Electronic Journal of Statistics
Vol. 14 (2020) 2551–2599
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1728

Detangling robustness in high dimensions: Composite versus model-averaged estimation

Jing Zhou and Gerda Claeskens

ORStat and Leuven Statistics Research Center, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium
e-mail: Jing.Zhou@kuleuven.be; Gerda.Claeskens@kuleuven.be

Jelena Bradic

Department of Mathematics and Halicioğlu Data Science Institute, University of California San Diego, 9500 Gilman Drive 0120, La Jolla, California
e-mail: jbradic@ucsd.edu

Abstract: Robust methods, though ubiquitous in practice, are yet to be fully understood in the context of regularized estimation and high dimensions. Even simple questions become challenging very quickly. For example, classical statistical theory identifies equivalence between model-averaged and composite quantile estimation. However, little to nothing is known about such equivalence between methods that encourage sparsity. This paper provides a toolbox to further study robustness in these settings and focuses on prediction. In particular, we study optimally weighted model-averaged as well as composite l_1-regularized estimation. Optimal weights are determined by minimizing the asymptotic mean squared error. This approach incorporates the effects of regularization, without the assumption of perfect selection, as is often used in practice. Such weights are then optimal for prediction quality. Through an extensive simulation study, we show that no single method systematically outperforms others. We find, however, that model-averaged and composite quantile estimators often outperform least-squares methods, even in the case of Gaussian model noise. Real data application witnesses the method's practical use through the reconstruction of compressed audio signals.

MSC 2010 subject classifications: Primary 62J07; secondary 62F12.
Keywords and phrases: Mean squared error, l_1-regularization, approximate message passing, quantile regression.

Received March 2020.

Contents

1. Introduction

We investigate the benefits of model-averaged as well as composite estimators in high-dimensional problems where the underlying goal is superior prediction quality. Robustness in data analysis with potentially more parameters than samples is a critical practical question and is of particular interest in constructing recoveries of compressed images and signals which should have high precision.

Model averaging, often used as a first tool to improve estimation quality, forms a weighted average of estimators and is here utilized for regularized sparsity-encouraging estimation in a high-dimensional regression setting. Model averaging is also well-known in the Bayesian setting [22], though we focus on

its frequentist version in which a user determines the weights assigned to the separate estimators [e.g., see 12, 21, 37]. Model averaging enjoys a wide application, see, for example, the recent overview paper for model averaging in ecology by Dormann et al. [15] and for application to hydrology by Höge et al. [20]. In econometrics, the terminology "forecast combinations" appears [e.g., in 11, 3]; whereas "multimodel inference" is another commonly used term for this procedure [10].

While the technique is quite thoroughly investigated for low-dimensional models, far fewer results have been obtained in high dimensions. Ando and Li [1] consider high-dimensional linear regression. By computing the marginal correlation between each covariate and the response and forming groups according to the obtained values, regularized estimation is avoided. The authors fit a fixed number of low-dimensional models by the least squares method and subsequently average them. Zhao et al. [38] extend this method to dependent data, while Ando and Li [2] extend this approach to generalized linear models, again by only fitting low-dimensional models, this time via maximum likelihood estimation. In these papers, the weights are obtained via cross-validation; see also Hansen [18] and Hansen and Racine [19] for similar weight finding approaches in low-dimensional models.

Our setting is different and is theoretically valid (see Theorem 3 below). We explicitly work with l_1-regularized estimators that are averaged, and we do not rely on the correct low-dimensional representation of the model. When designing the optimal weights, we explicitly take variable selection effects (of regularization itself) into account. Is the dependence among regularized estimators an impediment or a hidden benefit in obtaining robust predictions, i.e., predictions that do not change much when the data is changed a little?

A second approach to robustness is through composite estimation. While model averaging combines estimators after optimization of their respective loss functions, composite estimation weights the loss functions directly (before optimization). For quantile regression in low dimensions, Koenker [29, Theorem 5.2] stated the asymptotic equivalence of model-averaged and composite quantile regression estimators, provided each method uses its own, optimal set of weights that minimize the asymptotic variance. Hence, with optimal weights, there is no asymptotic preference between the two methods in low dimensions. For high-dimensional quantile regression, when one restricts the attention to inference regarding the true nonzero part of the regression coefficient and ignores the variable selection effect, Bloznelis et al. [7] obtained the same equivalence for high-dimensional quantile regression using different types of regularizations (SCAD, lasso, adaptive lasso).

In practice, however, one works with an estimated coefficient vector for which one is not sure that the regularization has led to the correct selection. Therefore, incorporating imperfections of variable selection is especially important for achieving robustness. This is where our approach differs from Bloznelis et al. [7] or Bradic et al. [9], where an irrepresentable condition (needed for consistent model or asymptotically perfect selection) has been used to specify weights and analyze robustness.

The approximate message passing (AMP) algorithm is crucial in our approach to take the variable selection into account when studying the estimators' asymptotic mean squared errors. The use of such algorithms has been investigated by Donoho et al. [13] and Bayati and Montanari [5] for compressed sensing. Donoho and Montanari [14] explain the use of AMP algorithms for obtaining the variance of high-dimensional M-estimators for which $n/p \to \delta \in (1, \infty)$. Here, n denotes the sample size and p the number of regression coefficients. However, the robustness of sparsity encouraging AMP estimators is still largely unknown.

In this paper, we first extend the robust AMP (RAMP) of Bradic [8] to regularized composite estimation. Second, we construct estimators and develop new theory for the asymptotic mean squared error (AMSE) both for model-averaged and for composite estimators. Note that model-averaged AMSE required an extension of AMP theory for a challenging case of dependent estimates. Besides, we establish new Stein-type risk estimates of the AMSE in both cases.

The new estimates of the AMSE of the model-averaged and composite estimators enable a theoretically justified and data-driven optimal weight choice by minimizing the estimated AMSE (without relying on perfect variable selection). The estimated AMSE provides more information regarding the estimators than merely considering which variables have been selected.

Organization of the paper. First, in Section 2, we detail the model-averaged and composite estimators in a high-dimensional setup. Next, we explain the model-averaged robust message passing algorithm in Section 3. The limiting behavior of the estimators in the algorithm is studied by state evolution parameters in Section 4. We obtain the estimators' asymptotic mean squared error as well as an estimator of that quantity in Section 5. We showcase the procedure for high-dimensional regularized quantile regression in Section 6 and present numerical results in Section 7. Section 8 concludes. All proofs, together with the assumptions and some technical lemmas, are collected in the Appendix.

2. Model-averaged and composite estimation

We consider a high-dimensional linear model $Y = X\beta + \varepsilon$ with $Y \in \mathbb{R}^n$, the design matrix $X \in \mathbb{R}^{n \times p}$ and the parameter vector $\beta \in \mathbb{R}^p$. The ith row of X is denoted $X_{i\cdot}$, $i = 1, \ldots, n$, the jth column of X is denoted by $X_{\cdot j}$, $j = 1, \ldots, p$. We assume the components of ε to be independent and identically distributed with mean zero, cumulative distribution function F_ε and probability density function f_ε. We allow for a sparse high-dimensional setup. Denote by s the l_0 norm of the parameter vector, $s = \|\beta\|_0$, which counts the number of nonzero components of the vector β. We assume that the ratios $n/p \to \delta \in (0, 1)$ and $n/s \to a \in (1, \infty)$ when p, n, s tend to ∞.

We consider two types of weighted estimation methods. First, model-averaged estimation where estimators from different models or estimation methods are weighted and summed to arrive at a final estimator, see (2). Second, composite estimation where a weighted average of loss functions is minimized; see (3).

For model-averaged estimation of the parameter β, define for $k = 1, \ldots, K$ the regularized estimators

$$\widehat{\beta}_k(\lambda_k) = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} \rho_k(Y_i - X_{i\cdot}\beta) + \lambda_k \|\beta\|_1 \right\}, \tag{1}$$

where $\rho_1, \ldots, \rho_K$ are nonnegative convex loss functions and $\lambda = (\lambda_1, \ldots, \lambda_K)^\top$ is a vector of possibly different nonnegative regularization parameters. For a set of weights $w = (w_1, \ldots, w_K)^\top$, the model-averaged estimator is defined as

$$\widehat{\beta}_{\mathrm{MA}}(\lambda) = \sum_{k=1}^{K} w_k \widehat{\beta}_k(\lambda_k). \tag{2}$$

Often one assumes that the weights $w_1, \ldots, w_K$ are all nonnegative and sum to 1, although this is not necessary for the computation of the estimator.

For composite estimation we consider again K loss functions, though only with a single nonnegative regularization parameter λ, such that the regularized composite estimator is defined as

$$\widehat{\beta}_{\mathrm{C}}(\lambda) = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{k=1}^{K} \sum_{i=1}^{n} w_k \rho_k(Y_i - X_{i\cdot}\beta) + \lambda \|\beta\|_1 \right\}. \tag{3}$$

Computationally, composite estimation is harder than model-averaged estimation and requires that all weights are positive to ensure a nonnegative and convex weighted loss function, even when all ρ_k are nonnegative and convex. Hence, for composite estimation it is required that the weight vector $w \in [0,1]^K$ such that $\sum_{k=1}^{K} w_k = 1$.

As a worked-out scenario throughout the paper, we consider quantile loss functions $\rho_k(\cdot), k = 1, \ldots, K$ that are defined below. For more information about quantile regression with i.i.d. errors, see Koenker [29, Sec. 3.2.2]. In this paper, we assume that the design matrix X does not contain a column of ones; see assumption (A1) in the Appendix. With $\tau \in (0,1)$, the τ-quantile of the response Y is obtained as $X\beta + F_\varepsilon^{-1}(\tau) = X\beta + u_\tau$.

Figure 1 presents first a single quantile loss function with $\tau = 0.3$,

$$\rho(x) = (x - u_\tau)(\tau - I\{x \le u_\tau\}).$$

For model averaging we specify K different quantile levels and use K different such quantile loss functions for estimation of β:

$$\rho_k(x) = (x - u_{\tau_k})(\tau_k - I\{x \le u_{\tau_k}\}), \quad k \in \{1, \ldots, K\}.$$

For composite quantile estimation we assume that the K quantile levels $\tau_1 < \cdots < \tau_K$, then also the quantiles of ε are sorted $u_{\tau_1} < \cdots < u_{\tau_K}$. Define $u_{\tau_0} = -\infty$ and $u_{\tau_{K+1}} = \infty$.

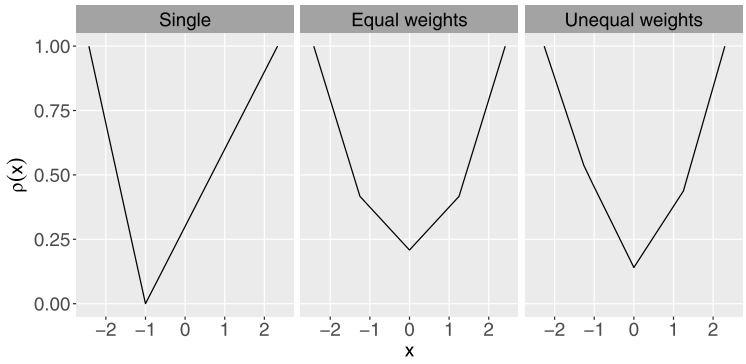

FIG 1. *Examples of quantile loss functions. Left:* $\tau = 0.3$ *quantile loss function. Middle: Composite quantile loss function at quantile levels 0.25, 0.5, 0.75 with equal weights* $w = (1/3, 1/3, 1/3)^\top$. *Right: Composite quantile loss function at quantile levels 0.25, 0.5, 0.75 with weights* $w = (0.15, 0.55, 0.3)^\top$.

The middle panel of Figure 1 depicts such a composite quantile loss function $\rho_C = \sum_{k=1}^K w_k \rho_k$ for $K = 3$ quantile levels 0.25, 0.5 and 0.75 with equal weights $w = (1/3, 1/3, 1/3)^\top$. The panel on the right in Figure 1 uses the same quantile levels but depicts the quantile loss function ρ_C with weights $w = (0.15, 0.55, 0.3)^\top$.

In general, the composite quantile loss function can be rewritten in the following way,

$$\rho_{\mathrm{C}}(x) = \begin{cases} \sum_{k=1}^K w_k(1-\tau_k)(u_{\tau_k} - x), & x < u_{\tau_1} \\ \sum_{k=1}^K w_k \tau_k (x - u_{\tau_k}), & x \geq u_{\tau_K} \\ \sum_{k=1}^{\ell} w_k \tau_k |x - u_{\tau_k}| + \sum_{k=\ell+1}^K w_k(1-\tau_k)|x - u_{\tau_k}|, & x \in [u_{\tau_\ell}, u_{\tau_{\ell+1}}) \\ \quad \text{for } \ell = 1, \ldots, K-1. & \end{cases} \tag{4}$$

Note that a single quantile loss function can be seen as a particular case of a composite loss function: take $K = 1$ and the single weight $w_1 = 1$. Theoretical results regarding regularized estimation for a single quantile loss function can be found in Bradic [8]. Henceforth, we concentrate on the example of the composite case.

One aim of this paper it to investigate the weight choice w by minimizing the asymptotic mean squared error of the estimators $\widehat{\beta}_{\mathrm{MA}}(\lambda)$ and $\widehat{\beta}_{\mathrm{C}}(\lambda)$.

3. Robust approximate message passing

The idea behind approximate message passing algorithms is to provide an iterative procedure that has as its fixed point the estimator of interest; in this case the minimizer (2) of the regularized loss function in the case of model averaging, and the estimator (3) in the case of composite estimation. Due to a convergence in the mean square between the solution of the approximate message passing algorithm and the estimator (2), respectively (3), the asymptotic

mean squared error that holds for the solution of the approximate message-passing algorithm, is also the asymptotic MSE of the other estimator. Studying effects of regularization while allowing $n/p \to \delta \in (0,1)$ is challenging. The AMP provides theoretical advantages in these cases as it enables a complete and tractable, albeit challenging, structure for obtaining AMSE. This paper is the first to obtain and use the asymptotic mean square error of the regularized estimators to optimize the weight choice of both the model-averaged estimator and the composite estimator. We extend the theory of the RAMP to apply to the model-averaged estimator; see Theorem 1. Challenges arise with incorporating dependence into the AMSE expression; see Theorem 2. Theorem 2, in turn, leads to a new Stein-type estimator of RAMPs asymptotic MSE. While we focus on the weight choice, the availability of an estimated AMSE may be used in other contexts, for instance, for the construction of confidence intervals.

3.1. Notation

When the composite loss function $\rho_C = \sum_{k=1}^K w_k \rho_k$ is used in the RAMP algorithm with tuning parameter α we denote the estimator at iteration number t by $\widehat{\beta}_{C,(t)}(\alpha)$. When the value of the tuning parameter is clear from the context, we also denote the RAMP estimator by $\widehat{\beta}_{C,(t)}$.

For constructing the model averaging estimator we denote the separate estimators from the RAMP algorithm using regularity parameters α_k, $k = 1, \ldots, K$ by $\widehat{\beta}_{k,(t)}(\alpha_k)$ and the model-averaged estimator is denoted by $\widehat{\beta}_{\mathrm{MA},(t)}(\alpha) = \sum_{k=1}^K w_k \widehat{\beta}_{k,(t)}(\alpha_k)$ with $\alpha = (\alpha_1, \ldots, \alpha_K)^\top$. When the value of the tuning parameters is clear from the context, we denote the model averaging RAMP estimator by $\widehat{\beta}_{\mathrm{MA},(t)}$.

A generic estimator, without referring to a specific loss function or construction, is denoted by $\widehat{\beta}_{(t)}$, using tuning parameter α; the subscript (t) refers to the iteration number.

3.2. The robust approximate message passing algorithm

We first revise the (robust) approximate message passing algorithm, which consists of three steps iterated until convergence. In comparison with the more straightforward AMP for the case with a differentiable convex loss function [13], this procedure for robust high-dimensional parameter estimation [14, 8] adjusts the residuals to incorporate the valid score directly. While more details are given in Algorithm 1, which is applied to the different loss functions $\rho_1, \ldots, \rho_K$ and to their weighted sum $\rho_C = \sum_{k=1}^K w_k \rho_k$, we here provide the main outline. The used notation does not explicitly indicate a dependence on the number of coefficients p to not overcomplicate the formulas.

Donoho and Montanari [14] proposed to use the following proximal mapping operator to adjust the residuals. With $b > 0$,

$$\mathrm{Prox}(z, b) = \arg\min_{x \in \mathbb{R}} \{ b\rho(x) + \frac{1}{2}(x - z)^2 \}$$

which minimizes the square loss regularized by the non-differentiable loss, ρ. The parameter b controls how the proximal operator map points to the minimum of the non-differentiable loss, where small values correspond to a small movement towards the minimum of ρ. The fixed point solution of the proximal operator coincides with the minimum of the loss function ρ. For more information, see Parikh and Boyd [33].

We continue with the worked out example on quantile regression, see (4). For $\ell = 0, \dots, K$, define

$$h(\ell) = \sum_{k=1}^{\ell} w_k \tau_k - \sum_{k=\ell+1}^{K} w_k (1 - \tau_k), \tag{5}$$

where we define a summation sign to be equal to zero in the case where the upper summation index is smaller than the lower one, that is, $\sum_{i=a}^{b} x_i = 0$ if $b < a$. The proximal operator for the composite quantile case, see (4), is

$$\operatorname{Prox}(z; b) = \begin{cases} z - bh(\ell), & z \in (u_{\tau_\ell} + bh(\ell), u_{\tau_{\ell+1}} + bh(\ell)),\ \ell = 0, \dots, K \\ u_{\tau_\ell} & z \in [u_{\tau_\ell} + bh(\ell - 1), u_{\tau_\ell} + bh(\ell)],\ \ell = 1, \dots, K. \end{cases} \tag{6}$$

See Section B.2.1 for the derivation of the algorithm.

We now describe the three steps in more detail.

Step 1: Create adjusted residuals

We use the estimates $\widehat{\beta}_{(t-1)}$ and $\widehat{\beta}_{(t)}$ from iteration steps $t-1$ and t to compute the adjusted residuals

$$z_{(t)} = Y - X\widehat{\beta}_{(t)} + \tag{7}$$
$$n^{-1} G(z_{(t-1)}; b_{(t-1)}) \sum_{j=1}^{p} I\left\{ \eta\big(\widehat{\beta}_{(t-1),j} + X_{\cdot j} G(z_{(t-1)}; b_{(t-1)}); \theta_{t-1}\big) \neq 0 \right\},$$

where the soft-thresholding function $\eta(x; \theta) = \operatorname{sign}(x) \max(|x| - \theta, 0)$ and the score function G is defined in (11).

In Algorithm 1, see Section 4, we give details on how to set the soft-thresholding parameter θ, which might change in each iteration, and we explain that a proper choice of θ as a function of the regularity constant λ leads to an equivalence of the RAMP estimator and the regularized estimator.

The effective score function used in Donoho and Montanari [14] is

$$\widetilde{G}(z; b) = b \cdot \partial\rho(x)|_{x = \operatorname{Prox}(z;b)}, \text{ with } b > 0; \tag{8}$$

a subgradient is used in case of nondifferentiability. That is, for a value x where ρ is non-differentiable

$$\partial\rho(x) = \{y : \rho(u) \geq \rho(x) + y(u - x), \forall u\}.$$

Throughout, we use ∂_1 as the notation for the partial derivative or partial subgradient of a function with respect to its first argument. Functions (e.g. $\widetilde{G}$) are applied componentwise to vectors.

For the example on composite quantile regression the subgradient of ρ_C is computed as,

$$\partial\rho_C(x)\begin{cases} = h(\ell), & x \in (u_{\tau_\ell},\ u_{\tau_{\ell+1}}),\ \text{for } \ell = 0,\dots,K, \\ \in [h(\ell-1), h(\ell)], & x = u_{\tau_\ell},\ \text{for } \ell = 1,\dots,K, \end{cases} \tag{9}$$

where $h(\ell)$ is defined in (5). The effective score function for composite quantile regression, see Section B.2.2, is

$$\widetilde{G}(z;b) = \begin{cases} bh(\ell), & z \in (u_{\tau_\ell} + bh(\ell), u_{\tau_{\ell+1}} + bh(\ell)),\ \ell = 0,\dots,K \\ z - u_{\tau_\ell}, & z \in [u_{\tau_\ell} + bh(\ell-1), u_{\tau_\ell} + bh(\ell)],\ \ell = 1,\dots,K. \end{cases} \tag{10}$$

To incorporate the sparsity, Bradic [8], see also Bayati and Montanari [5], used the rescaled, min regularized effective score function,

$$G(z;b) = \delta\omega^{-1}\widetilde{G}(z;b) \tag{11}$$

where $\omega = P(B_0 \neq 0)$, see assumption (A2) in the Appendix, which corresponds to the limit of s/p, with $s = \|\beta\|_0$, the true number of nonzero components, as p tends to infinity.

Step 2: Use the effective score function to set b

We choose the scalar $b_{(t)}$ such that the empirical average of the effective score function $G(z;b)$ has slope 1, thus $n^{-1}\sum_{i=1}^n \partial_1 G(z_{i,(t)}; b_{(t)}) = 1$. In the case of a non-differentiable loss function, Bradic [8] proposed to solve $\widehat{\nu}(b_{(t)}) = 1$ with

$$\widehat{\nu}(b_{(t)}) = \frac{b_{(t)}\delta}{\omega}\Big(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{2}\partial v_j\{z_{(t),i}\} + \sum_{l=1}^{L-1}\gamma_l\{\widehat{f}_P(r_{l+1}) - \widehat{f}_P(r_l)\}\Big). \tag{12}$$

Assumption (A3) in the Appendix defines γ_l, r_l and the differentiable functions v_1 and v_2 [See also Condition (R) of 8], $\widehat{f}_P$ is the estimated density of $\text{Prox}(z_{i,(t)}; b_{(t)})$ for $i = 1,\dots,n$.

The derivation of the estimator $\widehat{\nu}(b_{(t)})$, see also Section B.2.4, relies on the limiting behaviour of the system, see Section 4.

For the composite quantile loss, see (4), we clearly see the dependence on the quantiles. The estimator of ν in (12) uses $v_1(z) = 0$ and $v_2(z) = z - u_{\tau_\ell}$ $z \in [u_{\tau_\ell} + bh(\ell-1), u_{\tau_\ell} + bh(\ell)]$, $\ell = 1,\dots,K$, corresponding to the differentiable pieces in (10). The step functions $v_3(z) = bh(\ell)$ when $z \in (u_{\tau_\ell} + bh(\ell), u_{\tau_\ell+1} + bh(\ell))$, $\ell = 0,\dots,K$. Solving for b in the equation $\widehat{\nu}(b) = 1$ is equivalent to solving for b in the following equation,

$$\begin{aligned} \frac{s}{n} \quad = \quad & b\Big[\sum_{k=0}^{K-1} h(k) f_z\{u_{\tau_{k+1}} + bh(k)\} - \sum_{k=1}^{K} h(k) f_z\{u_{\tau_k} + bh(k)\}\Big] \\ & + F_z\{bh(K)\} - F_z\{bh(0)\}, \end{aligned} \tag{13}$$

where F_z is the cumulative distribution function and f_z the density function of the adjusted residuals. In practice, a grid search is performed to approximate

the solution $\widehat{b}_t$. For each b in the grid, we use the empirical cumulative distribution, that is, $\widehat{F}_z(bh(K)) = n^{-1}\sum_{i=1}^n I\{z_{i;(t)} \leq bh(K)\}$. A kernel density estimator of f_z with the Gaussian kernel estimates defined as $\widehat{f}_z\{u_{\tau_k} + bh(k)\} = (nh)^{-1}\sum_{i=1}^n \phi\{(z_{i;(t)} - u_{\tau_k} - bh(k))/h\}$ with ϕ being the standard normal density function. The solution $\widehat{b}_t$ is taken to be the average of the smallest b in the grid that makes the righthand side of (13) smaller than $\frac{s}{n}$ and the next value in the grid.

Step 3: Update the estimator of β

Use the estimated $b_{(t)}$ from the previous step to update the estimate of β to

$$\widehat{\beta}_{(t+1)} = \eta(\widetilde{\beta}_{(t)}; \theta_{(t)}), \text{ where } \widetilde{\beta}_{(t)} = \widehat{\beta}_{(t)} + X^\top G(z_{(t)}; b_{(t)}). \tag{14}$$

The estimator $\widetilde{\beta}_{(t)}$, before applying the soft-thresholding function, is of interest too since it can be interpreted as a debiased estimator [25, 26, 36, 27]; a thorough study of which, however, is beyond the current work.

4. State evolution

Within each iteration step t of the approximate message passing algorithm, state evolution studies the limiting behaviour of the estimators when the sample size goes to infinity. We now define the state evolution parameter $\bar{\zeta}^2_{(t)}$ which is critical for Algorithm 1. We start by defining the empirical version as follows

$$\bar{\zeta}^2_{\text{emp,(t)}} = \frac{1}{n}\sum_{i=1}^n G(z_{i,(t)}; b_{(t)})^2. \tag{15}$$

This quantity is linked to the state evolution recursion which describes the limiting behaviour of large systems, see Theorem 2 in Bayati and Montanari [5] and Lemma 1 in Bradic [8]. It holds that, see Section B.2.3 for details,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n G\big(z_{i,(t)}; b_{(t)}\big)^2 \overset{a.s.}{=} E[G(\varepsilon - \bar{\sigma}_{(t)}Z; b_{(t)})^2] = \bar{\zeta}^2_{(t)}, \tag{16}$$

where $\bar{\zeta}_{(t)}$ is the state evolution parameter for the large system, Z is a random variable with standard normal distribution independent of everything else and $\bar{\sigma}(t)$ is defined in (18).

Due to the symmetry of Z, the state evolution parameter is formally defined as

$$\bar{\zeta}^2_{(t)} = E[G(\varepsilon + \bar{\sigma}_{(t)}Z; b_{(t)})^2]. \tag{17}$$

This definition explicitly features the extra Gaussian component $\bar{\sigma}_{(t)}Z$ in the limiting version, with variance

$$\bar{\sigma}^2_{(t)} = \delta^{-1}E[(\eta(B_0 + \bar{\zeta}_{(t-1)}Z; \theta_{(t-1)}) - B_0)^2] \tag{18}$$

Algorithm 1: RAMP algorithm for a single loss function with tuning parameter α

Function `singleRAMP(`α`):`

Initialization: $\widehat{\beta}_{(0)} \leftarrow 0 \in \mathbb{R}^p$,
iteration index $t \leftarrow 0$, final iteration $t_{\text{final}} \leftarrow 0$,
adjusted residuals $z_{(0)} \leftarrow Y \in \mathbb{R}^n$,
empirical state evolution $\bar{\zeta}^2_{(0)}$ using (15),
tuning parameter of the soft-thresholding function $\theta_{(0)} = \alpha\bar{\zeta}_{(0)}$

while *iteration* $t \leq T$ *and tolerance* $tol > \varepsilon_{\text{tol}}$ **do**

1. *Adjust residuals:* adjust the residuals $z_{(t)} \in \mathbb{R}^n$:

$$z_{(t)} \leftarrow Y - X\widehat{\beta}_{(t)} + \frac{1}{n}G(z_{(t-1)}; b_{(t-1)}) \sum_{j=1}^{p} I\left\{\eta(\widehat{\beta}_{j,(t-1)} + X_{\cdot j}^{\top} G(z_{(t-1)}; b_{(t-1)}); \theta_{(t-1)}) \neq 0\right\}.$$

2. *Effective score:*
   (a) choose the scalar $b_{(t)}$ satisfying
   if G differentiable **then** $1 = \frac{1}{n}\sum_{i=1}^{n} \partial_1 G(z_{i,(t)}; b_{(t)})$
   else $1 = \widehat{\nu}(b_{(t)})$, see (12);
   (b) update the state evolution parameter $\bar{\zeta}^2_{(t)}$ using (15)
   (c) update the tuning parameter $\theta_{(t)} \leftarrow \alpha\bar{\zeta}_{(t)}$.
3. *Estimation:* Update the coefficient estimation

$$\widetilde{\beta}_{(t)} \leftarrow \widehat{\beta}_{(t)} + X^{\top} G(z_{(t)}; b_{(t)}) \text{ and } \widehat{\beta}_{(t+1)} \leftarrow \eta(\widetilde{\beta}_{(t)}; \theta_{(t)}),$$

4. *Adjust iteration index:* $t \leftarrow t+1$; $t_{\text{final}} \leftarrow t$.
5. *Calculate tolerance:* $tol = \|\widehat{\beta}_{(t)} - \widehat{\beta}_{(t-1)}\|^2/p$

end

return $\widehat{\beta} \leftarrow \widehat{\beta}_{(t_{\text{final}})}$, $\widetilde{\beta} \leftarrow \widetilde{\beta}_{t_{\text{final}}}$,
the estimated AMSE$(\widehat{\beta}; \beta)$ *for* $\widehat{\beta}$, *see Theorems 1 and 2.*

with B_0 defined in (A2). To connect the theoretical expression of $\bar{\sigma}^2_{(t)}$ to Algorithm 1, we apply Eq. (3.6) in Bayati and Montanari [5], and Eqs. (7.10) and (7.19) in Bradic [8]. This leads to

$$\delta^{-1} \lim_{p\to\infty} \frac{1}{p}\sum_{j=1}^{p} \{\eta(\widehat{\beta}_{(t),j} + X_{\cdot j}^{\top} G(z_{i,(t)}; b_{(t)}); \theta_{(t)}) - \beta_j\}^2 \overset{a.s.}{=} \bar{\sigma}^2_{(t)}. \tag{19}$$

Note that (19) features the debiased estimator from (14).

We now explain the connection between the estimators that explicitly use an l_1-regularization and the corresponding estimators from the RAMP algorithm.

By applying Theorem 2 of [8], we get the immediate connection between the regularized estimators $\widehat{\beta}_k(\lambda_k)$ for $k = 1, \ldots, K$ and the corresponding estimators obtained by applying the RAMP algorithm with a suitable choice of its regularity parameter α. We explain this below. Since the regularized estima-

tors $\widehat{\beta}_k(\lambda_k)$ for $k = 1, \ldots, K$ are used for $\widehat{\beta}_{\mathrm{MA}}(\lambda)$, (2), the connection between the model-averaged estimators from regularization and from application of the RAMP algorithm, follows immediately from the connections between the K separate estimators. The composite estimator $\widehat{\beta}_{\mathrm{C}}(\lambda)$, (3), is a special case of a model-averaged estimator with $K = 1$, weight equal to one, and loss function $\rho_{\mathrm{C}} = \sum_{k=1}^{K} w_k \rho_k$.

Denote $(\bar{\zeta}^2, b)$ as the fixed point solution when the iteration number $t \to \infty$ of the following equations,

$$\bar{\zeta}_{(t)}^2 = E[G(\varepsilon + \bar{\zeta}_{(t)} Z; b_{(t)})^2] \quad = \quad (\delta/\omega)^2 E[\widetilde{G}(\varepsilon + \bar{\zeta}_{(t)} Z; b_{(t)})^2] \tag{20}$$

$$1 = E[\partial_1 G(\varepsilon + \bar{\zeta}_{(t)} Z; b_{(t)})] \quad = \quad (\delta/\omega) E[\partial_1 \widetilde{G}(\varepsilon + \bar{\zeta}_{(t)} Z_k; b_{(t)})]. \tag{21}$$

Note that (20) is the state evolution recursion for the large system while in (21) the first equality is the population version of the requirement in step 2 in Algorithm 1 which states that $n^{-1} \sum_{i=1}^{n} \partial_1 G(z_{i,(t)}; b_{(t)}) = 1$. The second equalities of both (20) and (21) follow by using the definition of G in (11), with $\widetilde{G}$ being defined in (8).

Then, under assumptions (A1)–(A5) (see the Appendix), for the RAMP algorithm with $\theta = \alpha\bar{\zeta}$, where the tuning parameter $\alpha > 0$ (which motivates the definition of $\theta_{(t)} = \alpha\bar{\zeta}_{(t)}$ in Algorithm 1), and for the l_1-optimization with

$$\lambda = \frac{\alpha\bar{\zeta}}{b\delta} P(|B_0 + \bar{\zeta} Z| \geq \alpha\bar{\zeta}), \tag{22}$$

it follows by Theorem 2 of Bradic [8] that

$$\lim_{t\to\infty} \lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \{\widehat{\beta}_{\mathrm{C},j}(\lambda) - \widehat{\beta}_{\mathrm{C},(t),j}(\alpha)\}^2 = 0 \text{ a.s.} \tag{23}$$

The convergence in (23) explicitly connects the two composite estimators: one estimator uses an explicit l_1-regularization as in (3), the other estimator is obtained via the RAMP algorithm. Similar results can be found in Huang [23, Theorem 2.2] for a generalized AMP algorithm with non-negative convex loss function, and in Bayati and Montanari [6, Theorem 1.8] for the AMP algorithm with least squares loss function.

For the model averaging estimator we use such an equivalence for estimation with each separate loss function ρ_k, $k = 1, \ldots, K$. When using explicit l_1-regularization as in (1) with the regularization constants λ_k matching as in (22) the values $\theta_k = \alpha_k \bar{\zeta}_k$, for $k = 1, \ldots, K$ that are used in the RAMP algorithm, again Theorem 2 of Bradic [8] applies. It hence follows that

$$\lim_{t\to\infty} \lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \{\widehat{\beta}_{\mathrm{MA},j}(\lambda) - \widehat{\beta}_{\mathrm{MA},(t),j}(\alpha)\}^2 = 0, \text{a.s.}$$

5. Theoretical contributions

This section contains detailed theoretical developments for the composite as well as the model-averaged AMP estimators in high-dimensions.

5.1. Asymptotic mean squared error

We first define the asymptotic mean squared error as

$$\text{AMSE}(\widehat{\beta}_{(t)}, \beta) = \lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} (\widehat{\beta}_{(t),j} - \beta_j)^2. \tag{24}$$

Combining (19) and (16), we obtain

$$\begin{aligned} \text{AMSE}(\widehat{\beta}_{(t)}, \beta) &= \lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \left(\eta(\widetilde{\beta}_{(t-1),j}; \theta_{(t-1)}) - \beta_j\right)^2 \\ &\overset{a.s.}{=} E[\{\eta(B_0 - \bar{\zeta}_{(t-1)} Z; \theta_{(t-1)}) - B_0\}^2], \end{aligned} \tag{25}$$

which corresponds to Eq. (3.4) in Bradic [8] with $\widetilde{\beta}_{(t),j}$ the debiased estimator in (14).

In Section 4, we defined the empirical state evolution parameter $\bar{\zeta}^2_{\text{emp},(t)}$, and we described the connections between the empirical updates in Algorithm 1 and the theoretical state evolution recursion, which connects to the theoretical expression of the AMSE. While Algorithm 1 and the theoretical state evolution recursion involve only a single estimator, the model-averaged estimator, on the other hand, is the weighted sum of K such estimators $\widehat{\beta}_k$, $k = 1, \ldots, K$, each obtained by Algorithm 1. Consequently, the estimators $\widehat{\beta}_k$, $k = 1, \ldots, K$ are correlated.

Lemma 1 extends Theorem 2 in Bayati and Montanari [5] and (3.16) in Lemma 1(b) in Bayati and Montanari [5] to the almost sure convergence of the product for any two recursions among K paralleled recursions. All proofs are contained in Appendix B.2.

Lemma 1. *Let the sequences of design matrices $\{X(p)\}$, coefficient vectors $\{\beta(p)\}$, error vectors $\{\varepsilon(p)\}$, initial condition vectors $\{q_0(p)\}$ be the common sequences for K recursions satisfying assumptions (A1)–(A4) in the Appendix. Let $\{\bar{\sigma}^2_{k,(t)}, \bar{\zeta}^2_{k,(t)}\}$ be defined uniquely by the recursions in (17) and (18). These are the state evolution parameters for the kth estimation with initialization $\bar{\sigma}^2_{k,(0)} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} q^2_{(0),i}/\delta$. Then Lemma 1 in Bayati and Montanari [5] holds individually for each of the K recursions; additionally, for all pseudo-Lipschitz functions $\tilde{\psi}_c : \mathbb{R}^{t+2} \to \mathbb{R}$ of order κ_c for some $1 \le \kappa_c \le \kappa/2$ with κ as in (A4) and t a natural number larger than or equal to 0,*

$$\lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \tilde{\psi}_c(h_{k_1,(1),j}, \ldots, h_{k_1,(t+1),j}, \beta_j) \tilde{\psi}_c(h_{k_2,(1),j}, \ldots, h_{k_2,(t+1),j}, \beta_j) \overset{a.s.}{=}$$

$$E[\tilde{\psi}_c(\bar{\zeta}_{k_1,(0)} Z_{k_1,(0)}, \ldots, \bar{\zeta}_{k_1,(t)} Z_{k_1,(t)}, B_0) \tilde{\psi}_c(\bar{\zeta}_{k_2,(0)} Z_{k_2,(0)}, \ldots, \bar{\zeta}_{k_2,(t)} Z_{k_2,(t)}, B_0)]$$

where $(Z_{k,(0)}, \ldots, Z_{k,(t)}) \sim \mathcal{N}(0, I_{t+1})$, $k = k_1, k_2$, is a $(t+1)$-dimensional zero-mean multivariate standard normal vector independent of B_0, ε; at iteration t, $(Z_{k_1,(t)}, Z_{k_2,(t)})$ is a bivariate standard normal vector with covariance not necessarily equal to zero.

Note that Algorithm 1 belongs to the general recursion in Bayati and Montanari [5], the initial condition takes $q_{(0)} = -\beta$ and the kth estimator calculated by Algorithm 1 takes $h_{k,(t+1)} = \beta - X^\top G(z_{k,(t)}; b_{k,(t)}) - \beta_{k,(t)}$.

We obtain at iteration t, for $k_1, k_2 \in \{1, \ldots, K\}$,

$$\lim_{p\to\infty} \frac{1}{p}\sum_{j=1}^{p}(\widehat{\beta}_{k_1,(t),j} - \beta_j)(\widehat{\beta}_{k_2,(t),j} - \beta_j)$$
$$\stackrel{a.s.}{=} E\Big[\prod_{r=1}^{2}\big\{\eta(B_0 + \bar{\zeta}_{k_r,(t-1)}Z_{k_r}; \theta_{k_r,(t-1)}) - B_0\big\}\Big],$$

where Z_{k_1} and Z_{k_2} are possibly dependent standard normal random variables.

Since the estimators $\widehat{\beta}_{k_r}$, $r = 1, 2$ use the same design matrix, a correlation between Z_{k_1} and Z_{k_2} exists (see Corollary 2) and contributes to the correlation between $\widehat{\beta}_{k_1}$ and $\widehat{\beta}_{k_2}$. Using Lemma 1, we obtain the theoretical AMSE for the regularized model-averaged estimator.

Theorem 1. *Assume assumptions (A1)–(A5) in the Appendix. At Algorithm 1's iteration step t for the estimator $\widehat{\beta}_{k,(t)}$, for each $k = 1, \ldots, K$, and for a weight vector $w = (w_1, \ldots, w_K)^\top$, the model-averaged estimator $\widehat{\beta}_{\mathrm{MA},(t)} = \sum_{k=1}^{K} w_k \widehat{\beta}_{k,(t)}$ has asymptotic mean squared error*

$$\begin{aligned}\mathrm{AMSE}(\widehat{\beta}_{\mathrm{MA},(t)}, \beta) &= \lim_{p\to\infty}\frac{1}{p}\sum_{j=1}^{p}(\widehat{\beta}_{\mathrm{MA},(t),j} - \beta_j)^2 \\ &= \lim_{p\to\infty} w^\top \Sigma_{0,(t)}(p) w \stackrel{a.s.}{=} w^\top \Sigma_{(t)} w \end{aligned} \tag{26}$$

where $\Sigma_{0,(t)}(p)$ is a $K \times K$ matrix with (k_1, k_2)th component

$$(\Sigma_{0,(t)})_{(k_1,k_2)}(p) = p^{-1}\sum_{j=1}^{p}(\widehat{\beta}_{k_1,(t),j} - \beta_j)(\widehat{\beta}_{k_2,(t),j} - \beta_j); \tag{27}$$

similarly, $\Sigma_{(t)}$ is a $K \times K$ matrix with the (k_1, k_2)th component

$$(\Sigma_{(t)})_{(k_1,k_2)} = E\Big[\prod_{r=1}^{2}\{\eta(B_0 + \bar{\zeta}_{k_r,(t-1)}Z_{k_r}; \theta_{k_r,(t-1)}) - B_0\}\Big].$$

Since the AMSE expression of the regularized model-averaged estimator is a quadratic function of the weight vector w, Corollary 1 readily provides the lower bound of the AMSE as well as the weight vector reaching this lower bound. The K-vector $\mathbf{1}_K$ consists of ones only.

Corollary 1. *Constraining the weights to sum to one, the lower bound of the AMSE at iteration t for the model-averaged estimator as in (26) is equal to $\big(\mathbf{1}_K^\top(\Sigma_{(t)})^{-1}\mathbf{1}_K\big)^{-1}$. This lower bound is attained for the theoretical optimal weights $w_{\mathrm{MA}} = \big(\Sigma_{(t)}\big)^{-1}\mathbf{1}_K\big(\mathbf{1}_K^\top(\Sigma_{(t)})^{-1}\mathbf{1}_K\big)^{-1}$.*

5.2. Estimating optimal weights

The expression of the core matrix $\Sigma_{(t)}$, which is the limit matrix for $n, p \to \infty$, contains the random variable B_0 which satisfies assumption (A2) in the Appendix. Likewise, $\Sigma_{0,(t)}$ which is the limit matrix for fixed p while $n \to \infty$, contains the true coefficient β (see (27)). In practice, neither the true coefficient vector β nor the random variable B_0 is known. To make practical use of the expressions of the AMSE, we derive an estimator of the matrix $\Sigma_{0,(t)}$ relying only on sequences generated in Algorithm 1.

5.2.1. Model-averaged estimator

Before deriving the estimator of the AMSE for the model-averaged estimator, we first define $\bar{\zeta}_{\text{emp},(k_1,k_2),(t)}$ which is an estimator of the parameter $\bar{\zeta}_{(k_1,k_2),(t)}$, a quantity similar to the state evolution parameter $\bar{\zeta}^2_{k,(t)}$, which records the covariance between the unbiased sequences $\widetilde{\beta}_{k_1,(t)}$ and $\widetilde{\beta}_{k_2,(t)}$ generated in (14) in Algorithm 1 when $p \to \infty$. Since model-averaged estimators combine estimators constructed from the same data into one weighted average, the correlation between $\widehat{\beta}_{k_1}$ and $\widehat{\beta}_{k_2}$ is needed to understand the AMSE of the model-averaged estimator.

Notice that the unbiasedness of the sequence $\widetilde{\beta}_{k,(t)}$ follows from the argument that $\widetilde{\beta}_{k,j,(t)}$ converges weakly to $B_0 + \bar{\zeta}_{k,(t)} Z_k$ when $p \to \infty$, while assigning $1/p$ point mass to each entry of the vector. Then, $\widetilde{\beta}_{k,j,(t)} | (B_0 = \beta_j) \sim N(\beta_j, \bar{\zeta}^2_{k,(t)})$ for large p, indicating that $\widetilde{\beta}_{k,j,(t)}$ centers at β_j ensuring the unbiasedness. Moreover, the vector $\widetilde{\beta}_{k,(t)}$ has Gaussian distribution. By applying the soft-thresholding function η on $\widetilde{\beta}_{k,j,(t)}$ in Lemma 4, we avoid the usage of the true coefficient vector β in $\Sigma_{0,(t)}$ resulting in a Stein-type risk estimator requiring only observables from Algorithm 1. A Gaussianity argument has also been used in Bayati and Montanari [6], Bayati et al. [4], Mousavi et al. [31, 32] to derive a similar Stein-type risk estimator for the Lasso. Details can be found in Section B.2.8. The bias of the estimator $\widehat{\beta}_{k,(t)}$ is introduced in Algorithm 1 by applying the soft-thresholding function componentwise to the unbiased sequence $\widetilde{\beta}_{k,(t)}$.

Corollary 2. *Assume assumptions (A1)–(A5) in the Appendix. For any* $k_1, k_2 = 1, \ldots, K$*, at iteration* t*,*

$$\lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} (\widetilde{\beta}_{k_1,(t),j} - \beta)(\widetilde{\beta}_{k_2,(t),j} - \beta) \overset{a.s.}{=} \bar{\zeta}_{k_1,(t)} \bar{\zeta}_{k_2,(t)} \text{Cov}(Z_{k_1}, Z_{k_2}),$$

where $\bar{\zeta}_{k,(t)}, k = k_1, k_2$ *is the state evolution parameter corresponding to* $\widehat{\beta}_k$*.*

Corollary 2 indicates both the existence and a feasible estimation of the covariance between Z_{k_1} and Z_{k_2}. As an estimator for

$$\bar{\zeta}_{(k_1,k_2),(t)} = \bar{\zeta}_{k_1,(t)} \bar{\zeta}_{k_2,(t)} \text{Cov}(Z_{k_1}, Z_{k_2})$$

we define

$$\bar{\zeta}_{\text{emp},(k_1,k_2),(t)} = \frac{1}{p-1}\sum_{j=1}^{p}\big(\widetilde{\beta}_{k_1,(t),j} - \frac{1}{p}\sum_{j=1}^{p}\widetilde{\beta}_{k_1,(t),j}\big)\big(\widetilde{\beta}_{k_2,(t),j} - \frac{1}{p}\sum_{j=1}^{p}\widetilde{\beta}_{k_2,(t),j}\big). \tag{28}$$

We now state an unbiased estimator for the matrix $\Sigma_{0,(t)}$, and a consistent estimator for the matrix $\Sigma_{(t)}$ upon convergence of Algorithm 1.

Theorem 2. *Assume assumptions (A1)–(A5) in the Appendix, and that the state evolution parameter in (15) satisfies* $\bar{\zeta}^2_{\text{emp},(t)} - \bar{\zeta}^2_{\text{emp},(t-1)} = o(1)$. *For any* $k_1, k_2 = 1, \ldots, K$, *define*

$$\begin{aligned}(\widehat{\Sigma}_0)_{(k_1,k_2),(t)} = &-\bar{\zeta}_{\text{emp},(k_1,k_2),(t-1)} \\ &+ \frac{1}{p}\sum_{j=1}^{p}\prod_{r=1}^{2}\big\{\eta(\widetilde{\beta}_{k_r,(t-1),j};\theta_{k_r,(t-1)}) - \widetilde{\beta}_{k_r,(t-1),j}\big\} \\ &+ \bar{\zeta}_{\text{emp},(k_1,k_2),(t-1)}\cdot\frac{1}{p}\sum_{j=1}^{p}\sum_{r=1}^{2} I\{|\widetilde{\beta}_{k_r,(t-1),j}| \geq \theta_{k_r,(t-1)}\},\end{aligned}$$

with $\widetilde{\beta}_{k_1,(t-1)}$, $\widetilde{\beta}_{k_2,(t-1)}$ *in (14) Then,* $(\widehat{\Sigma}_0)_{(k_1,k_2),(t)}$ *is an unbiased estimator of component* (k_1, k_2) *of the matrix* $\Sigma_{0,(t)}$ *at iteration* t. *Further,* $(\widehat{\Sigma}_0)_{(k_1,k_2),(t)}$ *is a consistent estimator of the matrix* $\Sigma_{(t)}$ *in Theorem 1.*

This new estimator can be compared to the estimator used in Bayati et al. [4, Def. 2] and Mousavi et al. [32, Eq. (9)] for the case of a single estimator ($K = 1$). The proof of Theorem 2, see Section B.2.8 uses Stein's lemma (see Lemma 4) to estimate the covariances that appear in the matrix $\Sigma_{0,(t)}$. The soft-thresholding function $\eta(\cdot;\theta)$ that appears in the estimator $\widehat{\Sigma}_{0,(t)}$ links the estimator $\widehat{\beta}_k$ to the estimator $\widetilde{\beta}_k$. The proof also uses the joint asymptotic normality of the jth components of the vectors $\widetilde{\beta}_{k_1}$ and $\widetilde{\beta}_{k_2}$. The obtained estimator for $\Sigma_{0,(t)}$ in the case $K > 1$ is nontrivial and new to the literature.

Estimated AMSE-optimal weights for the model-averaged estimator are obtained by using the estimator $\widehat{\Sigma}_{0,(t)}$ at the final iteration in Theorem 2. In combination with the sum-to-one constrained weights this gives the estimated weights that minimize the estimated AMSE for the model-averaged estimator

$$\widehat{w}_{\text{MA}} = \big(\widehat{\Sigma}_{(t)}\big)^{-1}\mathbf{1}_K\big(\mathbf{1}_K^{\top}(\widehat{\Sigma}_{(t)})^{-1}\mathbf{1}_K\big)^{-1}.$$

When additional constraints such as positivity are needed, the optimal weights no longer have an explicit formula, but they are straightforward to compute, see (31).

5.2.2. Composite estimator

The AMSE of a composite estimator can be obtained from Theorem 1 as a special case, treating the composite loss function as a single loss function with

weight one, thus $\rho_{\mathrm{C}} = \sum_{k=1}^{K} w_k \rho_k$ as in (3). At iteration t,

$$\Sigma_{(t)} = E[\{\eta(B_0 + \bar{\zeta}_{(t-1)} Z; \theta_{(t-1)}) - B_0\}]^2, \text{ and } \Sigma_{0,(t)} = p^{-1} \sum_{j=1}^{p} (\widehat{\beta}_{(t),j} - \beta_j)^2.$$

The matrices $\Sigma_{(t)}, \Sigma_{0,(t)}$ are now real numbers and coincide with the AMSE of the estimator in (25). We obtain the corresponding estimator for the AMSE

$$\begin{aligned} \widehat{\Sigma}_{\mathrm{C},0} &= \widehat{\mathrm{AMSE}}_{\mathrm{C}}(w) \qquad (29) \\ &= -\bar{\zeta}^2_{\mathrm{emp}}(w) + \frac{1}{p} \sum_{j=1}^{p} \Big[\{\eta(\widetilde{\beta}_j(w); \theta) - \widetilde{\beta}_j(w)\}^2 + 2\bar{\zeta}^2_{\mathrm{emp}}(w) I\{|\widetilde{\beta}_j(w)| \geq \theta\} \Big]. \end{aligned}$$

For the single loss function, ρ_{C}, the estimator of AMSE in (29) can be compared to the Stein-type estimator that has been obtained in Definition 2 in Bayati et al. [4] for the AMP algorithm using the least squares loss, which is a particular case of Algorithm 1.

Finding optimal weights for the composite estimator is complicated. Indeed, while the model-averaged estimator has an AMSE, which is a quadratic function in the weights, see (26), the composite estimator and its AMSE depend on the weights in a highly nonlinear fashion; e.g., observe that the soft-thresholding function in (29) depends on w.

Therefore, optimization of the estimated AMSE with respect to the weights proceeds numerically;

$$w_{\mathrm{C},1} = \arg\min_{w} \widehat{\mathrm{AMSE}}_{\mathrm{C}}(w).$$

See Section 6.2 for more details.

5.3. The case of dense (non-sparse) linear models with $n/p \to \delta \geq 1$: asymptotic variance optimality

Donoho and Montanari [14] and El Karoui et al. [16] showed that the asymptotic variance of the M-estimators in the case where $p, n \to \infty$ and $n/p \to \delta \in [1, \infty)$ contains an extra Gaussian component. Recently, Lei et al. [30] obtained the coordinate-wise asymptotic normality of regression M-estimators in the moderate p/n regime for a fixed design matrix. In the sparse high-dimensional linear model setting where $\delta \in (0, 1)$, it was shown that the sequence $\widetilde{\beta}_{(t)}$ in (14) follows for the Lasso estimator [4] a similar normal distribution with the variance containing an extra Gaussian component. The above-mentioned literature focuses on the asymptotics for a single M-estimator; we extend the asymptotic result to the model-averaged estimator. In this section, we only characterize the asymptotic variance of the model-averaged estimator for dense linear models with $n/p \to \delta \geq 1$, following Donoho and Montanari [14].

Under the dense linear model with $n \geq p$, the soft-thresholding function $\eta(\cdot; \theta)$ is replaced by the identity function and the ratio $\omega = P(B_0 \neq 0) = 1$.

Consequently, Algorithm 1 is adjusted to estimate

$$\widehat{\beta}_k = \arg\min_{\beta\in\mathbb{R}^p} \Big\{ \sum_{i=1}^{n} \rho_k(Y_i - X_{i\cdot}\beta) \Big\},$$

where β is dense. It is trivial to show that Algorithm 1 still belongs to the general recursion in Bayati and Montanari [5]. For a single estimator at iteration t denoted as $\widehat{\beta}_{k,(t)}$, the two state evolution parameters $\bar{\zeta}^2_{k,(t)}$ and $\bar{\sigma}^2_{k,(t)}$ coincide and Theorem 4.1 in Donoho and Montanari [14] holds.

Theorem 3. *Assume assumptions (A1)–(A5) in the Appendix. Let $n/p \to \delta \geq 1$ when $n, p \to \infty$. For the asymptotic variance of the model-averaged estimator $\widehat{\beta}_{\mathrm{MA}}$ holds that*

$$\lim_{n,p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \mathrm{Var}(\widehat{\beta}_{\mathrm{MA},j}) \overset{a.s}{=} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \mathrm{Cov}(Z_{k_1}, Z_{k_2}) \prod_{r=1}^{2} \{ w_{k_r} V^{1/2}(\widetilde{G}_{k_r}; \widetilde{F}_{k_r}) \} \tag{30}$$

for differentiable $\widetilde{G}$, where $V(\widetilde{G}_k; F_k) = (\int \widetilde{G}_k^2 dF_k)/(\int \partial_1 \widetilde{G}_k dF_k)^2$ denotes the Huber asymptotic variance formula for M-estimators. For non-differentiable $\widetilde{G}$, we replace V in (30) by the consistent estimator $\widehat{V}(\widetilde{G}_k; F_k) = (\int \widetilde{G}_k^2 dF_k)/\widehat{\nu}(b_k)^2$. The extra Gaussian component is identified in the convolution of the regression noise distribution and a Gaussian distribution: $\widetilde{F}_k = F_\varepsilon \star N(0, \bar{\zeta}_k^2)$.

Recall that the componentwise empirical distribution of $\widehat{\beta}_k(p)$, when $p \to \infty$, converges weakly to $B_0 + \bar{\zeta}_k Z_k$ following Bayati and Montanari [5] and Donoho and Montanari [14]. Then for large p, while the iteration $t \to \infty$, $\widehat{\beta}_k(p) \sim N(\beta, \bar{\zeta}_k^2 I_p)$ [14, 31] with I_p the $p \times p$ identity matrix. The (k_1, k_2)th component of the empirical variance matrix is denoted by $\big(\Sigma_{\mathrm{emp}}(p)\big)_{(k_1,k_2)} = p^{-1} \sum_{j=1}^{p} (\widehat{\beta}_{k_1,j} - \beta_j)(\widehat{\beta}_{k_2,j} - \beta_j)$, which is unbiasedly estimated by

$$\big(\widehat{\Sigma}_{\mathrm{emp}}(p)\big)_{(k_1,k_2)} = \sum_{j=1}^{p} (\widehat{\beta}_{k_1,j} - \frac{1}{p}\sum_{j=1}^{p} \widehat{\beta}_{k_1,j})(\widehat{\beta}_{k_2,j} - \frac{1}{p}\sum_{j=1}^{p} \widehat{\beta}_{k_2,j})/(p-1).$$

Note that this estimator coincides with (28) for the special case that $n \geq p$ and the soft-thresholding function is replaced by the identity function.

6. Computational details

6.1. Regularized model-averaged quantile estimation

The estimation of the quantile $u_{\tau_k} = F_\varepsilon^{-1}(\tau_k)$ follows a two-step procedure.

1. Obtain an initial slope estimate $\widehat{\beta}_{\mathrm{init}}$ and calculate the residuals. Example initial slope estimates are the Lasso or regularized quantile estimation with a single quantile level.

2. For $k = 1, \ldots, K$, estimate the quantile intercepts $\widehat{u}_{\tau_k}$ by taking the corresponding $\tau_k \times 100\%$ quantile of the residuals from the previous step.

The regularized model-averaged estimator is obtained by averaging over K paralleled estimators. See Algorithm 2 for the pseudo-code, of which the core is Algorithm 1; there the effective score function G is that of a single quantile loss function with $K = 1$, see also Example 2 in Bradic [8]. In our numerical work, the upper bound for the number of iteration steps T is set to be 50 in both the simulation and the data analysis sections. With $K = 1$, this algorithm applies to the regularized composite estimator too.

Algorithm 2: RAMP algorithm for K paralleled estimations with tuned α's

Function `KparallelRAMP(`K`)`**:**
  for k *in* $\{1, \ldots, K\}$ **do**
    Initialization: $\widehat{\beta}(\alpha_{k,\text{opt}}) \leftarrow 0 \in \mathbb{R}^p$, $\widetilde{\beta}_k(\alpha_{k,\text{opt}}) \leftarrow 0 \in \mathbb{R}^p$ and $\text{AMSE}(\widehat{\beta}_k(\alpha_{k,\text{opt}}); \beta) \leftarrow 0$
    for α in candidate set $\mathcal{A}$ **do**
      `singleRAMP(`α`)` in Algorithm 1
      if $\text{AMSE}(\widehat{\beta}_k(\alpha); \beta) \leq \text{AMSE}(\widehat{\beta}_k(\alpha_{k,\text{opt}}); \beta)$ **then**
        $\widehat{\beta}_k(\alpha_{k,\text{opt}}) \leftarrow \widehat{\beta}_k(\alpha)$, $\widetilde{\beta}_k(\alpha_{k,\text{opt}}) \leftarrow \widetilde{\beta}_k(\alpha)$, $\text{AMSE}(\widehat{\beta}_k(\alpha_{k,\text{opt}}); \beta) \leftarrow \text{AMSE}(\widehat{\beta}_k(\alpha); \beta)$
      end
    end
  end
  return $(\widehat{\beta}_1(\alpha_{1,\text{opt}}), \ldots, \widehat{\beta}_K(\alpha_{K,\text{opt}}))$, $(\widetilde{\beta}_1(\alpha_{1,\text{opt}}), \ldots, \widetilde{\beta}_K(\alpha_{K,\text{opt}}))$, *and* $(\text{AMSE}(\widehat{\beta}_1(\alpha_{1,\text{opt}}); \beta), \ldots, \text{AMSE}(\widehat{\beta}_K(\alpha_{K,\text{opt}}); \beta))$

AMSE refers to the estimated version. The $\widetilde{\beta}_k$s are recorded for calculating the weights in Corollary 1.

The tuning parameter α of Algorithm 2 controls the sparsity of the estimators and requires a tuning procedure to choose it in practice. In Section 7, we consider the one dimensional Golden-section search algorithm [28] for tuning the value α in the range $[\alpha_{\min}, \alpha_{\max}]$ that minimize the estimated MSE of $\widehat{\beta}$ using the estimator derived in Section 5.2. The upper bound $\alpha_{\max}$ is chosen to be 2.3 for the simulations and data analysis. The lower bound $\alpha_{\min}$ in the data analysis follows the lower bound in Proposition 9.2 in Eldar and Kutyniok [17] and is chosen to be the unique non-negative solution to the equation $(1+\alpha^2)\Phi(-\alpha) - \alpha\phi(\alpha) = \delta/2$, where $\phi(x)$ and $\Phi(x)$ denote the p.d.f and c.d.f of the standard normal distribution respectively. In the simulation study, the lower bound $\alpha_{\min}$ is chosen to be 1.3 for computational efficiency purposes, since the optimal tuning parameter for those settings was rarely less than 1.3.

6.2. Optimization of the weights

To obtain the regularized model-averaged quantile estimations with the AMSE-type weight derived in Corollary 1, we follow the following procedure:

1. Obtain optimally tuned paralleled regularized quantile estimates, see (14), $(\widehat{\beta}_{\tau_1}(\alpha_{1,\mathrm{opt}}), \ldots, \widehat{\beta}_{\tau_K}(\alpha_{K,\mathrm{opt}}))$, and the additional K estimates $(\widetilde{\beta}_{\tau_1}(\alpha_{1,\mathrm{opt}}), \ldots, \widetilde{\beta}_{\tau_K}(\alpha_{K,\mathrm{opt}}))$ from the converged iterations using Algorithm 2.
2. Estimate the AMSE-type optimal weight $\widehat{w}_{\mathrm{MA},1}$ with constraints by

$$\widehat{w}_{\mathrm{MA},1} = \arg \min_{w \geq 0, \mathbf{1}_K^\top w = 1} w^\top \widehat{\Sigma}_0 w \tag{31}$$

   where the $K \times K$ matrix $\widehat{\Sigma}_0$ is the consistent estimator of Theorem 2.
3. Obtain the regularized model-averaged estimate (2) with the estimated AMSE-type optimal weight.

It is worth mentioning that $\widehat{w}_{\mathrm{MA},1}$ is a constrained version of w_{MA} attaining the lower bound of the AMSE in Corollary 1. $\widehat{w}_{\mathrm{MA},1}$ focuses on approximating the lower bound of the AMSE of the sparse coefficient vector β without assuming that the nonzero entries are selected perfectly; whereas another type of weight choice derived in Bradic et al. [9], Bloznelis et al. [7] aims at the lower bound of the variance of the nonzero part of β by imposing the perfect selection assumption. A numerical comparison of these two types of weight choices is presented in Section 7.

To equip the regularized composite quantile estimator with the weight minimizing the estimated AMSE, we cannot make use of an analytical solution to the weight minimization problem. Instead, a numerical search for a better weight choice in the neighbourhood of an initial weight proposal is employed. The basic idea is that the estimator $\widehat{\beta}_{\mathrm{C}}(w_{\mathrm{C}})$ is treated as a function of the weights. We propose a collection of candidate weight vectors in the neighborhood of the weight chosen in the previous step. The weight for $\widehat{\beta}_{\mathrm{C}}(w_{\mathrm{C}})$ is updated in each step by the one having the lowest estimated AMSE, i.e.,

$$w_{\mathrm{C},1} = \arg \min_{w_{\mathrm{cand}}} \widehat{\mathrm{AMSE}}(\widehat{\beta}_{\mathrm{C}}(\alpha_{\mathrm{opt}}; w_{\mathrm{cand}}); \beta).$$

A more detailed search procedure is as follows.

1. Propose a reasonable initial weight vector $w_{\mathrm{C,init}}$, e.g. the vector of equal weights; estimate $\widehat{\beta}_{\mathrm{C}}$ at the initial weight $w_{\mathrm{C,init}}$ and obtain the estimate of $\mathrm{AMSE}\big(\widehat{\beta}_{\mathrm{C}}(\alpha_{\mathrm{opt}}; w_{\mathrm{C,init}}); \beta\big)$.
2. Initiate the searching step calculator $s_{\mathcal{D}} = 0$, the candidate optimal weight $w_{\mathrm{C},1} = w_{\mathrm{C,init}}$, and the corresponding candidate minimum MSE

$$\mathrm{AMSE}(w_{\mathrm{C},1}) = \mathrm{AMSE}\big(\widehat{\beta}_{\mathrm{C}}(\alpha_{\mathrm{opt}}; w_{\mathrm{C,init}}); \beta\big)$$

   estimated by the AMSE estimator in Theorem 2 for $K = 1$, the collection of the used weight vectors $\mathcal{V}_w = \{w_{\mathrm{C},1}\}$.
3. Propose a set of candidate weight vectors $\mathcal{V}_{w_{\mathrm{cand}}}$. This is to exclude those recorded in the collection of the used weight vectors $\mathcal{V}_w$. In addition, $\mathcal{V}_{w_{\mathrm{cand}}}$ should be in the neighborhood of the current optimal weight $w_{\mathrm{C},1}$. Rules of proposing candidate weight vectors are user-decided; here, we consider a $(K-1)$-dimensional grid search centering at $w_{\mathrm{C},1}$.

4. Obtain the regularized composite quantile estimates at all candidate weight vectors in $\mathcal{V}_{w_{\text{cand}}}$ with Algorithm 2. Update the used weight vector collection $\mathcal{V}_w$, increase the counter $s_{\mathcal{V}} = s_{\mathcal{V}} + 1$, update the candidate optimal weight $w_{\text{C},1}$ by the weight with the lowest estimated AMSE in $\mathcal{V}_w = \{w_{\text{C},1}\}$, and update the candidate minimum AMSE value $\text{AMSE}(w_{\text{C},1})$.
5. Stop the iteration if the searching step calculator $s_{\mathcal{V}} > S_{\mathcal{V}}$ or the candidate weight vector collection $\mathcal{V}_{w_{\text{cand}}} = \emptyset$; otherwise repeat steps 3 and 4.

The pseudocode of the search procedure is stated in Algorithm 3.

Algorithm 3: Weight search for regularized composite estimator

Function `Weight Search`**:**
  Initialization: Better weight recorder $w_{\text{C},1} \leftarrow w_{\text{C,init}}$, step calculator $s_{\mathcal{V}} \leftarrow 0$, MSE recorder $\text{AMSE}(w_{\text{C},1}) \leftarrow \widehat{\text{AMSE}}\big(\widehat{\beta}_{\text{C}}(\alpha_{\text{opt}}; w_{\text{C,init}}); \beta\big)$, and the collection of the used weight vectors $\mathcal{V}_w = \{w_{\text{C},1}\}$.
  while *searching step* $s_{\mathcal{V}} \le S_{\mathcal{V}}$ *or candidate weight collection* $\mathcal{V}_{w_{\text{cand}}} = \emptyset$ **do**
    1. Propose a new $\mathcal{V}_{w_{\text{cand}}}$ in the neighbourhood of $w_{\text{C},1}$. Rules of proposing candidate weight vectors are user-decided; here, we consider a $(K-1)$-dimensional grid search centering at $w_{\text{C},1}$.
    2. **for** w_{cand} *in* $\mathcal{V}_{w_{\text{cand}}} \cap \mathcal{V}_w^{\complement}$ **do**
      Estimate $\widehat{\beta}_{\text{C}}(\alpha_{\text{opt}}; w_{\text{cand}})$ and $\widehat{\text{AMSE}}\big(\widehat{\beta}_{\text{C}}(\alpha_{\text{opt}}; w_{\text{cand}}); \beta\big)$
      if $\widehat{\text{AMSE}}\big(\widehat{\beta}_{\text{C}}(\alpha_{\text{opt}}; w_{\text{cand}}); \beta\big) < \text{AMSE}(\text{w}_{\text{C},1})$ **then**
        $w_{\text{C},1} \leftarrow w_{\text{cand}}$, $\text{AMSE}(w_{\text{C},1}) \leftarrow \widehat{\text{AMSE}}\big(\widehat{\beta}_{\text{C}}(\alpha_{\text{opt}}; w_{\text{cand}}); \beta\big)$
      end
    end
    3. Update $s_{\mathcal{V}} = s_{\mathcal{V}} + 1$.
  end
  return $w_{\text{C},1}$, $\widehat{\beta}(\alpha_{\text{opt}}, w_{\text{C},1})$, *and* $\widehat{\text{AMSE}}\big(\widehat{\beta}_{\text{C}}(\alpha_{\text{opt}}; w_{\text{cand}}); \beta\big)$

A possible initial weight vector $w_{\text{C,init}}$ is the vector of equal weights or the weight proposed in Bradic et al. [9]; $\widehat{\beta}_{\text{C}}$ is estimated by Algorithm 2, and $\text{AMSE}(\widehat{\beta}_{\text{C}}; \beta)$ is estimated by (29).

7. Numerical results

7.1. Simulation study

In this section, we consider the following setup under the high-dimensional linear model setting.

1. Fix the dimension $p = 500$, the sample size $n = 250$, the ratio $\delta = 0.5$. The number of non-zero components s is taken to be 5 for the high-sparsity setting and 50 for the medium-sparsity setting; the non-zero part is generated from the Dirac distribution with a point mass equally distributed on -1 and 1, or a standard normal distribution.
2. In each repetition, we generate a new dataset by randomly generating a sensing matrix X, a coefficient vector β, and an error vector ε. The components of the sensing matrix X are independent and generated from $N(0, 1/250)$.

3. As error distributions, we take the standard normal $N(0,1)$, student-t with degrees of freedom 3, and the mixture of normal distributions $0.5N(0,1)+0.5N(5,9)$; errors generated in Step 2 are centered and rescaled to have standard deviation 0.2.

The objective is to compare the performance of the regularized model-averaged estimator and the composite estimator with different weights, with emphasis on the weights where the selection uncertainty is taken into account. The simulation is repeated to get 500 estimates for each setup. For both the regularized model-averaged and composite quantile estimator, the weights considered are (1) the estimated AMSE-type weights (i.e. $w_{\mathrm{MA},1}$ for the model-averaged quantile estimator and $w_{\mathrm{C},1}$ for the composite quantile estimator), (2) the estimated weights based on minimising the asymptotic variance of the estimators of only the active set of coefficients, denoted by $w_{\mathrm{MA},2}$ [7] and $w_{\mathrm{C},2}$ [9] where, with the (k_1,k_2)th component of A equal to $A_{k_1,k_2}=\min(\tau_{k_1},\tau_{k_2})\{1-\max(\tau_{k_1},\tau_{k_2})\}$, $A_\varepsilon=\mathrm{diag}(f_\varepsilon(u_{\tau_1}),\ldots,f_\varepsilon(u_{\tau_K}))$, and $a_\varepsilon=(f_\varepsilon(u_{\tau_1}),\ldots,f_\varepsilon(u_{\tau_K}))^\top$

$$w_{\mathrm{MA},2}=\arg\min_{w,\mathbf{1}_K^\top w=1,w_k\geq 0}\left\{w^\top A_\varepsilon^{-1}AA_\varepsilon^{-1}w\right\} \tag{32}$$

and

$$w_{\mathrm{C},2}=\arg\min_{w,a_\varepsilon^\top w=1,w_k\geq 0}\left[w^\top Aw\right].$$

Only considering the variance has been the standard practice so far. (3) Equal weights $1/K$ for each component.

The number of quantiles K for both estimators is taken to be 3, with quantile levels $25\%,50\%,75\%$.

We present the empirical MSEs of the abovementioned estimators for estimation of three vectors of coefficients. First, we consider the estimator of the subvector of the full coefficient that consists of only the non-zero true coefficients, we refer to this as the "non-zero part". Second, we consider the estimator of the subvector of the coefficients that are truly zero. This is referred to as the "zero part". Third, we consider the full vector of estimated coefficients. Note that some truly zero coefficients might have a non-zero estimate, while some truly non-zero coefficients might be estimated as zero. For each of these three vectors, "parts", we compare the estimated values with the true values to get

$$\mathrm{MSE}(\widehat{\beta}_{\mathrm{part}})=\sum_{j_{\mathrm{part}}=1}^{p_{\mathrm{part}}}(\widehat{\beta}_{j_{\mathrm{part}}}-\beta_{j_{\mathrm{part}}})^2/p_{\mathrm{part}}$$

for the appropriate part of the full vectors. Results for the regularized model-averaged quantile estimator with different weights are presented in Table 1. We observe that the model-averaged quantile estimator using the weight in (31) has lower MSEs for estimating the non-zero part of β and for the full vector β, and this for t_3 and the mixture of normally distributed errors in the high-sparse case where the number of non-zero components $s=5$. Using equal weights leads to a fair performance of the model-averaged quantile estimator,

especially for estimating the all-zero part of β. The Lasso estimator is considered as the baseline comparison, which from Table 1 seems to have a competitive performance, especially in the medium sparsity settings. However, the Lasso mostly gives over-sparse estimations, which can be observed in the top half of Table 2 summarizing the averaged true positive (TP) and true negative (TN) recovery rates which are defined as

$$\text{TP (TN)} = \frac{\text{number of correctly identified as non-zeros (zeros)}}{\text{number of true non-zeros (zeros)}}$$

The Lasso has the highest TN rate consistently and mostly the lowest TP rate. Further, while increasing the standard deviation of the errors, the Lasso's overly-sparse estimation becomes clearer, i.e., Lasso gives sparser estimations and becomes all-zeros eventually. The regularized model-averaged estimator with equal weights mostly has the highest TP rate, except for the medium sparsity settings where the non-zero part of the true regression coefficient is sampled from a Dirac distribution at -1 and 1, and the errors are sampled from $N(0,1)$ or $0.5N(0,1)+0.5N(5,9)$. The model-averaged estimator with the weight in (31) has the second-highest TN rate consistently.

Since there is no analytical expression for the selection incorporated weight of the regularized composite quantile estimator $w_{\mathrm{C},1}$, the choice of weights can only be determined numerically by an exhaustive search. To reduce the searching time of the composite quantile estimator, we set the stopping criterion $S_{\mathcal{V}}$ to be five and only randomly select 4 points in the neighborhood $\mathcal{V}_{w_{\text{cand}}}$; the tuning parameter α of the soft-thresholding function is tuned once for the regularized composite quantile estimator with the weight $w_{\mathrm{C},2}$, then fixed after that.

Table 3 summarizes the empirical MSEs of the regularized composite quantile estimator with different weights. Since the tuning parameter, α is selected for $w_{\mathrm{C},2}$ and a fixed tuning parameter is used for obtaining the regularized composite quantile estimates with other weights, it is not surprising that using $w_{\mathrm{C},2}$ leads to lower MSEs in most cases. However, it is worth noticing that using equal weights, while α is not optimally tuned, leads to the regularized composite quantile estimator's fair performances. The Lasso estimator consistently has the lowest empirical MSEs recovering the all-zero parts, through the largest empirical MSEs recovering the non-zero parts. This is caused by overly sparse estimations of the Lasso, which is indicated in the bottom half of Table 2. The regularized composite estimator with locally optimized $w_{\mathrm{C},1}$ consistently has the highest TP rate, and second-highest TN rate among all competitors, except the TN rate for t_3 distributed errors and TP rate for $0.5N(0,1)+0.5N(5,9)$ distributed errors. At the same time, the non-zero parts of β are generated from Dirac distribution at -1 and 1.

Tables 3 and 2 illustrate that the regularized composite quantile estimator mostly improves the performance of regularized single quantile estimator. For the same simulations settings, we compare the averaged empirical MSEs, true positive and true negative rates of the regularized composite quantile estimator, see Table 2, column 7 and 12, and Table 3, column 7, with the single regularized quantile estimator at the median $\tau = 0.5$. For settings where $s = 5$, the

TABLE 1

The mean, over 500 simulation repetitions, of the empirical MSE of the regularized model-averaged *quantile estimator with $K = 3$ for three error distributions. Empirical MSEs are calculated for the non-zero parts, all-zero parts, and the full vector of the true coefficient β. The non-zero part of the true coefficient vector is generated from Dirac distribution with point mass equally distributed on -1 and 1 (top half), or standard normal distribution (bottom half). Smaller values of MSE among competitors indicate more accurate estimations.*

f_ε	part	MSE($\widehat{\beta}_{\widehat{w}_{\mathrm{MA},1}}$)	MSE($\widehat{\beta}_{\widehat{w}_{\mathrm{MA},2}}$)	MSE($\widehat{\beta}_{\widehat{w}^{\mathrm{MA}}_{\mathrm{eq}}}$)	MSE($\widehat{\beta}_{\mathrm{Lasso}}$)
Non-zero part of β: **Dirac distribution at -1 and 1** ($*$: $\times 10^{-2}$, †: $\times 10^{-3}$, ‡: $\times 10^{-4}$)					
$\mathbf{s = 5}$	Non-zero	0.312	0.299	0.306	0.480
$N(0,1)$	Zero (‡)	6.812	6.436	5.276	0.526
	Full vec (†)	3.790	3.630	3.585	4.854
	Non-zero	0.167	0.168	0.182	0.681
t_3	Zero (‡)	4.051	3.579	3.041	0.106
	Full vec (†)	2.078	2.039	2.121	6.816
	Non-zero	0.247	0.355	0.314	0.412
$0.5N(0,1)+$	Zero (‡)	4.593	7.516	5.418	0.791
$0.5N(5,9)$	Full vec (†)	2.920	4.294	3.680	4.207
$\mathbf{s = 50}$	Non-zero	0.487	0.502	0.526	0.376
$N(0,1)$	Zero (†)	5.498	4.438	3.675	5.710
	Full vec ($*$)	5.364	5.419	5.590	4.275
	Non-zero	0.399	0.427	0.452	0.384
t_3	Zero (†)	4.976	3.945	3.412	5.317
	Full vec ($*$)	4.436	4.630	4.832	4.318
	Non-zero	0.504	0.517	0.540	0.371
$0.5N(0,1)+$	Zero (†)	5.303	4.386	3.635	5.913
$0.5N(5,9)$	Full vec ($*$)	5.514	5.566	5.724	4.241
Non-zero part of β: $\mathbf{N(0,1)}$ ($*$: $\times 10^{-2}$, †: $\times 10^{-3}$, ‡: $\times 10^{-4}$)					
$\mathbf{s = 5}$	Non-zero	0.206	0.197	0.203	0.378
$N(0,1)$	Zero (‡)	5.624	5.683	4.439	0.158
	Full vec (†)	2.613	2.537	2.465	3.800
	Non-zero	0.123	0.126	0.132	0.540
t_3	Zero (‡)	3.727	3.153	2.752	0.017
	Full vec (†)	1.601	1.574	1.590	5.403
	Non-zero	0.159	0.230	0.204	0.313
$0.5N(0,1)+$	Zero (‡)	3.788	6.723	4.720	0.348
$0.5N(5,9)$	Full vec (†)	1.969	2.970	2.511	3.162
$\mathbf{s = 50}$	Non-zero	0.257	0.256	0.265	0.216
$N(0,1)$	Zero (†)	3.377	2.835	2.401	2.445
	Full vec ($*$)	2.870	2.819	2.870	2.376
	Non-zero	0.201	0.207	0.216	0.244
t_3	Zero (†)	2.831	2.275	2.009	1.859
	Full vec ($*$)	2.264	2.278	2.336	2.611
	Non-zero	0.275	0.278	0.285	0.220
$0.5N(0,1)+$	Zero (†)	3.571	2.945	2.536	2.530
$0.5N(5,9)$	Full vec ($*$)	3.076	3.049	3.077	2.423

TABLE 2

The mean, over 500 simulation repetitions, of the true positive (TP) and true negative (TN) rate of the regularized model-averaged *(top half) and* composite *(bottom half) quantile estimator with $K = 3$ for three error distributions. The TP and TN rates of the regularized single quantile estimator at quantile level 0.5 are presented in the 7th and 12th columns. The non-zero part of the true coefficient vector is generated from Dirac distribution with point mass equally distributed on -1 and 1 (left), or standard normal distribution (right). Larger values of TP and TN indicate a better identification power; the largest values among competitors are highlighted in green, whereas the second largest values are highlighted in yellow.*

Non-zero part of β:		**Dirac distribution at -1 and 1**					**N(0, 1)**				
f_ε	rate	$\widehat{\beta}_{\widehat{w}_{\mathrm{MA},1}}$	$\widehat{\beta}_{\widehat{w}_{\mathrm{MA},2}}$	$\widehat{\beta}_{\widehat{w}^{\mathrm{MA}}_{\mathrm{eq}}}$	$\widehat{\beta}_{\mathrm{Lasso}}$	$\widehat{\beta}_{0.5}$	$\widehat{\beta}_{\widehat{w}_{\mathrm{MA},1}}$	$\widehat{\beta}_{\widehat{w}_{\mathrm{MA},2}}$	$\widehat{\beta}_{\widehat{w}^{\mathrm{MA}}_{\mathrm{eq}}}$	$\widehat{\beta}_{\mathrm{Lasso}}$	$\widehat{\beta}_{0.5}$
s = 5 $N(0,1)$	TP	0.992	0.991	0.993	0.906	0.982	0.677	0.683	0.688	0.419	0.660
	TN	0.904	0.903	0.896	0.995	0.940	0.916	0.912	0.907	0.998	0.945
t_3	TP	0.999	0.999	0.999	0.663	1.000	0.754	0.762	0.765	0.294	0.739
	TN	0.910	0.903	0.896	0.999	0.941	0.913	0.905	0.899	1.000	0.943
$0.5N(0,1)+$	TP	0.992	0.984	0.992	0.942	0.820	0.719	0.711	0.724	0.486	0.482
$0.5N(5,9)$	TN	0.922	0.912	0.906	0.992	0.942	0.927	0.916	0.911	0.997	0.946
s = 50 $N(0,1)$	TP	0.836	0.847	0.854	0.889	0.548	0.647	0.658	0.666	0.619	0.600
	TN	0.843	0.830	0.823	0.868	0.606	0.843	0.832	0.807	0.883	0.894
t_3	TP	0.892	0.899	0.904	0.882	0.453	0.696	0.706	0.715	0.590	0.622
	TN	0.833	0.816	0.807	0.873	0.707	0.839	0.822	0.811	0.931	0.842
$0.5N(0,1)+$	TP	0.822	0.837	0.843	0.892	0.531	0.633	0.643	0.650	0.621	0.539
$0.5N(5,9)$	TN	0.845	0.833	0.826	0.864	0.601	0.846	0.834	0.827	0.911	0.834
f_ε	rate	$\widehat{\beta}_{\widehat{w}_{\mathrm{C},1}}$	$\widehat{\beta}_{\widehat{w}_{\mathrm{C},2}}$	$\widehat{\beta}_{\widehat{w}^{\mathrm{C}}_{\mathrm{eq}}}$	$\widehat{\beta}_{\mathrm{Lasso}}$	$\widehat{\beta}_{0.5}$	$\widehat{\beta}_{\widehat{w}_{\mathrm{C},1}}$	$\widehat{\beta}_{\widehat{w}_{\mathrm{C},2}}$	$\widehat{\beta}_{\widehat{w}^{\mathrm{C}}_{\mathrm{eq}}}$	$\widehat{\beta}_{\mathrm{Lasso}}$	$\widehat{\beta}_{0.5}$
s = 5 $\mathrm{N}(0,\,1)$	TP	0.993	0.991	0.991	0.911	0.982	0.664	0.656	0.657	0.444	0.660
	TN	0.946	0.946	0.946	0.994	0.940	0.963	0.963	0.963	0.998	0.945
t_3	TP	1.000	1.000	1.000	0.675	1.000	0.740	0.729	0.736	0.303	0.739
	TN	0.945	0.944	0.944	0.998	0.941	0.957	0.957	0.956	1.000	0.943
$0.5N(0,1)+$	TP	0.990	0.982	0.968	0.939	0.820	0.699	0.650	0.626	0.485	0.482
$0.5N(5,9)$	TN	0.955	0.953	0.951	0.991	0.942	0.965	0.966	0.967	0.993	0.946
s = 50 $N(0,1)$	TP	0.903	0.895	0.891	0.883	0.548	0.669	0.668	0.665	0.613	0.600
	TN	0.824	0.823	0.824	0.872	0.606	0.848	0.846	0.846	0.916	0.894
t_3	TP	0.939	0.933	0.934	0.871	0.453	0.724	0.722	0.720	0.590	0.622
	TN	0.819	0.817	0.820	0.879	0.707	0.835	0.834	0.835	0.929	0.842
$0.5N(0,1)+$	TP	0.886	0.881	0.875	0.891	0.531	0.651	0.648	0.642	0.615	0.539
$0.5N(5,9)$	TN	0.827	0.824	0.825	0.867	0.601	0.855	0.852	0.852	0.914	0.834

TABLE 3

The mean, over 500 simulation repetitions, of the empirical MSE of the regularized composite *quantile estimator with $K = 3$ and the regularized single quantile estimator at quantile level 0.5 for three error distributions. Empirical MSEs are calculated for the non-zero parts, all-zero parts, and the full vector of the true coefficient β. The non-zero part of the true coefficient vector is generated from Dirac distribution with point mass equally distributed on -1 and 1 (top half), or standard normal distribution (bottom half). Smaller values of MSE among competitors indicate more accurate estimations.*

f_ε	part	MSE($\widehat{\beta}_{\widehat{w}_{\mathrm{C},1}}$)	MSE($\widehat{\beta}_{\widehat{w}_{\mathrm{C},2}}$)	MSE($\widehat{\beta}_{\widehat{w}^{\mathrm{C}}_{\mathrm{eq}}}$)	MSE($\widehat{\beta}_{\mathrm{Lasso}}$)	MSE($\widehat{\beta}_{0.5}$)
Non-zero part of β: **Dirac distribution at -1 and 1** ($*$: $\times 10^{-2}$, †: $\times 10^{-3}$, ‡: $\times 10^{-4}$)						
s = 5	Non-zero	0.226	0.246	0.249	0.479	0.272
$N(0,1)$	Zero (‡)	6.566	7.119	7.199	0.571	11.641
	Full vec (†)	2.906	3.163	3.202	4.847	3.752
	Non-zero	0.122	0.135	0.133	0.674	0.142
t_3	Zero (‡)	3.782	4.148	4.109	0.112	5.650
	Full vec (†)	1.593	1.756	1.740	6.747	1.911
	Non-zero	0.184	0.246	0.311	0.420	0.461
$0.5N(0,1)+$	Zero (‡)	4.635	6.165	7.353	1.015	20.016
$0.5N(5,9)$	Full vec (†)	2.301	3.068	3.839	4.303	6.011
s = 50	Non-zero	0.342	0.359	0.367	0.384	0.310
$N(0,1)$	Zero (†)	8.722	9.273	9.536	5.572	4.368
	Full vec ($*$)	4.203	4.423	4.524	4.339	4.308
	Non-zero	0.280	0.294	0.298	0.398	0.598
t_3	Zero (†)	7.026	7.511	7.618	5.159	19.415
	Full vec ($*$)	3.429	3.617	3.663	4.443	5.709
	Non-zero	0.358	0.376	0.385	0.375	0.318
$0.5N(0,1)+$	Zero (†)	9.099	9.644	10.007	5.750	4.301
$0.5N(5,9)$	Full vec ($*$)	4.403	4.631	4.751	4.268	4.360
Non-zero part of β: $\mathbf{N(0,1)}$ ($*$: $\times 10^{-2}$, †: $\times 10^{-3}$, ‡: $\times 10^{-4}$)						
s = 5	Non-zero	0.157	0.173	0.175	0.363	0.177
$N(0,1)$	Zero (‡)	3.782	4.311	4.327	0.220	9.528
	Full vec (†)	1.946	2.153	2.178	3.655	2.555
	Non-zero	0.099	0.110	0.108	0.527	0.107
t_3	Zero (‡)	2.575	2.861	2.826	0.043	5.169
	Full vec (†)	1.245	1.382	1.360	5.273	1.492
	Non-zero	0.134	0.176	0.212	0.317	0.406
$0.5N(0,1)+$	Zero (‡)	2.787	3.772	4.550	0.476	23.379
$0.5N(5,9)$	Full vec (†)	1.615	2.135	2.568	3.213	4.469
s = 50	Non-zero	0.174	0.182	0.186	0.216	0.230
$N(0,1)$	Zero (†)	4.925	5.220	5.369	2.299	3.970
	Full vec($*$)	2.181	2.289	2.342	2.371	2.571
	Non-zero	0.130	0.138	0.139	0.236	0.168
t_3	Zero (†)	3.849	4.077	4.152	1.888	2.887
	Full vec ($*$)	1.645	1.744	1.765	2.531	1.925
	Non-zero	0.189	0.198	0.205	0.214	0.236
$0.5N(0,1)+$	Zero (†)	5.149	5.507	5.738	2.386	3.984
$0.5N(5,9)$	Full vec ($*$)	2.354	2.476	2.562	2.353	2.776

composite quantile estimator clearly dominates the single quantile estimator for all three error distributions. For settings where $s = 50$, the composite estimator still mostly outperforms the single quantile estimator, except for the following cases: (1) the MSE for the non-zero and zero estimated subvector of β in settings where errors are generated from $N(0,1)$ and $0.5N(0,1)+0.5N(5,9)$ distribution and the true non-zero subvector of β is generated from a Dirac distribution; (2) TN rates in settings where errors are generated from $N(0,1)$ and t_3 distribution and the true non-zero subvector of β is generated from $N(0,1)$.

The percentage of converged cases for the model-averaged and composite estimator, while setting the tolerance ε_{tol} to be 10^{-6} for different error distributions, are included in Table 4, where we define a estimator to have converged when the needed number of iterations was less than 50.

TABLE 4

Percentage of converged cases of both regularized model-averaged and composite quantile estimators with the convergence tolerance $\varepsilon_{\text{tol}} = 10^{-6}$. The convergence percentage of the regularized model-averaged estimator is calculated by including only those cases of which all single quantile component estimates converge in less than 50 iterations.

(%)	$s = 10$		$s = 50$	
f_ε	model-averaged	composite	model-averaged	composite
$N(0,1)$	76	90	69	86
t_3	78	78	71	82
$0.5N(0,1)+0.5N(5,9)$	77	86	71	85

Assumption (A1) restricts Algorithm 1 to a special design matrix that does not allow correlations between the $X_{\cdot j}$'s. However, since such correlation might be present in reality, it is of interest to see if Algorithm 1 is still numerically robust while Assumption (A1) is relaxed in practice. We consider a similar simulation setup as used before with $p = 500$, the sample size $n = 250$, and $\delta = 0.5$. The number of non-zero components s is taken to be 5 or 50; the non-zero components are generated from the Dirac distribution with point mass equally distributed on -1 or 1, or a standard normal distribution. In each simulation replication, a design matrix is first generated from a multivariate Gaussian distribution $N(0, \Sigma_X)$, then the components X_{ij} are centered and scaled such that the components of the rescaled matrix X have sample variance $1/n$. Here, we allow for a Toeplitz covariance matrix Σ_X of which its (i,j)th component $(\Sigma_X)_{i,j} = \sigma_X^{|i-j|}, i, j = 1, \ldots, p$. We consider $\sigma_X = 0, 0.1, 0.3$.

To investigate the effect of the correlation on the RAMP algorithm we consider the regularized single quantile estimator at quantile level 0.5. The error distribution considered is t_3. Table 5 records the performance of Algorithm 1 with tolerance $\varepsilon_{\text{tol}} = 10^{-6}$ for such a correlated design matrix; the performance is evaluated by the empirical MSEs, the TP and TN rates, and the percentage of convergence.

We see from Table 5 that parameter estimation using Algorithm 1 remains accurate and stable when weak correlations such as with $\sigma_X = 0.1$ exist between the $X_{\cdot j}$'s; the accuracy drops when we further increase the correlations as with

TABLE 5

The mean, over 500 simulation repetitions, of the empirical MSEs, the true positive (TP), the true negative (TN), and convergence percentages of the regularized model-averaged quantile estimators for t_3 distributed errors. Empirical MSEs are calculated for the non-zero parts, all-zero parts, and the full vector of the true coefficient β. The non-zero part of the true coefficient vector is generated from Dirac distribution with point mass equally distributed on -1 and 1 (top), or standard normal distribution (bottom).

$f_\varepsilon : t_3$	$\widehat{\beta}_{\widehat{w}_{\mathrm{MA},1}}$	MSE($\widehat{\beta}_{\mathrm{vec}}$)			TP	TN	Convergence %
		Non-zero	Zero	Full vec			
Non-zero part of β: **Dirac distribution at -1 and 1** ($*$: $\times 10^{-2}$, †: $\times 10^{-3}$, ‡: $\times 10^{-4}$)							
$\sigma_X = 0$	$s = 5$	0.142	5.650 (‡)	1.911 (†)	1.000	0.941	98
	$s = 50$	0.598	19.415 (†)	5.709 ($*$)	0.453	0.707	87
$\sigma_X = 0.1$	$s = 5$	0.143	5.455(‡)	1.969 (†)	1.000	0.944	97
	$s = 50$	0.403	5.170 (†)	4.492 ($*$)	0.843	0.881	85
$\sigma_X = 0.3$	$s = 5$	0.145	5.923(‡)	2.037 (†)	1.000	0.940	87
	$s = 50$	0.461	5.159 (†)	5.074 ($*$)	0.793	0.895	41
Non-zero part of β: $\mathbf{N}(\mathbf{0}, \mathbf{1})$ ($*$: $\times 10^{-2}$, †: $\times 10^{-3}$, ‡: $\times 10^{-4}$)							
$\sigma_X = 0$	$s = 5$	0.107	5.169 (‡)	1.492 (†)	0.943	0.482	98
	$s = 50$	0.168	2.887 (†)	1.925 ($*$)	0.622	0.842	87
$\sigma_X = 0.1$	$s = 5$	0.106	4.842 (‡)	1.534 (†)	0.741	0.948	97
	$s = 50$	0.182	2.875 (†)	2.079 ($*$)	0.653	0.892	86
$\sigma_X = 0.3$	$s = 5$	0.109	4.636 (‡)	1.546 (†)	0.738	0.948	87
	$s = 50$	0.198	2.857 (†)	2.236 ($*$)	0.638	0.900	47

$\sigma_X = 0.3$; it is worth mentioning that the convergence percentages decrease when the correlation increases. Further research concerning correlated data is worth considering.

7.2. Data analysis

We consider the audio wave file of a waveshape from Octave in the R package `signal`. The dataset is a list of 3 elements; the audio wave sample is a vector of 17380 entries stored in the element "sound", the sample rate is 22050 Hz stored in the element "rate", and the resolution of the wave file is 16 bits recorded in the element "bits". To alleviate the computational burden of the signal compression and reconstruction, we only consider the signal from the 6145th entry to the 8192th entry of the original sound wave signal.

7.2.1. The preprocessing – discrete wavelet transform

Originated from the compressed sensing problem, the sparse linear model $Y = X\beta + \varepsilon$ describes the image or signal compression. The s-sparse p-dimensional input signal β is first compressed by a known sensing matrix $X \in \mathbb{R}^{n\times p}$ with $n < p$; the compressed signal vector $X\beta \in \mathbb{R}^n$ can be corrupted by the noise ε with ε_i's i.i.d. via transmission. Notice that the p-dimensional input signal vector β is assumed to be s-sparse which is usually unsatisfied by signals expressed

in the standard basis. To obtain the sparse representation of β in practice, an intermediate stage of expressing the natural non-sparse vector β^* in a proper orthonormal basis $\Psi^* = (\psi_1^*, \ldots, \psi_p^*)$ is required. Examples of such an orthonormal basis include the orthonormal wavelet basis, the Fourier basis, and so forth. To perform the discrete wavelet transform, we use the R package `wavethresh`. The collection of the coefficients at all resolution levels is used for further compression.

7.2.2. The artificially corrupted compression

To imitate the compressed sensing process, we process the audio wave signal vector as follows:

1. Perform the Daubechies' least asymmetric wavelet transform with 8 vanishing moments using the `wd` function in the R package `wavethresh` on the original signal $\beta^* \in \mathbb{R}^{2048}$ and obtain the corresponding wavelet coefficient vector $\beta \in \mathbb{R}^{2047}$ with $p = 2047$.
2. Randomly generate the sensing matrix X with i.i.d components $X_{ij} \sim N(0, 1/n)$, where $n = \lfloor \delta' p \rfloor$ and δ' is the undersampling ratio chosen to be 0.5 here; compress the corresponding wavelet coefficients β by computing $X\beta$.
3. Corrupt the compressed wavelet coefficients by the error vector ε with i.i.d. components ε_i having p.d.f f_ε; obtain the artificial observed signal vector $Y = X\beta + \varepsilon$. Additionally, the standard normal $N(0, 1)$, student-t with 3 degrees of freedom, and the bimodal mixed normal $0.5N(0, 1) + 0.5N(5, 9)$ are used as the corruption error distributions; the errors are sampled according to the distributions first, then centered and rescaled to have standard deviation 0.03.

In practice, the artificial vector Y and the sensing matrix X are observed. The accurate recovery of the original wavelet coefficient vector β is of practical interest. To obtain an impression on the performance of the AMSE-type optimal weight, we generate the sensing matrix X under a fixed seed number, which is set to be 1 in our case, then generate the error vector ε under various seed numbers. However, we only present the reconstructions under one seed for each setting in Section 7.2.3 due to limited space.

7.2.3. Signal recovery

To reconstruct the signal vector β expressed in the wavelet basis from the sensing matrix X and the observed compressed signal vector Y corrupted by potentially non-Gaussian distributed error ε, we consider the regularized model-averaged and the composite quantile estimator weighting over three equally-spaced quantiles $(25\%, 50\%, 75\%)$ using equal weights, the oracle-type weights and the new AMSE-type weights. The tolerance in the RAMP algorithm is set as $\varepsilon_{\text{tol}} = 10^{-8}$. The Lasso estimator is considered as the baseline comparison. Notice that the

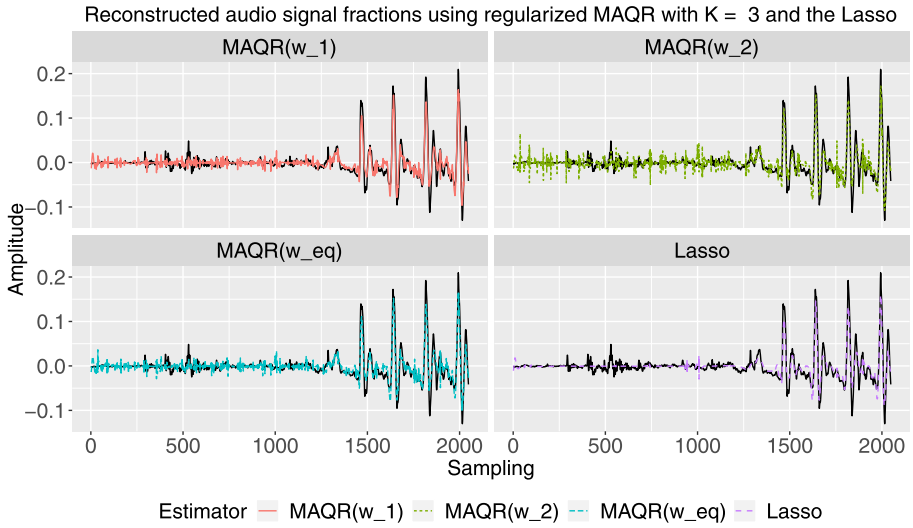

FIG 2. *Reconstructed audio signal from using the regularized* model-averaged *estimator with the estimated AMSE-type weights in (31), oracle-type optimal weights in (32), and equal weights. The original audio curve is depicted in black. The Lasso reconstruction is presented at bottom-right. The error used for corruption follows the mixture of normals distribution* $0.5N(0,1) + 0.5N(5,9)$.

regularized estimates $\widehat{\beta}_{\mathrm{MA}}$ and $\widehat{\beta}_{\mathrm{C}}$ after reconstruction are the representations in the wavelet domain. To compare the accuracy of the reconstruction, we perform a back-transform on the estimates and obtain the corresponding signal vectors $\widehat{\beta}^*_{\mathrm{MA}}$ and $\widehat{\beta}^*_{\mathrm{C}}$ with representations in the natural basis.

Example reconstructions of the audio signal for $K = 3$ using the regularized model-averaged estimator equipped with different weights, with the baseline recovery from the Lasso represented in the natural basis are presented in Figure 2 for the mixture of normals distributed error, and in Figure 3 for the t_3 distributed error. We observe that the strong signals corresponding to large values located at the end of the sound signal are well captured by the model-averaged quantile estimator using different weights for both error distributions. For the weak signals clustering at the front of the signal, the model-averaged estimators using $\widehat{w}_{\mathrm{MA},1}$ and equal weights outperform the counterpart with $\widehat{w}_{\mathrm{MA},2}$ for $0.5N(0,1) + 0.5N(5,9)$ distributed errors; recovery differences for the weak signals of the model-averaged estimator using different weights are hardly observable for the t_3 distributed errors. Recovery using the Lasso is competitive to the model-averaged estimator using $w_{\mathrm{MA},1}$ for strong signals. However, the Lasso estimates the signals in an over-sparse way with too many zeros entries; one can observe the almost flat recovery for the weak signals for both error distributions.

Bates and Granger [3] provide an alternative weight choice for the model-averaged estimator obtained by considering only the variances of $\widehat{\beta}_k$'s and ig-

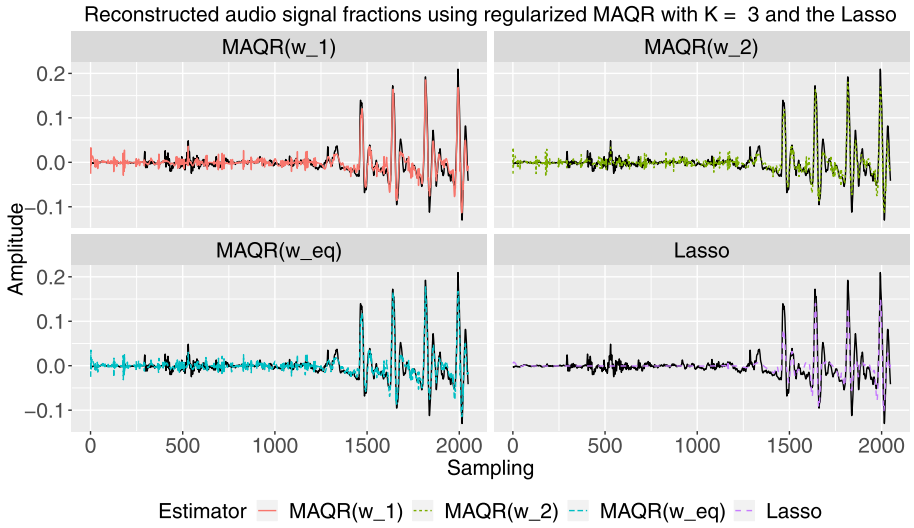

FIG 3. *Reconstructed audio signal using the regularized* model-averaged *estimator with the estimated AMSE-type weights in (31), oracle-type optimal weights in (32), and equal weights. The original audio curve is depicted in black. The Lasso reconstruction is presented at bottom-right. The error used for corruption is* t_3 *distributed.*

noring the covariances. This leads to

$$\widehat{w}_{\mathrm{MA},3} = \arg \min_{w \geq 0, \mathbf{1}_K^\top w = 1} w^\top \mathrm{diag}(\widehat{\Sigma}_{0,(t)}) w, \tag{33}$$

where $\mathrm{diag}(\widehat{\Sigma}_{0,(t)})$ denotes the diagonal matrix obtained from $\widehat{\Sigma}_{0,(t)}$ which keeps the diagonal and has zeros in all off-diagonal entries. Figure 4 contains the recovery of the audio signal using the model-averaged estimator using this weight.

For the composite quantile estimator $\widehat{\beta}_{\mathrm{C}}$, we performed the same weight searching method as for the simulation study. This is, $S_{\mathcal{V}} = 5$ and randomly select 4 candidate weights in the neighbourhood of the previous value. We select the tuning parameter α once for the starting weight $\widehat{w}_{\mathrm{C},2}$, it remains unchanged thereafter. The recovered signals by the composite estimator with different weights are very similar in all cases.

To compare the recovery of the regularized model-averaged and composite estimator combined with different weights, as well as the Lasso estimator, we present the mean absolute percentage error (MAPE) in Table 6 where the MAPE is defined as

$$\mathrm{MAPE}(\widehat{\beta}, \beta) = \frac{1}{p} \sum_{j=1}^{p} \left| \left(\widehat{\beta}_j - \beta_j\right) \Big/ \beta_j \right| \tag{34}$$

Table 7 reports the MSE.

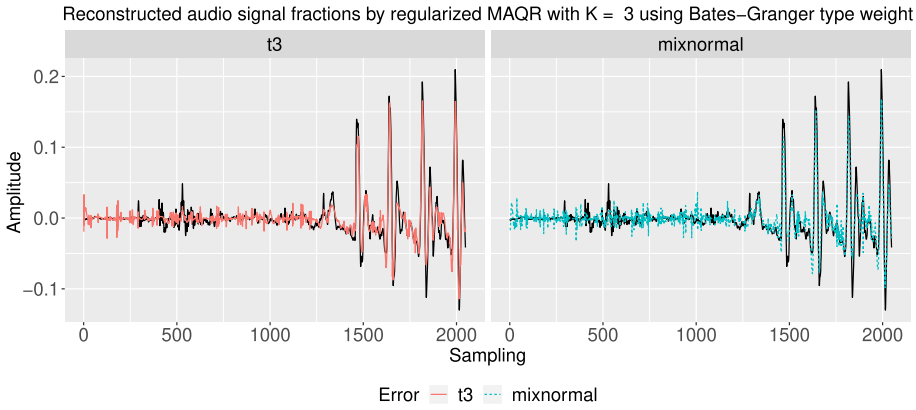

FIG 4. *Reconstructed audio signal using the regularized* model-averaged *estimator with Bates-Granger type weight in* (33). *The original audio curve is depicted in black. The left figure uses* t_3 *distributed corruption error, while the figure on the right used* $0.5N(0,1)+0.5N(5,9)$ *distributed corruption error.*

TABLE 6

The MAPE defined in (34) *of the audio signal recovered by the regularized model-averaged and composite estimators with different weights, and the Lasso estimator. The seed number used to generated the errors for corrupting the compressed signal vector is 37 for both* t_3 *and mixed normal distributed errors.*

f_ε	t_3				0.5N(0,1) + 0.5 N(5, 9)			
est: MA / C	$w_{\text{est},1}$	$w_{\text{est},2}$	w_{eq}	$w_{\text{est},3}$	$w_{\text{est},1}$	$w_{\text{est},2}$	w_{eq}	$w_{\text{est},3}$
MAQR	3.177	3.339	3.341	3.005	2.934	5.746	3.722	4.152
CQR	2.798	2.730	2.671	–	6.100	6.090	6.462	–
Lasso	1.346				1.536			

TABLE 7

The MSE of the audio signal recovered by the regularized model-averaged and composite estimators with different weights, and the Lasso estimator. The seed number used to generated the errors for corrupting the compressed signal vector is 37 for both t_3 *and mixed normal distributed errors.*

f_ε	t_3				0.5N(0,1) + 0.5 N(5, 9)			
est: MA / C ($\times 10^{-4}$)	$w_{\text{est},1}$	$w_{\text{est},2}$	w_{eq}	$w_{\text{est},3}$	$w_{\text{est},1}$	$w_{\text{est},2}$	w_{eq}	$w_{\text{est},3}$
MAQR	1.286	1.288	1.274	1.304	2.044	2.417	2.051	2.006
CQR	1.279	1.273	1.271	–	2.363	2.068	2.120	–
Lasso	2.566				2.003			

We see that the Lasso has the lowest MAPE for both t_3 and mixed normal distributed errors; at the same time, it estimates the weak signals in an over-sparse way and is not capable of capturing the weak signals. Comparing the effect of different weight choices on the regularized model-averaged quantile estimator with its composite quantile counterpart, we see that the MAPEs of the composite quantile estimators are relatively stable using different weights.

The model-averaged estimator with the AMSE-type weight $\widehat{w}_{\mathrm{MA},1}$ has excellent performance compared to the composite estimator, especially for the mixed normal distributed error. The Bates-Granger weighting provides good results regarding MAPE for the t_3 error case, but not for the mixed normal. Regarding MSE, it performs well for the mixed normal case but is worst for the t_3 errors, wherein this example the equal weights perform best, although all results are close. Searching for the selection incorporated weight $\widehat{w}_{\mathrm{C},1}$ for the regularized composite quantile estimator is computationally infeasible for large p (2047 in our case). Estimating the regularized model-averaged quantile estimator averaging three quantiles here takes approximately $4-5$ hours whereas estimating the regularized composite quantile estimator takes more than 16 hours with only five steps in a nearby search with four surrounding candidate weights, and the tuning parameter α tuned only once for the starting weight.

Additionally, we present the estimated weights for both regularized model-averaged and composite estimators in Table 8. An interesting observation is made by comparing the estimated weights $\widehat{w}_{\mathrm{MA},1}$ and $\widehat{w}_{\mathrm{MA},2}$ for the mixed normal distributed error. The weight $\widehat{w}_{\mathrm{MA},1}$ presented here is quite representative; it assigns weight 0 to the quantile estimate at 50% quantile level suggesting the final model-averaged estimate is obtained by averaging estimates at 25% and 75% quantile levels. On the contrary, $\widehat{w}_{\mathrm{MA},2}$ assigns the largest weight to the estimate at a 50% quantile level indicating the most significant contribution to the final model-averaged estimate.

TABLE 8

The estimated weights $\widehat{w}_{\mathrm{MA},1}$ and $\widehat{w}_{\mathrm{MA},2}$ for the model-averaged estimator, and $\widehat{w}_{\mathrm{C},1}$ and $\widehat{w}_{\mathrm{C},2}$ for the composite estimator. The seed number used to generated the errors for corrupting the compressed signal vector is 37 for both t_3 and mixed normal distributed errors.

f_ε	est: MA / C	MAQR	CQR
t_3	$w_{\mathrm{est},1}$	(0.156, 0.725, 0.119)	(0.089, 0.492, 0.419)
	$w_{\mathrm{est},2}$	(0.077, 0.650, 0.273)	(0.314, 0.267, 0.467)
$0.5N(0,1)+$	$w_{\mathrm{est},1}$	(0.548, 0, 0.452)	(0.469, 0.495, 0.036)
$0.5N(5,9)$	$w_{\mathrm{est},2}$	(0.147, 0.843, 0.010)	(0.369, 0.345, 0.286)

8. Discussion

This paper is the first to take the selection uncertainty due to regularization into account when computing the weights used in model-averaged and composite estimation. While we have studied both composite estimation and model-averaged estimation, the flexibility of allowing for parallel computation and a component-specific choice of regularization, combined with an explicit expression of the optimal weights for model averaging, places this method in a preferred position from a computational point of view.

It would be interesting to investigate whether AMSE expressions for other types of regularization may be obtained similarly. Going yet one step further

would be incorporating the effect of data-driven values of the regularization parameters λ (for composite estimation) and $\lambda_1, \ldots, \lambda_K$ (for model-averaged estimation) on the choice of the weights. To further study the weight selection and the effect of using data-driven weights, one should study the joint distribution of the estimated weights and the estimators of interest. To simplify such matters, sample splitting could be used such that the weights are computed on a hold-out sample and the estimation using those weights proceeds on the rest of the sample. In this paper, we used the same dataset for estimating both β and w.

To avoid overly complicated mathematical expressions, we followed earlier literature in the use of a design matrix where $X_{ij} \sim N(0, 1/n)$. Other applications might require studying, for example, fixed designs, which are beyond the scope of the current paper.

Appendix

Appendix A: Assumptions

(A1) Design: The elements of the design matrix X, that is X_{ij} for $i = 1, \ldots, p$ and $j = 1, \ldots, n$, are independent and identically distributed according to a $N(0, 1/n)$ which is also called a standard Gaussian design.

(A2) Coefficients: The p-vector β is such that the sequence of uniform distributions that is placed on its components converges, for p tending to infinity, to a distribution with a bounded $(2k-2)$th moment for $k \geq 2$. Denote by B_0 a random variable with this limiting distribution function F_{B_0}.

(A3) Loss function: (i) The subgradient $\partial\rho(u) = \sum_{j=1}^{3} v_j(u)$ where v_1 has an absolutely continuous derivative, v_2 is continuous and consists of piecewise linear parts and is constant outside a bounded interval, and v_3 is a non-decreasing step function. Denote $v_2'(u) = \alpha_l$ and $v_3(u) = \gamma_l$ when $u \in (r_l, r_{l+1}]$ where $\alpha_0 = \alpha_L = 0$, $-\infty = r_0 < r_1 < \ldots < r_L < r_{L+1} = \infty$ and $-\infty = \gamma_0 < \gamma_1 < \ldots < \gamma_L < \gamma_{L+1} = \infty$. (ii) The subgradient's absolute value $|\partial\rho(u)|$ is bounded for all $u \in \mathbb{R}$. (iii) $h(t) = \int \rho(z-t) dF_\varepsilon(z)$ has a unique minimum at $t = 0$. (iv) There exists a $\delta > 0$ and $\eta > 1$ such that $E[\{\sup_{|u| \leq \delta} |v_1''(z+u)|\}^\eta]$ is finite.

(A4) We assume that for some $\kappa > 1$,

(a) $\lim_{p\to\infty} E_{\widehat{f}_\beta}(B_0^{2\kappa-2}) = E_{f_{B_0}}(B_0^{2\kappa-2}) < \infty$

(b) $\lim_{p\to\infty} E_{\widehat{f}_\varepsilon}(\varepsilon^{2\kappa-2}) = E_{f_\varepsilon}(\varepsilon^{2\kappa-2}) < \infty$

(c) $\lim_{p\to\infty} E_{\widehat{f}_{q_0}}(B_0^{2\kappa-2}) < \infty$.

(A5) The regression errors $\varepsilon_1, \ldots, \varepsilon_n$ and ε are i.i.d. random variables with mean zero and finite 2nd moment. Assume ε has cumulative distribution function F_ε and probability density function f_ε. Let F_ε have bounded derivatives f_ε and ∂f_ε; further, let $f_\varepsilon > 0$ in the neighbourhood of $r_1, \ldots, r_L$ in (A3).

Assumption (A1) has been used by Bayati and Montanari [5], Donoho and Montanari [14], Bradic [8], assumption (A2) has been used by Bayati and Montanari [5], Bradic [8]; while assumptions (A3) and (A5) correspond to conditions (R) and (D) of [8]. Assumption (A4) is used in Lemma 1, in addition to the moment condition stated in (A2) and (A5). We take $\kappa = 2$ for Algorithm 1.

Appendix B: Lemmas and Proofs

B.1. Auxiliary definitions and lemmas

Definition 1 (Pseudo Lipschitz function)**.** A function $\phi : \mathbb{R}^m \to \mathbb{R}$ is pseudo-Lipschitz of order $\kappa \geq 1$, if there exists a constant $L > 0$, such that $\forall x, y \in \mathbb{R}^m$

$$|\phi(x) - \phi(y)| \leq L(1 + \|x\|^{\kappa-1} + \|y\|^{\kappa-1})\|x - y\|.$$

It follows that if ϕ is a pseudo-Lipschitz function of order κ, then there exists a constant L' such that $\forall x \in \mathbb{R}^m : |\phi(x)| \leq L'(1 + \|x\|^{\kappa})$.

Lemma 2 (Theorem 1 in Jameson [24])**.** *If $x_i \geq 0$ where $i = 1, \ldots, n$ and $p \geq 1$, then*

$$\sum_{i=1}^{n} x_i^p \leq (\sum_{i=1}^{n} x_i)^p \leq n^{p-1} \sum_{i=1}^{n} x_i^p.$$

The reversed inequality holds for $p \in (0, 1)$

Lemma 3 (Extrema of quadratic forms in Rao [34])**.** *Let A be a $m \times m$ matrix, B be a $m \times k$ matrix, and U be a k-vector. Denote by S^- any generalized inverse of $B^\top A^{-1} B$, then*

$$\inf_{B^\top X = U} X^\top A X = U^\top S^- U$$

where X is a column vector and the infimum is attained at $A^{-1} B S^- U$.

Lemma 4 (Stein's lemma in Stein [35])**.** *Let X_1, X_2 jointly Gaussian distributed. Let $g : \mathbb{R} \to \mathbb{R}$ be absolutely continuous with derivative ∂g and $E|\partial g(X_1)| < \infty$. Then*

$$\mathrm{Cov}(g(X_1), X_2) = \mathrm{Cov}(X_1, X_2) E[\partial g(X_1)].$$

Lemma 5 (Lemma 4 in Bayati and Montanari [5])**.** *Let $\kappa \geq 2$ and a sequence of vectors $\{\beta(p)\}_{p \geq 0}$ whose empirical distribution converges weakly to probability measure f_{B_0} on $\mathbb{R}$ with bounded κth moment; additionally, assume that $\lim_{p\to\infty} E_{\widehat{f}_\beta}(B_0^\kappa) = E_{f_{B_0}}(B_0^\kappa)$. Then for any pseudo-Lipschitz function $\psi : \mathbb{R} \to \mathbb{R}$ of order κ:*

$$\lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \psi(\beta_j) \overset{\text{a.s.}}{=} E[\psi(B_0)].$$

B.2. Proofs

B.2.1. Proof of (6)

Proof. By definition, the proximal mapping operator is the minimizer of the function $b\rho_{\mathrm{C}}(x) + 0.5(x-z)^2$ which is non-differentiable but subdifferentiable, with subgradient $b \cdot \partial\rho_{\mathrm{C}}(x) + x - z$. $\mathrm{Prox}(z;b)$ is the minimizer if and only if $0 \in \{b \cdot \partial\rho_{\mathrm{C}}(x)|_{x=\mathrm{Prox(z;b)}} + \mathrm{Prox}(z;b) - z\}$. We distinguish between intervals where ρ_{C} is differentiable and non-differentiable points. For $x \in (u_{\tau_\ell}, u_{\tau_{\ell+1}}), \ell = 0, \ldots, K$ the function ρ_C is differentiable. Using the expression of the subgradient in (9), we obtain $0 = bh(\ell) + x - z$, which is solved for x to get that $\mathrm{Prox}(z;b) = z - bh(\ell)$. From $\mathrm{Prox}(z;b) \in (u_{\tau_\ell}, u_{\tau_{\ell+1}})$ it follows that $z \in (u_{\tau_\ell} + bh(\ell), u_{\tau_{\ell+1}} + bh(\ell))$. For the non-differentiable points, that is $x = u_{\tau_\ell}, \ell = 1, \ldots, K$, having $0 \in \{b[h(\ell-1), h(\ell)] + x - z\}$ leads to $u_{\tau_\ell} = \mathrm{Prox}(z;b) \in [z - bh(\ell), z - bh(\ell-1)]$. This implies that $z \in [u_{\tau_\ell} + bh(\ell-1), b_{\tau_\ell} + bh(\ell)]$. □

B.2.2. Proof of (10)

Proof. By definition, $\widetilde{G}(z;b) = b \cdot \partial\rho(x)|_{x=\mathrm{Prox}(z;b)}$, and, see the Proof of (6) in Section B.2.1, $0 \in \{b \cdot \partial\rho(x)|_{x=\mathrm{Prox}(z;b)} + \mathrm{Prox}(z,b) - z\}$. Without loss of generality, we show the calculation for the cases where $z < u_{\tau_1} + bh(0)$ and where $z \in [u_{\tau_1} + bh(0), u_{\tau_1} + bh(1)]$.

For $z < u_{\tau_1} + bh(0)$ it holds that $\mathrm{Prox}(z;b) = z - bh(0) < u_{\tau_1}$, which leads to $\partial\rho(x)|_{x=\mathrm{Prox(z;b)}} = h(0)$. Hence, $\widetilde{G}(z;b) = b \cdot \partial\rho(x)|_{x=\mathrm{Prox(z;b)}} = bh(0)$.

Having $z \in [u_{\tau_1} + bh(0), u_{\tau_1} + bh(1)]$ corresponds to taking the nondifferentiable point $u_{\tau_1} = \mathrm{Prox}(z;b)$, see the proof of (6). We have $\partial\rho(x)|_{x=\mathrm{Prox}(z;b)} \in [h(0), h(1)]$. The subgradient $\partial\rho_{\mathrm{C}}$ is non-decreasing (Assumption (A3)) and linear. From (6) the proximal operator is also a linear function. An intuitive choice for $\widetilde{G}(z;b) = b \cdot \partial\rho(x)|_{x=Prox(z;b)}$ with $z \in [u_{\tau_1} + bh(0), u_{\tau_1} + bh(1)]$ is $\widetilde{G}(z;b) = z - b_{\tau_1}$ which keeps the linearity of the composition of the two functions $\partial\rho$ and $\mathrm{Prox}(\cdot;b)$. □

B.2.3. Proof of (16)

Proof. Theorem 2, Eq. (3.7) of Bayati and Montanari [5] states in our notation that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \psi(\varepsilon_i - z_{(t),i}, \varepsilon_i) \overset{a.s.}{=} E[\psi(\bar{\sigma}_{(t)} Z, \varepsilon)], \tag{35}$$

where $\psi : \mathbb{R}^2 \to \mathbb{R}$ is any pseudo-Lipschitz function, $Z \sim N(0,1)$, $\bar{\sigma}_{(t)}$ from (19), and ε as in (A5). Motivated by Eqs. (7.16) and (7.18) in Bradic [8] we take $\psi(d, \varepsilon) = \{G(\varepsilon - d; b_{(t)})\}^2$, with $b_{(t)}$ as in Algorithm 1, step 2. Applying (35) we obtain that as $n \to \infty$

$$\frac{1}{n} \sum_{i=1}^{n} G(\varepsilon_i - (\varepsilon_i - z_{i,(t)}); b_{(t)})^2 = \frac{1}{n} \sum_{i=1}^{n} G(z_i; b_{(t)})^2 \overset{a.s.}{\to} E[G(\varepsilon - \bar{\sigma}_{(t)} Z; b_{(t)})^2]. \quad \square$$

B.2.4. Estimation of $\nu(b)$

The effective score step in Section 3.2, in cases where $G(\cdot; b_{(t)})$ is non-differentiable, requires a solution $b_{(t)}$ to the equation $1 = \widehat{\nu}(b_{(t)})$ where $\widehat{\nu}(b_{(t)})$ is a consistent estimator of a population parameter $\nu(b_{(t)})$ defined as

$$\nu(b_{(t)}) = E[\partial_1 \widetilde{G}(C_{(t)}; b_{(t)})] = b_{(t)}(\delta/\omega)E(\partial[\partial\rho\{\mathrm{Prox}(C_{(t)}; b_{(t)})\}])$$

with $C_{(t)} = \varepsilon - \bar{\sigma}_{(t)}Z$ the random variable characterizing the limit distribution of the adjusted residuals $z_{(t)}$ when $p \to \infty$.

Using Assumption (A3) and Lemma 3 of Bradic [8], $\partial\rho$ can be written as a sum of three functions of which v_1 and v_2 are differentiable. For the step function v_3, we use Assumption (A3) on ρ, where γ_l is the step height on the interval $(r_l, r_{l+1}]$. Let $f_{C_{(t)} - \widetilde{G}(C_{(t)}; b)}$ denote the density of the variable $C_{(t)} - \widetilde{G}(C_{(t)}; b)$ which is equivalent to $\mathrm{Prox}(C_{(t)}; b)$. The equivalence is obtained by setting the derivative of the $b\rho(x) + \frac{1}{2}(x - C_{(t)})^2$ w.r.t. x to zero and evaluate at $\mathrm{Prox}(C_{(t)}; b)$, due to the fact that the proximal operator is the minimizer of the function $b\rho(x) + \frac{1}{2}(x - C_{(t)})^2$. Then we arrive at

$$\frac{\omega\nu(b_{(t)})}{\delta b_{(t)}} = \sum_{j=1}^{2} E[\partial v_j(C_{(t)})] + \sum_{l=1}^{L-1} \gamma_l \{ f_{C_{(t)} - \widetilde{G}(C_{(t)}; b_{(t)})}(r_{l+1}) - f_{C_{(t)} - \widetilde{G}(C_{(t)}; b)}(r_l) \}.$$

The consistent estimator in (12) is obtained by replacing the expectation above with the empirical mean and replacing the density of the proximal operator $\mathrm{Prox}(C_{(t)}; b)$ with its kernel density estimator.

B.2.5. Proof of Lemma 1

Since this proof is based on the general recursion and Lemma 1 in Bayati and Montanari [5], we first restate the general recursion to which Algorithm 1 belongs with slight changes in the notations. Given the noise $\varepsilon \in \mathbb{R}^n$ and the coefficient vector $\beta \in \mathbb{R}^p$, the general recursion is defined

$$\begin{aligned} h_{(t+1)} &= X^\top m_{(t)} - \xi_{1,(t)} q_{(t)}, \quad m_{(t)} = g_{1,t}(d_{(t)}, \varepsilon) \\ d_{(t)} &= X q_{(t)} - \xi_{2,(t)} m_{(t-1)}, \quad q_{(t)} = g_{2,(t)}(h_{(t)}, \beta) \end{aligned}$$

where $\xi_{1,(t)} = n^{-1} \sum_{i=1}^{n} \partial_1 g_{1,(t)}(d_{(t),i}, \varepsilon_i)$, $\xi_{2,(t)} = (\delta p)^{-1} \sum_{j=1}^{p} \partial_1 g_{2,(t)}(h_{(t),j}, \beta_j)$. Further, to connect the general recursion to Algorithm 1, we also state the exact form of $h_{(t+1)}, m_{(t)}, d_{(t)}, q_{(t)}$ taken in Algorithm 1. Lemma 1 in Bradic [8] states that Algorithm 1 takes $h_{(t+1)} = \beta - X^\top G(z_{(t)}; b_{(t)}) - \beta_{(t)}$, $q_{(t)} = \beta_{(t)} - \beta$, from (7) $z_{(t)} = \varepsilon - d_{(t)}$, which defines $d_{(t)}$, $m_{(t)} = -G(z_{(t)}; b_{(t)})$ with the functions $g_{1,(t)}(x_1, x_2) = -G(x_2 - x_1; b_{(t)})$, and $g_{2,(t)}(x_1) = \eta(\beta - x_1; \theta) - \beta$. To proceed with the proof of Lemma 1, we first recall the technique used for proving Lemma 1 in Bayati and Montanari [5], which uses induction on the iteration t. To not

fully repeat the long proof and all notations we only give details about where our proof differs from theirs.

1. $\mathcal{B}_{(0)}$: show properties (3.15), (3.17), (3.19), (3.21), (3.23) and (3.23) of Bayati and Montanari [5] which are related to the vectors $b_{(0)}$ and $m_{(0)}$, by conditioning on the σ-algebra $\mathcal{D}_{(0),(0)}$ generated by $\{\beta, \varepsilon, q_{(0)}\}$; obtain the σ-algebra $\mathcal{D}_{(1),(0)}$ by adding $b_{(0)}$ and $m_{(0)}$ to the set $S_{(0),(0)} = \{\beta, \varepsilon, q_{(0)}\}$.
2. $\mathcal{H}_1$: show that the properties (3.14), (3.16), (3.18), (3.20), (3.22), (3.24) and (3.25), which are related to the vectors $h_{(1)}$ and $q_{(1)}$, hold by conditioning on the σ-algebra $\mathcal{D}_{(1),(0)}$; obtain the σ-algebra $\mathcal{D}_{(1),(1)}$ by adding $h_{(1)}$ and $m_{(1)}$ to the set $S_{(1),(0)} = \{\beta, \varepsilon, q_{(0)}, d_{(0)}, m_{(0)}\}$
3. $\mathcal{B}_{(t)}$: Similar to $\mathcal{B}_{(0)}$; the proof is conditioning on the σ-algebra $\mathcal{D}_{(t),(t)}$ for the set containing $\beta, \varepsilon, q_{(0)}$ and all previous obtained vectors; obtain the new σ-algebra $\mathcal{D}_{(t+1),(t)}$ by adding $b_{(t+1)}$ and $m_{(t+1)}$ to the set.
4. $\mathcal{H}_{(t+1)}$: Similar to $\mathcal{H}_{(1)}$; conditioning on the σ-algebra $\mathcal{D}_{(t+1),(t)}$ for the set containing $\beta, \varepsilon, q_{(0)}$ and all previous obtained vectors.

Assuming Lemma 1 in Bayati and Montanari [5] holds for all K estimators $\widehat{\beta}_k, k = 1, \ldots, K$ in (2), we add an additional step considering the correlations between the estimators. The main technique is conditioning on the σ-algebra generated by $\cup_{k=1}^{K} \mathcal{S}_{k,(1),(0)}$ and $\cup_{k=1}^{K} \mathcal{S}_{k,(t+1),(t)}$, where $\mathcal{S}_{k,(1),(0)}$ and $\mathcal{S}_{k,(t+1),(t)}$ are the sets described in step 2 and 4 above for the kth estimator. The proof is similar to that of (3.16) in Lemma 1(b) of Bayati and Montanari [5], with different mathematical techniques in order to adjust the original proof from a single sequence of iterations to K paralleled sequences of iterations.

Proof. Idea of the construction: The construction of $\mathcal{B}_{(0)}$, $\mathcal{H}_{(0)}$, $\mathcal{B}_{(t+1)}$ and $\mathcal{H}_{(t+1)}$ depends on the space $\mathcal{D}_{(t+1),(t)}$ which is the space generated by the true coefficient β, the noise ε, the initial condition $q_{(0)}$, and the subsequent terms generated from Algorithm 1. The proof by induction is similar to the proof of Lemma 1(b) in Bayati and Montanari [5]. We prove that $\mathcal{H}_{(1)}$ holds and if $\mathcal{B}_{(r)}, \mathcal{H}_{(s)}$ holds for all $r \leq t$ and $s \leq t$, then $\mathcal{H}_{(t+1)}$ holds. Let $o_{k,(t)}(1)$ denote a vector in $\mathbb{R}^t$ for the kth estimator such that all of its entries converge to 0 almost surely for $p \to \infty$.

Step 2 from Bayati and Montanari [5]: $\mathcal{H}_{(1)}$: We know from Eq. (3.35) in Bayati and Montanari [5] that for each k and a Gaussian matrix $\tilde{X}_k$ with the same distribution as the design matrix X, see also Bayati and Montanari [5] Lemma 2 (1),

$$h_{k,(1)}|_{\mathcal{D}_{k,(1),(0)}} \stackrel{d}{=} (\tilde{X}_k)^\top m_{k,(0)} + o_{k,(1)}(1) q_{k,(0)}.$$

Let $a_{k,j} = ([(\tilde{X}_k)^\top m_{k,(0)}]_j + o_{1,k}(1) q_{k,(0),j}, \beta_j)$ and $c_{k,j} = ([(\tilde{X}_k)^\top m_{k,(0)}]_j, \beta_j)$ where $k = k_1, k_2$. We first show that for any two k_1, $k_2 \in \{1, \ldots, K\}$.

$$\lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \left[\tilde{\psi}_{\mathrm{c}}(a_{k_1,j}) \tilde{\psi}_{\mathrm{c}}(a_{k_2,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_1,j}) \tilde{\psi}_{\mathrm{c}}(c_{k_2,j}) \right] = 0. \tag{36}$$

Since $\tilde{\psi}_{\mathrm{c}}$ is κ_{c} order pseudo-Lipschitz, hence we have

$$|\tilde{\psi}_{\mathrm{c}}(a_{k,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k,j})| \le L\{1 + \max(\|a_{k,j}\|^{\kappa_{\mathrm{c}}-1}, \|c_{k,j}\|^{\kappa_{\mathrm{c}}-1})\}|q^0_{k,j}|o_{1,k}(1);$$
$$|\tilde{\psi}_{\mathrm{c}}(a_{k,j})| \le L'(1 + \|a_{k,j}\|^{\kappa_{\mathrm{c}}}), \quad |\tilde{\psi}_{\mathrm{c}}(c_{k,j})| \le L''(1 + \|c_{k,j}\|^{\kappa_{\mathrm{c}}});$$

meanwhile, from the proof in $\mathcal{H}_0$ in Lemma 1 in Bayati and Montanari [5], we have for an arbitrary κ_{c} order pseudo-Lipschitz function $\tilde{\psi}_{\mathrm{c}}$

$$\lim_{p\to\infty} \frac{1}{p}\sum_{j=1}^{p} |\tilde{\psi}_{\mathrm{c}}(a_{k,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k,j})| = 0. \tag{37}$$

Notice that

$$\begin{aligned}
&|\tilde{\psi}_{\mathrm{c}}(a_{k_1,j})\tilde{\psi}_{\mathrm{c}}(a_{k_2,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_1,j})\tilde{\psi}_{\mathrm{c}}(c_{k_2,j})| \\
&\quad = |\tilde{\psi}_{\mathrm{c}}(a_{k_1,j})\tilde{\psi}_{\mathrm{c}}(a_{k_2,j}) - \tilde{\psi}_{\mathrm{c}}(a_{k_2,j})\tilde{\psi}_{\mathrm{c}}(c_{k_1,j}) + \tilde{\psi}_{\mathrm{c}}(a_{k_2,j})\tilde{\psi}_{\mathrm{c}}(c_{k_1,j}) \\
&\qquad - \tilde{\psi}_{\mathrm{c}}(c_{k_1,j})\tilde{\psi}_{\mathrm{c}}(c_{k_2,j})| \\
&\quad \le |\tilde{\psi}_{\mathrm{c}}(a_{k_2,j})||\tilde{\psi}_{\mathrm{c}}(a_{k_1,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_1,j})| + |\tilde{\psi}_{\mathrm{c}}(c_{k_1,j})||\tilde{\psi}_{\mathrm{c}}(a_{k_2,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_2,j})|.
\end{aligned}$$

Then we have

$$\begin{aligned}
&\frac{1}{p}\sum_{j=1}^{p} |\tilde{\psi}_{\mathrm{c}}(a_{k_1,j})\tilde{\psi}_{\mathrm{c}}(a_{k_2,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_1,j})\tilde{\psi}_{\mathrm{c}}(c_{k_2,j})| \\
&\quad \le \frac{1}{p}\sum_{j=1}^{p} |\tilde{\psi}_{\mathrm{c}}(a_{k_2,j})||\tilde{\psi}_{\mathrm{c}}(a_{k_1,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_1,j})| + |\tilde{\psi}_{\mathrm{c}}(c_{k_1,j})||\tilde{\psi}_{\mathrm{c}}(a_{k_2,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_2,j})| \\
&\quad \le \max_j |\tilde{\psi}_{\mathrm{c}}(a_{k_2,j})| \cdot \frac{1}{p}\sum_{j=1}^{p} |\tilde{\psi}_{\mathrm{c}}(a_{k_1,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_1,j})| \\
&\qquad + \max_j |\tilde{\psi}_{\mathrm{c}}(c_{k_1,j})| \cdot \frac{1}{p}\sum_{j=1}^{p} |\tilde{\psi}_{\mathrm{c}}(a_{k_2,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_2,j})| \\
&\quad \le L'_2\{1 + \max_j(\|a_{k_2,j}\|^{\kappa_{\mathrm{c}}})\}\frac{1}{p}\sum_{j=1}^{p} |\tilde{\psi}_{\mathrm{c}}(a_{k_1,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_1,j})| \\
&\qquad + L''_1\{1 + \max_j(\|c_{k_1,j}\|^{\kappa_{\mathrm{c}}})\}\frac{1}{p}\sum_{j=1}^{p} |\tilde{\psi}_{\mathrm{c}}(a_{k_2,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k_2,j})|.
\end{aligned} \tag{38}$$

By (37), for $k = k_1, k_2$, $p^{-1}\sum_{j=1}^{p} |\tilde{\psi}_{\mathrm{c}}(a_{k,j}) - \tilde{\psi}_{\mathrm{c}}(c_{k,j})|$ tends to 0 as $p \to +\infty$. The remaining two factors are finite almost surely: $[(\tilde{X}^k)^\top m_{k,(0)}]_j$ is a Gaussian random variable which is finite almost surely; $\beta_{0,j}$ is finite almost surely since its limiting distribution has bounded moments up to $(2\kappa - 2)$ by assumption (A2). Hence, for any pairs $k_1, k_2 \in \{1, \ldots, K\}$ (36) holds.

From here, we consider $\tilde{h}_{k,(1)}|_{\mathcal{D}_{k,(1),(0)}} \stackrel{d}{=} (\tilde{X}_k)^\top m_{k,(0)}$ of which the components have the same distribution as $\|m_{k,(0)}\|Z_k/\sqrt{n}$ for $Z_k \sim N(0,1)$. Conditioning on $\mathcal{D}_{k_1,(1),(0)}$ and $\mathcal{D}_{k_2,(1),(0)}$, we use the strong law of large numbers for

triangular arrays in Theorem 3 of Bayati and Montanari [5] to obtain that

$$\lim_{p\to\infty}\frac{1}{p}\sum_{j=1}^{p}\Big\{\prod_{r=1}^{2}\tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_r,(1),j},\beta_j)-E_{(\tilde{X}_{k_1},\tilde{X}_{k_2})}[\prod_{r=1}^{2}\tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_r,(1),j},\beta_j)]\Big\}\overset{\text{a.s.}}{=}0. \quad (39)$$

We first prove (39). For $k_1\neq k_2$, we show that the condition in Theorem 3 of Bayati and Montanari [5] holds. To simplify the notation, we denote the independent copies of the matrices $\tilde{X}_{k_1},\tilde{X}_{k_2}$ to be X_{k_1},X_{k_2}. We take the random variables in the triangular array to be

$$\tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_1,(1),j},\beta_j)\tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_2,(1),j},\beta_j)-E_{(\tilde{X}_{k_1},\tilde{X}_{k_2})}[\tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_1,(1),j},\beta_j)\tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_2,(1),j},\beta_j)] \quad (40)$$

and let $0<\rho<1$ then

$$\begin{aligned}
&\frac{1}{p}\sum_{j=1}^{p}E\Big|\prod_{r=1}^{2}\tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_r,(1),j},\beta_j)-E_{(\tilde{X}_{k_1},\tilde{X}_{k_2})}[\prod_{r=1}^{2}\tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_r,(1),j},\beta_j)]\Big|^{2+\rho}\\
&\quad=\frac{1}{p}\sum_{j=1}^{p}E_{(X_{k_1},X_{k_2},\tilde{X}_{k_1},\tilde{X}_{k_2})}\Big[\Big|\tilde{\psi}_{\mathrm{c}}([X_{k_1}^{\top}m_{k_1,(0)}]_j,\beta_j)\tilde{\psi}_{\mathrm{c}}([X_{k_2}^{\top}m_{k_2,(0)}]_j,\beta_j)\\
&\qquad-\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_1}^{\top}m_{k_1,(0)}]_j,\beta_j)\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^{\top}m_{k_2,(0)}]_j,\beta_j)\Big|^{2+\rho}\Big]\\
&\quad=\frac{1}{p}\sum_{j=1}^{p}E\Big[\Big|\tilde{\psi}_{\mathrm{c}}([X_{k_1}^{\top}m_{k_1,(0)}]_j,\beta_j)\tilde{\psi}_{\mathrm{c}}([X_{k_2}^{\top}m_{k_2,(0)}]_j,\beta_j)\\
&\qquad\quad-\tilde{\psi}_{\mathrm{c}}([X_{k_1}^{\top}m_{k_1,(0)}]_j,\beta_j)\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^{\top}m_{k_2,(0)}]_j,\beta_j)\\
&\qquad\quad+\tilde{\psi}_{\mathrm{c}}([X_{k_1}^{\top}m_{k_1,(0)}]_j,\beta_j)\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^{\top}m_{k_2,(0)}]_j,\beta_j)\\
&\qquad\quad-\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_1}^{\top}m_{k_1,(0)}]_j,\beta_j)\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^{\top}m_{k_2,(0)}]_j,\beta_j)\Big|^{2+\rho}\Big]\\
&\quad\le\frac{1}{p}\sum_{j=1}^{p}E\Big[\Big|\tilde{\psi}_{\mathrm{c}}([X_{k_1}^{\top}m_{k_1,(0)}]_j,\beta_j)\tilde{\psi}_{\mathrm{c}}([X_{k_2}^{\top}m_{k_2,(0)}]_j,\beta_j)\\
&\qquad\quad-\tilde{\psi}_{\mathrm{c}}([X_{k_1}^{\top}m_{k_1,(0)}]_j,\beta_j)\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^{\top}m_{k_2,(0)}]_j,\beta_j)\Big|^{2+\rho}\\
&\qquad+\frac{1}{p}\sum_{j=1}^{p}E\Big|\tilde{\psi}_{\mathrm{c}}([X_{k_1}^{\top}m_{k_1,(0)}]_j,\beta_j)\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^{\top}m_{k_2,(0)}]_j,\beta_j)\\
&\qquad\quad-\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_1}^{\top}m_{k_1,(0)}]_j,\beta_j)\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^{\top}m_{k_2,(0)}]_j,\beta_j)\Big|^{2+\rho}\Big]\\
&\quad\le\frac{1}{p}\sum_{j=1}^{p}E\Big[\Big|\tilde{\psi}_{\mathrm{c}}([X_{k_1}^{\top}m_{k_1,(0)}]_j,\beta_j)\Big|^{2+\rho}\\
&\qquad\times\quad\Big|\tilde{\psi}_{\mathrm{c}}([X_{k_2}^{\top}m_{k_2,(0)}]_j,\beta_j)-\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^{\top}m_{k_2,(0)}]_j,\beta_j)\Big|^{2+\rho}\Big]\\
&\qquad+\frac{1}{p}\sum_{j=1}^{p}E\Big[\Big|\tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^{\top}m_{k_2,(0)}]_j,\beta_j)\Big|^{2+\rho}
\end{aligned}$$

$$
\begin{aligned}
&\times \quad \left|\tilde{\psi}_{\mathrm{c}}([X_{k_1}^\top m_{k_1,(0)}]_j, \beta_j) - \tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_1}^\top m_{k_1,(0)}]_j, \beta_j)\right|^{2+\rho}\Big] \\
&\le \frac{1}{p}\sum_{j=1}^{p} E\Big[\Big|L'\Big(1 + |[X_{k_1}^\top m_{k_1,(0)}]_j|^{\kappa_c} + |\beta_j|^{\kappa_c}\Big)\Big|^{2+\rho} \\
&\quad \times \quad \left|\tilde{\psi}_{\mathrm{c}}([X_{k_2}^\top m_{k_2,(0)}]_j, \beta_j) - \tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^\top m_{k_2,(0)}]_j, \beta_j)\right|^{2+\rho}\Big] \\
&\quad + \frac{1}{p}\sum_{j=1}^{p} E\Big[\Big|L'\Big(1 + |[\tilde{X}_{k_2}^\top m_{k_2,(0)}]_j|^{\kappa_c} + |\beta_j|^{\kappa_c}\Big)\Big|^{2+\rho} \\
&\quad \times \quad \left|\tilde{\psi}_{\mathrm{c}}([X_{k_1}^\top m_{k_1,(0)}]_j, \beta_j) - \tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_1}^\top m_{k_1,(0)}]_j, \beta_j)\right|^{2+\rho}\Big] \\
&\le \max_{j=1,\ldots,p} E\Big[\Big|L'\Big(1 + |[X_{k_1}^\top m_{k_1,(0)}]_j|^{\kappa_c} + |\beta_j|^{\kappa_c}\Big)\Big|^{2+\rho}\Big] \\
&\quad \times \quad \frac{1}{p}\sum_{j=1}^{p} E\Big[\left|\tilde{\psi}_{\mathrm{c}}([X_{k_2}^\top m_{k_2,(0)}]_j, \beta_j) - \tilde{\psi}_{\mathrm{c}}([(\tilde{X}_{k_2})^\top m_{k_2,(0)}]_j, \beta_j)\right|^{2+\rho}\Big] \\
&\quad + \max_{j=1,\ldots,p} E\Big[\Big|L'\Big(1 + |[\tilde{X}_{k_2}^\top m_{k_2,(0)}]_j|^{\kappa_c} + |\beta_j|^{\kappa_c}\Big)\Big|^{2+\rho}\Big] \\
&\quad \times \quad \frac{1}{p}\sum_{j=1}^{p} E\Big[\left|\tilde{\psi}_{\mathrm{c}}([X_{k_1}^\top m_{k_1,(0)}]_j, \beta_j) - \tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_1}^\top m_{k_1,(0)}]_j, \beta_j)\right|^{2+\rho}\Big].
\end{aligned}
$$

For the first term in the last inequality above, we see that the expectation $E[|L'(1 + |[\tilde{X}_{k_2}^\top m_{k_2,(0)}]_j|^{\kappa_c} + |\beta_j|^{\kappa_c})|^{2+\rho}]$ is bounded by some constant, since the expectation is with respect to the matrices $X_{k_1}, X_{k_2}, \tilde{X}_{k_1}, \tilde{X}_{k_2}$ of which the components are Gaussian distributed with mean 0 and variance $1/n$; the rest terms are bounded by a constant; the moments of Gaussian distributed r.v. are all finite. Let us denote the upper bound of this expectation by L'', then the first term of the inequality above is bounded by

$$
L''\frac{1}{p}\sum_{j=1}^{p} E\Big[\left|\tilde{\psi}_{\mathrm{c}}([X_{k_2}^\top m_{k_2,(0)}]_j, \beta_j) - \tilde{\psi}_{\mathrm{c}}([\tilde{X}_{k_2}^\top m_{k_2,(0)}]_j, \beta_j)\right|^{2+\rho}\Big],
$$

which can be shown to be bounded by $cp^{\rho/2}$ following a similar argument as in Lemma 1(b) in Bayati and Montanari [5]. The second term similarly can be shown to be bounded by $c'p^{\rho/2}$. Hence the variable defined in (40) satisfies the condition in Theorem 3 in Bayati and Montanari [5]; thus the a.s. convergence holds.

In the special case where $k_1 = k_2$, we show that the square of ψ_c is still pseudo-Lipschitz of order $2\kappa_c \le \kappa$, then the almost sure convergence hold by directly applying the result in Lemma 1 in Bayati and Montanari [5].

To simplify the notation, we use ψ to denote any pseudo-Lipschitz function here. For any pairs $x, y \in \mathbb{R}^m$, we have

$$
\begin{aligned}
&|\psi^2(x) - \psi^2(y)| \\
&\quad \le |\psi(x) + \psi(y)||\psi(x) - \psi(y)| \le (|\psi(x)| + |\psi(y)|)|\psi(x) - \psi(y)|
\end{aligned}
$$

$$\begin{aligned}
&\leq L'(1+\|x\|^{\kappa}+1+\|y\|^{\kappa})\cdot L(1+\|x\|^{\kappa-1}+\|y\|^{\kappa-1})\|x-y\| \\
&\leq LL''(1+\|x\|^{\kappa}+\|y\|^{\kappa})(1+\|x\|^{\kappa-1}+\|y\|^{\kappa-1})\|x-y\| \\
&\leq LL''(1+\|x\|+\|y\|)^{2\kappa-1}\|x-y\| \\
&\leq LL''3^{\kappa-1}(1+\|x\|^{2\kappa-1}+\|y\|^{\kappa-1})\|x-y\|.
\end{aligned}$$

Since $\kappa \geq 1$, $\|x\|, \|y\| \geq 0$, the last two inequalities are obtained by applying the first and second inequality in Lemma 2, respectively. Hence, the square of any arbitrary pseudo-Lipschitz function of order κ is still pseudo-Lipschitz with order 2κ. This proves (39).

Using Lemma 5 for $v = \beta$ and

$$\psi(\beta_j) = E_{(\tilde{X}_{k_1}, \tilde{X}_{k_2})} \tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_1,(1),j}, \beta_j) \tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_2,(1),j}, \beta_j),$$

the following convergence holds

$$\begin{aligned}
&\lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} E_{(\tilde{X}_{k_1}, \tilde{X}_{k_2})} \Big[\tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_1,(1),j}, \beta_j) \tilde{\psi}_{\mathrm{c}}(\tilde{h}_{k_2,(1),j}, \beta_j) \Big] \\
&\quad \overset{\text{a.s.}}{=} E_{B_0} \Big[E_{(Z_{k_1,(0)}, Z_{k_2,(0)})} \Big[\prod_{r=1}^{2} \tilde{\psi}_{\mathrm{c}}(\| \frac{m_{k_r,(0)}}{\sqrt{n}} \| Z_{k_r,(0)}, B_0) \Big] \Big] \\
&\quad \overset{\text{a.s.}}{=} E \Big[\prod_{r=1}^{2} \tilde{\psi}_{\mathrm{c}}(\bar{\zeta}_{k_r,(0)} Z_{k_r,(0)}, B_0) \Big].
\end{aligned}$$

Step 4 from Bayati and Montanari [5]: $\mathcal{H}_{t+1}$: Following the first expression in the proof of Lemma 1(b) in step 4 in Bayati and Montanari [5], for any index $k = 1, \ldots, K$

$$\tilde{\psi}_{\mathrm{c}}(h_{k,(1),j}, \ldots, h_{k,(t+1),j}, \beta_j)|_{\mathcal{D}_{k,(t+1),(t)}} \overset{d}{=}$$

$$\tilde{\psi}_{\mathrm{c}}\Big(h_{k,(1),j}, \ldots, h_{k,(t),j}, \Big[\sum_{r=0}^{t-1} \alpha_r h_{k,(r+1)} + (\tilde{X}_k)^{\top} m_{k,(t)} + \tilde{Q}_{k,(t+1)} o_{k,(t+1)}(1)\Big]_j, \beta_j\Big).$$

The columns of $\tilde{Q}_{k,(t+1)}$ form an orthogonal basis for the column space of $Q_{k,(t+1)} = [q_{k,(0)} \ldots q_{k,(t)}]$. Define the matrix $M_{k,(t)} = [m_{k,(0)} \ldots m_{k,(t-1)}]$, the vector $(m_{k,(t)})_{\|} = \sum_{r=0}^{t-1} \delta_r m_{k,(r)}$ as the projection of $m_{k,(t)}$ on the column space of $M_{k,(t)}$ and the vector $(m_{k,(t)})_{\perp} = m_{k,(t)} - (m_{k,(t)})_{\|}$. Similar to the proof in $\mathcal{H}_1$, we first show that the error term $\tilde{Q}_{k,(t+1)} o_{k,(t+1)}(1)$ can be dropped. Let $a_{k,j} =$

$$\Big(h_{k,(1),j}, \ldots, h_{k,(t),j}, \Big[\sum_{r=0}^{t-1} \delta_r h_{k,(r+1)} + (\tilde{X}_k)^{\top} (m_{k,(t)})_{\perp} + \tilde{Q}_{k',(t+1)} o_{k,(t+1)}(1)\Big]_j, \beta_j\Big)$$

and $c_{k,j} = \Big(h_{k,(1),j}, \ldots, h_{k,(t),j}, \Big[\sum_{r=0}^{t-1} \delta_r h_{k,(r+1)} + (\tilde{X}_k)^{\top} (m_{k,(t)})_{\perp}\Big]_j, \beta_j\Big)$.

To show that the left hand-side of (38) is finite for the new $a_{k,j}$ and $c_{k,j}$, it suffices to show that both $\max_j(\|a_{k_2,j}\|^{\kappa_{\mathrm{c}}})$ and $\max_j(\|c_{k_1,j}\|^{\kappa_{\mathrm{c}}})$ are finite almost

surely. By Lemma 2, we obtain the following inequality

$$\begin{aligned}\max_j(\|a_{k_2,j}\|^{\kappa_c}) &= \max_j\Big(C(\sum_{r=0}^{t}|h_{k_2,(r+1),j}|^{\kappa_c}+|\beta_j|^{\kappa_c})\Big)\\ &\le C\Big(\sum_{r=0}^{t}\max_j|h_{k_2,(r+1),j}|^{\kappa_c}+\max_j|\beta_j|^{\kappa_c}\Big)\end{aligned}$$

for some constant C. The finiteness of $\max_j|\beta_j|^{\kappa_c}$ has been discussed in $\mathcal{H}_1$; $\max_j|h_{k_2,(r+1),j}|$ is finite almost surely since Lemma 1 in Bayati and Montanari [5] states that for a higher order $l=k-1$, $\lim\sum_{p\to\infty}p^{-1}\sum_{j=1}^{p}(h_{k_2,(t+1),j})^{2l}<\infty$. The almost-sure finiteness of $\max_j|h_{k_2,(r+1),j}|$ follows by a simple contradiction: assume $P(\max_j|h_{k_2,(r+1),j}|=\infty)=P(|h_{k_2,(r+1),j_{\max}}|=\infty)>0$, then

$$\begin{aligned}&P\Big(\sup_{p'\ge p}\frac{1}{p'}\sum_{j=1}^{p'}(h_{k_2,(t+1),j})^{2l}<\infty\Big)\\ &\quad=P\Big(\sup_{p'\ge p}\frac{p'-1}{p'}\{\frac{1}{p'-1}\sum_{j\neq j_{\max}}(h_{k_2,(t+1),j})^{2l}\}+\frac{1}{p'}(h_{k_2,(t+1),j_{\max}})^{2l}<\infty\Big)<1.\end{aligned}$$

The above equation contradicts the result in Lemma 1(e) in Bayati and Montanari [5]. Follow similar arguments, we have $\max_j(\|c_{k_1,j}\|^{\kappa_c})$ finite almost surely. Now we consider the random variable

$$\begin{aligned}\tilde{A}_{k,j}=\tilde{\psi}_{\mathrm{c}}\Big(&h_{k,(1),j},\dots,h_{k,(t),j},\\ &\Big[\sum_{r=0}^{t-1}\delta_r h_{k,(r+1)}+(\tilde{X}_k)^{\top}(m_{k,(t)})_{\perp}+\tilde{Q}_{k,(t+1)}o_{k,(t+1)}(1)\Big]_j,\beta_j\Big).\end{aligned}$$

Following arguments as in $\mathcal{H}_1$, it is easy to show that

$$\lim_{p\to\infty}\frac{1}{p}\sum_{j=1}^{p}\Big[\tilde{A}_{k_1,j}\tilde{A}_{k_2,j}-E_{(\tilde{X}_{k_1},\tilde{X}_{k_2})}\tilde{A}_{k_1,j}\tilde{A}_{k_2,j}\Big]\overset{\text{a.s.}}{=}0. \tag{41}$$

By Lemma 5 and arguments as in the proof of Lemma 1 (b) in Bayati and Montanari [5],

$$\begin{aligned}&\lim_{p\to\infty}\frac{1}{p}\sum_{j=1}^{p}\tilde{\psi}_{\mathrm{c}}\Big(h_{k_1,(1),j},\dots,h_{k_1,(t),j},[\sum_{r=0}^{t-1}\delta_{k_1,(r)}h_{k_1,(r+1)}+(\tilde{X}_{k_1})^{\top}(m_{k_1,(t)})_{\perp}]_j,\beta_j\Big)\\ &\qquad\times\tilde{\psi}_{\mathrm{c}}\Big(h_{k_2,(1),j},\dots,h_{k_2,(t),j},[\sum_{r=0}^{t-1}\delta_{k_2,(r)}h_{k_2,(r+1)}+(\tilde{X}_{k_2})^{\top}(m_{k_2,(t)})_{\perp}]_j,\beta_j\Big)\\ &\overset{\text{a.s.}}{=}E_{B_0}E_{(Z_{k_1,(0)},\dots,Z_{k_1,(t)},Z_{k_2,(0)},\dots,Z_{k_2,(t)})}\\ &\qquad\Big[\prod_{r=1}^{2}\tilde{\psi}_{\mathrm{c}}\Big(\bar{\zeta}_{k_r,(0)}Z_{k_r,(0)},\dots,\bar{\zeta}_{k_r,(t)}Z_{k_r,(t)},B_0\Big)\Big]\end{aligned}$$

$$= E\left[\prod_{r=1}^{2} \tilde{\psi}_{\mathrm{c}}\left(\bar{\zeta}_{k_r,(0)} Z_{k_r,(0)}, \ldots, \bar{\zeta}_{k_r,(t)} Z_{k_r,(t)}, B_0\right)\right]. \qquad \square$$

B.2.6. Proof of Corollary 2

Proof. The almost sure convergence holds by choosing $\tilde{\psi}_{\mathrm{c}}(y_{(0)}, \ldots, y_{(t)}, \beta_j) = \psi_{\mathrm{c}}(y_{(t)}, \beta_j) = \left(\beta_j - y_{(t)}\right) - \beta_j$ in Lemma 1. $\square$

B.2.7. Proof of Theorem 1

Proof. By Lemma 1 and choosing $\tilde{\psi}_{\mathrm{c}}(y_{(0)}, \ldots, y_{(t)}, \beta_j) = \psi_{\mathrm{c}}(y_{(t)}, \beta_j) = \eta(\beta_j - y_{(t)}; \theta_{(t)}) - \beta_j$ which is a pseudo-Lipschitz function of order $\kappa_c = 1$ the convergence in (26) is obtained. $\square$

B.2.8. Proof of Theorem 2

Proof. Theorem 2 in Bayati and Montanari [5] showed that when assigning $1/p$ point mass to each entry of the vector, $\widetilde{\beta}_{k,j,(t-1)}(p)$ converges weakly to $B_0 + \bar{\zeta}_{k,(t-1)} Z_k$ for $p \to \infty$ where $Z_k \sim N(0,1)$ and B_0 has p.d.f. f_{B_0}. When p is large, $\widetilde{\beta}_{k,(t-1)} \mid (B_0 = \beta) \approx N(\beta, \bar{\zeta}^2_{k,(t-1)} I_p)$; the normality comes from $Z_k \sim N(0,1)$. Similar results for the Lasso estimator can be found in Bayati et al. [4] and Donoho and Montanari [14]. The normality of $\widetilde{\beta}_{k,(t-1)}$ ensures that the Stein's unbiased risk estimate is applicable for constructing the AMSE estimator.

Next, consider any pair (k_1, k_2) with $k_1, k_2 \in \{1, \ldots, K\}$ at iteration $t-1$. The conditional normality holds for $\widetilde{\beta}_{k_r,(t-1)}$ $(r = 1, 2)$. Each component of the estimator $\widetilde{\beta}_{k_r,(t-1)}$ is independent of the remaining entries. Hence, the dependence between $\widetilde{\beta}_{k_1,(t-1)}$ and $\widetilde{\beta}_{k_2,(t-1)}$ comes from the entry-wise dependence of the two variables. In other words, there is only dependence between $\widetilde{\beta}_{k_1,(t-1),j_1}$ and $\widetilde{\beta}_{k_2,(t-1),j_2}$ when $j_1 = j_2$. The covariance between the two estimators is

$$\bar{\zeta}_{(k_1,k_2),(t-1)} = \mathrm{Cov}(\widetilde{\beta}_{k_1,(t-1)}, \widetilde{\beta}_{k_2,(t-1)}).$$

Meanwhile, since $\bar{\zeta}^2_{\mathrm{emp},(t)} = \bar{\zeta}^2_{\mathrm{emp},(t-1)} + o(1)$ by assumption, $\theta_{k_r,(t)} = \alpha \bar{\zeta}_{k_r,(t)}$ where α is fixed for the different iterations, we obtain

$$\begin{aligned}
&\lim_{p\to\infty} \frac{1}{p} \sum_{j=1}^{p} |\widehat{\beta}_{k_r,(t),j} - \widehat{\beta}_{k_r,(t-1),j}| \\
&\overset{a.s}{=} E\left|\eta(B_0 + \bar{\zeta}_{k_r,(t)} Z_{k_r,(t),j}; \theta_{k_r,(t)}) - \eta(B_0 + \bar{\zeta}_{k_r,(t-1)} Z_{k_r,(t-1),j}; \theta_{k_r,(t-1)})\right| \\
&= E\left|\eta(B_0 + \bar{\zeta}_{k_r,(t)} Z_{k_r,(t),j}; \theta_{k_r,(t)}) - \eta(B_0 + \bar{\zeta}_{k_r,(t)} Z_{k_r,(t-1),j}; \theta_{k_r,(t)}) + o(1)\right| \\
&= 0.
\end{aligned}$$

The almost sure convergence holds by Lemma 1(b) [5]. The next equality holds by $\bar{\zeta}^2_{(t)} = \bar{\zeta}^2_{(t-1)} + o(1)$ and the definition of $\theta_{k_r,(t)}$. The last equality holds because both $Z_{k_r,(t-1),j}$ and $Z_{k_r,(t),j}$ are standard Gaussian distributed. Thus, $\widehat{\beta}_{k_r,(t),j} - \widehat{\beta}_{k_r,(t-1),j} | (B_0 = \beta_j)$ converges to 0 almost surely.

Further, by (14), $\widehat{\beta}_{k_r,(t-1),j} - \widetilde{\beta}_{k_r,(t-1),j} \overset{d}{=} \bar{\zeta}_{k_r,(t-1)} Z_{k,j}$ where $Z_{k_r,j} \sim N(0,1)$. Then, $\widehat{\beta}_{k_r,(t),j} - \widetilde{\beta}_{k_r,(t-1),j} = (\widehat{\beta}_{k_r,(t),j} - \widehat{\beta}_{k_r,(t-1),j}) + (\widehat{\beta}_{k_r,(t-1),j} - \widetilde{\beta}_{k_r,(t-1),j}) \overset{d}{=} \bar{\zeta}_{k_r,(t-1)} Z_{k,j}$, where $\widehat{\beta}_{k_r,(t),j} = \eta(\widetilde{\beta}_{k_r,(t-1),j}; \theta_{k_r,(t-1)})$, by Slutsky's theorem. This suggests that $\widehat{\beta}_{k_r,(t),j} - \widetilde{\beta}_{k_r,(t-1),j}$ converges in distribution to $\bar{\zeta}_{k_r,(t-1)} Z_{k,j}$ which is Gaussian distributed. Next, Stein's lemma stated in Lemma 4 [35] is applied. Notice that $\widehat{\beta}_{k_r,(t),j} - \widetilde{\beta}_{k_r,(t-1),j}$ and $\widetilde{\beta}_{k_2,(t-1),j} - \beta_j$ are jointly Gaussian distributed; further, the univariate function $g : x \to \big(\eta(x;\theta) - x\big)$ satisfies the condition in Lemma 4 [35]. We apply Lemma 4 to the jointly Gaussian distributed pairs $\widehat{\beta}_{k_r,(t),j} - \widetilde{\beta}_{k_r,(t-1),j}$ and $\widetilde{\beta}_{k_2,(t-1),j} - \beta_j$, $j = 1, \ldots, p$ with the univariate function g. We denote by A_j, conditioning on $(B_0 = \beta_j)$ and $\widetilde{\beta}_{k_r,(t-1),-j}, r = 1, 2$. It holds that

$$
\begin{aligned}
&E\Big[\{\eta(\widetilde{\beta}_{k_1,(t-1),j}; \theta_{k_1,(t-1)}) - \widetilde{\beta}_{k_1,(t-1),j}\}(\widetilde{\beta}_{k_2,(t-1),j} - \beta_j) | A_j\Big] \\
&\quad = \mathrm{Cov}\Big(\eta(\widetilde{\beta}_{k_1,(t-1),j}; \theta_{k_1,(t-1)}) - \widetilde{\beta}_{k_1,(t-1),j}, \widetilde{\beta}_{k_2,(t-1),j} | A_j\Big) \\
&\quad = \mathrm{Cov}(\widetilde{\beta}_{k_1,(t-1),j}, \widetilde{\beta}_{k_2,(t-1),j} | A_j) E[\partial_1 \eta(\widetilde{\beta}_{k_1,(t-1),j}; \theta_{k_1,(t-1)}) - 1 | A_j].
\end{aligned}
$$

Below we condition everywhere on B which denotes the event that $B_{0,j} = \beta_j$ for $j = 1, \ldots, p$ where $B_{0,j}$ are independent copies of B_0. Taking expectation w.r.t. $\widetilde{\beta}_{k_r,(t-1),-j}$, we obtain for the whole vector,

$$
\begin{aligned}
&E\Big[\{\eta(\widetilde{\beta}_{k_1,(t-1)}; \theta_{k_1,(t-1)}) - \widetilde{\beta}_{k_1,(t-1)}\}(\widetilde{\beta}_{k_2,(t-1)} - \beta) | B\Big] \\
&\quad = \bar{\zeta}_{(k_1,k_2),(t-1)} E[\partial_1 \eta(\widetilde{\beta}_{k_1,(t-1)}; \theta_{k_1,(t-1)}) - \mathbf{1}_p | B] \\
&E\Big[\{\eta(\widetilde{\beta}_{k_1,(t-1)}; \theta_{k_1,(t-1)}) - \widetilde{\beta}_{k_1,(t-1)}\}(\widetilde{\beta}_{k_2,(t-1)} - \beta) | B\Big] \\
&\quad = \bar{\zeta}_{(k_1,k_2),(t-1)} E[\partial_1 \eta(\widetilde{\beta}_{k_2,(t-1)}; \theta_{k_2,(t-1)}) - \mathbf{1}_p | B].
\end{aligned}
$$

Next, we show the construction of the estimator for $(\Sigma_0)_{(k_1,k_2),(t)}$ at iteration t. The product-sign notation $\prod_{r=1}^2 v_r = v_1^\top v_2$.

$$
\begin{aligned}
&E[(\widehat{\beta}_{k_1,(t)} - \beta)^\top (\widehat{\beta}_{k_2,(t)} - \beta) | B] = E\big[\prod_{r=1}^2 \{\eta(\widetilde{\beta}_{k_r,(t-1)}; \theta_{k_r,(t-1)}) - \beta\} | B\big] \\
&= E\big[\prod_{r=1}^2 \{\eta(\widetilde{\beta}_{k_r,(t-1)}; \theta_{k_r,(t-1)}) - \widetilde{\beta}_{k_r,(t-1)}\} | B\big] + E\big[\prod_{r=1}^2 (\widetilde{\beta}_{k_r,(t-1)} - \beta) | B\big] \\
&\quad + E\big[\{\eta(\widetilde{\beta}_{k_1,(t-1)}; \theta_{k_1,(t-1)}) - \widetilde{\beta}_{k_1,(t-1)}\}^\top (\widetilde{\beta}_{k_2,(t-1)} - \beta) | B\big] \\
&\quad + E\big[(\widetilde{\beta}_{k_1,(t-1)} - \beta)^\top \{\eta(\widetilde{\beta}_{k_2,(t-1)}; \theta_{k_2,(t-1)}) - \widetilde{\beta}_{k_2,(t-1)}\} | B\big]
\end{aligned}
$$

$$
\begin{aligned}
&= E\Big[\prod_{r=1}^{2}\{\eta(\widetilde{\beta}_{k_r,(t-1)};\theta_{k_r,(t-1)}) - \widetilde{\beta}_{k_r,(t-1)}\}|B\Big] + \bar{\zeta}_{(k_1,k_2),(t-1)} \\
&\quad + \bar{\zeta}_{(k_1,k_2),(t-1)} \sum_{r=1}^{2} E\big[\partial_1\eta(\widetilde{\beta}_{k_r,(t-1)};\theta_{k_r,(t-1)}) - \mathbf{1}_p|B\big] \\
&= -\bar{\zeta}_{(k_1,k_2),(t-1)} + E\Big[\prod_{r=1}^{2}\{\eta(\widetilde{\beta}_{k_r,(t-1)};\theta_{k_r,(t-1)}) - \widetilde{\beta}_{k_r,(t-1)}\}|B\Big] \\
&\quad + \bar{\zeta}_{(k_1,k_2),(t-1)} \sum_{r=1}^{2} E\big[\partial_1\eta(\widetilde{\beta}_{k_r,(t-1)};\theta_{k_r,(t-1)})|B\big].
\end{aligned}
$$

Replacing the expectations and the covariance $\bar{\zeta}_{(k_1,k_2),(t-1)}$ with their corresponding empirical versions leads to the unbiased estimator of $(\Sigma_0)_{(k_1,k_2),(t)}$,

$$
\begin{aligned}
(\widehat{\Sigma}_{0,(t)}(p))_{(k_1,k_2)} &= -\bar{\zeta}_{\mathrm{emp},(k_1,k_2),(t-1)} \\
&\quad + \frac{1}{p}\sum_{j=1}^{p}\prod_{r=1}^{2}\{\eta(\widetilde{\beta}_{k_r,(t-1),j};\theta_{k_r,(t-1)}) - \widetilde{\beta}_{k_r,(t-1),j}\} \\
&\quad + \frac{\bar{\zeta}_{\mathrm{emp},(k_1,k_2),(t-1)}}{p}\sum_{j=1}^{p}\sum_{r=1}^{2} I\{|\widetilde{\beta}_{k_r,(t-1),j}| \geq \theta_{k_r,(t-1)}\}.
\end{aligned}
$$

The consistency of the estimator $(\widehat{\Sigma}_{0,(t)}(p))_{(k_1,k_2)}$ follows since

$$
\lim_{p\to\infty}(\widehat{\Sigma}_{0,(t)}(p))_{(k_1,k_2)} = \lim_{p\to\infty}(\Sigma_{0,(t)}(p))_{(k_1,k_2)} = (\Sigma_{(t)})_{(k_1,k_2)}
$$

holds with probability one for for all $k_1, k_2 = 1, \ldots, K$. The first equality follows by the unbiasedness of $(\widehat{\Sigma}_{0,(t)}(p))_{(k_1,k_2)}$ for $(\Sigma_{0,(t)}(p))_{(k_1,k_2)}$, and the second equality holds by Lemma 1. The proof is completed by realizing the above equality shows almost sure convergence which indicates convergence in probability. □

B.2.9. Proof of Theorem 3

Proof. Under the assumption that $n > p$, the model-averaged estimator is asymptotically unbiased. Hence

$$
\mathrm{AMSE}(\widehat{\beta}_{\mathrm{MA}},\beta) = \lim_{n,p\to\infty}\frac{1}{p}\sum_{j=1}^{p}\mathrm{Var}(\widehat{\beta}_{\mathrm{MA},j}) \overset{a.s.}{=} w^{\top}\Sigma_{(\infty)}w, \tag{42}
$$

where $\Sigma_{(\infty)}$ is a $K \times K$ matrix with (k_1, k_2)th component

$$
\begin{aligned}
\Sigma_{(k_1,k_2)} &= E\big[\{I(B_0 + \bar{\zeta}_{k_1}Z_{k_1}) - B_0\}\{I(B_0 + \bar{\zeta}_{k_2}Z_{k_2}) - B_0\}\big] \\
&= \mathrm{Cov}\big(\bar{\zeta}_{k_1}Z_{k_1}, \bar{\zeta}_{k_2}Z_{k_2}\big) = \mathrm{Cov}(Z_{k_1}, Z_{k_2})\bar{\zeta}_{k_1}\bar{\zeta}_{k_2}.
\end{aligned} \tag{43}
$$

Combining (20) and (21), we obtain that

$$\begin{aligned}\bar{\zeta}_k &= \delta\{E[\widetilde{G}(\varepsilon + \bar{\zeta}_k Z_k; b_k)^2]\}^{1/2} \\ &= \{E[\widetilde{G}(\varepsilon + \bar{\zeta}_k Z_k; b_k)^2]\}^{1/2}\{E[\partial_1 \widetilde{G}(\varepsilon + \bar{\zeta}_k Z_k; b_k)]\}^{-1}. \end{aligned} \tag{44}$$

The expressions of the asymptotic variance of the model-averaged estimator in Theorem 3 hold by combining (42), (43), and (44). □

Acknowledgements

The authors thank the reviewers for the useful comments which helped improve the paper. Gerda Claeskens and Jing Zhou acknowledge the support of the Research Foundation Flanders and KU Leuven grant GOA/12/14. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Hercules Foundation and the Flemish Government – department EWI. Jelena Bradic acknowledges the support of the National Science Foundation's Division of Mathematical Sciences grant #1712481.

References

[1] Ando, T. and Li, K.-C. (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association*, 109(505):254–265. MR3180561

[2] Ando, T. and Li, K.-C. (2017). A weight-relaxed model averaging approach for high-dimensional generalized linear models. *The Annals of Statistics*, 45(6):2654–2679. MR3737905

[3] Bates, J. M. and Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly*, 20:451–468. MR0295497

[4] Bayati, M., Erdogdu, M., and Montanari, A. (2013). Estimating lasso risk and noise level. In *Advances in Neural Information Processing Systems*, pages 944–952.

[5] Bayati, M. and Montanari, A. (2011a). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785. MR2810285

[6] Bayati, M. and Montanari, A. (2011b). The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017. MR2951312

[7] Bloznelis, D., Claeskens, G., and Zhou, J. (2019). Composite versus model-averaged quantile regression. *Journal of Statistical Planning and Inference*, 200:32–46. MR3907267

[8] Bradic, J. (2016). Robustness in sparse high-dimensional linear models: Relative efficiency and robust approximate message passing. *Electronic Journal of Statistics*, 10(2):3894–3944. MR3581957

[9] Bradic, J., Fan, J., and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the*

Royal Statistical Society: Series B (Statistical Methodology), 73(3):325–349. MR2815779

[10] Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York. MR1919620

[11] Cheng, G., Wang, S., and Yang, Y. (2015). Forecast combination under heavy-tailed errors. *Econometrics*, 3:797–824.

[12] Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge. MR2431297

[13] Donoho, D., Maleki, A., and Montanari, A. (2009). Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919. MR2241189

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

[15] Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., Beale, C. M., Ciuti, S., Elith, J., Gerstner, K., Guelat, J., Keil, P., Lahoz-Monfort, J. J., Pollock, L. J., Reineking, B., Roberts, D. R., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Wood, S. N., Wüest, R. O., and Hartig, F. (2018). Model averaging in ecology: a review of bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs*, 88(4):485–504.

[16] El Karoui, N., Bean, D., Bickel, P. J., Lim, C., and Yu, B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562.

[17] Eldar, Y. C. and Kutyniok, G. (2012). *Compressed sensing: theory and applications*. Cambridge University Press. MR3095915

[18] Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, 75:1175–1189. MR2333497

[19] Hansen, B. E. and Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, 167:38–46. MR2885437

[20] Höge, M., Guthke, A., and Nowak, W. (2019). The hydrologist's guide to bayesian model selection, averaging and combination. *Journal of Hydrology*, 572:96–107.

[21] Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *J. Am. Statist. Assoc.*, 98:879–899. With discussion and a rejoinder by the authors. MR2041481

[22] Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14:382–417. With discussion and a rejoinder by the authors. MR1765176

[23] Huang, H. (2020). Asymptotic risk and phase transition of l_1-penalized robust estimator. *The Annals of Statistics, to appear*.

[24] Jameson, G. (2014). Some inequalities for (a + b) p and (a + b) p+(a- b) p. *The Mathematical Gazette*, 98(541):96–103. MR4022630

[25] Javanmard, A. and Montanari, A. (2014a). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine*

Learning Research, 15(1):2869–2909. MR3277152

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

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

[28] Kiefer, J. (1953). Sequential minimax search for a maximum. *Proceedings of the American mathematical society*, 4(3):502–506. MR0055639

[29] Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press. MR2268657

[30] Lei, L., Bickel, P. J., and El Karoui, N. (2018). Asymptotics for high dimensional regression m-estimates: fixed design results. *Probability Theory and Related Fields*, 172(3-4):983–1079. MR3877551

[31] Mousavi, A., Maleki, A., and Baraniuk, R. G. (2013). Parameterless optimal approximate message passing. *arXiv preprint* 1311.0035.

[32] Mousavi, A., Maleki, A., Baraniuk, R. G., et al. (2018). Consistent parameter estimation for lasso and approximate message passing. *The Annals of Statistics*, 46(1):119–148. MR3766948

[33] Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239.

[34] Rao, R. C. (1973). *Linear statistical inference and its applications*, volume 2. Wiley New York. MR0346957

[35] Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, pages 1135–1151. MR0630098

[36] Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., et al. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202. MR3224285

[37] Yuan, Z. and Yang, Y. (2005). Combining linear regression models: When and how? *Journal of the American Statistical Association*, 100(472):1202–1214. MR2236435

[38] Zhao, S., Zhou, J., and Li, H. (2016). Model averaging with high-dimensional dependent data. *Economics Letters*, 148:68–71. MR3566055