-zeros as 1194. For the non-banded case, we selected $k_0 = 1010$ so that the total number of zero coefficients is 4040 when $p = 100$. For the case $p = 300$, then the number of zero coefficients is defined as 36120 so that $k_0 = 9030$. We report the variable selection performance through the number of zero coefficients correctly estimated, denoted as C, the number of zero coefficients incorrectly estimated (i.e. an estimated zero coefficient whereas the true parameter is non-zero), denoted as $IC1$, the number of nonzero coefficients incorrectly estimated (i.e. an estimated non-zero coefficient whereas the true parameter is zero), denoted $IC2$, in Table 1 (resp. Table 2) for both banded and non-banded cases when $p = 100$ (resp. $p = 300$), averaged for these

[2] The Principal Orthogonal complEment Thresholding (POET) is based on the spectral decomposition of the sample variance covariance $\hat{S} = \sum_{i=1}^{K} \hat{\lambda}_i\hat{\zeta}_i\hat{\zeta}_i' + \hat{R}_K$, where $\hat{R}_K = \sum_{i=1}^{p} \hat{\lambda}_i\hat{\zeta}_i\hat{\zeta}_i'$, with $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$ are the ordered eigenvalues of $\hat{S}$, $\hat{\zeta}_i$ the corresponding eigenvalues. A hard-thresholding procedure is then proposed to penalise $\hat{R}_K$. Note that K is the number of of diverging eigenvalues.



(a) $\|.\|_2$-consistency

(b) $\|.\|_1$-consistency

(c) $\|.\|_\infty$-consistency

Fig 5. $\|.\|_2, \|.\|_1, \|.\|_\infty$ *consistencies for the setting of Subsection 6.4. For each case, the SCAD, MCP, Lasso in the two-step approach, the adaptive Lasso PML and POET are represented in red, blue, gray, black and cyan. For the two-step approach, the Gaussian (resp. least squares) loss case is represented in solid (resp. dashed) line. Each point represents an average of* 200 *trials for each sample size.*

hundred batches. The mean squared error is reported as an estimation accuracy measure.

We considered our proposed penalised two-step approach for both Gaussian and least squares losses, named as 2S-Lasso, 2S-SCAD and 2S-MCP. Although support recovery is not established in [6] and [15], we reported their proposed estimators, named as PML-aLasso and POET respectively. In the simulations, we selected the penalisation parameter $\gamma_n = c\sqrt{\log(p(p+1)/2)/n}$ and the threshold $\tau_{ij} = c\gamma_n\sqrt{\hat{\theta}_{ij}}$ $(\gamma_n = 1/\sqrt{p} + \sqrt{\log(p(p+1)/2)/n})$ in the POET case, with $c = 0.5$ when $p = 100, 300$. For the two-step method, for the least squares loss, we selected $b_{scad} = 120, b_{mcp} = 80$ (resp. $b_{scad} = 260, b_{mcp} = 230$) for both Ψ_0 cases when $p = 100$ (resp. $p = 300$). For the Gaussian loss, in both cases these parameters are set sufficiently large to satisfy the condition $4\alpha_1 > \mu$.

Our simulation results indicate the challenge to perfectly recover the true sparse model for all sparse estimator candidates. In the banded case, the PML, the Gaussian and the least squares based penalised methods perform well. In small sample sizes, the PML method of [6] provides good performance results, whereas the difference tends to mitigate with our two-step methods when n increases. Interestingly, in the non-banded case, the PML method tends to excessively shrink the parameters, which translates into a large number of incorrectly identified zero coefficients $IC2$. Compared to the latter method, our proposed two-step approach offers better results. For all cases, the two-step Gaussian MCP/SCAD penalised method provide similar results: since b_{scad}, b_{mcp} are large to satisfy the constraint $4\alpha_1 > 3\mu$, these penalisation methods consequently behave similarly. We note that $IC2$ increases with the sample size since a less severe penalisation is applied on the parameters: more non-zero coefficients tend to be estimated. Interestingly, although the POET method provides good performances in the non-banded case, its ability to correctly identify zero coefficients significantly diminishes in the banded case.

One key reason for the difficulty to obtain perfect support recovery lies in the accurate estimation of the loading matrix. In both the proposed method and the PML one, large values of p are required for accurate estimation of the factor loadings. Indeed, we cannot ensure the consistency of the loading estimators without the assumption $p \to \infty$ [5, 6].

6.6. A real data example

In this section, we propose to assess the relevance of our methodology through a portfolio allocation analysis. Portfolio allocation models may suffer from instability, which results from the instability of the variance covariance estimate. One application of the sparse approximate factor model based variance covariance matrix consists in optimal portfolio allocation since it is a key input of the investment problem as a risk measure. We carry out an out-of-sample analysis of the portfolio forecasting performances using the global minimum variance portfolio approach (GMVP). The GMVP is an investment strategy, whose explicit solution is given by the portfolio vector of weights $\omega = \Sigma^{-1}\iota/\iota'\Sigma^{-1}\iota$, where Σ is

TABLE 1
Model selection and precision accuracy for $p = 100$ based on 100 replications. The penalised two step Gaussian (resp. least squares) case is reported on the left (resp. right) side.

	Truth	2S-Lasso	2S-SCAD	2S-MCP	PML-aLasso	POET
Banded case						
$n = 5000$						
C	4656	4656 − 4656	4656 − 4655.8	4656 − 4654.8	4656	4312.2
IC1	0	273 − 195.2	233.8 − 157.5	233.8 − 131.8	67	40.4
IC2	0	0 − 0	0 − 0.2	0 − 1.2	0	343.8
MSE		567.8 − 297.4	394.8 − 138.9	391.3 − 89.1	198.2	1342.2
$n = 10000$						
C	4656	4655.8 − 4654.7	4656 − 4642.5	4656 − 4624.7	4636	4342.2
IC1	0	250.5 − 139.3	170.8 − 85	170.8 − 72.3	30.8	36.4
IC2	0	0.25 − 1.25	0 − 13.5	0 − 31.25	38.4	313.8
MSE		461.3 − 188.6	253.1 − 64.5	247.35 − 49.8	165.83	1367.0
$n = 30000$						
C	4656	4646 − 4572.9	4568.5 − 4502.4	4568.5 − 4517.1	4574.7	4364.1
IC1	0	140.1 − 60.7	101.9 − 46.2	101.9 − 42.7	34.6	32.4
IC2	0	376.4 − 77.5	175.5 − 198.2	175.5 − 201.4	93.6	291.9
MSE		211.1 − 79.6	173.6 − 39.2	171.3 − 38.7	216.5	1384.5
$n = 50000$						
C	4656	4635.5 − 4422.3	4351.7 − 4286.7	4351.7 − 4252.4	4240.2	4366.4
IC1	0	74.9 − 45.3	99.5 − 35.5	99.5 − 33	27.1	31.4
IC2	0	90.5 − 233.7	324.3 − 339.3	324.3 − 323.6	276	289.7
MSE		96.9 − 57.7	111.9 − 33.6	109.6 − 36.3	79.1	1387.8
Non-banded case						
$n = 5000$						
C	4040	4040 − 4040	4040 − 4040	4040 − 4040	4040	3787.7
IC1	0	802.3 − 650.2	747.3 − 592.6	747.3 − 569.2	826.8	633.5
IC2	0	0 − 0	0 − 0	0 − 0	0	252.3
MSE		241.1 − 220.8	223.1 − 198.4	222.9 − 179.2	222.6	242.6
$n = 10000$						
C	4040	4040 − 4040	4040 − 4040	4040 − 4040	4040	3766.4
IC1	0	684.1 − 507.8	604.1 − 385.9	604.1 − 377.2	786.6	608.9
IC2	0	0 − 0	0 − 0	0 − 0	0	273.6
MSE		230.2 − 183.7	193.2 − 140.9	192.6 − 111.7	199.6	240.9
$n = 30000$						
C	4040	4040 − 4040	4039.7 − 4020.8	4039.7 − 3979.2	4040	3732.1
IC1	0	482.3 − 387	395.3 − 288.9	395.3 − 273.1	714.8	580.3
IC2	0	0 − 1.1	0.4 − 17.7	0.4 − 35.8	0	307.9
MSE		188.1 − 102.7	109.9 − 47.5	107.81 − 33.2	119.93	157.5
$n = 50000$						
C	4040	4040 − 4017.1	4004.8 − 3963.6	4004.8 − 3911.2	4034.1	3716.8
IC1	0	347.8 − 284.2	313.1 − 237.4	313.1 − 204.2	618.2	570.9
IC2	0	0 − 21.5	19 − 28.1	19 − 59.2	0.1	323.2
MSE		153.5 − 67.0	61.7 − 27.8	60.2 − 22.3	92.75	138.5

Table 2

Model selection and precision accuracy for $p = 300$ based on 100 replications. The penalised two step Gaussian (resp. least squares) case is reported on the left (resp. right) side.

	Truth	2S-Lasso	2S-SCAD	2S-MCP	PML-aLasso	POET
Banded case						
$n = 5000$						
C	43956	43956 − 43956	43956 − 43956	43956 − 43956	43956	40158.0
IC1	0	860.5 − 750.5	847.6 − 828.2	847.6 − 824	282	90.0
IC2	0	0 − 0	0 − 0	0 − 0	0	3798.0
MSE		1767.5 − 1195.9	1738.6 − 1293.5	1711.6 − 1258.8	944.6	1203.3
$n = 10000$						
C	43956	43956 − 43956	43954.3 − 43956	43954.3 − 43956	43952	39821.9
IC1	0	821.5 − 617.7	801.6 − 743.5	801.6 − 732.3	255.8	80.1
IC2	0	0 − 0	2.5 − 0	2.5 − 0	3.7	4134.2
MSE		1456.4 − 811.9	1392.8 − 924.6	1374.1 − 865.2	831.1	1225.6
$n = 30000$						
C	43956	43925 − 43948.8	43889.5 − 43956	43889.5 − 43955.1	43955.2	39101.1
IC1	0	630.5 − 311.4	619 − 435.5	619 − 407.6	232.5	72.4
IC2	0	30.4 − 5	13.8 − 0	13.8 − 0	2.5	4854.9
MSE		1340.9 − 779.8	843.2 − 254.1	812.4 − 211.9	982	1246.0
$n = 50000$						
C	43956	43891.5 − 43750.2	43857.6 − 43917.9	43857.6 − 43898.2	43601.1	38749.2
IC1	0	377.1 − 217.8	431.9 − 263.3	431.9 − 253.1	212.8	70.1
IC2	0	48.1 − 210.8	21.5 − 39.8	21.5 − 58.6	13.6	5206.8
MSE		455.9 − 162.5	454.9 − 102.6	451.6 − 91.1	561.8	1250.2
Non-banded case						
$n = 5000$						
C	36120	36120 − 36120	36120 − 36120	36120 − 36120	36120	35663.9
IC1	0	8545.1 − 8234.4	8156.4 − 7992.3	8156.4 − 7937.5	8558.9	4815.80
IC2	0	0 − 0	0 − 0.1	0 − 0.2	0	456.1
MSE		1500.1 − 1407	1386.8 − 1234.8	1329.5 − 1198.2	1452	647.0
$n = 10000$						
C	36120	36120 − 36120	36120 − 36114	36120 − 36111.3	36120	35806.1
IC1	0	8357.3 − 7584.9	7586.1 − 6883.4	7586.1 − 6727.4	8382.9	4426.1
IC2	0	0 − 0.1	0.1 − 6.4	0.1 − 8.6	0	313.9
MSE		1439.7 − 1229.5	1229.1 − 858.6	1215.7 − 796.4	1359.5	558.5
$n = 30000$						
C	36120	36120 − 36013.6	36011.1 − 34936.9	36011.1 − 34704.5	34521.6	35837.5
IC1	0	7601 − 5417.4	5376.5 − 4070.8	5376.5 − 3952.3	7476.6	3982.4
IC2	0	0 − 106.8	109.4 − 867.2	109.4 − 956.5	2.6	282.5
MSE		1226.8 − 725	718.8 − 323.3	711.2 − 304.5	1035.3	468.2
$n = 50000$						
C	36120	36119.1 − 35305.9	35073.9 − 34229.3	35073.9 − 33915.5	35742.1	35817.1
IC1	0	5843.3 − 2987.5	4049.2 − 2651.5	4049.2 − 2510.3	7047.9	3836.5
IC2	0	0.6 − 756.5	957.5 − 1018.1	957.5 − 1204.5	32.7	302.8
MSE		838.9 − 508.6	438.8 − 256.7	436.3 − 255.8	725.3	441.1

the $p \times p$ variance covariance matrix of the p-dimensional vector of asset returns and ι is a $p \times 1$ vector of 1's. As a function depending only on Σ, the GMVP performance essentially depends on the precise measurement of Σ.

To evaluate the out-of-sample forecasting performances of Σ, we consider a portfolio of monthly financial returns composed of the MSCI stock index based on the sample December 1998-March 2018, which yields a total sample size $T = 231$, and for the following 23 countries: Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Greece, Hong Kong, Ireland, Italy, Japan, Netherlands, New Zealand, Norway, Portugal, Singapore, Spain, Sweden, Switzerland, the United-Kingdom, the United-States. For each variance covariance model, we estimate the portfolio weights using a 60 month rolling window, idest corresponding to 5 years of observed returns. For each time t, we use the observed returns based on the last $h = 60$ months, which then allows to compute the GMVP $\hat{\omega}$ based on an estimate of Σ. This estimate is based on the factor decomposition approach, where we consider the POET, the adaptive Lasso PML, our MCP/SCAD/Lasso two-step approach denoted by 2S-MCP, 2S-SCAD, 2S-Lasso, respectively, and a 60 month rolling window sample variance covariance matrix, denoted by Sample. The tuning parameters for the penalised methods are set to obtain the best performance for each method (for the POET, $(m, c) = (2, 0.2)$; for PML, $(m, c) = (2, 0.4)$; for the proposed two-step method, $(m, c) = (3, 0.3)$; here m is the number of factors and c the constant scaling the rate $1/\sqrt{p} + \sqrt{\log(p(p+1)/2)/n}$ for the POET and $\sqrt{\log(p(p+1)/2)/n}$ for the other cases). We propose an out-of-sample analysis only for the least squares based method for the following two reasons: first, in light of the performances in terms of variable selection and MSE of Tables 1 and 2, the least squares based sparse estimator of Ψ is more desirable; due to the lack of curvature of the Gaussian based loss function, which translates into an extremely low parameter α_1 - whose analytical expression is given in 3.3 -, then the Gaussian loss based estimator $\hat{\Psi}^g$ is highly sensitive to the initial value $\Psi^{(0)}$ of the algorithm, which is in line with Figures 5.(c) and 5.(d), Section 5 of [21], and Figures 2.(c) and 2.(d), Section 5 of [22]. Finally, the equally weighted portfolio approach, denoted by $1/p$, is reported as an alternative investment strategy.

Let r_t be the vector of monthly returns based on the 23 stocks at time t. For each method, we compute the $t+1$ out-of-sample portfolio return as $\hat{r}_{t+1} := \hat{w}' r_{t+1}$. Thus, using the $m = T - h = 231 - 60$ out-of-sample portfolio returns, we then compute the empirical average return $\hat{\mu}$ and the empirical variance $\hat{\sigma}^2$, for each method, as follows:

$$\hat{\mu} = \frac{1}{m} \sum_{t=60+1}^{231} \hat{r}_t, \quad \hat{\sigma}^2 = \frac{1}{m-1} \sum_{t=60+1}^{231} (\hat{r}_t - \hat{\mu})^2.$$

The out-of-sample average return (AVG) and standard deviation (SD) are used as the criteria for the performance evaluation.

The annualised results of the portfolio analysis are reported in Table 3. First, although the minimum variance method may not always provide better AVG performances with respect to the $1/p$ approach - the expected return does not

TABLE 3
Estimated GMVP performances of the proposed 2S-methods and competing methods.

Method	Sample	$1/p$	POET	PML	2S-MCP	2S-SCAD	2S-Lasso
AVG	0.0996	0.0540	0.1067	0.1153	0.1332	0.1337	0.1329
SD	0.1138	0.1719	0.0976	0.1272	0.1261	0.1258	0.1261

intervene in the GMVP strategy optimisation -, our proposed method significantly outperforms the competing models in terms of AVG. The performances are more mixed in terms of SD: although our proposed approach outperforms the PML, the likelihood based methods (both two-step and PML) are outperformed by the POET in the SD sense.

7. Discussion

The focus of this paper is devoted to the sparse estimation of Ψ using the two-step approach (2.2) and as an alternative (2.3), where the diagonal assumption on the latter matrix is relaxed. Our main contributions consisted in the derivation of error bounds in the $\ell_2, \ell_1, \ell_\infty$ senses as well as the conditions for support recovery of the true sparse support of Ψ_0. To obtain such bounds, the restricted strong convexity of the loss functions $\mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0), \mathbb{F}_{n,p}(\tilde{\Lambda}; \Psi_0)$ is a key property: Corollaries 3.3 and 3.4 partly focused on verifying such regularity condition. Furthermore, non-convex penalty functions are more desirable in terms of support recovery since the so-called incoherence condition can be relaxed as emphasized in Corollaries 4.1 and 4.2.

Various issues and extensions can be further considered. Rather than only penalising the idiosyncratic variance covariance matrix, penalising both the loading factor matrix Λ and Ψ would be relevant. However, a careful treatment of the rotational indeterminacy should be carried out when regularizing Λ. To avoid the rotational indeterminacy, we need to consider an appropriate constraint on Λ (e.g., IC5 in [6]). However, the sparsity of Λ also highly depends on the constraint on Λ. We may obtain a sparse solution under the constraint but not the simplest and sparsest solution, which could be easily interpreted. Without constraints on Λ, we cannot ensure the positive definiteness of the Hessian of the loss function, and thus obtaining theoretical guarantees is challenging. Addressing this issue can be part of a future work.

Acknowledgements

This work was supported by the Japan Society for the Promotion of Science (B.Poignard was supported by Grant 19K23193; Y. Terada was supported by Grants 16K16024 and 20K19756). The authors gratefully acknowledge the helpful comments of the Associate Editor and referees.

References

[1] Abadir, K.M. and Magnus, J.R. (2005). *Matrix algebra.* Cambridge University Press. MR2408356

[2] Anderson, T.W. and Amemiya, Y. (1988). *The asymptotic normal distribution of estimators in factor analysis under general conditions.* The Annals of Statistics, Vol. 16, No. 2, 759-771. MR0947576

[3] Bai, J. (2003). *Inferential theory for factor models of large dimensions.* Econometrica, Vol. 71, 135–171. MR1956857

[4] Bai, J. and Li, K. (2012). *Statistical analysis of factor models of high dimension.* The Annals of Statistics, Vol. 40, No. 1, 436-465. MR3014313

[5] Bai, J. and Li, K. (2016). *Maximum likelihood estimation and inference for approximate factor models of high dimension.* The Review of Economics and Statistics, Vol. 98, No. 2.

[6] Bai, J. and Liao, K. (2016). *Efficient estimation of approximate factor models via penalised maximum likelihood.* Journal of Econometrics, Vol. 191, 1-18. MR3434432

[7] Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications.* Berlin: Springer Series in Statistics. MR2807761

[8] Candès, E.J and Plan, Y. (2009). *Near-ideal model selection by ℓ_1 minimization.* The Annals of Statistics, Vol. 37, No. 5A, 2145-2177. MR2543688

[9] Chamberlain, G., and Rothschild, M. (1983). *Arbitrage, factor structure and mean–variance analysis in large asset markets.* Econometrica, Vol. 51, No. 5, 1305–1324 MR0736050

[10] Fan, J., Fan, Y. and Lv, J. (2008). *Large dimensional covariance matrix estimation using a factor model.* Journal of Econometrics, Vol. 147, 186–197. MR2472991

[11] Fan, J., Feng, Y. and Wu, Y. (2009). *Network exploration via the adaptive lasso and scad penalties.* The Annals of Applied Statistics, Vol. 3, No. 2, 521-541. MR2750671

[12] Fan, J. and Li, R. (2001). *Variable selection via nonconcave penalised likelihood and its oracle properties.* Journal of the American Statistical Association, Vol. 96, 1348-1360. MR1946581

[13] Fan, J., Liao, Y. and Liu, H. (2016). *An overview of the estimation of large covariance and precision matrices.* The Econometrics Journal, Vol. 19, 1-32. MR3501529

[14] Fan, J., Liao, Y. and Mincheva, M. (2011). *High-dimensional covariance matrix estimation in approximate factor models.* The Annals of Statistics, Vol. 39, No. 6, 3320-3356. MR3012410

[15] Fan, J., Liao, Y. and Mincheva, M. (2013). *Large covariance estimation by thresholding principal orthogonal complements.* Journal of the Royal Statistical Society Series B, Statistical Methodology, Vol. 75, No. 4. MR3091653

[16] Fan, J., Zhang, J. and Yu, K. (2012). *Vast portfolio selection with gross-exposure constraints.* Journal of the American Statistical Association, Vol.

107, 592-606. MR2980070

[17] Goldberg, L.R. (1992) *The development of markers for the Big-Five factor structure.* Psychological assessment Vol. 4 No. 1, 26–42.

[18] Harman, H.H. (1967) Modern factor analysis (2nd ed.), University of Chicago Press. MR0229335

[19] Lawley, D.N.and Maxwell, A.E. (1971) Factor Analysis as a Statistical Method (2nd ed.), Elsevier. MR0343471

[20] Loh, P.L. (2017). *Statistical consistency and asymptotic normality for high-dimensional robust M-estimators.* The Annals of Statistics, Vol. 45, No. 2, 866-896. MR3650403

[21] Loh, P.L. and Wainwright, M.J. (2015). *Regularised M-estimators with non-convexity: statistical and algorithmic theory for local optima.* Journal of Machine Learning Research, Vol. 16, 559-616. MR3335800

[22] Loh, P.L. and Wainwright, M.J. (2017). *Support recovery without incoherence: a case for non-convex regularisation.* The Annals of Statistics, Vol. 45, No. 6, 2455-2482. MR3737898

[23] Merlevède, F., Peligrad, M. and Rio, E. (2009). *Bernstein inequality and moderate deviations under strong mixing conditions.* Institute of Mathematical Statistics Collections, High Dimensional Probability, Vol. 5, 273-292. MR2797953

[24] Negahban, S.N, Ravikumar, P., Wainwright, M.J., and Yu, B. (2012). *A unified framework for high-dimensional analysis of M-estimators with decomposable regularisers.* Statistical Science, Vol. 27, No. 4, 538-557. MR3025133

[25] Poignard, B. and Fermanian, J.D. (2018). *Finite sample properties of Sparse M-estimators with Pseudo-Observations.* Working Paper CREST.

[26] Ravikumar, P., Wainwright, M.J. and Lafferty, J.D. (2010). *High-dimensional Ising model selection using ℓ_1-regularised logistic regression.* The Annals of Statistisc, Vol. 38, 1287–1319. MR2662343

[27] Ravikumar, P., Wainwright, M.J., Raskutti, G. and Yu, B. (2011). *High-dimensional covariance estimation by minimizing ℓ_1-penalised log-determinant divergence.* Electronic Journal of Statistics, Vol. 5, 935–980. MR2836766

[28] Ross, S. A. (1976). *The arbitrage theory of capital asset pricing.* Journal of Economic Theory, Vol. 13, 341–360. MR0429063

[29] van de Geer, S. (2016). *Estimation and testing under sparsity.* École d'Éte de Saint-Flour XLV, Springer. MR3526202

[30] Wainwright, M.J. (2009). *Sharpe thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso).* IEEE Transactions on Information Theory, Vol. 55, No. 5, 2183-2202. MR2729873

[31] Zhang, C.-H. (2010). *Nearly unbiased variable selection under minimax concave penalty.* The Annals of Statistics, Vol. 38, 894-942. MR2604701

[32] Zhao, P. and Yu, B. (2006). *On model selection consistency of Lasso.* Journal of Machine Learning Research, Vol. 7, 2541-2567. MR2274449

Appendix A: Technical appendix

A.1. Intermediary results

We provide the primal dual witness method (PWD), as in [22]. The parameter vector $\theta \in \mathbb{R}^d$ belongs to the convex set Θ. The PDW approach relies on the following steps.

Step 1. We define the estimator

$$\hat{\theta}_{\mathcal{A}} = \underset{\theta \in \mathbb{R}^{|\mathcal{A}|}: g(\theta) \leq R, \theta \in \Omega}{\arg\ \min} \Big\{ \mathbb{L}_n(\theta) + p(\gamma_n, \theta) \Big\}. \tag{A.1}$$

We solve problem (A.1), under the constraint $\hat{\mathcal{A}} \subseteq \mathcal{A}$ and prove $\|\hat{\theta}_{\mathcal{A}}\|_1 < R$.

Step 2. Defining $\hat{z}_{\mathcal{A}} \in \partial \|\hat{\theta}_{\mathcal{A}}\|$, we choose $\hat{z}_{\mathcal{A}^c}$ satisfying the orthogonality condition

$$\nabla_\theta \mathbb{L}_n(\hat{\theta}) - \nabla_\theta q(\gamma_n, \hat{\theta}) + \gamma_n \hat{z} = 0, \tag{A.2}$$

with $\hat{z} = (\hat{z}_{\mathcal{A}}, \hat{z}_{\mathcal{A}^c})$, $\hat{\theta} = (\hat{\theta}_{\mathcal{A}}, 0_{\mathcal{A}^c})$ and $q(\gamma_n, \rho) = \gamma_n |\rho| - p(\gamma_n, \rho)$ for $\rho \in \mathbb{R}$. Note that the vector version of q is given in assumption 5. We then prove the strict dual feasibility $\|\hat{z}_{\mathcal{A}^c}\|_\infty < 1$.

Step 3. We prove that $\hat{\theta}$ is a local optimum of (3.1) and that any stationary point of (3.1) satisfies $\text{supp}(\hat{\theta}) \subseteq \mathcal{A}$.

The PDW procedure does not allow for *practically* solving the regularisation problem (3.1) as step 1 requires to know the true subset model $\mathcal{A} = \text{supp}(\theta_0) = \{i : \theta_{0,i} \neq 0\}$. However, this approach is useful as a proof method to characterize the optimal solution $\hat{\Psi}$. In **Step 1**, the criterion (A.1) is striclty convex under the RSC condition. This implies that for $\|\hat{\theta}_{\mathcal{A}}\|_1 < 1$, the subgradient condition (A.2) must hold at $\hat{\theta}_{\mathcal{A}}$ for the restricted problem (A.1). [22] proves that, although problem A.1 may be non-convex, the RSC condition and regularity conditions on the penalty function allow them to prove that the optimum obtained in **Step 3** is a local optimum: see in particular their Lemma 10.

Using optimization reasoning, [22] provide conditions on γ_n, R to ensure the success of the PDW technique, which depends on **Step 3**, under the assumption that $\mathbb{L}_n(.)$ satisfies the RSC condition with parameters $(\alpha_k, \tau_k)_{k=1,2}$ and $4\alpha_1 > 3\mu$. Indeed, these conditions guarantee that the support of $\hat{\theta}$ satisfying (A.2) in **Step 2** is the unique stationary point of the criterion (3.1): to be precise, the first condition concerns the suitable scaling of γ_n and R; the second condition ensures strict dual feasibility - that is $\|\hat{z}_{\mathcal{A}^c}\|_\infty < 1$ in **Step 2**. This is the object of the following Theorem.

Theorem A.1 ([22])**.** *Suppose $\mathbb{L}_n(.)$ satisfies the RSC condition with $(\alpha_k, \tau_k)_{k=1,2}$ parameters and $p(\gamma_n, .)$ is a μ-amenable penalty, with $0 \leq \mu < \alpha_1$. Suppose*

(i) The parameters (γ_n, R) satisfy

$$4 \max \left\{ \|\nabla_\theta \mathbb{L}_n(\theta_0)\|_\infty, \alpha_2 \sqrt{\frac{\log k_0}{n}} \right\} \leq \gamma_n \leq \sqrt{\frac{(4\alpha_1 - 3\mu)\alpha_2}{384 k_0}}, \tag{A.3}$$

$$\max\Big\{2\|\theta_0\|_1, \frac{48k_0\gamma_n}{4\alpha_1 - 3\mu}\Big\} \leq R \leq \min\Big\{\frac{\alpha_2}{8\gamma_n}, \frac{\alpha_2}{\tau_2}\sqrt{\frac{n}{\log p}}\Big\}. \tag{A.4}$$

(ii) For some $\delta \in [\dfrac{4R\tau_1 \log d}{n\gamma_n}, 1]$, the vector $\hat{z}$ from the PDW construction satisfies the strict dual feasibility condition

$$\|\hat{z}_{\mathcal{A}^c}\|_\infty \leq 1 - \delta. \tag{A.5}$$

Then for any k_0-sparse vector θ_0, the program (3.1) with a sample size $n \geq \frac{2\tau_1}{2\alpha_1 - \mu} k_0 \log d$ has a unique stationary point given by the primal output $\hat{\theta}$ of the PDW construction.

Suitable calibrations of γ_n and R, and thus a proper scaling (n, d, k_0), are necessary. Using exponential bounds, it is possible to evaluate the probability of satisfying (A.3) and (A.4) and thus the probability of the PDW success. In all their applications of interest - linear model, generalized linear model, Gaussian graphical Lasso - [22] obtain the upper bound $\|\nabla_\theta \mathbb{L}_n(\theta_0)\|_\infty \leq C\sqrt{\log d/n}$ with high probability. This motivates the choice γ_n proportional to $\sqrt{\log d/n}$ to satisfy (A.3). Finally, it is worth noting that the trade-off between the curvature of the loss function through α_1 and the non-convexity degree of the penalty function through μ appears. As our simulations emphasize this trade-off for the Gaussian loss in particular, significantly large values for b_{scad}, b_{mcp} are necessary to ensure $4\alpha_1 > 3\mu$.

In their Theorem 2, [22] provide an additional error bound under the conditions of Theorem A.1. It also provides the guarantees that the unique optimum - local or global - (3.1) is the oracle estimator. The latter is defined as the non-penalised estimator obtained from minimizing the criterion $\mathbb{L}_n(\theta)$ over the true support $\mathcal{A}$. This is the object of the following Theorem.

Theorem A.2 ([22]). *Under the conditions of Theorem A.1, suppose strict dual feasibility (A.5) holds, suppose $p(\gamma_n, .)$ is μ-amenable with $\mu \in [0, \alpha_1)$. Then the unique stationary solution of (3.1) satisfies*

(i)

$$\|\hat{\theta} - \theta_0\|_\infty \leq \|\hat{K}^{-1}_{\mathcal{A}\mathcal{A}} \nabla_\theta \mathbb{L}_n(\theta_0)\|_\infty + \gamma_n \|\hat{K}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty,$$

with $\hat{K} = \int_0^1 \nabla^2_{\theta\theta'} \mathbb{L}_n(\theta_0 + u(\hat{\theta} - \theta_0))du$.

(ii) If $p(\gamma_n, .)$ is (μ, ζ)-amenable and if the lower bound

$$\min_{i \in \mathcal{A}} |\theta_{0,i}| \geq \gamma_n\big(\zeta + \|\hat{K}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty\big) + \|\hat{K}^{-1}_{\mathcal{A}\mathcal{A}} \nabla_\theta \mathbb{L}_n(\theta_0)\|_\infty,$$

holds, then $\hat{\theta}$ agrees with the oracle estimator $\hat{\theta}^{\mathcal{O}}$ and we have the bound

$$\|\hat{\theta} - \theta_0\|_\infty \leq \|\hat{K}^{-1}_{\mathcal{A}\mathcal{A}} \nabla_\theta \mathbb{L}_n(\theta_0)\|_\infty.$$

These inequalities are expressed in a deterministic manner. As in Theorem A.1, exponential bounds allow for upper bounding $\|\nabla_\theta \mathbb{L}_n(\theta_0)\|_\infty$, which will provide explicit convergence rates over the $\|.\|_\infty$-error. The application of Theorem A.2 requires that strict dual feasibility holds under the RSC condition. In their Proposition 1, [22] provide sufficient conditions to satisfy strict dual feasibility in the case of a (μ,ζ)-amenable regulariser, which thus allows for using Theorem A.2. These conditions are given in the following Proposition.

Proposition A.3 ([22])**.** *Under the conditions of Theorem A.1, suppose $p(\gamma_n,.)$ is (μ,ζ)-amenable. Suppose*

$$\theta_{0,\min} \geq \gamma_n(\zeta + \|\hat{K}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty) + \|\hat{K}^{-1}_{\mathcal{A}\mathcal{A}}\nabla_\theta \mathbb{L}_n(\theta_0)_{\mathcal{A}}\|_\infty,$$

with $\theta_{0,\min} = \min\limits_{i\in\mathcal{A}}|\theta_i|$ and $\hat{K} = \int_0^1 \nabla^2_{\theta\theta'}\mathbb{L}_n(\theta_0 + u(\hat{\theta}-\theta_0))du$. Then strict dual feasibility holds provided

$$\|\nabla_\theta \mathbb{L}_n(\theta_0)\|_\infty \leq \frac{1-\delta}{2}\gamma_n, \qquad \textit{and} \tag{A.6}$$

$$\|\hat{K}_{\mathcal{A}^c\mathcal{A}}\hat{K}^{-1}_{\mathcal{A}\mathcal{A}}\nabla_\theta \mathbb{L}_n(\theta_0)_{\mathcal{A}}\|_\infty \quad \leq \quad \frac{1-\delta}{2}\gamma_n. \tag{A.7}$$

A.2. Proofs

Proof of Lemma 3.2. Since we have

$$\begin{aligned} & \mathbb{E}\left[|X-\mathbb{E}[X]|^m\right] \\ \leq \quad & \mathbb{E}\left[\{|X|+|\mathbb{E}[X]|\}^m\right] \leq 2^{m-1}\{\mathbb{E}[|X|^m] + |\mathbb{E}[X]|^m\} \leq 2^m\mathbb{E}[|X|^m], \end{aligned}$$

we obtain

$$\mathbb{E}\left[|X_{ki}X_{kj}-\sigma_{ij}|^m\right] \leq 2^m\mathbb{E}\left[|X_{ki}X_{kj}|^m\right] \leq \left\{\mathbb{E}[|X_{ki}|^{2m}]\mathbb{E}[|X_{kj}|^{2m}]\right\}^{1/2}.$$

From $\mathbb{E}[\exp(t|X|)] - 1 \geq t^k\mathbb{E}[|X|^k]/k!$ for $t \geq 0$, we have

$$\mathbb{E}[|X|^{2m}]/K^{2m} \leq m!\left(\mathbb{E}[\exp(|X|^2/K^2)] - 1\right).$$

Combining these facts with the sub-Gaussian assumption, we obtain

$$\mathbb{E}\left[|X_{ki}X_{kj}-\sigma_{ij}|^m\right] \leq 2^m m!\sigma_0^2 K^{2(m-1)} = \frac{m!}{2}(2K^2)^{m-2}(2\sqrt{2}K\sigma_0)^2. \tag{A.8}$$

Let $\gamma_{ij}(X_k) := (X_{ki}X_{kj}-\sigma_{ij})/(2\sqrt{2}K\sigma_o)$ and we have $\mathbb{E}[\gamma_{ij}(X_k)] = 0$. Moreover, by (A.8)

$$\mathbb{E}\left[|\gamma_{ij}(X_k)|^m\right] \leq \frac{m!}{2}\left(\frac{K}{\sqrt{2}\sigma_0}\right)^{m-2} =: \frac{m!}{2}K_*^{m-2}.$$

By Lemma 14.3 in [7],

$$\mathbb{P}\left(\max_{i\leq j}\left|\frac{1}{n}\sum_{k=1}^n \gamma_{ij}(X_k)\right| \geq K_*t + \sqrt{2t} + \lambda\left(K_*, n, \binom{p}{2}\right)\right) \leq \exp(-nt),$$

where

$$\lambda(K,n,p) := \sqrt{\frac{2\log(2p)}{n}} + \frac{K\log(2p)}{n}.$$

Therefore, denoting

$$h(t;n,p,K,\sigma_0^2) = 2K^2t + 4K\sigma_0\sqrt{t} + 2\sqrt{2}K\sigma_0\lambda\left(K/(\sqrt{2}\sigma_0), n, \binom{p}{2}\right),$$

we obtain

$$\mathbb{P}\left(\|\hat{S}-\Sigma_0\|_{\max} \geq h(t;n,p,K,\sigma_0^2)\right) \leq \exp(-nt) \qquad \square$$

Proof of Corollary 3.3. We first establish the RSC property. To do so, we derive the first and second order derivatives of the Gaussian QML function $\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi)$ defined in (2.2). Using the differential operator applied with respect to Ψ, for any fixed $\tilde{\Lambda}$, we obtain

$$\begin{aligned} & d\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi) \\ = \; & \frac{1}{2p}\big(\text{tr}(\Sigma(\tilde{\Lambda},\Psi)^{-1}d\Sigma(\tilde{\Lambda},\Psi)) - \text{tr}(\Sigma(\tilde{\Lambda},\Psi)^{-1}d\Sigma(\tilde{\Lambda},\Psi)\Sigma(\tilde{\Lambda},\Psi)^{-1}\hat{S})\big). \end{aligned}$$

Moreover, we have

$$A := d\Sigma(\tilde{\Lambda},\Psi) = d\Psi.$$

Using the trace operator property $\text{tr}(X'Y) = \text{tr}(XY') = \text{tr}(Y'X)$, we obtain

$$d\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi) = \frac{1}{2p}\text{tr}(\Sigma(\tilde{\Lambda},\Psi)^{-1}\big(\Sigma(\tilde{\Lambda},\Psi)-\hat{S}\big)\Sigma(\tilde{\Lambda},\Psi)^{-1}(d\Psi)).$$

Hence in vech(.) form, the derivative becomes

$$\nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi) = \frac{1}{2p}\text{vech}(\Sigma(\tilde{\Lambda},\Psi)^{-1}\big(\Sigma(\tilde{\Lambda},\Psi)-\hat{S}\big)\Sigma(\tilde{\Lambda},\Psi)^{-1}).$$

Taking the $\|.\|_\infty$ norm, we have on the true parameter Ψ_0

$$\begin{aligned} & \|\nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0)\|_\infty = \|\nabla_\Psi\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0)\|_{\max} \\ \leq \; & \|\nabla_\Psi\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0)\|_s \leq \|\Sigma(\tilde{\Lambda},\Psi_0)^{-1}\|_s^2\|\Sigma(\tilde{\Lambda},\Psi_0)-\hat{S}\|_s/(2p). \end{aligned}$$

We also have $\lambda_{\max}(\Sigma(\tilde{\Lambda},\Psi_0)^{-1}) = \lambda_{\max}((\tilde{\Lambda}\tilde{\Lambda}'+\Psi_0)^{-1}) \leq \lambda_{\max}(\Psi_0^{-1})$. Hence, we obtain $\|\nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0)\|_\infty \leq \lambda_{\max}(\Psi_0^{-1})^2\|\Sigma(\tilde{\Lambda},\Psi_0)-\hat{S}\|_s/(2p)$. We now focus on the Hessian matrix. Omitting the arguments in Σ, the second order differential is given by

$$\begin{aligned} d^2\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi) \;=\; & \tfrac{1}{2p}\text{tr}\Big(-\Sigma^{-1}(d\Psi)\Sigma^{-1}(d\Psi) + \Sigma^{-1}(d\Psi)\Sigma^{-1}\hat{S}\Sigma^{-1}(d\Psi) \\ & +\Sigma^{-1}\hat{S}\Sigma^{-1}(d\Psi)\Sigma^{-1}(d\Psi)\Big) \end{aligned}$$

We aim at extracting the form $\text{tr}(L(d\Lambda)'M(d\Lambda))$ for L (resp. M) any square $m\times m$ matrix (resp. $p\times p$). Since $d\Psi = (d\Psi)'$, we have

$$\nabla^2_{\text{vec}(\Psi)\text{vec}(\Psi)'}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi)$$

$$= \quad \frac{1}{2p}\text{vec}\big(\{\Sigma^{-1}\otimes\Sigma^{-1}\big(\hat{S}-\Sigma\big)\Sigma^{-1}\}+\{\Sigma^{-1}\hat{S}\Sigma^{-1}\otimes\Sigma^{-1}\}\big),$$

which can be expressed in vech(.) form as

$$\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi)=\frac{1}{2p}D'_p\big(\{\Sigma^{-1}\otimes\Sigma^{-1}\big(\hat{S}-\Sigma\big)\Sigma^{-1}\}+\{\Sigma^{-1}\hat{S}\Sigma^{-1}\otimes\Sigma^{-1}\}\big)D_p,$$

where D_p is the $p^2\times p(p+1)/2$ duplication matrix, which allows for the treatment of redundant terms: see exercise 13.65 of [1]. For some $\Psi_1\in\Omega$ and $u\in[0,1]$, let us define $\Psi=\Psi_0+u\Gamma$ where $\Gamma=\Psi_1-\Psi_0$. Then $\Psi\in\Omega$ and

$$\begin{aligned} & f_n(\Psi):=\text{vech}(\Gamma)'\Big\{\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi)\Big\}\text{vech}(\Gamma)\\ \geq\quad & \text{vech}(\Gamma)'D'_p\Big(\{\Sigma^{-1}\otimes\Sigma^{-1}\big(\hat{S}-\Sigma\big)\Sigma^{-1}\}\\ & +\{\Sigma^{-1}\hat{S}\Sigma^{-1}\otimes\Sigma^{-1}\}\Big)D_p\text{vech}(\Gamma)/(2p)\\ \geq\quad & \|\text{vech}(\Gamma)\|_2^2\lambda_{\min}\big(2\hat{S}-\Sigma\big)\lambda_{\min}(\Sigma^{-1})^3/(2p), \end{aligned}$$

since the spectrum of $A\otimes B$ is the cross product of the spectrums of A and B, and $\lambda_{\min}(\Psi)=\inf_x x'\Psi x/\|x\|_2$. We now focus on $\lambda_{\min}(\Sigma^{-3})$, where we have

$$\lambda_{\max}(\Sigma)\leq\lambda_{\max}(\tilde{\Lambda}\tilde{\Lambda}')+\lambda_{\max}(\Psi)\leq\lambda_{\max}(\tilde{\Lambda}\tilde{\Lambda}')+\lambda_{\max}(\Psi_0)+\lambda_{\max}(\Psi_1-\Psi_0),$$

which implies

$$\lambda_{\min}(\Sigma^{-3})\geq\{\lambda_{\max}(\tilde{\Lambda}\tilde{\Lambda}')+\lambda_{\max}(\Psi)\}^{-3}\geq\{\lambda_{\max}(\tilde{\Lambda}\tilde{\Lambda}')+\lambda_{\max}(\Psi_0)+1\}^{-3}.$$

Therefore

$$\begin{aligned} & f_n(\Psi)\geq\|\text{vech}(\Gamma)\|_2^2\{\lambda_{\max}(\tilde{\Lambda}\tilde{\Lambda}')+b_2\}^{-3}\lambda_{\min}\big(2\hat{S}-\Sigma\big)\\ \geq\quad & \|\text{vech}(\Gamma)\|_2^2\{\lambda_{\max}(\tilde{\Lambda}\tilde{\Lambda}')+\lambda_{\max}(\Psi_0)+1\}^{-3}a/p. \end{aligned}$$

We thus deduce that

$$\begin{aligned} & \text{vech}(\Psi-\Psi_0)'\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi^*)\text{vech}(\Psi-\Psi_0)\\ \geq\quad & \|\text{vech}(\Gamma)\|_F^2\{\lambda_{\max}(\tilde{\Lambda}\tilde{\Lambda}')+\lambda_{\max}(\Psi_0)+1\}^{-3}a/(2p), \end{aligned}$$

where Ψ^* lies between Ψ and Ψ_0. The RSC condition would thus be satisfied for the parameters

$$\alpha_1=\{\lambda_{\max}(\tilde{\Lambda}\tilde{\Lambda}')+\lambda_{\max}(\Psi_0)+1\}^{-3}a/(2p),\ \alpha_2=\alpha_1,\ \tau_1=\tau_2=0.$$

This bound holds for all $\text{vech}(\Gamma)\in\mathbb{R}^{p(p+1)/2}$. The RSC condition also holds for $\|\text{vech}(\Gamma)\|_2\geq 1$ by Lemma 9 of [21]. Hence, based on Theorem 3.1, we obtain the desired upper bounds for the $\|.\|_1$ and $\|.\|_2$ errors.

We now evaluate the probability so that condition (3.2) is satisfied. The Jacobian is upper bounded by

$$\|\nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0)\|_\infty=\|\nabla_\Psi\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0)\|_{\max}$$

$$\leq \quad \lambda_{\max}(\Psi_0^{-1})^2 \|\Sigma(\tilde{\Lambda}, \Psi_0) - \hat{S}\|_s/(2p).$$

Moreover, $\Sigma(\tilde{\Lambda}, \Psi_0) - \hat{S} = \Sigma_0 - \hat{S} + \tilde{\Lambda}\tilde{\Lambda}' - \Lambda_0\Lambda_0'$, which implies

$$\|\Sigma(\tilde{\Lambda}, \Psi_0) - \hat{S}\|_s \leq \|\Sigma_0 - \hat{S}\|_s + \|\tilde{\Lambda}\tilde{\Lambda}' - \Lambda_0\Lambda_0'\|_s.$$

Besides

$$\|\tilde{\Lambda}\tilde{\Lambda}' - \Lambda_0\Lambda_0'\|_F = \|\tilde{\Lambda}(\tilde{\Lambda} - \Lambda_0)' + (\tilde{\Lambda} - \Lambda_0)\Lambda_0'\|_F \leq 2\max\{\|\tilde{\Lambda}\|_F, \|\Lambda_0\|_F\}\|\tilde{\Lambda} - \Lambda_0\|_F,$$

and under the first step probability bounds are $\|\tilde{\Lambda} - \Lambda_0\|_F = O_p(\sqrt{\frac{p}{n}}) + O_p(\sqrt{\frac{1}{p}})$, for $C > 0$ sufficiently large

$$\|\tilde{\Lambda}\|_F \leq \|\tilde{\Lambda} - \Lambda_0\|_F + \|\Lambda_0\|_F \leq C(\sqrt{\frac{p}{n}} + \sqrt{\frac{1}{p}}) + \|\Lambda_0\|_F.$$

We thus obtain for C sufficiently large

$$\begin{aligned} \|\tilde{\Lambda}\tilde{\Lambda}' - \Lambda_0\Lambda_0'\|_F \quad &\leq \quad 2(C(\sqrt{\tfrac{p}{n}} + \sqrt{\tfrac{1}{p}}) + \|\Lambda_0\|_F)C(\sqrt{\tfrac{p}{n}} + \sqrt{\tfrac{1}{p}}) \\ &\leq \quad 2\{C^2\big(\tfrac{p}{n} + 2\sqrt{\tfrac{1}{n}} + \tfrac{1}{p}\big) + C\|\Lambda_0\|_F(\sqrt{\tfrac{p}{n}} + \sqrt{\tfrac{1}{p}})\} \end{aligned}$$

Besides, $\|\Sigma_0 - \hat{S}\|_s \leq p\|\Sigma_0 - \hat{S}\|_{\max}$, thus for $C, K > 0$ sufficiently large, with probability at least $1 - \exp(-\log p)$, since $\|\Lambda_0\|_F$ is of p-order, we obtain

$$\begin{aligned} & \|\nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)\|_\infty = \|\nabla_\Psi\mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)\|_{\max} \leq \|\nabla_\Psi\mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)\|_s \\ \leq \quad & \frac{\|\Psi_0^{-1}\|_s^2}{2p}\Big(Kp\sqrt{\frac{\log p}{n}} + 2\{C^2\big(\frac{p}{n} + 2\sqrt{\frac{1}{n}} + \frac{1}{p}\big) + C\|\Lambda_0\|_F(\sqrt{\frac{p}{n}} + \sqrt{\frac{1}{p}})\}\Big) \\ \leq \quad & L\sqrt{\frac{p}{n}}, \end{aligned}$$

for $L > 0$ sufficiently large, where we used the eigenvalue constraint in Ω, and sample size $n > Mp\lambda_{\max}(\Psi_0^{-1})^4$ for $M > 0$. Consequently, under the scaling assumption $\gamma_n \geq L\sqrt{\frac{p}{n}}$, we obtain

$$\begin{aligned} \|\text{vech}(\hat{\Psi}^g) - \text{vech}(\Psi_0)\|_2 \quad &\leq \quad \frac{6\gamma_n\sqrt{k_0}}{4\alpha_1 - 3\mu}, \\ \|\text{vech}(\hat{\Psi}^g) - \text{vech}(\Psi_0)\|_1 \quad &\leq \quad \frac{6(16\alpha_1 - 9\mu)\gamma_n k_0}{(4\alpha_1 - 3\mu)^2}, \end{aligned}$$

with probability at least $1 - \exp(-\log p)$. □

Proof of Corollary 3.4. We first establish the RSC condition. Using the differential operator with respect to Ψ, for any fixed $\tilde{\Lambda}$, we have

$$d\mathbb{F}_{n,p}(\tilde{\Lambda}; \Psi) = -\frac{1}{p}\text{tr}(\hat{\Sigma} - \tilde{\Lambda}\tilde{\Lambda}' - \Psi)(d\Psi).$$

Hence

$$\nabla_{\theta_\Psi}\mathbb{F}_{n,p}(\tilde{\Lambda};\Psi) = -\frac{1}{p}\text{vech}\big(\hat{\Sigma}-\tilde{\Lambda}\tilde{\Lambda}'-\Psi\big).$$

As for the Hessian, by identification, we obtain

$$\nabla^2_{\text{vec}(\Psi)\text{vec}(\Psi)'}\mathbb{F}_{n,p}(\tilde{\Lambda};\Psi) = \frac{1}{p}\text{vec}(I_p \otimes I_p),$$

or expressed with respect to θ_Ψ

$$\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{F}_{n,p}(\tilde{\Lambda};\Psi) = \frac{1}{p}D'_p(I_p \otimes I_p)D_p.$$

Thus

$$\text{vech}(\Psi-\Psi_0)'\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{F}_{n,p}(\tilde{\Lambda};\Psi^*)\text{vech}(\Psi-\Psi_0) \geq \|\text{vech}(\Gamma)\|_2^2/p,$$

where Ψ^* lies between Ψ and Ψ_0. The RSC condition would thus be satisfied for the parameters

$$\alpha_1 = \frac{1}{p},\ \alpha_2 = \alpha_1,\ \tau_1 = \tau_2 = 0.$$

We thus deduce the bounds (3.5). Now using the sub-Gaussian assumption, and using the same development on $\hat{\Sigma}-\tilde{\Lambda}\tilde{\Lambda}'-\Psi_0$ as in the proof of Corollary 3.3, we have

$$\begin{aligned}
&\|\nabla_{\theta_\Psi}\mathbb{F}_{n,p}(\tilde{\Lambda};\Psi_0)\|_\infty = \|\hat{\Sigma}-\tilde{\Lambda}\tilde{\Lambda}'-\Psi_0\|_{\max}/p \leq \|\hat{\Sigma}-\tilde{\Lambda}\tilde{\Lambda}'-\Psi_0\|_s/p \\
\leq\ & \Big(Kp\sqrt{\frac{\log p}{n}} + 2\{C^2(\frac{p}{n}+2\sqrt{\frac{1}{n}}+\frac{1}{p}) + C\|\Lambda_0\|_F(\sqrt{\frac{p}{n}}+\sqrt{\frac{1}{p}})\}\Big)/p \\
\leq\ & L\sqrt{\frac{p}{n}}.
\end{aligned}$$

Consequently, under the scaling assumption $\gamma_n \geq L\sqrt{\frac{p}{n}}$, then (3.5) hold with probability $1-\exp(-\log p)$. □

Proof of Corollary 4.1. Point (i). We aim at proving the strict dual feasibility condition. To do so, following the PDW construction, we consider the estimator

$$\hat{\Psi} = \underset{\Psi:\Psi\in\Omega,\text{supp}(\Psi)\subseteq\text{supp}(\Psi_0)}{\arg\ \min}\ \Big\{\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi) + p(\gamma_n,\text{vec}(\Psi))\Big\}, \tag{A.9}$$

where we take the penalisation with respect to $\text{vec}(\Psi)$ for the sake of clarification of our arguments, so that we will consider gradient and Hessian quantities with respect to $\text{vec}(\Psi)$ from now on. In the rest of the proof, we denote $\theta_\Psi := \text{vec}(\Psi)$. By the zero gradient condition (A.2) of the PDW step, we obtain

$$\nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\hat{\Psi}) - \nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0) + \nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0) - \nabla_{\theta_\Psi}q(\gamma_n,\text{vec}(\hat{\Psi})) + \gamma_n\hat{z} = 0.$$

This implies

$$\hat{K}\text{vec}(\hat{\Psi}-\Psi_0) + \nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0) - \nabla_{\theta_\Psi}q(\gamma_n,\text{vec}(\hat{\Psi})) + \gamma_n\hat{z} = 0,$$

with $\hat{K} = \int_0^1 \nabla^2_{\theta_\Psi \theta'_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0 + u(\hat{\Psi} - \Psi_0))du$. Equivalently, we have

$$\begin{pmatrix} \hat{K}_{\mathcal{A}\mathcal{A}} & \hat{K}_{\mathcal{A}\mathcal{A}^c} \\ \hat{K}_{\mathcal{A}^c\mathcal{A}} & \hat{K}_{\mathcal{A}^c\mathcal{A}^c} \end{pmatrix} \begin{pmatrix} \text{vec}(\hat{\Psi} - \Psi_0)_{\mathcal{A}} \\ \mathbf{0} \end{pmatrix}$$
$$+ \begin{pmatrix} \nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}} - \nabla_{\theta_\Psi} q(\gamma_n, \text{vec}(\hat{\Psi})_{\mathcal{A}}) \\ \nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}^c} - \nabla_{\theta_\Psi} q(\gamma_n, \text{vec}(\hat{\Psi})_{\mathcal{A}^c}) \end{pmatrix} + \gamma_n \begin{pmatrix} \hat{z}_{\mathcal{A}} \\ \hat{z}_{\mathcal{A}^c} \end{pmatrix} = \mathbf{0}.$$

Consequently, we obtain

$$\hat{z}_{\mathcal{A}^c} = \frac{1}{\gamma_n} \Big\{ \nabla_{\theta_\Psi} q(\gamma_n, \text{vec}(\hat{\Psi})_{\mathcal{A}^c}) - \nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}^c}$$
$$+ \hat{K}_{\mathcal{A}^c\mathcal{A}} \hat{K}^{-1}_{\mathcal{A}\mathcal{A}} \Big(\nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}} - \nabla_{\theta_\Psi} q(\gamma_n, \text{vec}(\hat{\Psi})_{\mathcal{A}}) + \gamma_n \hat{z}_{\mathcal{A}} \Big) \Big\}.$$

Using the regularity condition (v), we have

$$\nabla_{\theta_\Psi} q(\gamma_n, \text{vec}(\hat{\Psi})_{\mathcal{A}^c}) = \nabla_{\theta_\Psi} q(\gamma_n, \mathbf{0}_{\mathcal{A}^c}) = \mathbf{0}_{\mathcal{A}^c}.$$

This implies

$$\begin{aligned} & \hat{z}_{\mathcal{A}^c} \\ = \; & \frac{1}{\gamma_n} \Big\{ - \nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}^c} + \hat{K}_{\mathcal{A}^c\mathcal{A}} \hat{K}^{-1}_{\mathcal{A}\mathcal{A}} \Big(\nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}} \\ & - \nabla_{\theta_\Psi} q(\gamma_n, \text{vec}(\hat{\Psi})_{\mathcal{A}}) + \gamma_n \hat{z}_{\mathcal{A}} \Big) \Big\}. \end{aligned}$$

Taking the ℓ_∞-norm, we obtain

$$\begin{aligned} & \|\hat{z}_{\mathcal{A}^c}\|_\infty \\ \leq \; & \frac{1}{\gamma_n} \| - \nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}^c} + \hat{K}_{\mathcal{A}^c\mathcal{A}} \hat{K}^{-1}_{\mathcal{A}\mathcal{A}} \nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}} \|_\infty \\ + \; & \frac{1}{\gamma_n} \| \hat{K}_{\mathcal{A}^c\mathcal{A}} \hat{K}^{-1}_{\mathcal{A}\mathcal{A}} \big(\gamma_n \hat{z}_{\mathcal{A}} - \nabla_{\theta_\Psi} q(\gamma_n, \text{vec}(\hat{\Psi})_{\mathcal{A}}) \big) \|_\infty \\ \leq \; & \frac{1}{\gamma_n} \| - \nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}^c} + \hat{K}_{\mathcal{A}^c\mathcal{A}} \hat{K}^{-1}_{\mathcal{A}\mathcal{A}} \nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}} \|_\infty \\ & + \| \hat{K}_{\mathcal{A}^c\mathcal{A}} \hat{K}^{-1}_{\mathcal{A}\mathcal{A}} \|_\infty, \end{aligned}$$

using $\|\gamma_n \hat{z}_{\mathcal{A}} - \nabla_{\theta_\Psi} q(\gamma_n, \text{vec}(\hat{\Psi})_{\mathcal{A}})\|_\infty = \|\nabla_{\theta_\Psi} q(\gamma_n, \text{vec}(\hat{\Psi})_{\mathcal{A}})\|_\infty \leq \gamma_n$ from Lemma 8 of [22]. Furthermore, we have

$$\begin{aligned} & \| - \nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}^c} + \hat{K}_{\mathcal{A}^c\mathcal{A}} \hat{K}^{-1}_{\mathcal{A}\mathcal{A}} \nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}} \|_\infty \\ \leq \; & \| \nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0) \|_\infty + \| \hat{K}_{\mathcal{A}^c\mathcal{A}} \hat{K}^{-1}_{\mathcal{A}\mathcal{A}} \nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}} \|_\infty. \end{aligned}$$

To verify inequality (A.7), we have

$$\hat{K} - \nabla^2_{\theta_\Psi \theta'_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}, \Psi_0)$$

$$
\begin{aligned}
&= \int_0^1 \big(\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0+s(\hat{\Psi}-\Psi_0)) - \nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0)\big)ds \\
&= \int_0^1 s\nabla_{\theta_\Psi}\big(\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\bar{\Psi})\big)\text{vec}(\hat{\Psi}-\Psi_0)ds,
\end{aligned}
$$

where we used the mean value theorem with the matrix parameter $\bar{\Psi}$ satisfying $\|\bar{\Psi}-\Psi_0\|_F \leq \|\hat{\Psi}-\Psi_0\|_F$. Let $w = p^2$, for any $(u,v) \in \mathbb{R}^w \times \mathbb{R}^w$, for any fixed first step estimate $\tilde{\Lambda}$, we have

$$
\begin{aligned}
& |u'\Big\{\hat{K} - \nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda},\Psi_0)\Big\}v| \\
= \;& |\int_0^1 \Big\{s\sum_{i,j,k=1}^{w}\big(\nabla^3_{ijk}\mathbb{G}_{n,p}(\tilde{\Lambda};\bar{\Psi})\text{vec}(\hat{\Psi}-\Psi_0)_k u_i v_j\big)\Big\}ds| \\
\leq \;& \int_0^1 s|\Big\{\sum_{i,j,k=1}^{w}\big(\nabla^3_{ijk}\mathbb{G}_{n,p}(\tilde{\Lambda};\bar{\Psi})\text{vec}(\hat{\Psi}-\Psi_0)_k u_i v_j\big)\Big\}|ds,
\end{aligned}
$$

where ∇^3_{ijk} refers to the third order derivative with respect to θ_Ψ. Element by element, for any fixed first-step estimate $\tilde{\Lambda}$, for any $i,j,k = 1,\cdots,w$, the third order derivative applied with respect to θ_Ψ is given by

$$
\begin{aligned}
& \partial^3_{ijk}\mathbb{G}_{n,p}(\tilde{\Lambda};\bar{\Psi}) \\
= \;& \frac{1}{2p}\Big\{\text{tr}(\Sigma^{-1}(\partial_k\Sigma)\Sigma^{-1}(\partial_j\Sigma)\Sigma^{-1}(\partial_i\Sigma) + \Sigma^{-1}(\partial_j\Sigma)\Sigma^{-1}(\partial_k\Sigma)\Sigma^{-1}(\partial_i\Sigma))\Big\} \\
& -\frac{1}{2p}\Big\{\text{tr}(\Sigma^{-1}(\partial_k\Sigma)\Sigma^{-1}(\partial_j\Sigma)\Sigma^{-1}(\partial_i\Sigma)\Sigma^{-1}\hat{S} \\
& +\Sigma^{-1}(\partial_j\Sigma)\Sigma^{-1}(\partial_k\Sigma)\Sigma^{-1}(\partial_i\Sigma)\Sigma^{-1}\hat{S}) \\
& +\text{tr}(\Sigma^{-1}(\partial_k\Sigma)\Sigma^{-1}(\partial_i\Sigma)\Sigma^{-1}(\partial_j\Sigma)\Sigma^{-1}\hat{S} \\
& +\Sigma^{-1}(\partial_i\Sigma)\Sigma^{-1}(\partial_k\Sigma)\Sigma^{-1}(\partial_j\Sigma)\Sigma^{-1}\hat{S}) \\
& +\text{tr}(\Sigma^{-1}(\partial_i\Sigma)\Sigma^{-1}(\partial_j\Sigma)\Sigma^{-1}(\partial_k\Sigma)\Sigma^{-1}\hat{S}) \\
& +\text{tr}(\Sigma^{-1}(\partial_j\Sigma)\Sigma^{-1}(\partial_i\Sigma)\Sigma^{-1}(\partial_k\Sigma)\Sigma^{-1}\hat{S})\Big\} \\
= \;& -\frac{1}{2p}\Big\{\text{tr}(\Sigma^{-1}(\partial_k\Sigma)\Sigma^{-1}(\partial_j\Sigma)\Sigma^{-1}(\partial_i\Sigma)\Sigma^{-1}(\hat{S}-\Sigma) \\
& +\Sigma^{-1}(\partial_j\Sigma)\Sigma^{-1}(\partial_k\Sigma)\Sigma^{-1}(\partial_i\Sigma)\Sigma^{-1}(\hat{S}-\Sigma)) \\
& +\text{tr}(\Sigma^{-1}(\partial_k\Sigma)\Sigma^{-1}(\partial_i\Sigma)\Sigma^{-1}(\partial_j\Sigma)\Sigma^{-1}\hat{S} \\
& +\Sigma^{-1}(\partial_i\Sigma)\Sigma^{-1}(\partial_k\Sigma)\Sigma^{-1}(\partial_j\Sigma)\Sigma^{-1}\hat{S}) \\
& +\text{tr}(\Sigma^{-1}(\partial_i\Sigma)\Sigma^{-1}(\partial_j\Sigma)\Sigma^{-1}(\partial_k\Sigma)\Sigma^{-1}\hat{S}) \\
& +\text{tr}(\Sigma^{-1}(\partial_j\Sigma)\Sigma^{-1}(\partial_i\Sigma)\Sigma^{-1}(\partial_k\Sigma)\Sigma^{-1}\hat{S})\Big\} \\
= \;& -\frac{1}{2p}\Big\{\text{tr}(\Sigma^{-1}(\partial_k\Sigma)\Sigma^{-1}(\partial_j\Sigma)\Sigma^{-1}(\partial_i\Sigma)\Sigma^{-1}(2\hat{S}-\Sigma) \\
& +\Sigma^{-1}(\partial_j\Sigma)\Sigma^{-1}(\partial_k\Sigma)\Sigma^{-1}(\partial_i\Sigma)\Sigma^{-1}(2\hat{S}-\Sigma))
\end{aligned}
$$

$$+\mathrm{tr}(\Sigma^{-1}(\partial_i\Sigma)\Sigma^{-1}(\partial_j\Sigma)\Sigma^{-1}(\partial_k\Sigma)\Sigma^{-1}\hat{S})$$
$$+\mathrm{tr}(\Sigma^{-1}(\partial_j\Sigma)\Sigma^{-1}(\partial_i\Sigma)\Sigma^{-1}(\partial_k\Sigma)\Sigma^{-1}\hat{S})\Big\}$$

where $\Sigma = \Sigma(\tilde{\Lambda}, \bar{\Psi})$ and $\partial_k\Sigma = \partial_k\Psi$. Importantly, $\mathrm{supp}(\hat{\Psi}) \subseteq \mathrm{supp}(\Psi_0)$ since we consider the estimator (A.9). This implies that each component of $\nabla^3_{ijk}\mathbb{G}_{n,p}(\tilde{\Lambda};\bar{\Psi})$ is multiplied $\mathrm{vec}(\hat{\Psi}-\Psi_0)$, whose non-zero components are of order k_0. Restricting to these elements, and by the Cauchy-Schwartz inequality, we have

$$\begin{aligned}&|\nabla_{\theta_\Psi}\Big\{u'\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\bar{\Psi})v\Big\}\mathrm{vec}(\hat{\Psi}-\Psi_0)|^2\\ \leq\ &\Big\{\sum_{i,j,k=1}^{w}\partial^3_{ijk}\mathbb{G}_{n,p}(\tilde{\Lambda};\bar{\Psi})^2u_i^2v_j^2\Big\}\|\mathrm{vec}(\hat{\Psi}-\Psi_0)\|_2^2.\end{aligned}$$

Moreover, taking the supremum on the unit sphere and restricting to matrices with respect to the $\mathcal{A}$ block, we have

$$\begin{aligned}&\sum_{i,j,k=1}^{w}\partial^3_{ijk}\mathbb{G}_{n,p}(\tilde{\Lambda};\bar{\Psi})^2u_i^2v_j^2\\ \leq\ & p^2\big(L_1\|\Sigma^{-1}\|_F^3\|\Sigma^{-1}(2\hat{S}-\Sigma)\|_F+L_2\|\Sigma^{-1}\|_F^4\frac{1}{n}\sum_{i=1}^{n}\|X_{i,\mathcal{A}}\|_2^2\big)^2/(4p^2),\end{aligned}$$

where L_1, L_2 are positive constants and $X_{i,\mathcal{A}}$ denotes the vector X_i restricted to $\mathcal{A}$. Besides

$$\|\Sigma^{-1}(2\hat{S}-\Sigma)\|_F \leq 2\|\Sigma^{-1}\|_F\|\hat{S}\|_F + k_0 \leq L_3\|\Psi_0^{-1}\|_F k_0 + k_0,$$

with $L_3 > 0$ when restricting to $\mathcal{A}$, where we used the sub-Gaussian assumption and $\|X_{i,\mathcal{A}}\|_2^2 \leq Mk_0$. Now using the consistency of $\|\mathrm{vec}(\hat{\Psi}-\Psi_0)\|_2^2$ obtained in Corollary 3.3 and the rate of γ_n obtained when bounding $\|\nabla_\Psi\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0)\|_{\max}$, we obtain

$$\begin{aligned}&\Big\{\sum_{i,j,k=1}^{w}\partial^3_{ijk}\mathbb{G}_{n,p}(\tilde{\Lambda};\bar{\Psi})^2u_i^2v_j^2\Big\}\|\mathrm{vec}(\hat{\Psi}-\Psi_0)\|_2^2\\ \leq\ &\big(L'_1\|\Psi_0^{-1}\|_F^4k_0+L_1\|\Psi_0^{-1}\|_F^3k_0+L'_2\|\Psi_0^{-1}\|_F^4k_0\big)^2\|\Psi_0^{-1}\|_F^4k_0\frac{p}{n}.\end{aligned}$$

As a consequence, taking the supremum with respect to unit vector $u, v \in \mathbb{R}^{\mathcal{A}}$, we obtain

$$\|\hat{K}_{\mathcal{A}\mathcal{A}} - \nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda},\Psi_0)_{\mathcal{A}\mathcal{A}}\|_s \leq L\sqrt{\|\Psi_0^{-1}\|_F^{12}\frac{k_0^3p}{n}}, \tag{A.10}$$

for L sufficiently large. Moreover, we have

$$\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0)_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]_{\mathcal{A}\mathcal{A}}$$

$$\begin{aligned} &= \nabla^2_{\theta_\Psi \theta'_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0)_{\mathcal{A}\mathcal{A}} - \nabla^2_{\theta_\Psi \theta'_\Psi} \mathbb{G}_{n,p}(\Lambda_0;\Psi_0)_{\mathcal{A}\mathcal{A}} \\ &\quad + \nabla^2_{\theta_\Psi \theta'_\Psi} \mathbb{G}_{n,p}(\Lambda_0;\Psi_0)_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta_\Psi \theta'_\Psi} \mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]_{\mathcal{A}\mathcal{A}}. \end{aligned}$$

By a Taylor expansion, we obtain

$$\begin{aligned} &\nabla^2_{\theta_\Psi \theta'_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0)_{\mathcal{A}\mathcal{A}} - \nabla^2_{\theta_\Psi \theta'_\Psi} \mathbb{G}_{n,p}(\Lambda_0;\Psi_0)_{\mathcal{A}\mathcal{A}} \\ &= \nabla_{\theta_\Lambda} \Big\{ \nabla^2_{\theta_\Psi \theta'_\Psi} \mathbb{G}_{n,p}(\bar{\Lambda};\Psi_0)_{\mathcal{A}\mathcal{A}} \Big\} \text{vec}(\tilde{\Lambda} - \Lambda_0), \end{aligned}$$

where $\|\bar{\Lambda} - \Lambda_0\|_F \leq \|\tilde{\Lambda} - \Lambda_0\|_F$. Let $\theta = (\theta'_\Lambda, \theta'_\Psi)'$ with $\theta_\Lambda = vec(\Lambda)$. Element-by-element, let $k = 1, \cdots, m(m+1)/2 + (p-m)^2$ and $i, j = 1, \cdots, p^2$, the derivative with respect to $\theta_{k,\Lambda}$ applied to the second derivative with respect to $\partial^2_{\theta_{i,\Psi}\theta_{j,\Psi}} \mathbb{G}_{n,p}(\bar{\Lambda};\Psi_0)$ is

$$\begin{aligned} &\partial^3_{ijk} \mathbb{G}_{n,p}(\bar{\Lambda};\Psi_0) \\ &= \frac{1}{2p} \text{tr}(\Sigma^{-1}(\partial_k \Sigma)\Sigma^{-1}(\partial_j \Sigma)\Sigma^{-1}(\partial_i \Sigma) + \Sigma^{-1}(\partial_j \Sigma)\Sigma^{-1}(\partial_k \Sigma)\Sigma^{-1}(\partial_i \Sigma)) \\ &\quad - \frac{1}{2p} \Big\{ \text{tr}(\Sigma^{-1}(\partial_k \Sigma)\Sigma^{-1}(\partial_j \Sigma)\Sigma^{-1}(\partial_i \Sigma)\Sigma^{-1}\hat{S}) \\ &\quad + \text{tr}(\Sigma^{-1}(\partial_j \Sigma)\Sigma^{-1}(\partial_k \Sigma)\Sigma^{-1}(\partial_i \Sigma)\Sigma^{-1}\hat{S}) \\ &\quad + \text{tr}(\Sigma^{-1}(\partial_k \Sigma)\Sigma^{-1}(\partial_i \Sigma)\Sigma^{-1}(\partial_j \Sigma)\Sigma^{-1}\hat{S}) \\ &\quad + \text{tr}(\Sigma^{-1}(\partial_i \Sigma)\Sigma^{-1}(\partial_k \Sigma)\Sigma^{-1}(\partial_j \Sigma)\Sigma^{-1}\hat{S}) \\ &\quad + \text{tr}(\Sigma^{-1}(\partial_i \Sigma)\Sigma^{-1}(\partial_j \Sigma)\Sigma^{-1}(\partial_k \Sigma)\Sigma^{-1}\hat{S}) \\ &\quad + \text{tr}(\Sigma^{-1}(\partial_j \Sigma)\Sigma^{-1}(\partial_i \Sigma)\Sigma^{-1}(\partial_k \Sigma)\Sigma^{-1}\hat{S}) \Big\} \\ &= -\frac{1}{2p} \Big\{ \text{tr}(\Sigma^{-1}(\partial_k \Sigma)\Sigma^{-1}(\partial_j \Sigma)\Sigma^{-1}(\partial_i \Sigma)\Sigma^{-1}(\hat{S} - \Sigma) \\ &\quad + \Sigma^{-1}(\partial_j \Sigma)\Sigma^{-1}(\partial_k \Sigma)\Sigma^{-1}(\partial_i \Sigma)\Sigma^{-1}(\hat{S} - \Sigma)) \\ &\quad + \text{tr}(\Sigma^{-1}(\partial_k \Sigma)\Sigma^{-1}(\partial_i \Sigma)\Sigma^{-1}(\partial_j \Sigma)\Sigma^{-1}\hat{S}) \\ &\quad + \text{tr}(\Sigma^{-1}(\partial_i \Sigma)\Sigma^{-1}(\partial_k \Sigma)\Sigma^{-1}(\partial_j \Sigma)\Sigma^{-1}\hat{S}) \\ &\quad + \text{tr}(\Sigma^{-1}(\partial_i \Sigma)\Sigma^{-1}(\partial_j \Sigma)\Sigma^{-1}(\partial_k \Sigma)\Sigma^{-1}\hat{S}) \\ &\quad + \text{tr}(\Sigma^{-1}(\partial_j \Sigma)\Sigma^{-1}(\partial_i \Sigma)\Sigma^{-1}(\partial_k \Sigma)\Sigma^{-1}\hat{S}) \Big\}, \end{aligned}$$

where $\forall k, \partial_k \Sigma(\bar{\Lambda}, \Psi_0) = (\partial_k \bar{\Lambda})\bar{\Lambda}' + \bar{\Lambda}(\partial_k \bar{\Lambda})'$. Thus, using the same approach when controlling for $\hat{K} - \nabla^2_{\theta_\Psi \theta'_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}, \Psi_0)$, by the Cauchy-Schwartz inequality and since $\|\tilde{\Lambda} - \Lambda_0\|_F = O_p(\sqrt{\frac{p}{n}}) + O_p(\sqrt{\frac{1}{p}})$, we obtain for a sufficiently large constant $C > 0$

$$\begin{aligned} &\|\nabla^2_{\theta_\Psi \theta'_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0)_{\mathcal{A}\mathcal{A}} - \nabla^2_{\theta_\Psi \theta'_\Psi} \mathbb{G}_{n,p}(\Lambda_0;\Psi_0)_{\mathcal{A}\mathcal{A}}\|_s \\ \leq\; &\|\nabla_{\theta_\Lambda} \{\nabla^2_{\theta_\Psi \theta'_\Psi} \mathbb{G}_{n,p}(\bar{\Lambda};\Psi_0)_{\mathcal{A}\mathcal{A}}\}_{\mathcal{A}}\|_2 \|\tilde{\Lambda} - \Lambda_0\|_F \leq C\sqrt{\|\Psi_0^{-1}\|_F^8 k_0^2 \frac{p}{n}}. \end{aligned}$$

Moreover, regarding the control over the Hessian evaluated at the true parameter, we obtain with probability $1-\exp(-\log p)$ that

$$\begin{aligned}&\|\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)_{\mathcal{A}\mathcal{A}}-\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]_{\mathcal{A}\mathcal{A}}\|_s\\ \leq\quad&\|\Sigma^{-1}\|^3_s\|\hat{S}-\mathbb{E}[X_iX'_i]\|_s/(2p)\leq\|\Psi_0^{-1}\|^3_F(Kp\sqrt{\frac{\log p}{n}})/(2p).\end{aligned}$$

Consequently, for $C,M>0$ sufficiently large,

$$\begin{aligned}&\|\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0)_{\mathcal{A}\mathcal{A}}-\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]_{\mathcal{A}\mathcal{A}}\|_s\\ \leq\quad&\|\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0)_{\mathcal{A}\mathcal{A}}-\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)_{\mathcal{A}\mathcal{A}}\|_s\\ &+\|\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)_{\mathcal{A}\mathcal{A}}-\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]_{\mathcal{A}\mathcal{A}}\|_s\\ \leq\quad&C\sqrt{\|\Psi_0^{-1}\|^8_F k_0^2\frac{p}{n}}+M\|\Psi_0^{-1}\|^3_F\sqrt{\frac{\log p}{n}},\end{aligned}\tag{A.11}$$

with probability at least $1-\exp(-\log p)$. Thus by inequalities (A.10) and (A.11), we have

$$\|\hat{K}_{\mathcal{A}\mathcal{A}}-\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0,\Psi_0)]_{\mathcal{A}\mathcal{A}}\|_s\leq L\sqrt{\|\Psi_0^{-1}\|^{12}_F\frac{k_0^3p}{n}},$$

for L a sufficiently large constant. Using Lemma 11 of [22], we obtain

$$\|\hat{K}^{-1}_{\mathcal{A}\mathcal{A}}-\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0,\Psi_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_s\leq L\sqrt{\|\Psi_0^{-1}\|^{12}_F\frac{k_0^3p}{n}}.\tag{A.12}$$

Based on the same arguments for deriving (A.10), we have

$$\max_{i\in\mathcal{A}^c}\|e'_i\big(\hat{K}_{\mathcal{A}^c\mathcal{A}}-\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda},\Psi_0)_{\mathcal{A}^c\mathcal{A}}\big)\|_2\leq L\sqrt{\|\Psi_0^{-1}\|^{12}_F\frac{k_0^3p}{n}},$$

and

$$\begin{aligned}&\max_{i\in\mathcal{A}^c}\|e'_i\big(\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0)_{\mathcal{A}^c\mathcal{A}}-\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]_{\mathcal{A}^c\mathcal{A}}\big)\|_2\\ &\qquad\leq\quad L\sqrt{\|\Psi_0^{-1}\|^{12}_F\frac{k_0^3p}{n}},\end{aligned}$$

which implies

$$\max_{i\in\mathcal{A}^c}\|e'_i\big(\hat{K}_{\mathcal{A}^c\mathcal{A}}-\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]_{\mathcal{A}^c\mathcal{A}}\big)\|_2\leq L\sqrt{\|\Psi_0^{-1}\|^{12}_F\frac{k_0^3p}{n}}.\tag{A.13}$$

Following inequality (A.7), we have

$$\|\hat{K}_{\mathcal{A}^c\mathcal{A}}\hat{K}^{-1}_{\mathcal{A}\mathcal{A}}\nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda},\Psi_0)_{\mathcal{A}}\|_\infty\leq M_1+M_2,$$

with

$$M_1 = \\ \|\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0,\Psi_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0,\Psi_0)]^{-1}_{\mathcal{A}^c\mathcal{A}}\nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde\Lambda,\Psi_0)_{\mathcal{A}}\|_\infty,$$

and

$$M_2 = \\ \|\Big\{\hat K_{\mathcal{A}^c\mathcal{A}}\hat K^{-1}_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0,\Psi_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0,\Psi_0)]^{-1}_{\mathcal{A}\mathcal{A}}\Big\} \\ \times\nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde\Lambda,\Psi_0)_{\mathcal{A}}\|_\infty.$$

By assumption, the population level Hessian is bounded. As for the gradient, using Corollary 3.1, we obtain with probability $1-\exp(-\log p)$ that $M_1 \leq C\sqrt{\frac{p}{n}}$. As for M_2, we have

$$M_2 \leq \\ \max_{i\in\mathcal{A}^c}\|e'_i\{\hat K_{\mathcal{A}^c\mathcal{A}}\hat K^{-1}_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0,\Psi_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0,\Psi_0)]^{-1}_{\mathcal{A}\mathcal{A}}\}\|_2 \\ \|\nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde\Lambda,\Psi_0)_{\mathcal{A}}\|_2. \tag{A.14}$$

We proved

$$\begin{aligned} & \|\nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde\Lambda,\Psi_0)_{\mathcal{A}}\|_2 \\ \leq\ & \sqrt{k_0}\|\nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde\Lambda,\Psi_0)_{\mathcal{A}}\|_\infty = \sqrt{k_0}\|\nabla_\Psi\mathbb{G}_{n,p}(\tilde\Lambda,\Psi_0)_{\mathcal{A}}\|_{\max} \\ \leq\ & \tilde L\|\Psi_0^{-1}\|_F^2\sqrt{k_0}\sqrt{\frac{p}{n}}, \end{aligned} \tag{A.15}$$

for $\tilde L$ a sufficiently large constant with high probability. Moreover

$$\begin{aligned} & \|e'_i\{\hat K_{\mathcal{A}^c\mathcal{A}}\hat K^{-1}_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0,\Psi_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0,\Psi_0)]^{-1}_{\mathcal{A}\mathcal{A}}\}\|_2 \\ \leq\ & \|e'_i\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]_{\mathcal{A}^c\mathcal{A}}\Upsilon_1\|_2 + \|e'_i\Upsilon_2\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_2 \\ & +\|e'_i\Upsilon_2\Upsilon_1\|_2, \end{aligned} \tag{A.16}$$

with

$$\begin{aligned} \Upsilon_1 &= \hat K^{-1}_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]^{-1}_{\mathcal{A}\mathcal{A}}, \\ \Upsilon_2 &= \hat K_{\mathcal{A}^c\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]_{\mathcal{A}^c\mathcal{A}}. \end{aligned}$$

By inequalities (A.12) and (A.13), we obtain

$$\|\Upsilon_1\|_s \leq L\sqrt{\|\Psi_0^{-1}\|_F^{12}\frac{k_0^3 p}{n}},\ \ \max_{i\in\mathcal{A}^c}\|e'_i\Upsilon_2\|_2 \leq L\sqrt{\|\Psi_0^{-1}\|_F^{12}\frac{k_0^3 p}{n}}.$$

Hence, using inequalities (A.14), (A.15) and (A.16), we obtain for a sufficiently large constant $\tilde C$ that

$$M_2 \leq \tilde C\sqrt{k_0}\sqrt{\|\Psi_0^{-1}\|_F^4\frac{p}{n}}\sqrt{\|\Psi_0^{-1}\|_F^{12}\frac{k_0^3 p}{n}}.$$

Moreover, using the incoherence condition,

$$\begin{aligned}
& \|\hat{K}_{\mathcal{A}^c\mathcal{A}}\hat{K}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty \\
\leq \; & \|\hat{K}_{\mathcal{A}^c\mathcal{A}}\hat{K}^{-1}_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty \\
+ \; & \|\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]_{\mathcal{A}^c\mathcal{A}}\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty \\
\leq \; & L\sqrt{\|\Psi_0^{-1}\|_F^{12}\frac{k_0^4 p}{n}} + \eta.
\end{aligned}$$

Thus, putting the pieces together, we have for $L_1, L_2, L_3 > 0$ sufficiently large

$$\|z_{\mathcal{A}^c}\|_\infty \leq \frac{1}{\gamma_n}\Big(L_1\sqrt{\|\Psi_0^{-1}\|_F^{12}\frac{k_0^4 p}{n}} + L_2\sqrt{\|\Psi_0^{-1}\|_F^{4}\frac{p}{n}}\Big) + L_3\sqrt{\|\Psi_0^{-1}\|_F^{12}\frac{k_0^4 p}{n}} + \eta.$$

Hence, then strict dual feasibility of Theorem A.1 is satisfied when

$$\frac{1}{1-\eta}L\sqrt{\frac{p}{n}} \leq \gamma_n,$$

under the scaling $n > C\|\Psi_0^{-1}\|_F^{12}k_0^4 p$.
We now turn to the ℓ_∞-bound. Under the scaling $n > C\|\Psi_0^{-1}\|_F^{12}k_0^4 p$, we have

$$\begin{aligned}
& \|\hat{K}^{-1}_{\mathcal{A}\mathcal{A}}\nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0)_{\mathcal{A}}\|_\infty \\
\leq \; & \|\{\hat{K}^{-1}_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]^{-1}_{\mathcal{A}\mathcal{A}}\}\nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0)_{\mathcal{A}}\|_\infty \\
& +\|\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]^{-1}_{\mathcal{A}\mathcal{A}}\nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0)_{\mathcal{A}}\|_\infty \\
\leq \; & \sqrt{k_0}\|\hat{K}^{-1}_{\mathcal{A}\mathcal{A}} - \mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]^{-1}_{\mathcal{A}\mathcal{A}}\|_s\|\nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0)_{\mathcal{A}}\|_\infty \\
& +\|\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]^{-1}_{\mathcal{A}\mathcal{A}}\nabla_{\theta_\Psi}\mathbb{G}_{n,p}(\tilde{\Lambda};\Psi_0)_{\mathcal{A}}\|_\infty \\
\leq \; & C_1\sqrt{k_0}\sqrt{\|\Psi_0^{-1}\|_F^{12}\frac{k_0^3 p}{n}}\sqrt{\|\Psi_0^{-1}\|_F^{4}\frac{p}{n}} + C_2\sqrt{\|\Psi_0^{-1}\|_F^{4}\frac{p}{n}},
\end{aligned}$$

for $C_1, C_2 > 0$. By inequality (A.12), we obtain

$$\begin{aligned}
& \|\hat{K}^{-1}_{\mathcal{A}\mathcal{A}} - (\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty \\
\leq \; & \sqrt{k_0}\|\hat{K}^{-1}_{\mathcal{A}\mathcal{A}} - (\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]_{\mathcal{A}\mathcal{A}})^{-1}\|_s \leq L\sqrt{\|\Psi_0^{-1}\|_F^{12}\frac{k_0^4 p}{n}} \\
\leq \; & \beta_\infty.
\end{aligned}$$

Hence

$$\begin{aligned}
& \|\hat{K}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty \leq \\
& \|\hat{K}^{-1}_{\mathcal{A}\mathcal{A}} - (\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty + \|(\mathbb{E}[\nabla^2_{\theta_\Psi\theta'_\Psi}\mathbb{G}_{n,p}(\Lambda_0;\Psi_0)]_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty \\
& \leq 2\beta_\infty.
\end{aligned}$$

Consequently, by part (i) of Theorem A.2, we obtain

$$\|\hat{\Psi} - \Psi_0\|_{\max} \leq \tilde{L}\sqrt{\frac{p}{n}} + \gamma_n\beta_\infty,$$

for $\tilde{L} > 0$.

Point (ii). For (μ, ζ)-amenable penalties, we apply Proposition A.3 of [22] and control for each norm quantities. To do so, the same approach as in the proof of (i) can be applied. Since the regulariser is assumed to be (μ, ζ)-amenable, we have by Lemma 5 of [22] that $\gamma_n \hat{z}_{\mathcal{A}} - \nabla_{\theta_\Psi} q(\gamma_n, \text{vec}(\hat{\Psi})_{\mathcal{A}}) = 0$. Hence we have

$$\|\hat{z}_{\mathcal{A}^c}\|_\infty \leq \frac{1}{\gamma_n} \| - \nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}^c} + \hat{K}_{\mathcal{A}^c\mathcal{A}} \hat{K}^{-1}_{\mathcal{A}\mathcal{A}} \nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}}\|_\infty.$$

Following the same steps as in the proof of part (i), we upper bound both $\|\nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}^c}\|_\infty$ and $\|\hat{K}_{\mathcal{A}^c\mathcal{A}} \hat{K}^{-1}_{\mathcal{A}\mathcal{A}} \nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}}\|_\infty$ and apply Proposition A.3, thus establishing strict dual feasibility. Then the remainder follows from part (ii) of Theorem A.2. □

Proof of Corollary 4.2. We first establish strict dual feasibility. We denote the parameter vector as $\theta_\Psi = \text{vec}(\Psi)$, we have

$$\hat{z}_{\mathcal{A}^c} = \frac{1}{\gamma_n} \Big\{ - \nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}^c} + \hat{K}_{\mathcal{A}^c\mathcal{A}} \hat{K}^{-1}_{\mathcal{A}\mathcal{A}} \Big(\nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}} \\ - \nabla_{\theta_\Psi} q(\gamma_n, \text{vec}(\hat{\Psi})_{\mathcal{A}}) + \gamma_n \hat{z}_{\mathcal{A}} \Big) \Big\}.$$

Taking the ℓ_∞-norm, we obtain

$$\begin{aligned} \|\hat{z}_{\mathcal{A}^c}\|_\infty &\leq \tfrac{1}{\gamma_n} \| - \nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}^c} + \hat{K}_{\mathcal{A}^c\mathcal{A}} \hat{K}^{-1}_{\mathcal{A}\mathcal{A}} \nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}}\|_\infty \\ &+ \tfrac{1}{\gamma_n} \|\hat{K}_{\mathcal{A}^c\mathcal{A}} \hat{K}^{-1}_{\mathcal{A}\mathcal{A}} \big(\gamma_n \hat{z}_{\mathcal{A}} - \nabla_{\theta_\Psi} q(\gamma_n, \text{vec}(\hat{\Psi})_{\mathcal{A}}) \big)\|_\infty \\ &\leq \tfrac{1}{\gamma_n} \| - \nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}^c} + \hat{K}_{\mathcal{A}^c\mathcal{A}} \hat{K}^{-1}_{\mathcal{A}\mathcal{A}} \nabla_{\theta_\Psi} \mathbb{G}_{n,p}(\tilde{\Lambda}; \Psi_0)_{\mathcal{A}}\|_\infty, \end{aligned}$$

using $\gamma_n \hat{z}_{\mathcal{A}} - \nabla_{\theta_\Psi} q(\gamma_n, \text{vec}(\hat{\Psi})_{\mathcal{A}}) = 0$ from Lemma 8 of [22]. Then we have

$$\|\hat{z}_{\mathcal{A}^c}\|_\infty \leq \frac{1}{p\gamma_n} \|\text{vec}(\hat{S} - \tilde{\Lambda}\tilde{\Lambda}' - \Psi_0)_{\mathcal{A}^c} - \hat{K}_{\mathcal{A}^c\mathcal{A}} \hat{K}^{-1}_{\mathcal{A}\mathcal{A}} \text{vec}(\hat{S} - \tilde{\Lambda}\tilde{\Lambda}' - \Psi_0)_{\mathcal{A}}\|_\infty,$$

which implies with probability $1 - \exp(-\log p)$ that

$$\|\hat{z}_{\mathcal{A}^c}\|_\infty \leq \frac{2}{p\gamma_n} \|\text{vec}(\hat{S} - \tilde{\Lambda}\tilde{\Lambda}' - \Psi_0)\|_\infty = \frac{2}{p\gamma_n} \|\hat{S} - \tilde{\Lambda}\tilde{\Lambda}' - \Psi_0\|_{\max} \leq \frac{1}{\gamma_n} L \sqrt{\frac{p}{n}},$$

using the arguments in the proof of Corollary 3.4. Provided $\gamma_n > L\sqrt{\frac{p}{n}}$, strict dual feasibility holds and support recovery is satisfied by Theorem A.1 of [22].

As for the ℓ_∞-bound, using $\|\hat{K}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty = 1$, we have

$$\begin{aligned} \|\hat{K}^{-1}_{\mathcal{A}\mathcal{A}} \nabla_{\theta_\Psi} \mathbb{F}_{n,p}(\tilde{\Lambda}; \Psi_0)\|_\infty &= \tfrac{1}{p} \|\hat{K}^{-1}_{\mathcal{A}\mathcal{A}} \text{vec}(\hat{S} - \tilde{\Lambda}\tilde{\Lambda}' - \Psi_0)\|_\infty \\ &\leq L\sqrt{\tfrac{p}{n}}, \end{aligned}$$

for $L > 0$ large enough, with probability $1 - \exp(-\log p)$. □