Electronic Journal of Statistics
Vol. 16 (2022) 5591–5622
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2065

Asymptotic normality of robust M-estimators with convex penalty*

Pierre C. Bellec, Yiwei Shen and Cun-Hui Zhang

Department of Statistics, Rutgers University
e-mail: pcb71@stat.rutgers.edu

Abstract: This paper develops asymptotic normality results for individual coordinates of robust M-estimators with convex penalty in high-dimensions, where the dimension p is at most of the same order as the sample size n, i.e, $p/n \leq \gamma$ for some fixed constant $\gamma > 0$. The asymptotic normality requires a bias correction and holds for most coordinates of the M-estimator for a large class of loss functions including the Huber loss and its smoothed versions regularized with a strongly convex penalty.

The asymptotic variance that characterizes the width of the resulting confidence intervals is estimated with data-driven quantities. This estimate of the variance adapts automatically to low ($p/n \to 0$) or high ($p/n \leq \gamma$) dimensions and does not involve the proximal operators seen in previous works on asymptotic normality of M-estimators. For the Huber loss, the estimated variance has a simple expression involving an effective degrees-of-freedom as well as an effective sample size. The case of the Huber loss with Elastic-Net penalty is studied in details and a simulation study confirms the theoretical findings. The asymptotic normality results follow from Stein formulae for high-dimensional random vectors on the sphere developed in the paper which are of independent interest.

Keywords and phrases: Robust estimation, M-estimator, asymptotic normality, confidence Intervals, high-dimensional statistics, bias-correction, Stein's formula.

Received June 2021.

1. Introduction

1.1. Robust inference

In his seminal paper on robustness, Huber [20] introduced M-estimators for an unknown location parameter $\mu \in \mathbb{R}$ from observations $Y_i = \mu + \varepsilon_i$, $i = 1, \ldots, n$, where ε_i are iid noise random variables distributed as a mixture $F = (1 - \epsilon)N(0,1) + \epsilon H$ with H being the distribution of the contaminated samples, possibly chosen by an adversary. Given a differentiable loss function $\rho : \mathbb{R} \to \mathbb{R}$ and its derivative $\psi = \rho'$, Huber defined M-estimators as minimizers $\hat{\mu} = \operatorname{argmin}_{b \in \mathbb{R}} \sum_{i=1}^n \rho(Y_i - b)$, or equivalently as solutions to $\sum_{i=1}^n \psi(Y_i - b) = 0$. Huber [20] went on to prove consistency and asymptotic normality of such M-estimators, obtaining among other results that if ρ is convex and ψ is absolutely

*P.C. Bellec was partially supported by the NSF Grants DMS-1811976 and DMS-1945428. C.-H. Zhang was partially supported by the NSF Grants DMS-1721495, IIS-1741390, CCF-1934924, DMS-2052949 and DMS-2210850.

continuous, then under mild assumptions on F, the convergence $\hat{\mu} \to \mu$ in probability holds as well as the asymptotic normality

$$n^{1/2}(\hat{\mu} - \mu) \to^d N\big(0, \mathbb{E}[\psi^2(\varepsilon_1)]/\mathbb{E}[\psi'(\varepsilon_1)]^2\big).$$

Huber's M-estimators were extended to regression models, where a design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is observed together with responses $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$ where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are the rows of $\boldsymbol{X}$ and $\varepsilon_1, ..., \varepsilon_n$ are possibly contaminated noise random variables as in the previous paragraph. For fixed or slowly growing dimension p as $n \to +\infty$, consistency and asymptotic normality of M-estimators of the form $\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} \sum_{i=1}^n \rho(y_i - \boldsymbol{x}_i^\top \boldsymbol{b})$ were obtained, see [21, Section 7] or [27] among others. Explicitly, if $\boldsymbol{e}_j \in \mathbb{R}^p$ is a canonical basis vector and one is interested in the asymptotic normality of $\hat{\beta}_j - \beta_j$ for the purpose of confidence intervals, [27] shows that

$$\big(\boldsymbol{e}_j^\top (\boldsymbol{X}^\top \boldsymbol{X}/n)^{-1} \boldsymbol{e}_j\big)^{-1/2} \sqrt{n}(\hat{\beta}_j - \beta_j) \ \to^d \ N\Big(0, \frac{\mathbb{E}[\psi^2(\varepsilon_1)]}{\mathbb{E}[\psi'(\varepsilon_1)]^2}\Big) \tag{1.1}$$

if $(p \log n)^{3/2}/n \to 0$ and under mild assumptions on $\boldsymbol{X}$. As in the location parameter problem of the previous paragraph, the asymptotic variance is characterized by the ratio $\mathbb{E}[\psi^2(\varepsilon_1)]/\mathbb{E}[\psi'(\varepsilon_1)]^2$.

The last decade has seen striking developments of similar asymptotic normality results in high-dimensions, where $p/n \to \gamma$ for some constant $\gamma < 1$, cf. [17, 3, 25, 15, 16]. In terms of asymptotic normality, these works show that if $\boldsymbol{X}$ has iid $N(0, \boldsymbol{\Sigma})$ rows, the unregularized M-estimator $\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} \sum_{i=1}^n \rho(y_i - \boldsymbol{x}_i^\top \boldsymbol{b})$ satisfies asymptotic normality of the form

$$(\boldsymbol{e}_j^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{e}_j)^{-1/2} \sqrt{p}(\hat{\beta}_j - \beta_j) \to^d N(0, r^2) \tag{1.2}$$

where $r > 0$ is a deterministic constant that captures the high-dimensionality of the problem [17, Lemma 1]. The constant $r > 0$ is determined by solving a system of nonlinear equations with two unknowns: In the unregularized setting, [17, S2] describes this system of nonlinear equations with unknowns (r, c) as

$$\begin{cases} \mathbb{E}\big[1 - [\operatorname{prox}_c(\rho)]'(\varepsilon_1 + rZ)\big] = \gamma, \\ \mathbb{E}\big[\big(\varepsilon_1 + rZ - [\operatorname{prox}_c(\rho)](\varepsilon_1 + rZ)\big)^2\big] = \gamma r^2 \end{cases} \quad \text{as} \quad p/n \to \gamma, \tag{1.3}$$

where $Z \sim N(0, 1)$ is independent of ε_1, and $\operatorname{prox}_c(\rho)(t) = \operatorname{argmin}_{u\in\mathbb{R}} \rho(u) + (t - u)^2/(2c)$ denotes the proximal operator of the convex function $t \to c\rho(t)$ with derivative $[\operatorname{prox}_c(\rho)]'(t)$. The optimality conditions

$$c^{-1}(t - [\operatorname{prox}_c(\rho)](t)) = \psi([\operatorname{prox}_c(\rho)](t))$$

of the proximal minimization problem leads to the expressions

$$c^{-2}\gamma r^2 = \mathbb{E}\big[\psi([\operatorname{prox}_c(\rho)](\varepsilon_1 + rZ))^2\big] \quad \text{and} \quad c^{-1}\gamma = \mathbb{E}\big[\tfrac{d}{dt}\psi([\operatorname{prox}_c(\rho)](t))\big|_{t=\varepsilon_1+rZ}\big]$$

for the solutions (r, c) to the above system. Hence (1.2) can be rewritten as

$$(\boldsymbol{e}_j^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{e}_j)^{-1/2} \sqrt{n}(\hat{\beta}_j - \beta_j) \ \to^d \ N\Big(0, \frac{\mathbb{E}[\psi([\mathrm{prox}_c(\rho)](\varepsilon_1 + rZ))^2]}{\mathbb{E}[\frac{d}{dt}\psi([\mathrm{prox}_c(\rho)](t))|_{t=\varepsilon_1+rZ}]^2}\Big), \quad (1.4)$$

see, e.g., [15, Theorem 4.1 and Corollary 4.6]. These results embody that when p and n are of the same order, the asymptotic variance in (1.1) must be modified to account for the high-dimensionality of the problem by (a) replacing ψ in the numerator and ψ' in the denominator by their compositions with the proximal operator $\mathrm{prox}_c(\rho)$, and (b) adding the extra Gaussian term rZ to the initial noise ε_1. The distribution of $\varepsilon_1 + rZ$ is sometimes referred to as the effective noise. The Gaussian assumption can be relaxed and some of the above results are still valid if $\boldsymbol{X}$ has iid centered entries with variance one [25, 16]. Despite the subtle introduction of the proximal operator and the constants (r, c), it is remarkable that the informal ratio $\frac{\mathrm{average}[\psi^2]}{\mathrm{average}[(d/dt)\psi]^2}$ unifies the results (1.1) and (1.4) in both low and high-dimensions.

All results of the previous section are applicable when $p/n \to \gamma$ with $\gamma < 1$. For $\gamma > 1$ regularization is required to ensure the uniqueness of $\hat{\boldsymbol{\beta}}$, for instance through an additive penalty which leads to regularized M-estimators of the form

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{b} \in \mathbb{R}^p}{\mathrm{argmin}} \frac{1}{n} \sum_{i=1}^n \rho(y_i - \boldsymbol{x}_i^\top \boldsymbol{b}) + g(\boldsymbol{b}) \quad (1.5)$$

for some convex penalty function $g : \mathbb{R}^p \to \mathbb{R}$. The case of Ridge regularization with $g(\boldsymbol{b}) = \tau \|\boldsymbol{b}\|_2^2/2$ for some constant $\tau > 0$ is treated in [25, 16]. In this case, the two solutions (r, c) of a system of two nonlinear equations similar to (1.3) characterize the error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$, asymptotic normality and asymptotic variance of $\sqrt{n}((1+a)\hat{\beta}_j - \beta_j)$ where a is a constant capturing the bias induced by regularization [16, Proposition 3.30] and a is a function of (γ, r, c). Thrampoulidis et al. [29] characterize the error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$ for a large class of (ρ, g) pairs using a technique known as the Convex Gordon Min-Max theorem pioneered by [28], and the recent paper [19] on Approximate Message Passing focused on $g(\boldsymbol{b}) = \lambda \|\boldsymbol{b}\|_1$ and ρ either the Huber loss or the absolute value. Little is known, however, on asymptotic normality of the regularized estimators (1.5) for penalty functions different from the square norm $\boldsymbol{b} \mapsto \tau \|\boldsymbol{b}\|_2^2$. The theories developed in [29, 19] do not readily provide asymptotic normality results and regularized M-estimators of the form (1.5) lack confidence interval capabilities. One goal of the present paper is to fill this gap.

A separate line of research develops asymptotic normality results and confidence intervals for regularized least-squares estimators of the form

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{b} \in \mathbb{R}^p}{\mathrm{argmin}} \frac{1}{2n} \sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top \boldsymbol{b})^2 + g(\boldsymbol{b}) = \underset{\boldsymbol{b} \in \mathbb{R}^p}{\mathrm{argmin}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2/(2n) + g(\boldsymbol{b}) \quad (1.6)$$

where $\boldsymbol{X}$ has rows $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$. Early results studied the Lasso with $g(\boldsymbol{b}) = \lambda \|\boldsymbol{b}\|_1$ [33, 22, 30] under a sparsity condition $s \log(p) = o(\sqrt{n})$ where $s = \|\boldsymbol{\beta}\|_0$, or

Ridge regression [11]. For the Lasso the sparsity condition was later improved to $s\log^2(p)/n \to 0$ [24], to $s\log(p/s)/n \to 0$ [8] and $p/n \to \gamma \in (0,\infty)$ with $s \lesssim n/\log(p/s)$ ([23, 26] for isotropic Gaussian design and [13] [7, Theorem 3.2] for non-isotropic Gaussian design). For penalty functions beyond the Lasso and Ridge regularization, [12, Proposition 4.3(iii)] provides asymptotic normality on average over the coordinates for permutation invariant penalty function g in (1.6), and [7, Theorem 3.1] proves asymptotic normality for individual coordinates of (1.6) under a strong convexity assumption. A high-level message of these works is that one must de-bias the regularized estimator (1.6) in order to obtain asymptotic normality at the $\sqrt{n}$-adjusted rate and construct confidence intervals. In the $p/n \to \gamma$ regime that is the focus of the present paper, this bias correction takes the following form. Under a strong convexity assumption and for $\boldsymbol{X}$ with iid $N(\boldsymbol{0}_p, \boldsymbol{\Sigma})$ rows, [7] proves that for most coordinates $j = 1, ..., p$,

$$\frac{(n-\hat{\mathsf{df}})(\hat{\beta}_j - \beta_j) + \boldsymbol{e}_j^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2}\Omega_{jj}^{-1/2} \to^d N(0,1) \tag{1.7}$$

where $\Omega_{jj} = \boldsymbol{e}_j^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j$ and $\hat{\mathsf{df}}$ is the effective degrees of freedom of $\hat{\boldsymbol{\beta}}$ defined as the Jacobian of the map $\boldsymbol{y} \mapsto \boldsymbol{X}\hat{\boldsymbol{\beta}}$ for fixed $\boldsymbol{X}$. For $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ and consequently $\Omega_{jj} = 1$, a similar bias correction proportional to $\boldsymbol{e}_j^\top \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ is visible in the asymptotic normality result [12, Proposition 4.3(iii)], although there the coefficients $(n - \hat{\mathsf{df}})$ and $\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2$ in (1.7) are replaced with deterministic scalar counterparts obtained by solving a system of nonlinear equations of a similar nature as (1.3). Another goal of the present paper is to equip robust M-estimators (1.5), with ρ different than the squared loss, with de-biasing and asymptotic normality results similar to the previous display, allowing for general robust loss functions ρ coupled with general convex penalty functions g.

1.2. Contributions

Our contributions are the following.

1. We provide de-biasing and asymptotic normality results for robust M-estimators with convex penalty functions when p and n are of the same order. This leads to confidence intervals for the j-th coordinate β_j of the unknown coefficient vector $\boldsymbol{\beta}$. Asymptotic normality holds for a large class of robust loss functions, including the Huber loss and its smoothed versions.
2. Although this bias correction required for asymptotic normality resembles the one-step estimators recommended in the theory of classical M-estimator to improve efficiency (e.g., [31, Eq. (1.11)]), a notable difference from the low-dimensional case is the requirement of a degrees-of-freedom adjustment to amplify the one-step correction. For the squared loss, this degrees-of-freedom adjustment takes the form of multiplication by $(n-\hat{\mathsf{df}})$ in (1.7); one contribution of this paper is to identify the degrees-of-freedom

adjustment that leads to asymptotic normality for robust and regularized M-estimators, beyond the squared loss.

3. The asymptotic variance is estimated by random, data-driven quantities, as opposed to the deterministic scalars (r, c) that determine the asymptotic variance for unregularized estimators in (1.4). The fact that the asymptotic variance is estimated by observable quantities makes this results more suitable for confidence intervals (case in point: computing the solution (r, c) of (1.3) and the asymptotic variance in (1.4) requires the knowledge of the noise distribution subject to contamination). The asymptotic normality result takes the form

$$\hat{V}^{-1/2}\Omega_{jj}^{-1/2}\sqrt{n}(\hat{\beta}_j - \beta_j) + [\text{observable bias correction}] \approx N(0,1)$$

where the data-driven variance estimate $\hat{V}$ again is a ratio of the form $\frac{\text{average}[\psi^2]}{\text{average}[(d/dt)\psi]^2}$ for a particular sense of average to be defined in (2.14) below. Interestingly, the expression for this average and $\hat{V}$ does not involve the proximal mapping in (1.4). This informal statement will be made precise in Section 2.4 below.

4. In order to derive these new asymptotic normality results, we develop in Appendix B new identities for random vectors uniformly distributed on the Euclidean sphere. Although the argument of the present paper for asymptotic normality is closely related to that used in [7] for the squared loss, this previous theory for the squared loss for functions of standard normal vector does not extend to robust loss functions due to the lack of strong convexity of ρ for robust losses, and consequently the lack of explicit lower bounds on $\frac{1}{n}\sum_{i=1}^n \psi(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})^2$. Developing these new identities and the corresponding asymptotic normality results for random vectors uniformly distributed on the sphere is a crucial step to overcome the lack of global strong convexity of ρ for robust losses and to obtain the asymptotic normality results. These new identities provide novel Stein formulae for random vectors on the sphere and may be used more broadly for elliptical distributions.

2. Model and main results

2.1. Model and assumptions

We consider a linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$, with a regularized M-estimator

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \rho(y_i - \boldsymbol{x}_i^\top \boldsymbol{b}) + g(\boldsymbol{b}), \tag{2.2}$$

where $\rho : \mathbb{R} \to \mathbb{R}$ is the loss and $g : \mathbb{R}^p \to \mathbb{R}$ is the penalty.

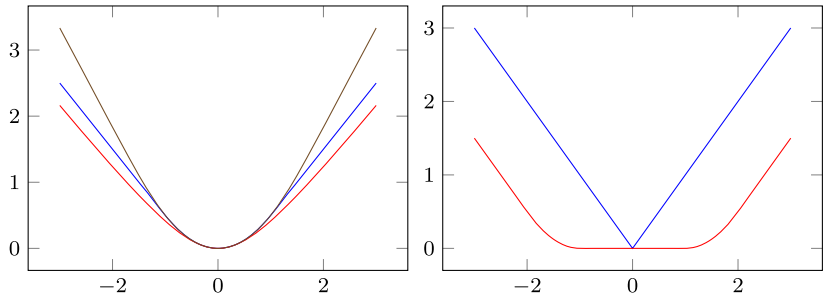


TABLE 1

Left: robust loss functions satisfying Assumption A: the Huber loss $x \mapsto H(x) = \int_0^{|x|} \min(t,1)dt$, *its smoothed versions* $x \mapsto \sqrt{1+x^2}$ *and* $x \mapsto \frac{x^2}{2} I\{|x| \le 1\} + (\frac{1}{6} - \frac{|x|}{2} + x^2 - \frac{|x|^3}{6}) I\{|x| \in (1,2)\} + (\frac{-7}{6} + \frac{3|x|}{2}) I\{(|x| \ge 2)\}$. *Right: two loss functions that do not satisfy Assumption A: the absolute deviation loss* $x \mapsto |x|$ *and the* 1*-insensitive loss* $x \mapsto H[(|x|-1)_+]$ *where* $H(\cdot)$ *is the Huber loss.*

Assumption A (Assumptions on the loss)**.** *Let* $\rho : \mathbb{R} \to \mathbb{R}$ *be convex and continuously differentiable on* $\mathbb{R}$, *with derivative* $\psi = \rho'$ *being* L*-Lipschitz with*

$$K^2 \le \psi'(x) + \psi(x)^2 \qquad \text{for almost every } x \in \mathbb{R} \tag{2.3}$$

for some positive constant $K > 0$ *independent of* n, p.

Two families of robust losses that do not satisfy Assumption A are non-differentiable losses such as the least absolute deviations $\rho(x) = |x|$, and δ-insensitive losses such as $\rho(x) = (|x| - \delta)_+^2$ as $\psi(x)^2 + \psi'(x) = 0$ in a neighborhood of 0. Assumption A is verified by the Huber loss $\rho(x) = \int_0^{|x|} \min(1,t)dt$ with $K = 1$, as well as by any smooth version of the Huber loss, for instance $\rho(x) = \sqrt{1+x^2}$ with $K^2 = \frac{23}{27} \approx 0.852$. The one-sided logistic loss $\rho(x) = \log(1+e^x)$ also satisfies Assumption A.

Assumption B (Strong convexity of g)**.** *For some constant* $\tau > 0$ *independent of* n, p, *the penalty* $g : \mathbb{R}^p \to \mathbb{R}$ *is* τ*-strongly convex in the sense that* $\boldsymbol{x} \mapsto g(\boldsymbol{x}) - \tau \|\boldsymbol{x}\|_2^2/2$ *is convex.*

Some useful characterizations of strong convexity (Assumption B) are the following. Throughout, $\partial g(\boldsymbol{b})$ denotes the subdifferential of g at $\boldsymbol{b} \in \mathbb{R}^p$. Then g is τ-strongly convex if and only if

$$g(\boldsymbol{a}) - g(\boldsymbol{b}) \ge \boldsymbol{u}^\top(\boldsymbol{a} - \boldsymbol{b}) + (\tau/2)\|\boldsymbol{a} - \boldsymbol{b}\|_2^2 \quad \text{for all } \boldsymbol{u} \in \partial g(\boldsymbol{a}) \text{ and } \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p. \tag{2.4}$$

Similarly g is τ-strongly convex if and only for any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$

$$(\boldsymbol{u} - \boldsymbol{v})^\top(\boldsymbol{a} - \boldsymbol{b}) \ge \tau \|\boldsymbol{a} - \boldsymbol{b}\|_2^2 \quad \text{for all } \boldsymbol{u} \in \partial g(\boldsymbol{a}), \boldsymbol{v} \in \partial g(\boldsymbol{b}). \tag{2.5}$$

As ψ in Assumption A is increasing, Assumption A implies that ρ is $K^2/2$-strongly convex in the interval $\{x \in \mathbb{R} : \psi(x)^2 \le K^2/2\}$, and conversely if ρ is μ-strongly convex in the interval $\{x \in \mathbb{R} : \psi(x)^2 \le C\}$ then Assumption A is satisfied with $K^2 = \min(\mu, C)$.

Assumption C. *The rows of the design matrix $\boldsymbol{X}$ are iid $N(\mathbf{0}, \boldsymbol{\Sigma})$ random vectors and all the eigenvalues of $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ are in $[\kappa, 1/\kappa]$ for some constant $\kappa \in (0,1)$ independent of n, p. The noise $\boldsymbol{\varepsilon}$ is independent of $\boldsymbol{X}$ and admits a density with respect to the Lebesgue measure.*

Assumption D. *$p/n \le \gamma$ for some constant $\gamma > 0$ independent of n, p.*

2.2. Notation

Throughout the paper $\Phi(t) = (2\pi)^{-1/2}\int_{-\infty}^{t}\exp(-u^2/2)du$ is the standard normal cumulative distribution function. Furthermore, $[n] = \{1, ..., n\}$ and we use the notation

$$\boldsymbol{\psi} = (\psi(y_i - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}))_{i\in[n]}, \qquad \boldsymbol{\psi}' = (\psi'(y_i - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}))_{i\in[n]}, \qquad \boldsymbol{h} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}. \quad (2.6)$$

For each $j \in [p]$, let $\boldsymbol{e}_j$ denote the j-th vector in the standard basis of $\mathbb{R}^p$, and let

$$\Omega_{jj} = \boldsymbol{e}_j^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j, \qquad \boldsymbol{z}_j = \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\Omega_{jj}^{-1}\boldsymbol{e}_j, \qquad \boldsymbol{Q}_j = \boldsymbol{I}_p - \boldsymbol{\Sigma}^{-1}\Omega_{jj}^{-1}\boldsymbol{e}_j\boldsymbol{e}_j^\top. \quad (2.7)$$

We remark that the above definition implies the following properties:

- $\boldsymbol{X} = \boldsymbol{X}\boldsymbol{Q}_j + \boldsymbol{z}_j\boldsymbol{e}_j^\top$ and $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{Q}_j\boldsymbol{\beta} + \boldsymbol{z}_j\beta_j$.
- $\boldsymbol{z}_j \sim N(\mathbf{0}, \Omega_{jj}^{-1}\boldsymbol{I}_n)$ is independent of $\boldsymbol{X}\boldsymbol{Q}_j$ (cf. Proposition D.1).
- Under Assumption C, $\Omega_{jj} \in [\kappa, 1/\kappa]$.

By construction of $\boldsymbol{z}_j$ and $\boldsymbol{Q}_j$, the response $\boldsymbol{y}$ can be decomposed as $\boldsymbol{y} = \beta_j\boldsymbol{z}_j + \boldsymbol{X}\boldsymbol{Q}_j\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\beta_j \in \mathbb{R}$ is the scalar parameter of interest for a fixed covariate $j \in [p]$, the vector $\boldsymbol{Q}_j\boldsymbol{\beta}$ is a high-dimensional nuisance parameter and $\boldsymbol{\varepsilon}$ is independent noise. Under the additional assumption of $\varepsilon_i \sim N(0,1)$, Ω_{jj}^{-1} is the Fisher information for the estimation of β_j.

In the proof, it will be useful to treat $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\varepsilon}, \boldsymbol{X})$ as a map from $\mathbb{R}^{n\times(p+1)} \to \mathbb{R}^n$, formally defined as

$$\begin{aligned}\hat{\boldsymbol{\beta}}(\boldsymbol{\varepsilon}, \boldsymbol{X}) &= \operatorname*{argmin}_{\boldsymbol{b}\in\mathbb{R}^p}\sum_{i\in[n]}\frac{\rho(\varepsilon_i - \boldsymbol{x}_i^\top(\boldsymbol{b}-\boldsymbol{\beta}))}{n} + g(\boldsymbol{b}),\\ \boldsymbol{\psi}(\boldsymbol{\varepsilon}, \boldsymbol{X}) &= \psi(\boldsymbol{\varepsilon} + \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\varepsilon}, \boldsymbol{X})).\end{aligned} \quad (2.8)$$

Since $(\boldsymbol{\beta}, \boldsymbol{\varepsilon})$ are unknown, we cannot compute the derivatives of $\boldsymbol{\psi}$ a priori. However, for a fixed $\boldsymbol{X}$, the quantity $\boldsymbol{\psi}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X})$ is observable since all terms in $(\boldsymbol{\beta}, \boldsymbol{\varepsilon})$ cancel out (indeed $\boldsymbol{\psi}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X})$ is simply $\psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ with $\hat{\boldsymbol{\beta}}$ in (2.2)). We can thus define the observable matrix of size $n \times n$

$$[\nabla_{\boldsymbol{y}}\boldsymbol{\psi}]^\top \stackrel{\text{def}}{=} (\partial/\partial\boldsymbol{y})\boldsymbol{\psi}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X}) \quad (2.9)$$

holding $\boldsymbol{X}$ fixed, at every point $\boldsymbol{y} \in \mathbb{R}^n$ where $\boldsymbol{y} \mapsto \boldsymbol{\psi}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X})$ is differentiable. By Proposition C.4 below, the map $\boldsymbol{y} \mapsto \boldsymbol{\psi}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X})$ is L-Lipschitz

and $\nabla_{\boldsymbol{y}}\boldsymbol{\psi}$ exists at Lebesgue almost every $\boldsymbol{y}$, and with probability one since $\boldsymbol{y}$ has continuous distribution under Assumption C. Furthermore, the gradient (2.9) at the currently observed data $(\boldsymbol{y}, \boldsymbol{X})$ does not depend on any unknown quantity. It can be computed from $(\boldsymbol{y}, \boldsymbol{X})$ either by finding a closed form expression for (2.9) for a given penalty function, or by approximation using finite difference or other numerical methods (e.g., [4, Section 2.7] in the same regularized M-estimation setting as the present paper). Note that in the concurrent paper [5] that focuses on asymptotic normality of residuals and estimation of the generalization error, the $n \times n$ matrix (2.9) is denoted $\boldsymbol{V}$ [5, (2.3)], and $\boldsymbol{V} = \operatorname{diag}(\boldsymbol{\psi}')(\boldsymbol{I}_n - \boldsymbol{X}\widehat{\boldsymbol{A}}\boldsymbol{X}^T \operatorname{diag}(\boldsymbol{\psi}'))$ for the matrix $\widehat{\boldsymbol{A}} \in \mathbb{R}^{p\times p}$ defined in [5, Theorem 1].

2.3. Main result

In the following result, we consider a sequence of integer pairs (n, p), regression problems (2.1) and M-estimator (2.2), without explicit reference to their dependence on n or p. For instance, one can think of $p = p_n$ as a nondecreasing function of n and $(g, \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}, \rho)$ are also implicitly indexed by n with values possibly changing with n.

Theorem 2.1 (Asymptotic Normality result for M-estimator)**.** *Consider the linear model (2.1) and the M-estimator $\hat{\boldsymbol{\beta}}$ in (2.2). Assume $\mathbb{E}[\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|_2^2] \le \mathscr{R} < +\infty$. Let Assumptions A, B, C and D be fulfilled for constants $\mathscr{R}, \tau, \kappa, K, L, \gamma > 0$ independent of n, p. Define the map $\boldsymbol{y} \mapsto \boldsymbol{\psi}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X})$ and its Jacobian $[\nabla_{\boldsymbol{y}}\boldsymbol{\psi}]^\top$ in (2.9) holding $\boldsymbol{X}$ fixed. For each $j \in [p]$ let*

$$\xi_j = \|\boldsymbol{\psi}\|^{-1}\big[\boldsymbol{\psi}^\top \boldsymbol{z}_j - n^{-1}\|\boldsymbol{z}_j\|_2^2(\beta_j - \widehat{\beta}_j)\operatorname{tr}(\nabla_{\boldsymbol{y}}\boldsymbol{\psi})\big], \tag{2.10}$$

$$\xi_j' = \|\boldsymbol{\psi}\|^{-1}\big[\boldsymbol{\psi}^\top \boldsymbol{z}_j - \Omega_{jj}^{-1}(\beta_j - \widehat{\beta}_j)\operatorname{tr}(\nabla_{\boldsymbol{y}}\boldsymbol{\psi})\big]. \tag{2.11}$$

Then for any positive sequence (a_p) with $\lim_{p\to+\infty} a_p = +\infty$,

$$\max_{j\in J_{n,p}} \Big|\mathbb{P}\big(\Omega_{jj}^{1/2}\xi_j \le t\big) - \Phi(t)\Big| + \Big|\mathbb{P}\big(\Omega_{jj}^{1/2}\xi_j' \le t\big) - \Phi(t)\Big| \to 0, \tag{2.12}$$

for some $J_{n,p} \subseteq [p]$ satisfying $|J_{n,p}|/p \ge 1 - a_p/p$.

The proof is given in Appendix A.2. To interpret the above result, we remark that for any slowly increasing sequence a_p such as $a_p = \log p$ or $a_p = \log\log p$, the asymptotic normality (2.12) holds uniformly over all coordinates $j = 1, ..., p$ except a_p of them, so that both ξ_j and ξ_j' are asymptotically pivotal for an overwhelming majority of β_j. Another interpretation is given in the following Corollary: For any given precision threshold $\upsilon > 0$, there exist at most a_* coordinates $j = 1, ..., p$ such that $|\mathbb{P}(\Omega_{jj}^{1/2}\xi_j \le t) - \Phi(t)| \ge \upsilon$ where a_* is a constant independent of n, p.

Corollary 2.2. *Let the setting and assumptions of Theorem 2.1 be fulfilled. For any arbitrarily small constant $\upsilon > 0$ independent of n, p, define*

$$J_{n,p}^{\upsilon} = \Big\{j \in [p] : \Big|\mathbb{P}(\Omega_{jj}^{1/2}\xi_j \le t) - \Phi(t)\Big| > \upsilon\Big\}.$$

Then, $\sup_{n,p}|J^{\upsilon}_{n,p}| \le a_*$ *for a certain constant* a_* *depending on* $\{\upsilon, \tau, \mathscr{R}, \gamma, L, K, t\}$ *only. In other words, for any* (n,p) *with* $p/n \le \gamma$ *there are at most a constant number of coordinates* $j = 1, ..., p$ *such that* $\left|\mathbb{P}(\Omega_{jj}^{1/2}\xi_j \le t) - \Phi(t)\right| > \upsilon$. *The same conclusion holds with* ξ_j *replaced by* ξ'_j.

Proof of Corollary 2.2. We proceed by contradiction. If the claim does not hold, there exists a constant $\upsilon_* > 0$ such that $|J^{\upsilon_*}_{n,p}| \ge 2a_p$ for an unbounded sequence a_p. By extracting a subsequence if necessary, we may assume without loss of generality that a_p is monotonically increasing with $a_p \to +\infty$. By Theorem 2.1 there exists $J_{n,p} \subset [p]$ such that (2.12) holds. By definition of $J^{\upsilon_*}_{n,p}$ and $J_{n,p}$, we have $J_{n,p} \cap J^{\upsilon_*}_{n,p} = \emptyset$ for p large enough. This implies that $p \ge |J_{n,p}| + |J^{\upsilon_*}_{n,p}| \ge (p - a_p) + 2a_p$ for p large enough, a contradiction. □

Remark 2.1. Theorem 2.1 requires the assumption $\mathbb{E}[\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2] \le \mathscr{R} < +\infty$ for some constant $\mathscr{R}$. This assumes that the expected risk of $\hat{\boldsymbol{\beta}}$ is bounded, which holds true under the following additional assumptions:

- The penalty is minimized at 0: $\mathbf{0} \in \operatorname{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} g(\boldsymbol{b})$;
- The loss is Lipschitz: $\sup_{t\in R}|\psi(t)| \le \mathscr{L}$;
- The noiseless signal has bounded variance: $\mathbb{E}[(\boldsymbol{x}_i^\top \boldsymbol{\beta})^2] = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|^2 \le \mathscr{V}$.

Above, $\mathscr{L}, \mathscr{V} > 0$ are constants. Then by the KKT conditions, $\boldsymbol{X}^\top \boldsymbol{\psi} \in n\partial g(\hat{\boldsymbol{\beta}})$ and thanks to (2.5) from Assumption B and the first bullet above,

$$(\hat{\boldsymbol{\beta}} - \mathbf{0})^\top \boldsymbol{X}^\top \boldsymbol{\psi} \in n(\hat{\boldsymbol{\beta}} - \mathbf{0})^\top (\partial g(\hat{\boldsymbol{\beta}}) - \partial g(\mathbf{0}) \ge n\tau\|\hat{\boldsymbol{\beta}}\|^2.$$

If follows by the Cauchy-Schwarz inequality that $n\tau\|\hat{\boldsymbol{\beta}}\| \le \|\boldsymbol{X}\|_{op}\sqrt{n}\mathscr{L}$. Thanks to Assumption C we thus obtain

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2] &\le 2\mathbb{E}[\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|_2^2] + 2\mathbb{E}[\|\boldsymbol{\Sigma}^{1/2}\hat{\boldsymbol{\beta}}\|_2^2] \\ &\le 2\mathscr{V} + 2(\mathscr{L}^2\kappa^{-2}\tau^{-2})\mathbb{E}[\|\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}\|_{op}^2/n].\end{aligned}$$

The upper bound (C.1) completes the proof that we can find a suitable constant $\mathscr{R}$ depending on $\gamma, \tau, \kappa, \mathscr{L}, \mathscr{V}$ only.

Remark 2.2. The set $J_{n,p}$ in (2.12) excludes a few coordinates $j \in \{1, ..., p\} \backslash J_{n,p}$. A natural question is whether the exclusion of some coordinates is necessary or an artefact of the current proofs. By Section 3.7 in [7], there are known examples where a "variance spike" occurs in at most a finite number of coordinates $j \in \{1, ..., p\}$: for these few coordinates, the asymptotic variance of ξ_j or ξ'_j is strictly larger than 1 and the convergence in distribution $\xi_j \to^d N(0,1)$ cannot hold.

2.4. Data-driven variance estimate

Except for at most a constant number of coordinates $j \in [p]$, the approximation

$$\hat{V}^{-1/2}\Omega_{jj}^{-1/2}\sqrt{n}(\hat{\beta}_j - \beta_j) + [\text{bias correction}] \approx N(0,1) \tag{2.13}$$

holds, where the bias correction is observable and determined by the first term in (2.11) and the data-driven variance estimate is

$$\hat{V} = \frac{\|\boldsymbol{\psi}\|_2^2/n}{(\mathrm{tr}(\nabla_{\boldsymbol{y}}\boldsymbol{\psi})/n)^2} = \frac{n^{-1}\sum_{i=1}^n \psi_i^2}{\big(n^{-1}\sum_{i=1}^n (\partial/\partial y_i)\psi_i\big)^2}. \tag{2.14}$$

which characterizes the length of confidence intervals for β_j. This ratio of an average of ψ^2 by a squared average of a derivative of ψ mirrors the robust asymptotic results in (1.1) and (1.4) as discussed in the introduction. Confidence intervals can be readily constructed from Theorem 2.1 or the informal approximation (2.13): a 95%-confidence interval for β_j is given by $\hat{\beta}_j + (\Omega_{jj}\hat{V}/n)^{1/2}([\text{bias correction}] \pm 1.96)$, that is,

$$\Big[\hat{\beta}_j + \frac{\Omega_{jj}\boldsymbol{\psi}^\top \boldsymbol{z}_j}{\mathrm{tr}[\nabla_{\boldsymbol{y}}\boldsymbol{\psi}]} - \Big(\frac{\Omega_{jj}\hat{V}}{n}\Big)^{1/2} 1.96, \quad \hat{\beta}_j + \frac{\Omega_{jj}\boldsymbol{\psi}^\top \boldsymbol{z}_j}{\mathrm{tr}[\nabla_{\boldsymbol{y}}\boldsymbol{\psi}]} + \Big(\frac{\Omega_{jj}\hat{V}}{n}\Big)^{1/2} 1.96\Big].$$

In contrast with the asymptotic variance in (1.4), the above $\hat{V}$ involves observable quantities. In particular, while the asymptotic variance in (1.4) depends on the distribution of the noise through the solutions (r, c) of the system (1.3), the knowledge of the noise distribution is not required to compute $\hat{V}$ and construct confidence intervals for β_j.

Theorem 2.1 is valid for $p/n \le \gamma$, without requiring a specific limit for the ratio p/n. Theorem 2.1 is also valid for $p = o(n)$, so that Theorem 2.1 and the estimated asymptotic variance (2.14) unifies both low- and high-dimensional asymptotic normality results.

For the Huber loss

$$H(u) = \int_0^{|u|} \min(1, t)dt = \begin{cases} u^2/2 & \text{if } |u| \le 1, \\ |u| - 1/2 & \text{if } |u| > 1, \end{cases} \tag{2.15}$$

the quantity $\mathrm{tr}[\nabla_{\boldsymbol{y}}\boldsymbol{\psi}]$ has a simpler form: By the chain rule

$$\mathrm{tr}[\nabla_{\boldsymbol{y}}\boldsymbol{\psi}] = \mathrm{tr}[\mathrm{diag}\,\boldsymbol{\psi}'] - \mathrm{tr}[\mathrm{diag}(\boldsymbol{\psi}')(\partial/\partial\boldsymbol{y})\boldsymbol{X}\hat{\boldsymbol{\beta}}]$$

where the differentiation is understood holding $\boldsymbol{X}$ fixed as in (2.9). With $\hat{n} = \mathrm{tr}[\mathrm{diag}\,\boldsymbol{\psi}']$ the number of observations such that $y_i - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}$ falls in the range where the Huber loss is quadratic and $\hat{\mathsf{df}} = \mathrm{tr}[\mathrm{diag}(\boldsymbol{\psi}')(\partial/\partial\boldsymbol{y})\boldsymbol{X}\hat{\boldsymbol{\beta}}]$ representing the degrees-of-freedom of the M-estimator $\hat{\boldsymbol{\beta}}$, the quantity $\mathrm{tr}[\nabla_{\boldsymbol{y}}\boldsymbol{\psi}]$ appearing in the dominator of $\hat{V}$ simplifies to $\mathrm{tr}[\nabla_{\boldsymbol{y}}\boldsymbol{\psi}] = \hat{n} - \hat{\mathsf{df}}$. In this case, the asymptotic normality for ξ_j' in Theorem 2.1 takes the form

$$\Omega_{jj}^{1/2}\xi_j' = \frac{(\hat{n} - \hat{\mathsf{df}})(\hat{\beta}_j - \beta_j) + \Omega_{jj}\boldsymbol{z}_j^\top\psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{\|\psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\|_2}\Omega_{jj}^{-1/2} \quad \to^d \quad N(0,1) \tag{2.16}$$

uniformly over all $j \in J_{n,p}$. This extends the asymptotic normality result (1.7) to the Huber loss with the following important modifications: the sample size

n is replaced by $\hat{n}$ and the residuals $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ are replaced by $\boldsymbol{\psi} = \psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$. The variance $\hat{V}$ that determines the length of the confidence interval resulting from (2.13) presents a trade-off among $\|\boldsymbol{\psi}\|^2/n$, an effective sample size $\hat{n}$ and the degrees-of-freedom $\hat{\mathsf{df}}$: For confidence intervals with small length, the M-estimator $\hat{\boldsymbol{\beta}}$ should have small residuals measured as $\|\psi(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})\|_2$, small degrees-of-freedom $\hat{\mathsf{df}}$, and large effective sample size $\hat{n}$.

2.5. Example

This section specializes the above results to scaled versions of the Huber loss (2.15) and the Elastic-Net penalty $g(\boldsymbol{b}) = \lambda\|\boldsymbol{b}\|_1 + \tau\|\boldsymbol{b}\|_2^2/2$ for tuning parameters $\lambda, \tau > 0$. We consider the M-estimator

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \sigma^2 H(\sigma^{-1}(y_i - \boldsymbol{x}_i^\top \boldsymbol{b})) + \lambda\|\boldsymbol{b}\|_1 + \tau\|\boldsymbol{b}\|_2^2/2, \tag{2.17}$$

which corresponds to the scaled Huber loss $\rho(u) = \sigma^2 H(\sigma^{-1}u)$ where $\sigma > 0$ is a scaling parameter. Since the derivative H' is 1-Lipschitz, so is $\psi = \rho'$. Furthermore, $\psi'(u) = H''(\sigma^{-1}u) = 1$ for $|u| \le \sigma$ and $|\psi(u)| = \sigma|H'(\sigma^{-1}u)| = \sigma$ for $|u| > \sigma$, so that $\min_{x\in\mathbb{R}}[\psi^2(x) + \psi'(x)] \ge \min(1, \sigma^2)$ and Assumption A is satisfied with $L = 1$ and $K^2 = \min(1, \sigma^2)$. Assumption B is also satisfied as the penalty is the sum of the ℓ_1 norm plus $\tau\|\boldsymbol{b}\|^2/2$. The quantity $\operatorname{tr}[\nabla_{\boldsymbol{y}}\boldsymbol{\psi}]$ appearing in Theorem 2.1 in the denominator of estimated variance $\hat{V}$ is computed in [4, Proposition 2.3]: For almost every $(\boldsymbol{X}, \boldsymbol{y})$,

$$\begin{aligned}
&\operatorname{tr}[\nabla_{\boldsymbol{y}}\boldsymbol{\psi}] = \hat{n} - \hat{\mathsf{df}},\\
&\qquad \hat{\mathsf{df}} = \operatorname{tr}\big[\operatorname{diag}(\boldsymbol{\psi}')\boldsymbol{X}_{\hat{S}}(\boldsymbol{X}_{\hat{S}}^\top \operatorname{diag}(\boldsymbol{\psi}')\boldsymbol{X}_{\hat{S}} + n\tau\boldsymbol{I}_{|\hat{S}|})^{-1}\boldsymbol{X}_{\hat{S}}^\top \operatorname{diag}(\boldsymbol{\psi}')\big],
\end{aligned}$$

where $\hat{n} = \operatorname{tr}[\operatorname{diag}(\boldsymbol{\psi}')] = |\{i \in [n] : \psi'(y_i - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}) = 1\}|$ is the number of observations $i = 1, ..., n$ such that $y_i - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}$ falls in the range where ρ is quadratic, $\hat{S} = \{j \in [p] : \hat{\beta}_j \ne 0\}$ and $\boldsymbol{X}_{\hat{S}} \in \mathbb{R}^{n\times|\hat{S}|}$ is the submatrix of $\boldsymbol{X}$ containing columns indexed in $\hat{S}$.

2.6. Simulation study

We provide here simulations to showcase the asymptotic normality in the Huber loss and Elastic-Net penalty example of the previous section.

We set $n = 200, p = 300$, and generate $\varepsilon_i \sim N(0, 1)$ and the coordinates of $\boldsymbol{\beta}$ from iid Bernoulli variables with parameter 0.1. We compute 1000 simulations of the Z-score $\Omega_{jj}^{1/2}\xi_j'$ from Theorem 2.1 for $j = 1$ for the M-estimator with the Huber loss and the Elastic-Net penalty (2.17) with $\sigma = 1$ and the four combinations $(\lambda, \tau) \in \{n^{-1/2}, 2n^{-1/2}\} \times \{0, 0.1\}$. The covariance matrix $\boldsymbol{\Sigma}$ is set as $\boldsymbol{R}^\top\boldsymbol{R}/(2p)$ where $\boldsymbol{R} \in \mathbb{R}^{2p\times p}$ has iid Rademacher entries; $\boldsymbol{\Sigma}$ is generated once and is the same across the 1000 simulations. The average value and standard

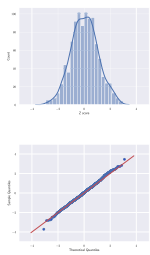
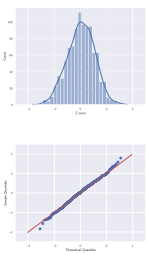
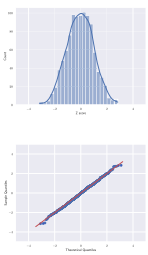
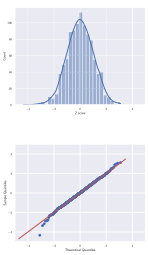
(λ, τ)	$(n^{-1/2}, 0.1)$	$(n^{-1/2}, 0)$	$(2n^{-1/2}, 0.1)$	$(2n^{-1/2}, 0)$
$\hat{n}$	81.2 $\pm$ 5.8	111.9 $\pm$ 6.4	38.3 $\pm$ 4.6	49.3 $\pm$ 5.5
$\hat{\mathsf{df}}$	53.6 $\pm$ 3.9	81.4 $\pm$ 5.1	14.2 $\pm$ 2.3	24.8 $\pm$ 4.1
$\lvert\hat{S}\rvert$	96.4 $\pm$ 6.8	81.4 $\pm$ 5.1	26.5 $\pm$ 4.6	24.8 $\pm$ 4.1
$\sqrt{\hat{V}/n}$	0.44 $\pm$ 0.06	0.37 $\pm$ 0.07	0.56 $\pm$ 0.09	0.55 $\pm$ 0.11
$\Omega_{jj}^{1/2}\xi'_j$				

TABLE 2

Averages and standard errors over 1000 simulations of $(\hat{n}, \hat{\mathsf{df}}, |\hat{S}|, \hat{V}^{1/2}n^{-1/2})$*, as well as histograms and QQ-plots for* $\Omega_{jj}^{1/2}\xi'_j$ *given in* (2.11)*. The coordinate* j *is always* $j = 1$*. The noise* ε_i *is set as iid standard Cauchy. The red line on the QQ-plots is the diagonal line with equation* $x = y$*.*

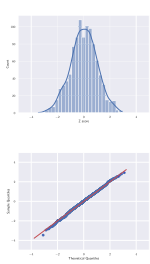
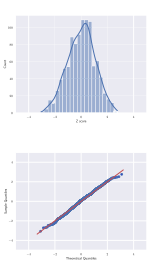
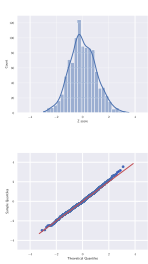
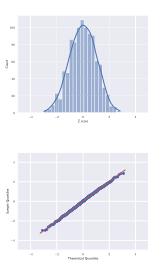
(λ, τ)	$(n^{-1/2}, 0.1)$	$(n^{-1/2}, 0)$	$(2n^{-1/2}, 0.1)$	$(2n^{-1/2}, 0)$
$\hat{n}$	90.0 $\pm$ 5.8	125.7 $\pm$ 6.4	42.5 $\pm$ 4.8	55.4 $\pm$ 5.7
$\hat{\mathsf{df}}$	57.8 $\pm$ 3.8	82.1 $\pm$ 4.9	16.5 $\pm$ 2.4	28.1 $\pm$ 4.3
$\lvert\hat{S}\rvert$	97.8 $\pm$ 6.5	82.1 $\pm$ 4.9	29.7 $\pm$ 4.5	28.1 $\pm$ 4.3
$\sqrt{\hat{V}/n}$	0.37 $\pm$ 0.04	0.25 $\pm$ 0.04	0.51 $\pm$ 0.08	0.48 $\pm$ 0.09
$\Omega_{jj}^{1/2}\xi'_j$				

TABLE 3

Averages and standard errors over 1000 simulations of $(\hat{n}, \hat{\mathsf{df}}, |\hat{S}|, \hat{V}^{1/2}n^{-1/2})$*, as well as histograms and QQ-plots for* $\Omega_{jj}^{1/2}\xi'_j$ *given in* (2.11)*. The coordinate* j *is always* $j = 1$*. The noise* ε_i *is set as iid t-distribution with degree of freedom 2. The red line on the QQ-plots is the diagonal line with equation* $x = y$*.*

error over the 1000 simulations of $\hat{n}$, $\hat{\mathsf{df}}$, $|\hat{S}|$ and $\hat{V}^{1/2}n^{-1/2}$ are presented in Table 2 for $\boldsymbol{\varepsilon}$ with iid standard Cauchy components and Table 3 for $\boldsymbol{\varepsilon}$ with iid components from the t-distribution with 2 degrees of freedom, together with

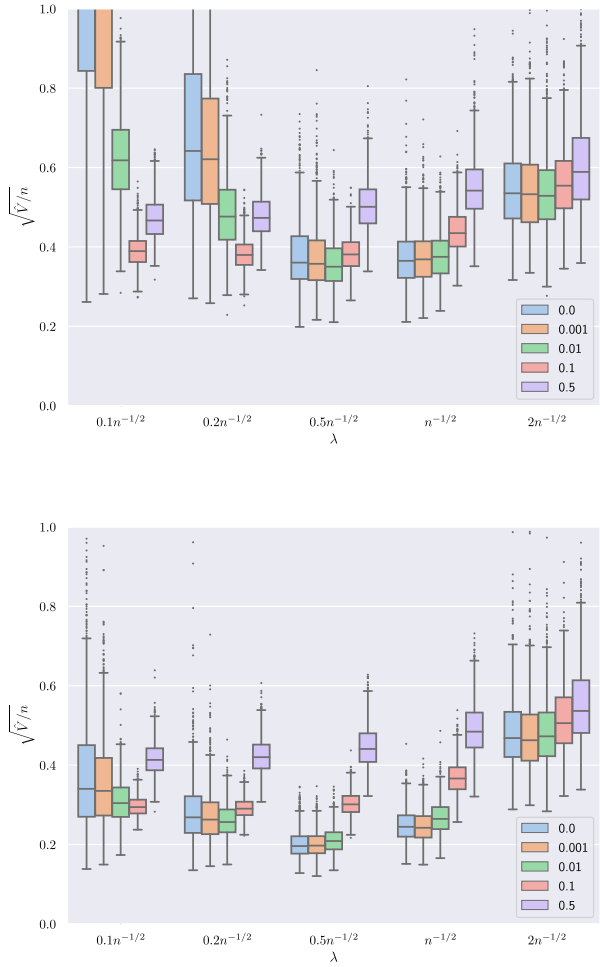


FIG 1. *Boxplots of simulated* $\hat{V}^{1/2}n^{-1/2}$. *The simulation setup is described in Section 2.6. The top plot corresponds to iid standard Cauchy noise* ε_i. *The noise in the second plot is iid from the t-distribution with 2 degrees of freedom. Different colors correspond to different values of* $\tau \in \{0, 10^{-3}, 10^{-2}, 0.1, 0.5\}$. *In the x-axis,* λ *takes values in* $\{0.1n^{-1/2}, 0.2n^{-1/2}, 0.5n^{-1/2}, n^{-1/2}, 2n^{-1/2}\}$.

histograms and QQ-plots against standard normal quantiles of $\Omega_{jj}^{1/2}\xi_j'$ in (2.16).

The quantity $\hat{V}^{1/2}n^{-1/2} = \|\boldsymbol{\psi}\|_2/(\hat{n} - \hat{\mathsf{df}})$ featured in the boxplots of Figure 1 characterizes the length of our confidence intervals in (2.13). Computing the values $\hat{V}^{1/2}$ for different tuning parameters lets the practitioner pick the tuning parameters leading to the smallest confidence interval width, although this process amounts to the construction of multiple confidence intervals and warrants a Bonferroni multiple testing correction.

The histograms and QQ-plots in Table 2 and Table 3 confirm the normality

of ξ_j' for these two heavy-tailed continuous noise distributions.

Since $\hat{n}-\hat{\mathsf{df}}$ appears in the denominator of $\hat{V}$, the length of confidence intervals can be large if $\hat{n}-\hat{\mathsf{df}}$ is nearly zero and the length is infinite if $\hat{n}-\hat{\mathsf{df}}=0$. This explains the large values and large variances observed in the boxplots of Figure 1 for small tuning parameters.

Appendix A: Proof of the main result

Throughout, $\phi_{\min}(\boldsymbol{M})$ and $\phi_{\max}(\boldsymbol{M})$ denote the smallest and largest eigenvalues of a positive definite matrix $\boldsymbol{M}$. We also recall the important notation defined in (2.6)-(2.7): $\boldsymbol{h}=\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}$ is the error vector, $\boldsymbol{\psi}$ and $\boldsymbol{\psi}'$ are the random vectors in $\mathbb{R}^n$ with coordinates $(\psi(y_i-\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}))_{i\in[n]}$ and $(\psi'(y_i-\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}))_{i\in[n]}$ respectively. Finally,

$$\Omega_{jj}=\boldsymbol{e}_j^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j,\qquad \boldsymbol{z}_j=\boldsymbol{X}\boldsymbol{\Sigma}^{-1}\Omega_{jj}^{-1}\boldsymbol{e}_j,\qquad \boldsymbol{Q}_j=\boldsymbol{I}_p-\boldsymbol{\Sigma}^{-1}\Omega_{jj}^{-1}\boldsymbol{e}_j\boldsymbol{e}_j^\top$$

and $\boldsymbol{z}_j\sim N(\boldsymbol{0},\Omega_{jj}^{-1}\boldsymbol{I}_n)$ is independent of $\boldsymbol{X}\boldsymbol{Q}_j$ (cf. Proposition D.1).

A.1. Supporting propositions

The proof of Theorem 2.1 relies on the two intermediary results given below. Proposition A.1 will be proved in Appendix B and Lemma A.2 in Appendix C.

Proposition A.1. *Let $n\geq 3$, $R>0$ and $\boldsymbol{z}\sim\mathrm{Unif}(\mathbb{S}^{n-1}(R))$, where $\mathbb{S}^{n-1}(R)$ is sphere of radius R in $\mathbb{R}^n$. Assume either:*

- *$\boldsymbol{f}$ is locally Lipschitz on $\mathbb{S}^{n-1}(R)$ with $\mathbb{E}[\|\nabla\boldsymbol{f}(\boldsymbol{z})^\top\|_F^2]<\infty$.*
- *$\boldsymbol{f}(\boldsymbol{z})$ is of the form $\widetilde{\boldsymbol{f}}(\boldsymbol{z})/\|\widetilde{\boldsymbol{f}}(\boldsymbol{z})\|_2$ where $\widetilde{\boldsymbol{f}}:\mathbb{R}^n\to\mathbb{R}^n$ is locally Lipschitz and satisfies $\mathbb{P}(\|\boldsymbol{f}(\boldsymbol{z})\|_2\neq 0)=1$ and $\mathbb{E}[\|\widetilde{\boldsymbol{f}}(\boldsymbol{z})\|_2^{-2}\|\nabla\widetilde{\boldsymbol{f}}(\boldsymbol{z})\|_F^2]<+\infty$.*

Define $\boldsymbol{P}_{\boldsymbol{z}}^\perp=\boldsymbol{I}_n-\boldsymbol{z}\boldsymbol{z}^\top/\|\boldsymbol{z}\|_2^2$ and $\xi_{\boldsymbol{f}}(\boldsymbol{z})=\boldsymbol{f}(\boldsymbol{z})^\top\boldsymbol{z}-R^2n^{-1}\operatorname{tr}(\nabla\boldsymbol{f}(\boldsymbol{z})^\top\boldsymbol{P}_{\boldsymbol{z}}^\perp)$. Then

$$\mathbb{E}\big[\big(\xi_{\boldsymbol{f}}(\boldsymbol{z})-\mathbb{E}[\boldsymbol{f}(\boldsymbol{z})]^\top\boldsymbol{z}\big)^2\big]\leq 2R^4(n^2-2n)^{-1}\mathbb{E}[\|\nabla\boldsymbol{f}(\boldsymbol{z})^\top\boldsymbol{P}_{\boldsymbol{z}}^\perp\|_F^2],\tag{A.1}$$

$$\mathbb{E}\big[\big|\|\boldsymbol{f}(\boldsymbol{z})\|_2-\|\mathbb{E}[\boldsymbol{f}(\boldsymbol{z})]\|_2\big|\big]^2\leq R^2(n-2)^{-1}\mathbb{E}[\|\nabla\boldsymbol{f}(\boldsymbol{z})^\top\boldsymbol{P}_{\boldsymbol{z}}^\perp\|_F^2].\tag{A.2}$$

Lemma A.2. *Let Assumptions A, B and C be fulfilled. Let $\eta_n=\sqrt{2\log(n)/n}+n^{-1/2}$, define the events*

$$\mathscr{E}_j=\{\|\boldsymbol{X}\boldsymbol{Q}_j\boldsymbol{\Sigma}^{-1/2}n^{-1/2}\|_{op}\leq 1+\sqrt{p/n}+\eta_n\}\cap\{\big|n^{-1/2}\Omega_{jj}^{1/2}\|\boldsymbol{z}_j\|_2-1\big|\leq\eta_n$$

for each $j\in[p]$, define $u\geq 0$ by

$$u_*=\big[K^2(1-1/n)(L\tau^{-1}\|\boldsymbol{\Sigma}\|_{op}(2+\sqrt{p/n}+2\eta_n)^2+1)^{-1}-3L/n\big]_+,$$

and let $\mathbb{E}_j$ *be the conditional expectation given* $(\|\boldsymbol{z}_j\|_2, \boldsymbol{X}\boldsymbol{Q}_j, \boldsymbol{\varepsilon})$. *Then, when* $u_* > 0$

$$\sup_{\delta>0} \mathbb{E}\Big[\frac{n}{\|\boldsymbol{\psi}\|_2^2+\delta}\sum_{j\in[p]} I_{\mathscr{E}_j} h_j^2\Big]^{1/2} \le \frac{\big[(1+\sqrt{\frac{p}{n}})^2+\frac{1}{n}\big]^{1/2}}{\phi_{\min}(\boldsymbol{\Sigma})^{1/2}(1-\eta_n)_+^2 u_*} + \frac{\big[\frac{2p}{n\tau}+\mathbb{E}[\|\boldsymbol{h}\|_2^2]\big]^{1/2}}{u_*^{1/2}}. \tag{A.3}$$

A.2. Proof of the main result

Proof of Theorem 2.1. Since $\boldsymbol{\psi} \neq \boldsymbol{0}_n$ for almost every $\boldsymbol{X}$ by Proposition D.2, ξ_j is well-defined with $\mathbb{P}$-probability 1. By Lemma A.2 and the monotone convergence theorem as $\delta \to 0$ for the left-hand side (A.3), when $u_* > 0$

$$\mathbb{E}\Big[\frac{n}{\|\boldsymbol{\psi}\|_2^2}\sum_{j\in[p]} I_{\mathscr{E}_j} h_j^2\Big] \le \frac{\big[(1+\sqrt{\frac{p}{n}})^2+\frac{1}{n}\big]^{1/2}}{\phi_{\min}(\boldsymbol{\Sigma})^{1/2}(1-\eta_n)_+^2 u_*} + \frac{\big[\frac{2p}{n\tau}+\mathbb{E}[\|\boldsymbol{h}\|_2^2]\big]^{1/2}}{u_*^{1/2}}.$$

Under our assumptions, $\|\boldsymbol{\Sigma}\|_{\mathrm{op}} \le 1/\kappa < +\infty$ and $\mathbb{E}[\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|_2^2] \le \mathscr{R} < +\infty$, so that there exists some finite constant $\mathscr{N} > 0$ and $\mathscr{A} < +\infty$ independent of n, p, such that for $n \ge \mathscr{N}$,

$$\sum_{j\in[p]} \mathbb{E}\Big[n\|\boldsymbol{\psi}\|_2^{-2} I_{\mathscr{E}_j} h_j^2\Big] \le \mathscr{A} < +\infty. \tag{A.4}$$

By Markov's inequality with respect to the uniform distribution on $[p] = \{1, ..., p\}$, the set

$$J_{n,p} := \big\{j \in [p] : \mathbb{E}\big[nI_{\mathscr{E}_j}\|\boldsymbol{\psi}\|_2^{-2} h_j^2\big] \le \mathscr{A}/a_p\big\} \quad \text{satisfies} \quad |J_{n,p}|/p \ge 1 - a_p/p. \tag{A.5}$$

While the function $\phi(\boldsymbol{\varepsilon}, \boldsymbol{X})$ in (2.8) is formally $\mathbb{R}^n \times \mathbb{R}^{n\times p} \to \mathbb{R}^n$, in this paragraph it is useful to only consider variations in $\boldsymbol{z}_j$, holding a fixed value of $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_j)$. Since $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_j)$ is independent of $\boldsymbol{z}_j$, the conditional probability distribution of $\boldsymbol{z}_j$ given $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_j)$ is still $N(\boldsymbol{0}, \Omega_{jj}^{-1}\boldsymbol{I}_n)$. To this end and with a slight abuse of notation, for a given, fixed value of $(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_j)$ we view $\boldsymbol{\psi}$ as a function of $\boldsymbol{z}_j$ only,

$$\boldsymbol{\psi} : \mathbb{R}^n \to \mathbb{R}^n, \qquad \boldsymbol{\psi} : \boldsymbol{z}_j \mapsto \boldsymbol{\psi}(\boldsymbol{z}_j) = \boldsymbol{\psi}(\boldsymbol{\varepsilon}, \boldsymbol{X}\boldsymbol{Q}_j + \boldsymbol{z}_j\boldsymbol{e}_j^\top) \tag{A.6}$$

and we denote its Jacobian by $\nabla\boldsymbol{\psi}(\boldsymbol{z}_j)^\top$ at any point $\boldsymbol{z}_j$ where (A.6) is Fréchet differentiable. Next, we argue conditionally on $(\boldsymbol{X}\boldsymbol{Q}_j, \|\boldsymbol{z}_j\|_2)$: Since $\boldsymbol{z}_j/\|\boldsymbol{z}_j\|_2$ is independent of $(\|\boldsymbol{z}_j\|_2, \boldsymbol{X}\boldsymbol{Q}_j)$ and by rotational invariance of the Gaussian distribution, conditionally on $(\|\boldsymbol{z}_j\|_2, \boldsymbol{X}\boldsymbol{Q}_j)$ the vector $\boldsymbol{z}_j$ is uniformly distributed on the sphere $\mathbb{S}^{n-1}(\|\boldsymbol{z}_j\|_2)$. Let $\mathbb{E}_j$ denote the conditional expectation of $\boldsymbol{z}_j$ given $(\boldsymbol{X}\boldsymbol{Q}_j, \|\boldsymbol{z}_j\|_2)$. By Proposition C.4, the above function is locally Lipschitz. From Proposition C.4 (iii) we have that

$$I_{\mathscr{E}_j}\mathbb{E}_j[\|\boldsymbol{\psi}\|_2^{-2}\|\nabla\boldsymbol{\psi}(\boldsymbol{z}_j)\|_F^2] \le L(n\tau)^{-1} + I_{\mathscr{E}_j}\mathbb{E}_j[nL^2\|\boldsymbol{\psi}\|_2^{-2}h_j^2] \tag{A.7}$$

and the right-hand side is finite with probability one with respect to $(\boldsymbol{X}\boldsymbol{Q}_j, \|\boldsymbol{z}_j\|_2)$ thanks to (A.4) and Tonelli's theorem for non-negative measurable functions. By Proposition D.2, conditional on almost every $(\boldsymbol{X}\boldsymbol{Q}_j, \|\boldsymbol{z}_j\|_2)$, $\boldsymbol{\psi}(\boldsymbol{z}_j) \neq \boldsymbol{0}_n$ for almost every $\boldsymbol{z}_j \in \mathbb{R}$. We are now in position to apply Proposition A.1 with $\boldsymbol{f} = \boldsymbol{\psi}/\|\boldsymbol{\psi}\|_2$, $\boldsymbol{z} = \boldsymbol{z}_j$ and

$$\xi_{\boldsymbol{f}}(\boldsymbol{z}_j) := \|\boldsymbol{\psi}\|_2^{-1}\big(\boldsymbol{\psi}^\top \boldsymbol{z}_j - (\|\boldsymbol{z}_j\|_2^2/n)\operatorname{tr}\big[\boldsymbol{P}_{\boldsymbol{\psi}}^\perp(\nabla\boldsymbol{\psi}(\boldsymbol{z}_j))^\top \boldsymbol{P}_{\boldsymbol{z}_j}^\perp\big]\big),$$

where $\boldsymbol{P}_{\boldsymbol{v}}^\perp := \boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{v}}$ with $\boldsymbol{P}_{\boldsymbol{v}} := \boldsymbol{v}\boldsymbol{v}^\top/\|\boldsymbol{v}\|_2^2$ for any $\boldsymbol{v} \in \mathbb{R}^n$. By Proposition A.1 and (A.7),

$$\begin{aligned}
&I_{\mathscr{E}_j}\mathbb{E}_j[(\xi_{\boldsymbol{f}}(\boldsymbol{z}_j) - \mathbb{E}_j[\boldsymbol{f}(\boldsymbol{z}_j)]^\top \boldsymbol{z}_j)^2] \\
&\le 2I_{\mathscr{E}_j}\|\boldsymbol{z}_j\|_2^4(n^2-2n)^{-1}\mathbb{E}_j[\|(\nabla\boldsymbol{f}(\boldsymbol{z}_j))^\top \boldsymbol{P}_{\boldsymbol{z}_j}^\perp\|_F^2] \\
&\le 2(1+\eta_n)^4(1-2/n)^{-1}I_{\mathscr{E}_j}\Omega_{jj}^{-2}\mathbb{E}_j[\|\boldsymbol{\psi}\|_2^{-2}\|\nabla\boldsymbol{\psi}(\boldsymbol{z}_j)\|_F^2] \\
&\le 2(1+\eta_n)^4(1-2/n)^{-1}\kappa^{-2}\big(L(n\tau)^{-1} + I_{\mathscr{E}_j}\mathbb{E}_j[nL^2\|\boldsymbol{\psi}\|_2^{-2}h_j^2]\big)
\end{aligned}$$

and

$$\begin{aligned}
&I_{\mathscr{E}_j}(\|\mathbb{E}_j[\boldsymbol{f}(\boldsymbol{z}_j)]\|_2 - 1)^2 \\
&\le I_{\mathscr{E}_j}\|\boldsymbol{z}_j\|_2^2(n-2)^{-1}\mathbb{E}_j[\|(\nabla\boldsymbol{f}(\boldsymbol{z}_j))^\top \boldsymbol{P}_{\boldsymbol{z}_j}^\perp\|_F^2]. \\
&\le (1+\eta_n)^2(1-2/n)^{-1}I_{\mathscr{E}_j}\Omega_{jj}^{-1}\mathbb{E}_j[\|\boldsymbol{\psi}\|_2^{-2}\|\nabla\boldsymbol{\psi}(\boldsymbol{z}_j)\|_F^2] \\
&\le (1+\eta_n)^2(1-2/n)^{-1}\kappa^{-1}\big(L(n\tau)^{-1} + I_{\mathscr{E}_j}\mathbb{E}_j[nL^2\|\boldsymbol{\psi}\|_2^{-2}h_j^2]\big),
\end{aligned}$$

where the upper bounds follow from $I_{\mathscr{E}_j}\|\boldsymbol{z}_j\|_2^4 \le (1+\eta_n)^4 n^2\Omega_{jj}^{-2}$. Taking $\mathbb{E}$ on both sides, we obtain $\max_{j\in J_{n,p}}\mathbb{E}\big[I_{\mathscr{E}_j}(\xi_{\boldsymbol{f}}(\boldsymbol{z}_j) - \mathbb{E}_j[\boldsymbol{f}(\boldsymbol{z}_j)]^\top \boldsymbol{z}_j)^2 + I_{\mathscr{E}_j}|\|\mathbb{E}_j[\boldsymbol{f}(\boldsymbol{z}_j)]\|_2 - 1|\big] \to 0$. Thanks to $\min_{j\in[p]}\mathbb{P}(\mathscr{E}_j^c) \to 1$ by Lemma C.1 this implies that both $|\|\mathbb{E}_j[\boldsymbol{f}(\boldsymbol{z}_j)]\|_2 - 1|$ and $|\xi_{\boldsymbol{f}}(\boldsymbol{z}_j)^\top \boldsymbol{z}_j - \mathbb{E}_j[\boldsymbol{f}(\boldsymbol{z}_j)]^\top \boldsymbol{z}_j|$ converge in probability to 0 uniformly over $j \in J_{n,p}$.

We now study the asymptotic distribution of $\xi_{\boldsymbol{f}}(\boldsymbol{z}_j)$. Since $\boldsymbol{z}_j \sim N(\boldsymbol{0}_n, \Omega_{jj}^{-1}\boldsymbol{I}_n)$ is independent with $\boldsymbol{X}\boldsymbol{Q}_j$ by Proposition D.1, without loss of generality, we can assume that $\boldsymbol{z}_j = \|\boldsymbol{z}_j\|_2\boldsymbol{\zeta}_j/\|\boldsymbol{\zeta}_j\|_2$ for some $\boldsymbol{\zeta}_j \sim N(\boldsymbol{0}_n, \boldsymbol{I}_n)$ independent of $(\boldsymbol{X}\boldsymbol{Q}_j, \|\boldsymbol{z}_j\|_2)$. Then $\mathbb{E}_j$ coincides with the conditional expectation of $\boldsymbol{\zeta}_j$ given $(\boldsymbol{X}\boldsymbol{Q}_j, \|\boldsymbol{z}_j\|_2)$. After some rearrangement,

$$\xi_{\boldsymbol{f}}(\boldsymbol{z}_j) = \xi_{\boldsymbol{f}}(\boldsymbol{z}_j) - \mathbb{E}_j[\boldsymbol{f}(\boldsymbol{z}_j)]^\top \boldsymbol{z}_j + \|\mathbb{E}_j[\boldsymbol{f}(\boldsymbol{z}_j)]\|_2\frac{\|\boldsymbol{z}_j n^{-1/2}\|_2}{\|\boldsymbol{\zeta}_j n^{-1/2}\|_2}\Big(\frac{\mathbb{E}_j[\boldsymbol{f}(\boldsymbol{z}_j)]}{\|\mathbb{E}_j[\boldsymbol{f}(\boldsymbol{z}_j)]\|_2}\Big)^\top \boldsymbol{\zeta}_j.$$

Uniformly over $j \in J_{n,p}$, we have (i) the limit $|\xi_{\boldsymbol{f}}(\boldsymbol{z}_j) - \mathbb{E}_j[\boldsymbol{f}(\boldsymbol{z}_j)]^\top \boldsymbol{z}_j| \to 0$ and the limit $\|\mathbb{E}_j[\boldsymbol{f}(\boldsymbol{z}_j)]\|_2 \to 1$ both in probability, (ii) $\frac{\mathbb{E}_j[\boldsymbol{f}(\boldsymbol{z}_j)]^\top}{\|\mathbb{E}_j[\boldsymbol{f}(\boldsymbol{z}_j)]\|_2}\boldsymbol{\zeta}_j \sim N(0,1)$ and (iii) the limit $\|\Omega_{jj}^{1/2}\boldsymbol{z}_j\|_2/\|\boldsymbol{\zeta}_j\|_2 \to 1$ in probability. By Slutsky's Theorem, $\Omega_{jj}^{1/2}\xi_{\boldsymbol{f}}(\boldsymbol{z}_j)$ converges in distribution to $N(0,1)$ uniformly over $j \in J_{n,p}$. That is, for any $t \in \mathbb{R}$, $\max_{j\in J_{n,p}}|\mathbb{P}(\Omega_{jj}^{1/2}\xi_{\boldsymbol{f}}(\boldsymbol{z}_j) \le t) - \Phi(t)| \to 0$.

It remains to relate $\xi_{\boldsymbol{f}}(\boldsymbol{z}_j)$ to ξ_j defined in (2.10). As the term $\|\boldsymbol{\psi}\|_2^{-1}\boldsymbol{z}_j^\top\boldsymbol{\psi}$ present in both ξ_j and $\xi_{\boldsymbol{f}}(\boldsymbol{z}_j)$ cancel out, we have the decomposition

$$\|\boldsymbol{\psi}\|_2\|\boldsymbol{z}_j\|_2^{-2}n\big(\xi_j - \xi_{\boldsymbol{f}}(\boldsymbol{z}_j)\big)$$

$$= (\widehat{\beta}_j - \beta_j)\operatorname{tr}[\nabla_{\boldsymbol{y}}\boldsymbol{\psi}^\top] + \operatorname{tr}(\boldsymbol{P}_{\boldsymbol{\psi}}^\perp(\nabla\boldsymbol{\psi}(\boldsymbol{z}_j))^\top\boldsymbol{P}_{\boldsymbol{z}_j}^\perp) \tag{A.8}$$

$$= \operatorname{tr}\big[\big((\widehat{\beta}_j - \beta_j)\nabla_{\boldsymbol{y}}\boldsymbol{\psi}^\top + \nabla\boldsymbol{\psi}(\boldsymbol{z}_j)^\top\big)\boldsymbol{P}_{\boldsymbol{\psi}}^\perp\big] \tag{A.9}$$

$$+ (\widehat{\beta}_j - \beta_j)\operatorname{tr}\big[\boldsymbol{P}_{\boldsymbol{\psi}}\nabla_{\boldsymbol{y}}\boldsymbol{\psi}^\top\big] + \operatorname{tr}\big[\boldsymbol{P}_{\boldsymbol{\psi}}^\perp\nabla\boldsymbol{\psi}(\boldsymbol{z}_j)^\top\boldsymbol{P}_{\boldsymbol{z}_j}\big]. \tag{A.10}$$

The matrix inside the trace in (A.8) is zero thanks to (C.24). It follows that only the two terms in (A.10) remain, hence

$$\mathbb{E}[I_{\mathscr{E}_j}(\xi_j - \xi_{\boldsymbol{f}}(\boldsymbol{z}_j))^2] \le \tfrac{2}{n}\mathbb{E}\big[I_{\mathscr{E}_h}\|\boldsymbol{z}_j\|_2^2\|\boldsymbol{\psi}\|_2^{-2}\big((\widehat{\beta}_j - \beta_j)^2\|\nabla_{\boldsymbol{y}}\boldsymbol{\psi}\|_{\mathrm{op}}^2 + \|\nabla\boldsymbol{\psi}(\boldsymbol{z}_j)\|_{\mathrm{op}}^2\big)\big].$$

Since $I_{\mathscr{E}_j}\|\boldsymbol{z}_j\|^2/n \le \Omega_{jj}^{-1}(1+\eta_n)^2$ and $\|\nabla_{\boldsymbol{y}}\boldsymbol{\psi}\|_{\mathrm{op}} \le L$ by Proposition C.4 (i) with $\boldsymbol{X} = \widetilde{\boldsymbol{X}}$, the first term is bounded from above by $L\Omega_{jj}^{-1}(1+\eta_n)^2\mathbb{E}[I_{\mathscr{E}_j}\|\boldsymbol{\psi}\|_2^{-2}h_j^2]$ which converges to 0 uniformly over $j \in J_{n,p}$ by definition of $J_{n,p}$ in (A.5). The second term also converges to 0 uniformly over $j \in J_{n,p}$ thanks to (A.7). Thus $\max_{j\in J_{n,p}}\mathbb{E}[I_{\mathscr{E}_j}(\xi_j - \xi_{\boldsymbol{f}}(\boldsymbol{z}_j))^2] \to 0$ which implies that $|\xi_j - \xi_{\boldsymbol{f}}(\boldsymbol{z}_j)| \to 0$ in probability uniformly over $j \in J_{n,p}$ and Slutsky's theorem completes the proof of (2.12) for ξ_j.

To prove a similar result for ξ_j', it is enough to prove $\Omega_j^{1/2}|\xi_j' - \xi_j| \to^{\mathbb{P}} 0$ uniformly over $j \in J_{n,p}$ by Slutsky's theorem. As

$$|\xi_j - \xi_j'| = \big|\operatorname{tr}[\nabla_{\boldsymbol{y}}\boldsymbol{\psi}]\big|\|\boldsymbol{\psi}\|_2^{-1}\big|h_j\big|\ \big|\|\boldsymbol{z}_j\|_2^2/n - \Omega_{jj}^{-1}\big|$$

and $|\operatorname{tr}[\nabla_{\boldsymbol{y}}\boldsymbol{\psi}]| \le nL$, by the Cauchy-Schwarz inequality we find

$$\mathbb{E}[I_{\mathscr{E}_j}|\xi_j - \xi_j'|] \le \Omega_{jj}^{-1}nL\mathbb{E}[I_{\mathscr{E}_j}h_j^2\|\boldsymbol{\psi}\|_2^{-2}]^{1/2}\mathbb{E}[(\Omega_{jj}\|\boldsymbol{z}_j\|_2^2/n - 1)^2]^{1/2}.$$

Since $\mathbb{E}[(\Omega_{jj}\|\boldsymbol{z}_j\|_2^2/n - 1)^2]^{1/2} = \sqrt{2/n}$ and $\Omega_{jj} \in [\kappa, 1/\kappa]$, the previous display converges to 0 uniformly over $j \in J_{n,p}$ by definition of $J_{n,p}$ in (A.5). □

Appendix B: Stein formulae on the sphere

The goal of this section is to prove Proposition A.1 and to develop Stein formulae for random vectors $\boldsymbol{z}$ uniformly distributed on the sphere.

Let $\mathbb{S}^{n-1}(R)$ be the sphere in $\mathbb{R}^n$ with center $\mathbf{0}$ and radius $R > 0$. We say that $\boldsymbol{z}$ is uniformly distributed in $\mathbb{S}^{n-1}(R)$ and write $\boldsymbol{z} \sim \mathrm{Unif}(\mathbb{S}^{n-1}(R))$ if $\boldsymbol{z}$ is equal in distribution to $R\boldsymbol{\zeta}/\|\boldsymbol{\zeta}\|_2$ where $\boldsymbol{\zeta} \sim N(\mathbf{0}, \boldsymbol{I}_n)$. We first develops Stein's formulae with respect to $\boldsymbol{z} \sim \mathrm{Unif}(\mathbb{S}^{n-1}(R))$ for functions $\boldsymbol{f} : \boldsymbol{z} \in \mathbb{S}^{n-1}(R) \mapsto \boldsymbol{f}(\boldsymbol{z}) \in \mathbb{R}^n$ in Sobolev spaces over $\mathbb{S}^{n-1}(R)$.

We derive next Stein formulae for functions in Sobolev spaces over $\mathbb{S}^{n-1}(R)$. One possible construction of such Sobolev spaces is obtained by completion of the space of infinitely differentiable functions $\mathbb{S}^{n-1}(R) \to \mathbb{R}$ with respect to the desired Sobolev norm as follows. Here, $\mathbb{S}^{n-1}(R)$ is viewed as a compact Riemannian manifold equipped with the canonical metric (the metric induced as a submanifold of $\mathbb{R}^n$ equipped with the Euclidean metric). As it will be convenient for compatibility with the rest of the paper to conserve the partial derivatives with respect to the canonical basis in $\mathbb{R}^n$, we adopt the

following notation. For a smooth function $h : \mathbb{S}^{n-1}(R) \to \mathbb{R}$ and $\Omega \subset \mathbb{R}^n$ an open neighborhood of $\mathbb{S}^{n-1}(R)$, define the smooth function $\check{h} : \Omega \to \mathbb{R}$ by $\check{h}(\boldsymbol{x}) = h(R\boldsymbol{x}/\|\boldsymbol{x}\|_2)$ and define the gradient of h as that of $\check{h}$, i.e., for $\boldsymbol{x} \in \mathbb{S}^{n-1}(R)$, set $\nabla h(\boldsymbol{x}) = ((\partial/\partial x_1)\check{h}(\boldsymbol{x}), ..., (\partial/\partial x_n)\check{h}(\boldsymbol{x}))$. For every $\boldsymbol{x} \in \mathbb{S}^{n-1}$, the gradient $\nabla h(\boldsymbol{x})$ belongs to the hyperplane orthogonal to $\boldsymbol{x}$ which is the tangent space of $\mathbb{S}^{n-1}$ at $\boldsymbol{x}$. Furthermore if $\boldsymbol{\gamma} : \mathbb{R} \to \mathbb{S}^{n-1}(R)$ is a smooth curve with $\boldsymbol{\gamma}(0) = \boldsymbol{x}, (d/dt)\boldsymbol{\gamma}(t)|_{t=0} = \boldsymbol{v}$ then $\boldsymbol{v}^\top \boldsymbol{x} = 0$ and $\nabla h(\boldsymbol{x})^\top \boldsymbol{v} = (d/dt)h(\boldsymbol{\gamma}(t))|_{t=0}$. For such smooth function h and equipped with its gradient, for $\alpha \in \{1, 2\}$ we define the Sobolev norm

$$\|h\|_{1,\alpha} = \mathbb{E}\big[|h(\boldsymbol{z})|^\alpha\big]^{1/\alpha} + \mathbb{E}\big[\|\nabla h(\boldsymbol{z})\|_2^\alpha\big]^{1/\alpha}, \qquad \boldsymbol{z} \sim \mathrm{Unif}(\mathbb{S}^{n-1}(R))$$

and the Sobolev space $W^{1,\alpha}(\mathbb{S}^{n-1}(R))$ as the completion of the space of smooth functions over $\mathbb{S}^{n-1}(R)$ with respect to the above norm. This definition is equivalent to the definition given in [18, Definition 2.1]. By [18, Proposition 2.3], since the manifold $\mathbb{S}^{n-1}(R)$ is compact the resulting Sobolev spaces do not depend on the chosen metric. Equivalently, the Sobolev space $W^{1,\alpha}(\mathbb{S}^{n-1}(R))$ is also the completion with respect to the above norm of the space of once continuously differentiable functions on $\mathbb{S}^{n-1}(R)$.

If h is locally Lipschitz on $\mathbb{S}^{n-1}(R)$ (i.e., every point has a neighborhood in $\mathbb{S}^{n-1}(R)$ on which h is Lipschitz), then again by considering an open neighborhood $\Omega \subset \mathbb{R}^n$ of $\mathbb{S}^{n-1}(R)$, the function $\check{h}(\boldsymbol{x}) = h(R\boldsymbol{x}/\|\boldsymbol{x}\|_2)$ is locally Lipschitz on Ω. Thus, in this case and by Rademacher's theorem, $\nabla h(\boldsymbol{z})$ is well defined almost everywhere in $\mathbb{S}^{n-1}(R)$ as the gradient of $\nabla\check{h}(\boldsymbol{x})$, and $h \in W^{1,\alpha}(\mathbb{S}^{n-1}(R))$ if and only if $\mathbb{E}\big[\|\nabla h(\boldsymbol{z})\|_2^\alpha\big] < \infty$. For example, $h \in W^{1,\alpha}(\mathbb{S}^{n-1}(R))$ when h is L-Lipschitz on $\mathbb{S}^{n-1}(R)$.

Finally, for $\alpha \in \{1, 2\}$ define $W^{1,\alpha}(\mathbb{S}^{n-1}(R))^n$ as the space of $\mathbb{R}^n$ valued functions $\boldsymbol{f} = (f_1, ..., f_n)$ with all coordinates f_i belonging to $W^{1,\alpha}(\mathbb{S}^{n-1}(R))$, equipped with the norm

$$\|\boldsymbol{f}\|_{1,\alpha} = \mathbb{E}[\|\boldsymbol{f}(\boldsymbol{z})\|_2^\alpha]^{1/\alpha} + \mathbb{E}[\|\nabla \boldsymbol{f}(\boldsymbol{z})\|_F^\alpha]^{1/\alpha}$$

where the gradient $\nabla \boldsymbol{f}$ is the matrix in $\mathbb{R}^{n\times n}$ with columns $\nabla f_1, ..., \nabla f_n$,

Lemma B.1 (Stein's formula on the sphere)**.** *Let $n \geq 3$, $R > 0$ and $\boldsymbol{z} \sim \mathrm{Unif}(\mathbb{S}^{n-1}(R))$. Let $\boldsymbol{P}_{\boldsymbol{x}}^\perp = \boldsymbol{I}_n - \boldsymbol{x}\boldsymbol{x}^\top/\|\boldsymbol{x}\|_2^2$ for $\boldsymbol{x} \neq \boldsymbol{0}$. Then, for all $\boldsymbol{f} = (f_1, \dots, f_n) \in W^{1,1}(\mathbb{S}^{n-1}(R))^n$,*

$$\mathbb{E}[\boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{z}] = (n-1)^{-1} R^2 \mathbb{E}[\mathrm{tr}((\nabla \boldsymbol{f}(\boldsymbol{z}))^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp)], \tag{B.1}$$

where $\nabla \boldsymbol{f} = (\nabla f_1, \dots, \nabla f_n)$. For all $\boldsymbol{f} = (f_1, \dots, f_n) \in W^{1,2}(\mathbb{S}^{n-1}(R))^n$ we also have

$$\mathbb{E}[\|\boldsymbol{f}(\boldsymbol{z}) - \mathbb{E}[\boldsymbol{f}(\boldsymbol{z})]\|_2^2] \leq (n-2)^{-1} R^2 \mathbb{E}[\|\nabla \boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp\|_F^2], \tag{B.2}$$

$$\begin{aligned}&\mathbb{E}\big[\big(nR^{-2}\boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{z} - \mathrm{tr}[\nabla \boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp]\big)^2\big] \\ &= nR^{-2}\mathbb{E}\big[\|\boldsymbol{f}\|_2^2\big] + (1-2/n)^{-1}\mathbb{E}\,\mathrm{tr}\big[\big(\nabla \boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp\big)^2\big] - \tfrac{2}{n-2}\mathbb{E}\big[\mathrm{tr}(\nabla f(\boldsymbol{z})^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp)^2\big]\end{aligned} \tag{B.3}$$

$$\leq nR^{-2}\mathbb{E}\big[\|\boldsymbol{f}\|_2^2\big] + (1-2/n)^{-1}\mathbb{E}\big[\|\nabla \boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp\|_F^2\big]. \tag{B.4}$$

Note that (B.2) is the classical Poincaré inequality on the sphere. A proof is provided for completeness.

Proof of Lemma B.1. As the operators T and T_{ij} defined by $T\boldsymbol{f}(\boldsymbol{z}) = \boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{z}$ and $T_{ij}\boldsymbol{f}(\boldsymbol{z}) = [\nabla \boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp]_{ij}$ for every $i, j = 1, ..., n$ are all continuous linear mappings from $W^{1,\alpha}(\mathbb{S}^{n-1}(R))^n$ to the L_α space with the norm $\|f\|_{L_\alpha} = (\mathbb{E}[|f(\boldsymbol{z})|^\alpha])^{1/\alpha}$, we assume without loss of generality that all coordinates of $\boldsymbol{f}$ are continuously differentiable. Indeed, if $(\boldsymbol{f}^{(q)})_{q\geq 1}$ is a sequence of smooth functions over the sphere converging to $\boldsymbol{f}$ in $W^{1,\alpha}(\mathbb{S}^{n-1}(R))^n$ for $\alpha = 1$ and (B.1) holds for $\boldsymbol{f}^{(q)}$ then (B.1) also holds for the limit by continuity; the same argument applies with $\alpha = 2$ for (B.2)-(B.3)-(B.4).

Let $\boldsymbol{\zeta} \sim N(\boldsymbol{0}, \boldsymbol{I}_n)$. Assume without loss of generality $\boldsymbol{z} = R\boldsymbol{\zeta}/\|\boldsymbol{\zeta}\|_2$ as $R\boldsymbol{\zeta}/\|\boldsymbol{\zeta}\|_2 \sim \mathrm{Unif}(\mathbb{S}^{n-1}(R))$. Let $\phi(t)$ be a continuously differentiable function in $\mathbb{R}$ with $\phi(t) = 0$ for $t \leq 0$ and $\phi(t) = 1$ for $t \geq 1$. For $\delta > 0$ define $\boldsymbol{\varphi}_\delta(\boldsymbol{x}) = \phi(\|\boldsymbol{x}\|_2/\delta)\boldsymbol{f}(R\boldsymbol{x}/\|\boldsymbol{x}\|_2)$. As $\|\boldsymbol{\zeta}\|_2$ is independent of $\boldsymbol{z}$ and $\boldsymbol{\varphi}_\delta(\boldsymbol{x})$ has uniformly bounded gradient, the first order Stein formula for $\boldsymbol{\varphi}_\delta$ yields

$$\mathbb{E}\big[\phi(\|\boldsymbol{\zeta}\|_2/\delta)\|\boldsymbol{\zeta}\|_2\big]\mathbb{E}[\boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{z}] = \mathbb{E}\big[R\boldsymbol{\varphi}_\delta(\boldsymbol{\zeta})^\top \boldsymbol{\zeta}\big] = R\,\mathbb{E}\big[\operatorname{tr}(\nabla \boldsymbol{\varphi}_\delta(\boldsymbol{\zeta}))\big].$$

By the product and chain rules, $\nabla \boldsymbol{\varphi}_\delta(\boldsymbol{x})$ is given by

$$\phi'(\|\boldsymbol{x}\|_2/\delta)(\boldsymbol{x}/(\|\boldsymbol{x}\|_2\delta))\boldsymbol{f}(R\boldsymbol{x}/\|\boldsymbol{x}\|_2)^\top + \phi(\|\boldsymbol{x}\|_2/\delta)(R/\|\boldsymbol{x}\|_2)\boldsymbol{P}_{\boldsymbol{x}}^\perp\big(\nabla \boldsymbol{f}(R\boldsymbol{x}/\|\boldsymbol{x}\|_2)\big)$$

with $\phi'(t) = (d/dt)\phi(t)$ and $\phi'(0) = \phi'(1) = 0$. As $\sup_{\delta>0, \boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{x}\|_2 \|\nabla \boldsymbol{\varphi}_\delta(\boldsymbol{x})\|_{\mathrm{F}} \leq C < \infty$ and $\mathbb{E}[\|\boldsymbol{\zeta}\|_2^{-2}] = 1/(n-2) < \infty$, the dominated convergence theorem gives

$$\mathbb{E}[\boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{z}] = \frac{R\lim_{\delta\to 0}\mathbb{E}\big[\operatorname{tr}(\nabla \boldsymbol{\varphi}_\delta(\boldsymbol{\zeta}))\big]}{\lim_{\delta\to 0}\mathbb{E}\big[\phi(\|\boldsymbol{\zeta}\|_2/\delta)\|\boldsymbol{\zeta}\|_2\big]} = \frac{R^2\mathbb{E}\big[1/\|\boldsymbol{\zeta}\|_2\big]\mathbb{E}\big[\operatorname{tr}(\nabla \boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp)\big]}{\mathbb{E}\big[\|\boldsymbol{\zeta}\|_2\big]},$$

which yields (B.1) due to $\mathbb{E}[\|\boldsymbol{\zeta}\|_2]/\mathbb{E}[\|\boldsymbol{\zeta}\|_2^{-1}] = (n-1)$.

Next, as the exchange of limit and expectation is allowed when $\boldsymbol{\varphi}_\delta \to \boldsymbol{\varphi}_{0+} = \boldsymbol{f}$, the Gaussian Poincaré inequality [9, Theorem 1.6.4] yields

$$\mathbb{E}[\|\boldsymbol{f}(\boldsymbol{z}) - \mathbb{E}[\boldsymbol{f}(\boldsymbol{z})]\|_2^2] \leq \lim_{\delta\to 0+}\mathbb{E}[\|\nabla \boldsymbol{\varphi}_\delta(\boldsymbol{\zeta})\|_F^2] = \mathbb{E}[R^2\|\nabla \boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp\|_F^2\|\boldsymbol{\zeta}\|_2^{-2}].$$

Since $\mathbb{E}[\|\boldsymbol{\zeta}\|_2^{-2}] = 1/(n-2)$ and $(\|\boldsymbol{\zeta}\|_2, \boldsymbol{z})$ are independent, we obtain (B.2). Finally by applying the Second Order Stein formula of [6] to $\boldsymbol{\varphi}_\delta(\boldsymbol{\zeta})$ we find by dominated convergence

$$\begin{aligned}
&\mathbb{E}\big[\big(R^{-1}\|\boldsymbol{\zeta}\|_2\boldsymbol{z}^\top \boldsymbol{f}(\boldsymbol{z}) - R\|\boldsymbol{\zeta}\|_2^{-1}\operatorname{tr}[\nabla \boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp]\big)^2\big] \\
&= \lim_{\delta\to 0+}\mathbb{E}\big[\big(\boldsymbol{\zeta}^\top \boldsymbol{\varphi}_\delta(\boldsymbol{\zeta}) - \operatorname{tr}[\nabla \boldsymbol{\varphi}_\delta(\boldsymbol{\zeta})]\big)^2\big] \\
&= \lim_{\delta\to 0+}\mathbb{E}\big[\|\boldsymbol{\varphi}_\delta(\boldsymbol{\zeta})\|_2^2\big] + \mathbb{E}\operatorname{tr}\big[\big(\nabla \boldsymbol{\varphi}_\delta(\boldsymbol{\zeta})^\top\big)^2\big] \\
&= \mathbb{E}\big[\|\boldsymbol{f}(\boldsymbol{z})\|_2^2\big] + R^2\mathbb{E}\operatorname{tr}\big[\big(\nabla \boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp\big)^2\big]\mathbb{E}\big[\|\boldsymbol{\zeta}\|_2^{-2}\big] \\
&= \mathbb{E}\big[\|\boldsymbol{f}(\boldsymbol{z})\|_2^2\big] + R^2\mathbb{E}\operatorname{tr}\big[\big(\nabla \boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp\big)^2\big]/(n-2),
\end{aligned}$$

where the last equality follows from the independence of $\boldsymbol{z}$ and $\|\boldsymbol{\zeta}\|_2$. We now simplify the left-hand side in order to get rid of $\|\boldsymbol{\zeta}\|_2$. By expanding the square, using the independence of $(\boldsymbol{z}, \|\boldsymbol{\zeta}\|_2)$ and $\mathbb{E}[\|\boldsymbol{\zeta}\|_2^{-2}] = \frac{1}{n-2}$, the above display is equal to

$$\begin{aligned}
&\begin{cases} R^{-2}\mathbb{E}\big[\|\boldsymbol{\zeta}\|_2^2\big]\mathbb{E}\big[(\boldsymbol{z}^\top \boldsymbol{f}(\boldsymbol{z}))^2\big] - 2\mathbb{E}\big[\boldsymbol{z}^\top \boldsymbol{f}(\boldsymbol{z}) \operatorname{tr}[\nabla \boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp]\big] \\ +R^2\mathbb{E}\big[\|\boldsymbol{\zeta}\|_2^{-2}\big]\mathbb{E}\big[(\operatorname{tr}[\nabla \boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp])^2\big] \end{cases} \\
&= \begin{cases} nR^{-2}\mathbb{E}\big[(\boldsymbol{z}^\top \boldsymbol{f}(\boldsymbol{z}))^2\big] - 2\mathbb{E}\big[\boldsymbol{z}^\top \boldsymbol{f}(\boldsymbol{z}) \operatorname{tr}[\nabla \boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp]\big] \\ +R^2(n-2)^{-1}\mathbb{E}\big[(\operatorname{tr}[\nabla \boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp])^2\big] \end{cases} \\
&= \begin{cases} \mathbb{E}\big[\big(\frac{\sqrt{n}}{R}\boldsymbol{z}^\top \boldsymbol{f}(\boldsymbol{z}) - \frac{R}{\sqrt{n}} \operatorname{tr}[\nabla \boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp]\big)^2\big] \\ +R^2\big((n-2)^{-1} - n^{-1}\big)\mathbb{E}\big[\big(\operatorname{tr}[\nabla \boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp]\big)^2\big]. \end{cases}
\end{aligned}$$

Since $1/(n-2) - 1/n = 2n^{-1}(n-2)^{-1}$, we obtain (B.3) after multiplying by n/R^2. The proof is complete since (B.4) follows directly from (B.3) by the Cauchy-Schwarz inequality. □

Proof of Proposition A.1. If $\boldsymbol{f}$ is locally Lipschitz and $\mathbb{E}[\|\nabla \boldsymbol{f}(\boldsymbol{z})^\top\|_F^2] < \infty$, then $\boldsymbol{f} \in W^{1,2}(\mathbb{S}^{n-1}(R))$. We consider the mapping $\boldsymbol{f}(\boldsymbol{z}) - \mathbb{E}\big[\boldsymbol{f}(\boldsymbol{z})\big]$ rather than $\boldsymbol{f}(\boldsymbol{z})$ in (B.3) and (B.4). Multiplying $R^4 n^{-2}$ on both sides of the inequality (B.4), it provides

$$\begin{aligned}
&\mathbb{E}[(\xi_{\boldsymbol{f}}(\boldsymbol{z}) - \mathbb{E}[\boldsymbol{f}(\boldsymbol{z})]^\top \boldsymbol{z})^2] \\
&\le R^2 n^{-1}\mathbb{E}[\|\boldsymbol{f} - \mathbb{E}[\boldsymbol{f}(\boldsymbol{z})]\|_2^2] + R^4(n^2-2n)^{-1}\mathbb{E}[\|\nabla \boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp\|_F^2] \\
&\le 2R^4(n^2-2n)^{-1}\mathbb{E}[\|\nabla \boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp\|_F^2],
\end{aligned}$$

where the second inequality follows from (B.2). By the triangle inequality and (B.2),

$$\begin{aligned}
\mathbb{E}[|\|\boldsymbol{f}(\boldsymbol{z})\|_2 - \|\mathbb{E}[\boldsymbol{f}(\boldsymbol{z})]\|_2|^2] &\le \mathbb{E}[\|\boldsymbol{f}(\boldsymbol{z}) - \mathbb{E}[\boldsymbol{f}(\boldsymbol{z})]\|_2^2] \\
&\le R^2(n-2)^{-1}\mathbb{E}[\|\nabla \boldsymbol{f}(\boldsymbol{z})^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp\|_F^2].
\end{aligned}$$

If $\boldsymbol{f}(\boldsymbol{z}) = \widetilde{\boldsymbol{f}}(\boldsymbol{z})/\|\widetilde{\boldsymbol{f}}(\boldsymbol{z})\|_2$ with locally Lipschitz $\widetilde{\boldsymbol{f}}$ on $\mathbb{S}^{n-1}(R)$ then $\widetilde{\boldsymbol{f}}(\boldsymbol{z})/(\delta \vee \|\widetilde{\boldsymbol{f}}(\boldsymbol{z})\|_2)$ is locally Lipschitz for $\delta > 0$ and converges to $\boldsymbol{f}$ in $W^{1,2}(\mathbb{S}^{n-1}(R))^n$ as $\delta \to 0+$ when $\mathbb{P}(\|\boldsymbol{f}(\boldsymbol{z})\|_2 \neq 0) = 1$ and $\mathbb{E}[\|\widetilde{\boldsymbol{f}}(\boldsymbol{z})\|_2^{-2}\|\nabla \widetilde{\boldsymbol{f}}(\boldsymbol{z})\|_F^2] < +\infty$. Thus, the proof still applies. □

Appendix C: Bounds on $(\widehat{\beta}_j - \beta_j)^2\|\psi\|_2^{-2}$

The goal of this section is to prove Lemma A.2.

C.1. High probability events $\mathscr{E}_j$

Lemma C.1 (high probability of $\mathscr{E}_j$)**.** *Assume that $\boldsymbol{X}$ has iid $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ rows. Then*

$$\mathbb{E}[\|\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}\|_{op}^2] \le (\sqrt{n} + \sqrt{p})^2 + 1. \tag{C.1}$$

Furthermore, with $\eta_n = \sqrt{2\log(n)/n} + n^{-1/2}$ *and for the events*

$$\mathscr{E}_j = \{\|\boldsymbol{X}\boldsymbol{Q}_j\boldsymbol{\Sigma}^{-1/2}n^{-1/2}\|_{op} \le 1 + \sqrt{p/n} + \eta_n\} \cap \{\left|n^{-1/2}\Omega_{jj}^{1/2}\|\boldsymbol{z}_j\|_2 - 1\right| \le \eta_n\},$$

we have $\mathbb{P}(\cap_{j\in[p]}\mathscr{E}_j) \ge 1 - (p+1/2)n^{-1}(\pi\log(n))^{-1/2}$.

Proof of Lemma C.1. Let us first notice that $\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$ is a random Gaussian matrix with iid standard normal entries. Theorem 7.3.1 in [32] provides that $\mathbb{E}[\|\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}\|_{\mathrm{op}}] \le \sqrt{n}+\sqrt{p}$. Since the operator norm of a matrix is a 1-Lipschitz function of the entries of the matrix, by the Gaussian Poincaré inequality [10, Theorem 3.20], $\mathrm{Var}(\|\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}\|_{\mathrm{op}}) \le 1$. This proves (C.1).

By Theorem II.13 in [14], we have for $t > 0$,

$$\mathbb{P}(\|\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}n^{-1/2}\|_{\mathrm{op}} \ge 1 + \sqrt{p/n} + t) \le \Phi(-t\sqrt{n}),$$

$$\mathbb{P}(\left|\Omega_{jj}^{1/2}\|\boldsymbol{z}_j n^{-1/2}\|_2 - 1\right| \ge n^{-1/2} + t) \le 2\Phi(-t\sqrt{n}).$$

Since $\boldsymbol{X}\boldsymbol{Q}_j\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}_j\boldsymbol{\Sigma}^{-1/2})$ and $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}_j\boldsymbol{\Sigma}^{-1/2}\|_{\mathrm{op}} \le 1$,

$$\|\boldsymbol{X}\boldsymbol{Q}_j\boldsymbol{\Sigma}^{-1/2}\|_{\mathrm{op}} \le \|\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}\|_{\mathrm{op}}$$

for all $j \in [p]$. Next, using a union bound over $j \in [p]$ when $t = \sqrt{2\log(n)/n}$ we have $\mathbb{P}(\cap_{j\in[p]}\mathscr{E}_j) \ge 1 - (2p+1)\Phi(-\sqrt{2\log(n)})$. We conclude the proof using $\Phi(-t) \le (2\pi)^{-1/2}\exp(-t^2/2)/t$ for $t > 0$, which provides

$$\mathbb{P}(\cap_{j\in[p]}\mathscr{E}_j) \ge 1 - (p+1/2)n^{-1}(\pi\log(n))^{-1/2}. \qquad \square$$

Remark C.1. Our specific construction of $\mathscr{E}_j$ satisfies the following properties:

1. $I_{\mathscr{E}_j}$ is a function of $\|\boldsymbol{X}\boldsymbol{Q}_j\|_{\mathrm{op}}$ and $\|\boldsymbol{z}_j\|_2$ only, consequently the event $\mathscr{E}_j$ is independent of $\boldsymbol{z}_j/\|\boldsymbol{z}_j\|_2$.
2. $\|\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}n^{-1/2}\|_{\mathrm{op}}I_{\mathscr{E}_j} \le 1 + \sqrt{p/n} + \eta_n + (1+\eta_n) = 2 + \sqrt{p/n} + 2\eta_n$.

C.2. Twice continuously differentiable penalty

Lemma C.2. *Let* L, τ *be such that Assumptions A and B are fulfilled. Further assume that the Hessian matrix* $\boldsymbol{G} = (\nabla^2 g(\boldsymbol{b}))\big|_{\boldsymbol{b}=\hat{\boldsymbol{\beta}}}$ *of* g *at* $\hat{\boldsymbol{\beta}}$ *exists and is symmetric, and define*

$$\boldsymbol{M} = (\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{\psi}')\boldsymbol{X} + n\boldsymbol{G})^{-1}, \tag{C.2}$$

$$\boldsymbol{V} = \operatorname{diag}(\boldsymbol{\psi}') - \operatorname{diag}(\boldsymbol{\psi}')\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^\top \operatorname{diag}(\boldsymbol{\psi}'). \tag{C.3}$$

Then with the partial order $\boldsymbol{A} \preceq \boldsymbol{B}$ *if and only if the matrix* $\boldsymbol{B} - \boldsymbol{A}$ *is positive semi-definite, we have*

$$\|\operatorname{diag}(\boldsymbol{\psi}')\boldsymbol{X}\boldsymbol{M}\|_{op} \le (1/2)L^{1/2}(n\tau)^{-1/2}, \tag{C.4}$$

$$\boldsymbol{M} \preceq (n\tau)^{-1}\boldsymbol{I}_p, \tag{C.5}$$

$$(L\tau^{-1}\|\boldsymbol{X}n^{-1/2}\|_{op}^2 + 1)^{-1}\operatorname{diag}(\boldsymbol{\psi}') \preceq \boldsymbol{V} \preceq \operatorname{diag}(\boldsymbol{\psi}') \preceq L\boldsymbol{I}_n. \tag{C.6}$$

Proof of Lemma C.2. Throughout the proof we use the notation

$$\boldsymbol{B} = \operatorname{diag}(\boldsymbol{\psi}')^{1/2}\boldsymbol{X}n^{-1/2}\boldsymbol{G}^{-1/2}.$$

By some algebra, we have $\boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{B}+\boldsymbol{I}_p)^{-1}\boldsymbol{B}^\top = \operatorname{diag}(\boldsymbol{\psi}')^{1/2}\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^\top\operatorname{diag}(\boldsymbol{\psi}')^{1/2}$. For an upper bound of $\operatorname{diag}(\boldsymbol{\psi}')\boldsymbol{X}\boldsymbol{M}$, we notice

$$\begin{aligned}\operatorname{diag}(\boldsymbol{\psi}')\boldsymbol{X}\boldsymbol{M} &= \operatorname{diag}(\boldsymbol{\psi}')\boldsymbol{X}(\boldsymbol{X}^\top\operatorname{diag}(\boldsymbol{\psi}')\boldsymbol{X} + n\boldsymbol{G})^{-1}\\ &= \operatorname{diag}(\boldsymbol{\psi}')^{1/2}\boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{B}+\boldsymbol{I}_p)^{-1}n^{-1/2}\boldsymbol{G}^{-1/2}.\end{aligned}$$

For any matrix $\boldsymbol{B}$, $\|\boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{B}+\boldsymbol{I}_p)^{-1}\|_{\mathrm{op}} \le \max_{t\ge 0} t/(t^2+1) = 1/2$. By strongly convexity of g, $\|\boldsymbol{G}^{-1/2}\|_{\mathrm{op}} \le \tau^{-1/2}$. Since ψ is L-Lipschitz, $\|\operatorname{diag}(\boldsymbol{\psi}')^{1/2}\|_{\mathrm{op}} \le L^{1/2}$. Combining those upper bounds, we obtain (C.4). For the upper bound of $\boldsymbol{M}$, we notice $\boldsymbol{M} = n^{-1}\boldsymbol{G}^{-1/2}(\boldsymbol{B}^\top\boldsymbol{B}+\boldsymbol{I}_p)^{-1}\boldsymbol{G}^{-1/2} \preceq n^{-1}\boldsymbol{G}^{-1} \preceq (n\tau)^{-1}\boldsymbol{I}_p$. This gives (C.5). For lower and upper bounds on $\boldsymbol{V}$, we first notice that by definition of $\boldsymbol{B}$, $\|\boldsymbol{B}\|_{\mathrm{op}} \le L^{1/2}\tau^{-1/2}\|\boldsymbol{X}n^{-1/2}\|_{\mathrm{op}}$, thus

$$(L\tau^{-1}\|\boldsymbol{X}n^{-1/2}\|_{\mathrm{op}}^2+1)^{-1}\boldsymbol{I}_n \preceq \boldsymbol{I}_n - \boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{B}+\boldsymbol{I}_p)^{-1}\boldsymbol{B}^\top \preceq \boldsymbol{I}_n.$$

Since $\boldsymbol{V} = \operatorname{diag}(\boldsymbol{\psi}')^{1/2}(\boldsymbol{I}_n - \boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{B}+\boldsymbol{I}_p)^{-1}\boldsymbol{B}^\top)\operatorname{diag}(\boldsymbol{\psi}')^{1/2}$, we have

$$(L\tau^{-1}\|\boldsymbol{X}n^{-1/2}\|_{\mathrm{op}}^2+1)^{-1}\operatorname{diag}(\boldsymbol{\psi}') \preceq \boldsymbol{V} \preceq \operatorname{diag}(\boldsymbol{\psi}').$$

By the L-Lipschitz property of ψ, we have $\operatorname{diag}(\boldsymbol{\psi}') \preceq L\boldsymbol{I}_n$. Thus (C.6) holds. □

Proposition C.3. *Assume that g is strongly convex with parameter $\tau > 0$ and $\psi = \rho'$ is L-Lipschitz. Let $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\boldsymbol{\varepsilon}, \boldsymbol{X})$, $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\varepsilon}, \boldsymbol{X})$ be as in (2.8) and set $\boldsymbol{h}(\boldsymbol{\varepsilon}, \boldsymbol{X}) = \hat{\boldsymbol{\beta}}(\boldsymbol{\varepsilon}, \boldsymbol{X}) - \boldsymbol{\beta}$. Define $\nabla_{\boldsymbol{z}}\boldsymbol{h} = (\partial/\partial\boldsymbol{z})\boldsymbol{h}(\boldsymbol{\varepsilon}, \boldsymbol{X}+\boldsymbol{z}\boldsymbol{a}^\top)\big|_{\boldsymbol{z}=0}$ and $\nabla_{\boldsymbol{z}}\boldsymbol{\psi} = (\partial/\partial\boldsymbol{z})\boldsymbol{\psi}(\boldsymbol{\varepsilon}, \boldsymbol{X}+\boldsymbol{z}\boldsymbol{a}^\top)\big|_{\boldsymbol{z}=0}$ for fixed $\boldsymbol{a}$, $\boldsymbol{\varepsilon}$ and $\boldsymbol{X}$. Let $\boldsymbol{P}_{\boldsymbol{x}}^\perp = \boldsymbol{I}_n - \boldsymbol{x}\boldsymbol{x}^\top/\|\boldsymbol{x}\|_2^2$ for $\boldsymbol{x} \ne \boldsymbol{0}$. Then*

(i) For fixed $\boldsymbol{\varepsilon}$, $\boldsymbol{h}(\boldsymbol{\varepsilon}, \boldsymbol{X})$ and $\boldsymbol{\psi}(\boldsymbol{\varepsilon}, \boldsymbol{X})$ are Lipschitz in $\boldsymbol{X}$ on every compact subset of $\mathbb{R}^{n\times p}$.

For (ii) and (iii), additionally assume that g is twice continuously differentiable.

(ii) For almost every $\boldsymbol{X}$ and every $\boldsymbol{a} \in \mathbb{R}^p$,

$$(\nabla_{\boldsymbol{z}}\boldsymbol{h})^\top = -(\boldsymbol{h}^\top\boldsymbol{a})\boldsymbol{M}\boldsymbol{X}^\top\operatorname{diag}(\boldsymbol{\psi}') + \boldsymbol{M}\boldsymbol{a}\boldsymbol{\psi}^\top, \tag{C.7}$$

$$(\nabla_{\boldsymbol{z}}\boldsymbol{\psi})^\top = -(\boldsymbol{h}^\top\boldsymbol{a})\boldsymbol{V} - \operatorname{diag}(\boldsymbol{\psi}')\boldsymbol{X}\boldsymbol{M}\boldsymbol{a}\boldsymbol{\psi}^\top. \tag{C.8}$$

(iii) For almost every $\boldsymbol{X}$ and every $\boldsymbol{a} \in \mathbb{R}^p$, if $\boldsymbol{\psi} \ne \boldsymbol{0}$, then

$$(\nabla_{\boldsymbol{z}}(\boldsymbol{\psi}/\|\boldsymbol{\psi}\|_2))^\top = \|\boldsymbol{\psi}\|_2^{-1}\boldsymbol{P}_{\boldsymbol{\psi}}^\perp\big[-(\boldsymbol{h}^\top\boldsymbol{a})\boldsymbol{V} - \operatorname{diag}(\boldsymbol{\psi}')\boldsymbol{X}\boldsymbol{M}\boldsymbol{a}\boldsymbol{\psi}^\top\big] \tag{C.9}$$

We remark that in view of (A.6), $\nabla_{\boldsymbol{z}}\boldsymbol{\psi} = \nabla\boldsymbol{\psi}(\boldsymbol{z}_j)$ when $\boldsymbol{a} = \boldsymbol{e}_j$.

Proof. (i) We refer our readers to the proof of Proposition C.4.

(ii) As the functions $\boldsymbol{h}$ and $\boldsymbol{\psi}$ are Lipschitz on every compact, their Fréchet derivatives exist almost everywhere by Rademacher's theorem. Moreover, the chain rule holds almost everywhere by [34, Theorem 2.1.11]. Let $\boldsymbol{G}, \boldsymbol{V}$ and $\boldsymbol{M}$ be as in Lemma C.2. By differentiating these KKT conditions $\boldsymbol{X}^\top \boldsymbol{\psi}(\boldsymbol{X}) = n(\nabla g(\boldsymbol{b}))\big|_{\boldsymbol{b}=\boldsymbol{\beta}+\boldsymbol{h}}$, and by the chain rule, we obtain that for almost every $\boldsymbol{X}$,

$$
\begin{aligned}
n\boldsymbol{G}\big(\nabla_{\boldsymbol{z}}\boldsymbol{h}\big)^\top &= \boldsymbol{a}\boldsymbol{\psi}^\top + \boldsymbol{X}^\top\big(\nabla_{\boldsymbol{z}}\boldsymbol{\psi}\big)^\top,\\
\big(\nabla_{\boldsymbol{z}}\boldsymbol{\psi}\big)^\top &= \operatorname{diag}(\boldsymbol{\psi}')\big\{-(\boldsymbol{a}^\top\boldsymbol{h})\boldsymbol{I}_n - \boldsymbol{X}\big(\nabla_{\boldsymbol{z}}\boldsymbol{h}\big)^\top\big\}.
\end{aligned}
$$

Solving the above system of equations gives (C.7) and (C.8).

(iii) Since the map $\boldsymbol{v} \mapsto \boldsymbol{v}/\|\boldsymbol{v}\|_2$ with $\boldsymbol{v} \in \mathbb{R}^n$ has Fréchet derivative $\|\boldsymbol{v}\|_2^{-1}\boldsymbol{P}_{\boldsymbol{v}}^\perp$ at every point $\boldsymbol{v} \neq \boldsymbol{0} \in \mathbb{R}^n$, by chain rule, (C.9) follows for almost every $\boldsymbol{X}$ if $\boldsymbol{\psi}(\boldsymbol{X}) \neq \boldsymbol{0}_n$. □

Proof of Lemma A.2 when g is twice continuously differentiable. Here, we further assume that g is twice differentiable so that $\boldsymbol{V}, \boldsymbol{M}$ in Lemma C.2 are well-defined.

By Proposition C.4, the map $\boldsymbol{z}_j \mapsto (\boldsymbol{h}, \boldsymbol{\psi})$ is locally Lipschitz, thus the map of the product $\boldsymbol{z}_j \mapsto h_j\boldsymbol{\psi}(\|\boldsymbol{\psi}\|_2^2+\delta)^{-1}$ is also locally Lipschitz. By Proposition D.2, without loss of generality, we can assume that $\boldsymbol{\psi} \neq \boldsymbol{0}_n$ at almost every point $\boldsymbol{z}_j \in \mathbb{R}^n$. By the first order Stein's formula on the sphere (B.1) for $(n-1)\|\boldsymbol{z}_j\|_2^{-2}\mathbb{E}_j[\boldsymbol{z}_j^\top \boldsymbol{f}(\boldsymbol{z}_j)]$ with $\boldsymbol{f}(\boldsymbol{z}_j) = h_j\boldsymbol{\psi}(\|\boldsymbol{\psi}\|_2^2+\delta)^{-1}$, we have

$$
\begin{aligned}
&\mathbb{E}_j[h_j^2(\|\boldsymbol{\psi}\|_2^2+\delta)^{-1}\operatorname{tr}(\boldsymbol{V}\boldsymbol{P}_{\boldsymbol{z}_j}^\perp)] && \text{(C.10)}\\
=&-\mathbb{E}_j[h_j(\|\boldsymbol{\psi}\|_2^2+\delta)^{-1}\boldsymbol{\psi}^\top\boldsymbol{z}_j](n-1)\|\boldsymbol{z}_j\|_2^{-2} && \text{(C.11)}\\
&-2\mathbb{E}_j[h_j(\|\boldsymbol{\psi}\|_2^2+\delta)^{-1}\boldsymbol{e}_j^\top(\operatorname{diag}(\boldsymbol{\psi}')\boldsymbol{X}\boldsymbol{M})^\top\boldsymbol{P}_{\boldsymbol{z}_j}^\perp\boldsymbol{\psi}] && \text{(C.12)}\\
&+\mathbb{E}_j[(\|\boldsymbol{\psi}\|_2^2+\delta)^{-1}\boldsymbol{e}_j^\top\boldsymbol{M}\boldsymbol{e}_j\boldsymbol{\psi}^\top\boldsymbol{P}_{\boldsymbol{z}_j}^\perp\boldsymbol{\psi}] && \text{(C.13)}\\
&+2\mathbb{E}_j[h_j^2(\|\boldsymbol{\psi}\|_2^2+\delta)^{-2}\boldsymbol{\psi}^\top\boldsymbol{V}\boldsymbol{P}_{\boldsymbol{z}_j}^\perp\boldsymbol{\psi}] && \text{(C.14)}\\
&+2\mathbb{E}_j[h_j(\|\boldsymbol{\psi}\|_2^2+\delta)^{-2}\boldsymbol{\psi}^\top\operatorname{diag}(\boldsymbol{\psi}')\boldsymbol{X}\boldsymbol{M}\boldsymbol{e}_j\boldsymbol{\psi}^\top\boldsymbol{P}_{\boldsymbol{z}_j}^\perp\boldsymbol{\psi}]. && \text{(C.15)}
\end{aligned}
$$

For the terms (C.12)-(C.14), by Lemma C.2 and $\|\boldsymbol{P}_{\boldsymbol{z}_j}^\perp\|_{\mathrm{op}} \le 1$ we find

$$
\begin{aligned}
\text{(C.12)} \vee \text{(C.15)} &\le \sqrt{L/(\tau n)}\,\mathbb{E}_j[|h_j|(\|\boldsymbol{\psi}\|_2^2+\delta)^{-1/2}] && \text{by (C.4)}\\
&\le \tfrac{1}{2}(\tau n)^{-1} + \tfrac{L}{2}\mathbb{E}_j[h_j^2(\|\boldsymbol{\psi}\|_2^2+\delta)^{-1}],\\
\text{(C.13)} &\le (\tau n)^{-1} && \text{by (C.5)},\\
\text{(C.14)} &\le 2L\mathbb{E}_j[h_j^2(\|\boldsymbol{\psi}\|_2^2+\delta)^{-1}] && \text{by (C.6)}.
\end{aligned}
$$

By leaving term (C.11) unchanged and using the above inequalities,

$$
\begin{aligned}
\text{(C.10)} \le &-(n-1)\|\boldsymbol{z}_j\|_2^{-2}\mathbb{E}_j[h_j(\|\boldsymbol{\psi}\|_2^2+\delta)^{-1}\boldsymbol{\psi}^\top\boldsymbol{z}_j]\\
&+2(\tau n)^{-1}\\
&+3L\mathbb{E}_j[h_j^2(\|\boldsymbol{\psi}\|_2^2+\delta)^{-1}].
\end{aligned} \tag{C.16}
$$

We now bound (C.10) from below. Let

$$U_j = (L\tau^{-1}(\|\boldsymbol{X}\boldsymbol{Q}_j n^{-1/2}\|_{\mathrm{op}} + \|\boldsymbol{z}_j n^{-1/2}\|_2)^2 + 1)^{-1}.$$

By $(L\tau^{-1}\|\boldsymbol{X}n^{-1/2}\|_{\mathrm{op}}^2 + 1)^{-1}\operatorname{diag}(\boldsymbol{\psi}') \preceq \boldsymbol{V} \preceq \operatorname{diag}(\boldsymbol{\psi}') \preceq L\boldsymbol{I}_n$ in (C.6) and by $\|\boldsymbol{X}\|_{\mathrm{op}} \le \|\boldsymbol{X}\boldsymbol{Q}_j\|_{\mathrm{op}} + \|\boldsymbol{z}_j\|_2$ for all $j \in [p]$, we have that $U_j \operatorname{diag}(\boldsymbol{\psi}') \preceq \boldsymbol{V} \preceq L\boldsymbol{I}_n$.

Since $U_j \operatorname{diag}(\boldsymbol{\psi}') \preceq \boldsymbol{V}$ and $K^2\boldsymbol{I}_n \preceq \operatorname{diag}(\boldsymbol{\psi}') + \operatorname{diag}\{\psi_i^2, i = 1, ..., n\}$ both holds, multiplying the second inequality by U_j and summing yields $K^2 U_j \boldsymbol{I}_n \preceq U_j \operatorname{diag}\psi_i^2 + \boldsymbol{V}$. Multiplying both sides $\boldsymbol{P}_{\boldsymbol{z}_j}^{\perp}$ to the left and to the right and using that $\boldsymbol{P}_{\boldsymbol{z}_j}^{\perp} \preceq \boldsymbol{I}_n$ and $\operatorname{tr}\boldsymbol{P}_{\boldsymbol{z}_j}^{\perp} = n-1$ we find

$$K^2 U_j \boldsymbol{P}_{\boldsymbol{z}_j}^{\perp} \preceq U_j \operatorname{diag}\psi_i^2 + \boldsymbol{P}_{\boldsymbol{z}_j}^{\perp}\boldsymbol{V}\boldsymbol{P}_{\boldsymbol{z}_j}^{\perp}, \qquad K^2 U_j(n-1) \le U_j\|\boldsymbol{\psi}\|^2 + \operatorname{tr}[\boldsymbol{P}_{\boldsymbol{z}_j}^{\perp}\boldsymbol{V}].$$

Multiplying by $(\|\boldsymbol{\psi}\|_2^2 + \delta)^{-1}h_j^2$ and taking the conditional expectation $\mathbb{E}_j$,

$$\begin{aligned} K^2 U_j(n-1)\mathbb{E}_j[(\|\boldsymbol{\psi}\|_2^2 + \delta)^{-1}h_j^2] &\le \mathbb{E}_j[h_j^2] + \mathbb{E}_j[\operatorname{tr}[\boldsymbol{P}_{\boldsymbol{z}_j}^{\perp}\boldsymbol{V}]h_j^2(\|\boldsymbol{\psi}\|_2^2 + \delta)^{-1}] \\ &= \mathbb{E}_j[h_j^2] + (\text{C.10}). \end{aligned} \tag{C.17}$$

Note that by definition of u_* and the events in Lemma C.1, we have when $u_* > 0$,

$$\mathscr{E}_j \subset \{nu_* \le K^2 U_j(n-1) - 3L\} \cap \{\Omega_{jj}\|\boldsymbol{z}_j\|_2^2 \ge n(1-\eta_n)_+^2\}. \tag{C.18}$$

Then combining (C.16) with the previous display, multiplying both sides by $I_{\mathscr{E}_j}$ and summing over $j \in [p]$ we find

$$nu_* \sum_{j=1}^p \mathbb{E}_j\Big[\frac{I_{\mathscr{E}_j}h_j^2}{\|\boldsymbol{\psi}\|_2^2 + \delta}\Big] \le \frac{2p}{\tau n} + \sum_{j=1}^p \mathbb{E}_j[I_{\mathscr{E}_j}h_j^2] - (n-1)\mathbb{E}_j\Big[\frac{\boldsymbol{\psi}^\top \boldsymbol{z}_j I_{\mathscr{E}_j}\boldsymbol{e}_j^\top \boldsymbol{h}}{(\|\boldsymbol{\psi}\|_2^2 + \delta)\|\boldsymbol{z}_j\|_2^2}\Big].$$

Taking expectations $\mathbb{E}$, letting $\operatorname{diag}\{I_{\mathscr{E}_j}\}$ denote the diagonal matrix with the j-th diagonal element $I_{\mathscr{E}_j}$, we find

$$\begin{aligned} &nu_*\mathbb{E}\big[(\|\boldsymbol{\psi}\|_2^2 + \delta)^{-1}\boldsymbol{h}^\top \operatorname{diag}\{I_{\mathscr{E}_j}\}\boldsymbol{h}\big] - 2p/(\tau n) - \mathbb{E}[\|\boldsymbol{h}\|_2^2] \\ &\le -(n-1)\mathbb{E}\big[(\|\boldsymbol{\psi}\|_2^2 + \delta)^{-1}\boldsymbol{\psi}^\top \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\operatorname{diag}\{\Omega_{jj}^{-1}I_{\mathscr{E}_j}\|\boldsymbol{z}_j\|_2^{-2}\}\boldsymbol{h}\big] \\ &\le (n-1)\mathbb{E}\big[\|\boldsymbol{X}\boldsymbol{\Sigma}^{-1}\operatorname{diag}\{\Omega_{jj}^{-1}I_{\mathscr{E}_j}\|\boldsymbol{z}_j\|_2^{-2}\}\|_{\mathrm{op}}^2\big]^{\frac{1}{2}}\mathbb{E}\big[(\|\boldsymbol{\psi}\|_2^2 + \delta)^{-1}\boldsymbol{h}^\top \operatorname{diag}\{I_{\mathscr{E}_j}\}\boldsymbol{h}\big]^{\frac{1}{2}} \\ &\le (1-\eta_n)_+^{-2}\mathbb{E}\big[\|\boldsymbol{X}\boldsymbol{\Sigma}^{-1}\|_{\mathrm{op}}^2\big]^{\frac{1}{2}}\mathbb{E}\big[(\|\boldsymbol{\psi}\|_2^2 + \delta)^{-1}\boldsymbol{h}^\top \operatorname{diag}\{I_{\mathscr{E}_j}\}\boldsymbol{h}\big]^{\frac{1}{2}} \end{aligned}$$

where the second inequality follows from the Cauchy-Schwarz inequality and $(\|\boldsymbol{\psi}\|_2^2 + \delta)^{-1/2}\|\boldsymbol{\psi}\|_2 \le 1$, and the third inequality follows from $\Omega_{jj}^{-1}I_{\mathscr{E}_j}(n-1)\|\boldsymbol{z}_j\|_2^{-2} \le (1-\eta_n)_+^{-2}$ thanks to (C.18). This implies that $x = (n\mathbb{E}[(\|\boldsymbol{\psi}\|_2^2 + \delta)^{-1}\boldsymbol{h}^\top \operatorname{diag}\{I_{\mathscr{E}_j}\}\boldsymbol{h}])^{1/2}$ satisfies $Ax^2 + Bx + C \le 0$ where the polynomial coefficients are

$$A = u_*, \;\; C = -2p(\tau n)^{-1} - \mathbb{E}[\|\boldsymbol{h}\|_2^2] \;\text{ and }\; B = -(1-\eta_n)_+^{-2}\mathbb{E}[\|\boldsymbol{X}\boldsymbol{\Sigma}^{-1}n^{-1}\|_{\mathrm{op}}^2]^{1/2}.$$

As $A > 0$, inequality $Ax^2 + Bx + C \le 0$ implies that x lies between the two real roots of the polynomial $AX^2 + BX + C$. In particular, x is smaller than the largest root, i.e., $x \le (-B + \sqrt{B^2 - 4AC})/(2A) \le |B|/A + (|C|/A)^{1/2}$. Here, $|C|/A = (2p(n\tau)^{-1} + \mathbb{E}[\|\boldsymbol{h}\|_2^2])/u_*$ and

$$|B|/A \le u_*^{-1}(1-\eta_n)_+^{-2}\big|\|\boldsymbol{\Sigma}^{-1/2}\|_{\mathrm{op}}\mathbb{E}[\|\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}n^{-1/2}\|_{\mathrm{op}}^2]^{1/2}.$$

The upper bound (C.1) on $\mathbb{E}[\|\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}\|_{\mathrm{op}}^2]$ completes the proof. □

C.3. Non-smooth strongly convex penalty g

C.3.1. Almost everywhere differentiability

In this section, we provide the almost everywhere existence of the Jacobian matrices. We also notice that if our penalty g is not twice differentiable, the matrices $\boldsymbol{M}$ and $\boldsymbol{V}$ in Lemma C.2 are not well defined. In this case we do not have explicit formula for the Jacobian matrices $(\partial/\partial \boldsymbol{z}_j)\boldsymbol{\psi}$ and $(\partial/\partial \boldsymbol{z}_j)\boldsymbol{h}$ such as those in terms of $\boldsymbol{V}, \mathrm{diag}(\boldsymbol{\psi}'), \boldsymbol{M}$ in Proposition C.3 (ii) and (iii). In this section we provide upper bounds of the norms of these Jacobian matrices without using Proposition C.3.

Proposition C.4. *Let $\rho : \mathbb{R} \to \mathbb{R}$ be convex and continuously differentiable with derivative $\psi = \rho'$ being L-Lipschitz. Let $g : \mathbb{R}^p \to \mathbb{R}$ be strongly convex with parameter $\tau > 0$. Let $\boldsymbol{X}, \widetilde{\boldsymbol{X}} \in \mathbb{R}^{n\times p}$, $\boldsymbol{\varepsilon}, \widetilde{\boldsymbol{\varepsilon}} \in \mathbb{R}^n$ and correspondingly,*

$$\begin{aligned}
&\hat{\boldsymbol{\beta}}(\boldsymbol{\varepsilon}, \boldsymbol{X}) = \operatorname*{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} \sum_{i\in[n]} \frac{\rho(\varepsilon_i - \boldsymbol{x}_i^\top(\boldsymbol{b}-\boldsymbol{\beta}))}{n} + g(\boldsymbol{b}), && \tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{\varepsilon}}, \widetilde{\boldsymbol{X}}),\\
&\boldsymbol{h}(\boldsymbol{\varepsilon}, \boldsymbol{X}) = \hat{\boldsymbol{\beta}}(\boldsymbol{\varepsilon}, \boldsymbol{X}) - \boldsymbol{\beta}, && \widetilde{\boldsymbol{h}} = \boldsymbol{h}(\widetilde{\boldsymbol{\varepsilon}}, \widetilde{\boldsymbol{X}}),\\
&\boldsymbol{\psi}(\boldsymbol{\varepsilon}, \boldsymbol{X}) = \psi(\boldsymbol{\varepsilon} - \boldsymbol{X}\boldsymbol{h}(\boldsymbol{\varepsilon}, \boldsymbol{X})), && \widetilde{\boldsymbol{\psi}} = \boldsymbol{\psi}(\widetilde{\boldsymbol{\varepsilon}}, \widetilde{\boldsymbol{X}}).
\end{aligned} \tag{C.19}$$

Then (i)

$$n\tau\|\boldsymbol{h}-\widetilde{\boldsymbol{h}}\|_2^2 + L^{-1}\|\boldsymbol{\psi}-\widetilde{\boldsymbol{\psi}}\|_2^2 \le (\boldsymbol{h}-\widetilde{\boldsymbol{h}})^\top(\boldsymbol{X}-\widetilde{\boldsymbol{X}})^\top\boldsymbol{\psi} + (\boldsymbol{\varepsilon}-\widetilde{\boldsymbol{\varepsilon}}+\widetilde{\boldsymbol{X}}\boldsymbol{h}-\boldsymbol{X}\boldsymbol{h})^\top(\boldsymbol{\psi}-\widetilde{\boldsymbol{\psi}}). \tag{C.20}$$

(ii) The map $(\boldsymbol{\varepsilon}, \boldsymbol{X}) \mapsto \big(\boldsymbol{h}(\boldsymbol{\varepsilon}, \boldsymbol{X}), \boldsymbol{\psi}(\boldsymbol{\varepsilon}, \boldsymbol{X})\big)$ is Lipschitz on every compact of $\mathbb{R}^n \times \mathbb{R}^{n\times p}$.

(iii) For any $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ fixed, and for any $\boldsymbol{\eta} \in \mathbb{R}^n$ and $\boldsymbol{a} \in \mathbb{R}^p$

$$\begin{aligned}
&n\tau\|\hat{\boldsymbol{\beta}}(\boldsymbol{\varepsilon}, \boldsymbol{X} + \boldsymbol{\eta}\boldsymbol{a}^\top) - \hat{\boldsymbol{\beta}}(\boldsymbol{\varepsilon}, \boldsymbol{X})\|_2^2 + L^{-1}\|\boldsymbol{\psi}(\boldsymbol{\varepsilon}, \boldsymbol{X} + \boldsymbol{\eta}\boldsymbol{a}^\top) - \boldsymbol{\psi}(\boldsymbol{\varepsilon}, \boldsymbol{X})\|_2^2\\
&\le (n\tau)^{-1}\|\boldsymbol{a}\|_2^2(\boldsymbol{\eta}^\top\boldsymbol{\psi})^2 + L(\boldsymbol{h}^\top\boldsymbol{a})^2\|\boldsymbol{\eta}\|_2^2.
\end{aligned}$$

Furthermore,

$$\sum_{i\in[n]} \|\boldsymbol{\psi}(\boldsymbol{\varepsilon}, \boldsymbol{X} + \boldsymbol{e}_i\boldsymbol{a}^\top) - \boldsymbol{\psi}(\boldsymbol{X})\|_2^2 \le L(n\tau)^{-1}\|\boldsymbol{a}\|_2^2\|\boldsymbol{\psi}\|_2^2 + nL^2(\boldsymbol{h}^\top\boldsymbol{a})^2. \tag{C.21}$$

(iv) If $\boldsymbol{\eta} \in \mathbb{R}^n$ *is such that* $\boldsymbol{\eta}^\top \boldsymbol{\psi}(\boldsymbol{\varepsilon}, \boldsymbol{X}) = 0$ *then* $\boldsymbol{\psi}(\boldsymbol{\varepsilon} + \boldsymbol{h}(\boldsymbol{\varepsilon}, \boldsymbol{X})^\top \boldsymbol{e}_j \boldsymbol{\eta}, \boldsymbol{X} + \boldsymbol{\eta}\boldsymbol{e}_j^\top) = \boldsymbol{\psi}(\boldsymbol{\varepsilon}, \boldsymbol{X})$*, so that if* $\boldsymbol{\psi}(\boldsymbol{\varepsilon}, \boldsymbol{X})$ *is Fréchet differentiable at* $(\boldsymbol{\varepsilon}, \boldsymbol{X})$ *then*

$$\sum_{i=1}^n \eta_i \Big[\frac{\partial \boldsymbol{\psi}}{\partial x_{ij}}(\boldsymbol{\varepsilon}, \boldsymbol{X}) + \Big(\boldsymbol{e}_j^\top \boldsymbol{h}(\boldsymbol{\varepsilon}, \boldsymbol{X}) \Big) \frac{\partial \boldsymbol{\psi}}{\partial \varepsilon_i}(\boldsymbol{\varepsilon}, \boldsymbol{X}) \Big] = 0 \text{ provided that } \boldsymbol{\eta}^\top \boldsymbol{\psi}(\boldsymbol{\varepsilon}, \boldsymbol{X}) = 0. \tag{C.22}$$

The content of the above proposition appeared in [4, Proposition 5.2] with variables $(\boldsymbol{y}, \boldsymbol{X})$ instead of $(\boldsymbol{\varepsilon}, \boldsymbol{X})$. It follows from strong convexity and the KKT conditions of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$. Its proof is provided below for completeness. An application of the above Proposition C.4 to normalized $\boldsymbol{\psi}$ yields the following.

Corollary C.5. *Under the conditions of Proposition C.4 and with the notation of Proposition C.3, at a point where* $\|\boldsymbol{\psi}\|_2 > 0$ *and* $\boldsymbol{\psi}$ *is Fréchet differentiable,*

$$\Big\| \nabla_{\boldsymbol{z}} \Big(\frac{\boldsymbol{\psi}}{\|\boldsymbol{\psi}\|_2} \Big) \Big\|_F^2 \le \frac{L\|\boldsymbol{a}\|_2^2}{n\tau} + \frac{nL^2(\boldsymbol{h}^\top \boldsymbol{a})^2}{\|\boldsymbol{\psi}\|_2^2}, \tag{C.23}$$

and with the $\nabla_{\boldsymbol{y}} \boldsymbol{\psi}$ *in* (2.9)

$$\Big(\nabla_{\boldsymbol{z}} \Big(\frac{\boldsymbol{\psi}}{\|\boldsymbol{\psi}\|_2} \Big) + (\boldsymbol{a}^\top \boldsymbol{h}) \frac{\nabla_{\boldsymbol{y}} \boldsymbol{\psi}}{\|\boldsymbol{\psi}\|_2} \Big)^\top \boldsymbol{P}_{\boldsymbol{\psi}}^\perp = \boldsymbol{0}. \tag{C.24}$$

Proof. For (C.23), by the chain rule

$$\|\nabla_{\boldsymbol{z}} (\boldsymbol{\psi}/\|\boldsymbol{\psi}\|_2)^\top\|_F^2 = \|\boldsymbol{\psi}\|_2^{-2} \|\boldsymbol{P}_{\boldsymbol{\psi}}^\perp (\nabla_{\boldsymbol{z}} \boldsymbol{\psi})^\top\|_F^2 \le \|\boldsymbol{\psi}\|_2^{-2} \|\nabla_{\boldsymbol{z}} \boldsymbol{\psi}\|_F^2.$$

By definition of the Frobenius norm $\|\nabla_{\boldsymbol{z}} \boldsymbol{\psi}\|_F^2 = \sum_{i \in [n]} \|\partial \boldsymbol{\psi}/\partial z_i\|_2^2$. By (C.21) with $\boldsymbol{a}$ replaced by $t\boldsymbol{a}$ and taking the limit as $t \to 0$ we obtain $\|\nabla_{\boldsymbol{z}} (\boldsymbol{\psi}/\|\boldsymbol{\psi}\|_2)\|_F^2 \le L(n\tau)^{-1}\|\boldsymbol{a}\|_2^2 + nL^2(\boldsymbol{h}^\top \boldsymbol{a})^2 \|\boldsymbol{\psi}\|_2^{-2}$.

For (C.24), we have

$$\Big(\Big(\frac{\partial}{\partial \boldsymbol{\eta}} + (\boldsymbol{a}^\top \boldsymbol{h}) \frac{\partial}{\partial \boldsymbol{y}} \Big) \frac{\boldsymbol{\psi}}{\|\boldsymbol{\psi}\|_2} \Big) \boldsymbol{P}_{\boldsymbol{\psi}}^\perp = \Big(\frac{\boldsymbol{P}_{\boldsymbol{\psi}}^\perp}{\|\boldsymbol{\psi}\|_2} \Big(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\eta}} + (\boldsymbol{a}^\top \boldsymbol{h}) \frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{y}} \Big) \Big) \boldsymbol{P}_{\boldsymbol{\psi}}^\perp = \boldsymbol{0}$$

due to $\boldsymbol{\psi}\big(\boldsymbol{\varepsilon} + \boldsymbol{h}(\boldsymbol{\varepsilon}, \boldsymbol{X})^\top \boldsymbol{a}(\boldsymbol{P}_{\boldsymbol{\psi}}^\perp \boldsymbol{\eta}), \boldsymbol{X} + (\boldsymbol{P}_{\boldsymbol{\psi}}^\perp \boldsymbol{\eta})\boldsymbol{a}^\top\big) = \boldsymbol{\psi}(\boldsymbol{\varepsilon}, \boldsymbol{X})$ by (C.20) as in the proof of Proposition C.4 (iv). Note that if $F(\boldsymbol{y}, \boldsymbol{X})$ and $G(\boldsymbol{\varepsilon}, \boldsymbol{X})$ are functions such that $G(\boldsymbol{\varepsilon}, \boldsymbol{X}) = F(\boldsymbol{\varepsilon} + \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X})$ and $F(\boldsymbol{y}, \boldsymbol{X}) = G(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X})$ then $(\partial/\partial y_i) F(\boldsymbol{y}, \boldsymbol{X}) = (\partial/\partial \varepsilon_i) G(\boldsymbol{\varepsilon}, \boldsymbol{X})$ whenever F is Fréchet differentiable at $(\boldsymbol{y}, \boldsymbol{X})$ and G is Fréchet differentiable at $(\boldsymbol{\varepsilon}, \boldsymbol{X})$ (i.e., translation by a constant in the variables does not change the derivatives). □

Proof of Proposition C.4 (i). Let $\partial g(\cdot)$ denote the subdifferential of g. The KKT conditions read $\boldsymbol{X}^\top \boldsymbol{\psi} \in n\partial g(\hat{\boldsymbol{\beta}})$ and $\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{\psi}} \in n\partial g(\tilde{\boldsymbol{\beta}})$. Taking the difference and by τ-strong convexity of g, we have

$$\begin{aligned} n\tau \|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_2^2 &\le (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top (\boldsymbol{X}^\top \boldsymbol{\psi} - \widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{\psi}}) \\ &= (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top \big[(\boldsymbol{X} - \widetilde{\boldsymbol{X}})^\top \boldsymbol{\psi} + \widetilde{\boldsymbol{X}}^\top (\boldsymbol{\psi} - \widetilde{\boldsymbol{\psi}}) \big] \end{aligned}$$

For the second term,

$$
\begin{aligned}
&(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top \widetilde{\boldsymbol{X}}^\top (\boldsymbol{\psi} - \widetilde{\boldsymbol{\psi}}) \\
&= (\widetilde{\boldsymbol{X}}\boldsymbol{h} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{h}})^\top (\boldsymbol{\psi} - \widetilde{\boldsymbol{\psi}}) \\
&= -(\boldsymbol{\varepsilon} - \boldsymbol{X}\boldsymbol{h} - (\widetilde{\boldsymbol{\varepsilon}} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{h}}))^\top (\boldsymbol{\psi} - \widetilde{\boldsymbol{\psi}}) + (\boldsymbol{\varepsilon} - \widetilde{\boldsymbol{\varepsilon}} + \widetilde{\boldsymbol{X}}\boldsymbol{h} - \boldsymbol{X}\boldsymbol{h})^\top (\boldsymbol{\psi} - \widetilde{\boldsymbol{\psi}}).
\end{aligned}
$$

Since ψ is non-decreasing and L-Lipschitz, $L^{-1}\|\boldsymbol{\psi} - \widetilde{\boldsymbol{\psi}}\|_2^2 \le \big(\boldsymbol{\varepsilon} - \boldsymbol{X}\boldsymbol{h} - (\widetilde{\boldsymbol{\varepsilon}} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{h}})\big)^\top (\boldsymbol{\psi} - \widetilde{\boldsymbol{\psi}})$ holds. Combining the above displays we obtain (C.20). □

Proof of Proposition C.4 (ii). For fixed values of $(\boldsymbol{\varepsilon}, \boldsymbol{X}, \boldsymbol{h}, \boldsymbol{\psi})$, inequality (C.20) divided by $1 + \|\boldsymbol{h} - \widetilde{\boldsymbol{h}}\|_2 + \|\boldsymbol{\psi} - \widetilde{\boldsymbol{\psi}}\|_2$ implies that $(\widetilde{\boldsymbol{h}}, \widetilde{\boldsymbol{\psi}}) \to (\boldsymbol{h}, \boldsymbol{\psi})$ as $(\widetilde{\boldsymbol{\varepsilon}}, \widetilde{\boldsymbol{X}}) \to (\boldsymbol{\varepsilon}, \boldsymbol{X})$, hence the function $(\boldsymbol{\varepsilon}, \boldsymbol{X}) \mapsto (\boldsymbol{h}(\boldsymbol{\varepsilon}, \boldsymbol{X}), \boldsymbol{\psi}(\boldsymbol{\varepsilon}, \boldsymbol{X}))$ is everywhere continuous. This implies that $S(K) = \sup_{(\boldsymbol{\varepsilon}, \boldsymbol{X}) \in K} \big((n\tau)^{-1}\|\boldsymbol{\psi}(\boldsymbol{\varepsilon}, \boldsymbol{X})\|_2^2 + L\|\boldsymbol{h}(\boldsymbol{\varepsilon}, \boldsymbol{X})\|_2^2\big)$ is finite for any compact $K \subset \mathbb{R}^n \times \mathbb{R}^{n \times p}$. The Cauchy-Schwarz inequality on the right hand side of (C.20) gives for any $(\boldsymbol{\varepsilon}, \boldsymbol{X}), (\widetilde{\boldsymbol{\varepsilon}}, \widetilde{\boldsymbol{X}}) \in K$

$$
\begin{aligned}
&n\tau\|\boldsymbol{h} - \widetilde{\boldsymbol{h}}\|_2^2 + L^{-1}\|\boldsymbol{\psi} - \widetilde{\boldsymbol{\psi}}\|_2^2 \\
&\le \|\boldsymbol{X} - \widetilde{\boldsymbol{X}}\|_{\mathrm{op}}(\|\boldsymbol{h} - \widetilde{\boldsymbol{h}}\|_2\|\boldsymbol{\psi}\|_2 + \|\boldsymbol{h}\|_2\|\boldsymbol{\psi} - \widetilde{\boldsymbol{\psi}}\|_2) + \|\boldsymbol{\varepsilon} - \widetilde{\boldsymbol{\varepsilon}}\|_2\|\boldsymbol{\psi} - \widetilde{\boldsymbol{\psi}}\|_2 \\
&\le \|\boldsymbol{X} - \widetilde{\boldsymbol{X}}\|_{\mathrm{op}}(n\tau\|\boldsymbol{h} - \widetilde{\boldsymbol{h}}\|_2^2 + L^{-1}\|\boldsymbol{\psi} - \widetilde{\boldsymbol{\psi}}\|_2^2)^{\frac{1}{2}} S(K)^{\frac{1}{2}} + \|\boldsymbol{\varepsilon} - \widetilde{\boldsymbol{\varepsilon}}\|_2\|\boldsymbol{\psi} - \widetilde{\boldsymbol{\psi}}\|_2,
\end{aligned}
$$

This implies that $(\boldsymbol{\varepsilon}, \boldsymbol{X}) \mapsto (\boldsymbol{h}, \boldsymbol{\psi})$ is Lipschitz on K. □

Proof of Proposition C.4 (iii). Combined with (C.20) with $\boldsymbol{\varepsilon} = \widetilde{\boldsymbol{\varepsilon}}$, we have

$$
\begin{aligned}
&n\tau\|\boldsymbol{h} - \widetilde{\boldsymbol{h}}\|_2^2 + L^{-1}\|\boldsymbol{\psi} - \widetilde{\boldsymbol{\psi}}\|_2^2 \\
&\le -(\boldsymbol{h} - \widetilde{\boldsymbol{h}})^\top \boldsymbol{a}(\boldsymbol{\eta}^\top \boldsymbol{\psi}) + (\boldsymbol{h}^\top \boldsymbol{a})\boldsymbol{\eta}^\top (\boldsymbol{\psi} - \widetilde{\boldsymbol{\psi}}) \\
&\le \|\boldsymbol{h} - \widetilde{\boldsymbol{h}}\|_2\|\boldsymbol{a}\|_2|\boldsymbol{\eta}^\top \boldsymbol{\psi}| + |\boldsymbol{h}^\top \boldsymbol{a}|\|\boldsymbol{\eta}\|_2\|\boldsymbol{\psi} - \widetilde{\boldsymbol{\psi}}\|_2 \\
&\le \big(n\tau\|\boldsymbol{h} - \widetilde{\boldsymbol{h}}\|_2^2 + L^{-1}\|\boldsymbol{\psi} - \widetilde{\boldsymbol{\psi}}\|_2^2\big)^{1/2}\big((n\tau)^{-1}\|\boldsymbol{a}\|_2^2(\boldsymbol{\eta}^\top \boldsymbol{\psi})^2 + L(\boldsymbol{h}^\top \boldsymbol{a})^2\|\boldsymbol{\eta}\|_2^2\big)^{1/2}
\end{aligned}
$$

so that the first inequality holds. Taking summation over $\boldsymbol{\eta} = \boldsymbol{e}_i$ for $i \in [n]$ gives (C.21). □

Proof of Proposition C.4 (iv). For $\widetilde{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon} + \boldsymbol{h}(\boldsymbol{\varepsilon}, \boldsymbol{X})^\top \boldsymbol{e}_j \boldsymbol{\eta}$ and $\widetilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{\eta}\boldsymbol{e}_j^\top$ we have $(\boldsymbol{X} - \widetilde{\boldsymbol{X}})^\top \boldsymbol{\psi} = 0$ thanks to $\boldsymbol{\eta}^\top \boldsymbol{\psi} = 0$ as well as $\boldsymbol{\varepsilon} - \widetilde{\boldsymbol{\varepsilon}} + (\widetilde{\boldsymbol{X}} - \boldsymbol{X})\boldsymbol{h} = 0$. Hence the two terms in the right-hand side of (C.20) are 0 and $\|\boldsymbol{\psi} - \widetilde{\boldsymbol{\psi}}\|_2^2 = 0$. Identity (C.22) then follows by definition of the Fréchet differentiability. □

C.3.2. Approximation using smooth penalty $\tilde{g}^\nu$

Lemma C.6 (Approximation of strongly convex functions)**.** *Let $g : \mathbb{R}^p \to \mathbb{R}$ be strongly convex with constant $\tau \ge 0$. Then for every $\nu > 0$ there exists a real-analytic strongly convex function $g_\nu : \mathbb{R}^p \to \mathbb{R}$ with constant τ such that $g_\nu - \nu \le g \le g_\nu$.*

Proof. Since g is proper, i.e., $-\infty \notin g(\mathbb{R}^p)$ and $\{\boldsymbol{b} \in \mathbb{R}^p \mid g(\boldsymbol{b}) < +\infty\} \neq \emptyset$, by Proposition 10.8 in [2], g is strongly convex with constant $\tau \geq 0$ if and only if $f := g - (\tau/2)\|\cdot\|_2^2$ is convex. By Theorem 1 in [1], there exists a function $f_\nu : \mathbb{R}^p \to \mathbb{R}$ real-analytic and convex that satisfies $f_\nu - \nu \leq f \leq f_\nu$. The conclusion follows by letting $g_\nu := f_\nu + (\tau/2)\|\cdot\|_2^2$. □

Lemma C.7. *Let $\rho : \mathbb{R} \to \mathbb{R}$ be convex and continuously differentiable with derivative $\psi = \rho'$ being L-Lipschitz. Let $g, \tilde{g} : \mathbb{R}^p \to \mathbb{R}$ be strongly convex with parameter $\tau, \tilde{\tau} \geq 0$. Let $\|g - \tilde{g}\|_\infty = \max_{\boldsymbol{x} \in \mathbb{R}^p} |g(\boldsymbol{x}) - \tilde{g}(\boldsymbol{x})|$. For $\boldsymbol{b} \in \mathbb{R}^p$, let $L(\boldsymbol{b}; g) = \frac{1}{n}\sum_{i\in[n]} \rho(y_i - \boldsymbol{x}_i^\top \boldsymbol{b}) + g(\boldsymbol{b})$ and define*

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} L(\boldsymbol{b}; g), \quad \tilde{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} L(\boldsymbol{b}; \tilde{g}), \quad \boldsymbol{\psi} = \psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}), \quad \widetilde{\boldsymbol{\psi}} = \psi(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}).$$

Then inequality $((\tau + \tilde{\tau})/2)\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\|_2^2 + (nL)^{-1}\|\boldsymbol{\psi} - \widetilde{\boldsymbol{\psi}}\|_2^2 \leq 2\|g - \tilde{g}\|_\infty$ holds.

Proof of Lemma C.7. Denote by $\partial g(\boldsymbol{b})$ subdifferential of g at $\boldsymbol{b} \in \mathbb{R}^p$. The KKT conditions read

$$(1/n)\boldsymbol{X}^\top \boldsymbol{\psi} \in \partial g(\hat{\boldsymbol{\beta}}), \qquad (1/n)\boldsymbol{X}^\top \widetilde{\boldsymbol{\psi}} \in \partial \tilde{g}(\tilde{\boldsymbol{\beta}}).$$

By the definition of the strongly convexity, the above display implies that

$$(\tau/2)\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\|_2^2 + (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^\top (1/n)\boldsymbol{X}^\top \boldsymbol{\psi} \leq g(\tilde{\boldsymbol{\beta}}) - g(\hat{\boldsymbol{\beta}}),$$
$$(\tilde{\tau}/2)\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_2^2 + (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top (1/n)\boldsymbol{X}^\top \widetilde{\boldsymbol{\psi}} \leq \tilde{g}(\hat{\boldsymbol{\beta}}) - \tilde{g}(\tilde{\boldsymbol{\beta}}).$$

Summing over the above displays, we have

$$\begin{aligned}\tfrac{\tau+\tilde{\tau}}{2}\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_2^2 + (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top (1/n)\boldsymbol{X}^\top (\widetilde{\boldsymbol{\psi}} - \boldsymbol{\psi}) &\leq g(\tilde{\boldsymbol{\beta}}) - \tilde{g}(\tilde{\boldsymbol{\beta}}) + \tilde{g}(\hat{\boldsymbol{\beta}}) - g(\hat{\boldsymbol{\beta}}) \\ &\leq 2\|g - \tilde{g}\|_\infty.\end{aligned}$$

We notice that the second term in the left hand side is can be rewritten

$$\frac{1}{n}\langle \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - (\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}), \psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) - \psi(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})\rangle = \sum_{i=1}^n \frac{(a_i - b_i)(\psi(a_i) - \psi(b_i))}{n}$$

where $a_i = y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}$ and $b_i = y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}$. Since ψ non-decreasing and L-Lipschitz, inequality $(a_i - b_i)(\psi(a_i) - \psi(b_i)) = |a_i - b_i||\psi(a_i) - \psi(b_i)| \geq \frac{1}{L}(\psi(a_i) - \psi(b_i))^2$ completes the proof. □

Proof of Lemma A.2 for g τ-strongly convex but not continuously differentiable. In this proof, we approximate the non-smooth g with smooth g using Lemma C.6. Let g be strongly convex with constant $\tau > 0$, not necessarily twice differentiable. By Lemma C.6, for all $\nu > 0$, there exists $\tilde{g}^\nu$ strongly convex with constant $\tau > 0$ such that $\|\tilde{g}^\nu - g\|_\infty \leq \nu$. Let $\tilde{\boldsymbol{\beta}}^\nu = \operatorname{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} L(\boldsymbol{b}; \tilde{g}^\nu)$ and $\widetilde{\boldsymbol{\psi}}^\nu = \psi(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}^\nu)$ be as in Lemma C.7. For any $\delta > 0$,

$$\mathbb{E}\Big[\frac{n}{\|\widetilde{\boldsymbol{\psi}}^\nu\|_2^2 + \delta}\sum_{j\in[p]} I_{\mathscr{E}_j}\tilde{h}_j^2\Big]^{1/2} \leq \frac{\big[(1 + \sqrt{\frac{p}{n}})^2 + \frac{1}{n}\big]^{1/2}}{\phi_{\min}(\boldsymbol{\Sigma})^{1/2}(1 - \eta_n)_+^2 u_*} + \frac{\big[\frac{2p}{n\tau} + \mathbb{E}[\|\widetilde{\boldsymbol{h}}^\nu\|_2^2]\big]^{1/2}}{u_*^{1/2}} \tag{C.25}$$

by (A.3) since $\tilde{g}^\nu$ is twice continuously differentiable. By Lemma C.7, we have

$$\tau\|\tilde{\boldsymbol{\beta}}^\nu - \hat{\boldsymbol{\beta}}\|_2^2 + (nL)^{-1}\|\widetilde{\boldsymbol{\psi}}^\nu - \boldsymbol{\psi}\|_2^2 \le 2\nu.$$

This implies that, as $\nu \to 0+$, the pointwise convergence $(\widetilde{\boldsymbol{h}}_j^\nu, \widetilde{\boldsymbol{\psi}}^\nu) \to (\boldsymbol{h}, \boldsymbol{\psi})$ holds. By the dominated convergence theorem, a sufficient condition that (C.25) holds with $(\widetilde{\boldsymbol{h}}^\nu, \widetilde{\boldsymbol{\psi}}^\nu)$ replaced by its pointwise limit $(\boldsymbol{h}, \boldsymbol{\psi})$ inside the two expectations in (C.25) is that $\mathbb{E}\sup_{\nu\in(0,1)} \|\tilde{\boldsymbol{h}}^\nu\|_2^2 < +\infty$. By Lemma C.7,

$$\|\widetilde{\boldsymbol{h}}^\nu\|_2^2 \le 2\|\widetilde{\boldsymbol{h}}^\nu - \boldsymbol{h}\|_2^2 + 2\|\boldsymbol{h}\|_2^2 < (2\nu/\tau) + 2\|\boldsymbol{h}\|_2^2 \tag{C.26}$$

which provides integrability of $\sup_{\nu\in(0,1)} \|\widetilde{\boldsymbol{h}}^\nu\|_2^2$ as $\mathbb{E}[\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}\|_2^2] < +\infty$ when the right-hand side of (A.3) is finite. □

Appendix D: Auxiliary propositions

D.1. Decomposition of the design matrix into independent components

Proposition D.1 (Independence between $\boldsymbol{X}(\boldsymbol{I}_p - \boldsymbol{b}\boldsymbol{a}^\top)$ and $\boldsymbol{X}\boldsymbol{b}$)**.** *Let each row $\boldsymbol{x}_i$ of $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ satisfy that $\boldsymbol{x}_i \sim^{iid} N(\boldsymbol{0}, \boldsymbol{\Sigma})$. Then for any deterministic vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$, $\boldsymbol{\Sigma}\boldsymbol{b} = (\boldsymbol{b}^\top\boldsymbol{\Sigma}\boldsymbol{b})\boldsymbol{a}$ holds if and only if $\boldsymbol{X}(\boldsymbol{I}_p - \boldsymbol{b}\boldsymbol{a}^\top)$ is independent with $\boldsymbol{X}\boldsymbol{b}$. Furthermore, if the above holds and the inverse matrix $\boldsymbol{\Sigma}^{-1}$ exists, then $(\boldsymbol{b}^\top\boldsymbol{\Sigma}\boldsymbol{b})(\boldsymbol{a}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{a}) = 1$.*

Proof. From the fact that $(\boldsymbol{X}(\boldsymbol{I}_p - \boldsymbol{b}\boldsymbol{a}^\top), \boldsymbol{X}\boldsymbol{b})$ can be represented in a linear transformation of $n \times p$ iid $N(0,1)$ random variable, the pair is distributed in a multivariate normal distribution. Since the rows of $\boldsymbol{X}$ are independent, the independence between $\boldsymbol{X}(\boldsymbol{I}_p - \boldsymbol{b}\boldsymbol{a}^\top)$ and $\boldsymbol{X}\boldsymbol{b}$ reduces to the independence between $\boldsymbol{x}_i^\top(\boldsymbol{I}_p - \boldsymbol{b}\boldsymbol{a}^\top)$ and $\boldsymbol{x}_i^\top\boldsymbol{b}$ for each $i \in [n]$, which holds if and only if the two random quantities are uncorrelated in the sense that

$$\mathbb{E}[(\boldsymbol{x}_i^\top(\boldsymbol{I}_p - \boldsymbol{b}\boldsymbol{a}^\top))(\boldsymbol{x}_i^\top\boldsymbol{b})] = \mathbb{E}[\boldsymbol{b}^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top(\boldsymbol{I}_p - \boldsymbol{b}\boldsymbol{a}^\top)] = \boldsymbol{b}^\top\boldsymbol{\Sigma}(\boldsymbol{I}_p - \boldsymbol{b}\boldsymbol{a}^\top) = \boldsymbol{0}.$$

If the inverse $\boldsymbol{\Sigma}^{-1}$ exists, the above display implies $(\boldsymbol{b}^\top\boldsymbol{\Sigma}\boldsymbol{b})(\boldsymbol{a}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{a}) = 1$. □

D.2. ψ at the residuals is almost surely nonzero

Proposition D.2. *If Assumptions A, B and C hold then $\mathbb{P}(\psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \neq \boldsymbol{0}) = 1$.*

Proof of Proposition D.2. If $\psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = 0$ then $\hat{\boldsymbol{\beta}}$ must be a minimizer of the penalty g. Let $\boldsymbol{b}_0$ be a minimizer of g, which is unique by strong convexity.

Our assumption on the convexity of ρ implies that $\psi(x)$ is non-decreasing in $x \in \mathbb{R}$. Combined with our assumption $\psi'(x) + \psi^2(x) \ge K^2 > 0$ for every point $x \in \mathbb{R}$, this implies that $\psi(x) = 0$ at no more than one point in $\mathbb{R}$. (Otherwise,

there exists an open interval on which $\psi(x) = 0$ and $\psi'(x) + \psi^2(x) = 0$. A contradiction then follows.)

Thus we have $\mathbb{P}(\psi(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{0}) \leq \mathbb{P}(\psi(\boldsymbol{\varepsilon} - \boldsymbol{X}(\boldsymbol{b}_0 - \boldsymbol{\beta})) = \boldsymbol{0}) = 0$ as $\boldsymbol{\varepsilon} - \boldsymbol{X}(\boldsymbol{b}_0 - \boldsymbol{\beta})$ has continuous distribution by Assumption C and $\{x \in \mathbb{R} : \psi(x) = 0\}$ has zero Lebesgue measure. □

References

[1] Daniel Azagra. Global and fine approximation of convex functions. *Proceedings of the London Mathematical Society*, 107(4):799–824, 2013. MR3108831

[2] Heinz H Bauschke and Patrick L Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2017. MR3616647

[3] Derek Bean, Peter J Bickel, Noureddine El Karoui, and Bin Yu. Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36):14563–14568, 2013.

[4] Pierre C Bellec. Out-of-sample error estimate for robust m-estimators with convex penalty. *arXiv preprint arXiv:2008.11840*, 2020.

[5] Pierre C Bellec and Yiwei Shen. Derivatives and residual distribution of regularized m-estimators with application to adaptive tuning. In *Conference on Learning Theory*, pages 1912–1947. PMLR, 2022.

[6] Pierre C Bellec and Cun-Hui Zhang. Second order stein: Sure for sure and other applications in high-dimensional inference. *arXiv preprint arXiv:1811.04121*, 2018. MR4319234

[7] Pierre C Bellec and Cun-Hui Zhang. De-biasing convex regularized estimators and interval estimation in linear models. *arXiv preprint arXiv:1912.11943*, 2019.

[8] Pierre C Bellec and Cun-Hui Zhang. De-biasing the lasso with degrees-of-freedom adjustment. *arXiv:1902.08885*, 2019. URL `https://arxiv.org/pdf/1902.08885.pdf`. MR4389062

[9] Vladimir Igorevich Bogachev. *Gaussian measures*. Number 62. American Mathematical Soc., 1998. MR1642391

[10] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013. MR3185193

[11] Peter Bühlmann et al. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242, 2013. MR3102549

[12] Michael Celentano and Andrea Montanari. Fundamental barriers to high-dimensional regression with convex penalties. *arXiv preprint arXiv:1903.10603*, 2019. MR4382013

[13] Michael Celentano, Andrea Montanari, and Yuting Wei. The lasso with general gaussian designs with applications to hypothesis testing. *arXiv preprint arXiv:2007.13716*, 2020.

[14] Kenneth R Davidson and Stanislaw J Szarek. Local operator theory, ran-

dom matrices and banach spaces. *Handbook of the geometry of Banach spaces*, 1(317-366):131, 2001. MR1863696

[15] David Donoho and Andrea Montanari. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3-4):935–969, 2016. MR3568043

[16] Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1-2):95–175, 2018. MR3748322

[17] Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[18] Emmanuel Hebey. *Sobolev spaces on Riemannian manifolds*, volume 1635. Springer Science & Business Media, 1996. MR1481970

[19] Hanwen Huang. Asymptotic risk and phase transition of l_1-penalized robust estimator. *Ann. Statist.*, 48(5):3090–3111, 10 2020. URL https://doi.org/10.1214/19-AOS1923. MR4152636

[20] Peter J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 03 1964. URL https://doi.org/10.1214/aoms/1177703732. MR0161415

[21] Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004. MR0606374

[22] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014. MR3277152

[23] Adel Javanmard and Andrea Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554, 2014. MR3265038

[24] Adel Javanmard and Andrea Montanari. Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622, 2018. MR3851749

[25] Noureddine El Karoui. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*, 2013.

[26] Léo Miolane and Andrea Montanari. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *arXiv preprint arXiv:1811.01212*, 2018. MR4319252

[27] Stephen Portnoy. Asymptotic behavior of m estimators of p regression parameters when p2/n is large; ii. normal approximation. *The Annals of Statistics*, pages 1403–1417, 1985. MR0811499

[28] Mihailo Stojnic. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013.

[29] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018. MR3832326

[30] Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014. MR3224285

[31] AW Van der Vaart. Estimating a real parameter in a class of semiparametric models. *The Annals of Statistics*, pages 1450–1474, 1988. MR0964933

[32] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018. MR3837109

[33] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014. MR3153940

[34] William P. Ziemer. *Weakly Differentiable Functions: Sobolev Spaces and Functions of Bounded Variation*. Graduate Texts in Mathematics. Springer, 1989. MR1014685