A general family of trimmed estimators for robust high-dimensional data analysis

Eunho Yang

School of Computing, Korea Advanced Institute of Science and Technology, South Korea
e-mail: eunhoy@kaist.ac.kr

Aurélie C. Lozano

Mathematical Sciences Department, IBM T.J. Watson Research Center, USA
e-mail: aclozano@us.ibm.com

Aleksandr Aravkin

Department of Applied Mathematics, University of Washington, USA
e-mail: saravkin@uw.edu

Abstract: We consider the problem of robustifying high-dimensional structured estimation. Robust techniques are key in real-world applications which often involve outliers and data corruption. We focus on trimmed versions of structurally regularized M-estimators in the high-dimensional setting, including the popular Least Trimmed Squares estimator, as well as analogous estimators for generalized linear models and graphical models, using convex and non-convex loss functions. We present a general analysis of their statistical convergence rates and consistency, and then take a closer look at the trimmed versions of the Lasso and Graphical Lasso estimators as special cases. On the optimization side, we show how to extend algorithms for M-estimators to fit trimmed variants and provide guarantees on their numerical convergence. The generality and competitive performance of high-dimensional trimmed estimators are illustrated numerically on both simulated and real-world genomics data.

Keywords and phrases: Lasso, robust estimation, high-dimensional variable selection, sparse learning.

1. Introduction

We consider the problem of high-dimensional estimation, where the number of variables p is much higher than the number of observations n. These problems arise in a variety of domains, including signal processing, computational biology and finance. The development and the statistical analysis of structurally constrained estimators for high-dimensional estimation has recently attracted considerable attention. These estimators seek to minimize the sum of a loss function and a weighted regularizer. The most popular example is that of

Lasso [43], which solves an ℓ_1-regularized (or equivalently ℓ_1-constrained) least squares problem. Under sub-Gaussian errors, Lasso has been shown to have strong statistical guarantees [44, 46]. Regularized maximum likelihood estimators (MLEs) have been developed for sparsity-structured Generalized Linear Models (GLMs), with theoretical guarantees such as ℓ_1 and ℓ_2-consistency [29], and model selection consistency [10]. For matrix-structured regression problems, estimators using nuclear-norm regularization have been studied e.g. by [36]. Another prime example is that of sparse inverse covariance estimation for graphical model selection [35].

In practice, however, the desirable theoretical properties of such regularized M-estimators can be compromised, since outliers and corruptions are often present in high-dimensional data problems. These challenges motivate the development of robust structured learning methods that can cope with observations deviating from the model assumptions. The problem of reliable high-dimensional estimation under possibly gross error has gained increasing attention. Relevant prior work includes the "extended" Lasso formulation [31, 53] which incorporates an additional sparse error vector to the original Lasso problem so as to account for corrupted observations, the LAD-Lasso [47] which uses the least absolute deviation combined with an ℓ_1 penalty, the "matrix uncertainty selectors" [38] which consider the case where the predictor matrix is observed with additive error, and the Robust Matching Pursuit method of Chen et al. [12] which performs feature selection based on a trimmed inner product of the features with the residuals to mitigate the impact of corrupted observations. Extending M-estimators beyond the least squares case along these directions is challenging. For example, Yang et al. [51], Tibshirani and Manning [42] extend the strategy in Nguyen and Tran [31] to generalized linear models in two ways: (1) modeling errors in the input space, using a convex objective with stringent conditions for consistency; and (2) modeling errors in the output space, breaking convexity but yielding milder conditions. Recent works [33, 24] consider estimation frameworks based on robust mean estimation however they are not able to guarantee consistent solutions (the estimation upper bound depends on the standard deviation of dense white noise, σ, which does not converge to zero even $n \to \infty$).

Another line of work aims at trimming outliers by minimizing the residuals over a selected subset. A key motivation for trimmed approaches is that convex loss functions with linear tail growth (such as the ℓ_1-norm and Huber loss) are not robust enough. As Alfons et al. [1] points out, these approaches have a breakdown point of $\epsilon = 0$, since even a single gross contamination can arbitrarily distort the regression coefficients. Remarkably, the median of least squares residual originally proposed by Rousseeuw [39] avoids this problem, reaching breakdown point of nearly 50%; the approach is equivalent to 'trimming' a portion of the largest residuals. This leads to the consideration of sparse Least Squares (sparse LTS) for robust high-dimensional linear models [1], who look for approximate solutions using alternating minimization schemes, since solving the trimming problem is combinatorial — computationally intractable as the number of features grows. While Alfons et al. [1] established high breakdown point property for sparse LTS, its statistical convergence has not been analyzed.

Statistical consistency under arbitrary corruption is certainly not a given. Consider a case of adversarial corruption, where some measurements are replaced by very high values, all exactly consistent with an estimator that is arbitrarily far away from the truth. Any likelihood formulation can easily be fooled and led to find the 'wrong' estimator. Some additional assumptions are thus needed as safeguard in the statistical analysis. In practice, extreme deviations might be easily detected in specific datasets; it is more difficult to catch outliers which masquerade as real data.

In this paper, we present a unified framework and statistical analysis for trimmed regularized M-estimators, generalizing the sparse least trimmed squares (Sparse LTS) estimator [1] to allow for a wide class of (possibly non-convex) loss functions as well as structured regularization. Using our analysis, we derive error bounds for the sparse LTS estimator. These require less stringent conditions for estimation consistency than those of Extended Lasso. [6] and follow-up work of Zhang et al. [54] consider this high dimensional linear model and propose a hard thresholding-based algorithm, which can be seen as a special case of the partial trimmed optimization discussed in Section 4. The analysis provided in [6] allows for a linear faction corruption but again involves the standard deviation of dense white noise σ in estimating the upper bound, which our analysis avoids. We also derive error bounds for sparse Gaussian graphical models (GGMs) as a specific example. In contrast, existing approaches for robust sparse GGMs estimation lack statistical guarantees.

We also show how to extend optimization algorithms for M-estimators to trimmed formulations, using partial minimization over auxiliary variables. An important example of the approach is a modified proximal gradient method. For convex M-estimators, we show that under moderate assumptions, the 'trimming' is completed in finitely many steps, and thereafter the method reduces to a descent method for a convex problem over a fixed set of identified 'inliers'. We use simulated data to compare with competing methods, and then apply our approach to real genomics datasets.

The manuscript is organized as follows. In Section 2 we introduce the general setup and present the family of High-Dimensional Trimmed estimators. Results on convergence and consistency of these estimators are given in Section 3, along with corollaries for linear models and Gaussian graphical models. Algorithms are developed in Section 4. Empirical results are presented in Section 5 for simulated data and Section 6 for the analysis of genomics datasets. All proofs are in the Appendix.

2. A general framework for high-dimensional trimmed estimators

Motivating example 1: Linear regression   We start with high-dimensional linear regression. The real-valued observation $y_i \in \mathbb{R}$ comes from the linear model

$$y_i = \langle x_i, \theta^* \rangle + o_i, \quad i = 1, \ldots, n, \tag{1}$$

where $x_i \in \mathbb{R}^p$ is a covariate, the true regression parameter vector is $\theta^* = (\theta_1^*, \ldots, \theta_p^*)^\top \in \mathbb{R}^p$, and o_i is the observation noise. Since outliers are commonly present in high-dimensional data problems, we assume p is substantially larger than n without loss of generality.

Let G be the set of "good" samples, and B denote the set of "bad" samples, corrupted arbitrarily. To deal with observations that deviate from the true model, Alfons et al. [1] proposed sparse LTS, an ℓ_1-penalized version of the classical least trimmed squares (LTS) estimator [39] solving

$$\underset{\theta}{\text{minimize}} \ \frac{1}{2h} \sum_{i=1}^h [r^2(\theta)]_{(i)} + \lambda \|\theta\|_1, \tag{2}$$

where $r^2(\theta) = (r_1^2, \ldots, r_n^2)^T$ with $r_i^2 = (y_i - \langle x_i, \theta \rangle)^2$, and $[r^2(\theta)]_{(1)} \leq \ldots \leq [r^2(\theta)]_{(n)}$ are the order statistics of the squared residuals $r^2(\theta)$. Alfons et al. [1] established the breakdown point of the resulting sparse LTS estimator, and proposed an iterative algorithm for its computation. At iteration t, the algorithm computes the Lasso solution based on the current subset H_t of observations with pre-defined size $h = |H_t|$, and constructs the next subset H_{t+1} from the observations corresponding to the h smallest squared residuals.

Our starting point is the following reformulation of regularized LTS problem:

$$\underset{w \in \Delta_h, \theta \in \rho \mathbb{B}_1}{\text{minimize}} \ \frac{1}{2h} \sum_{i=1}^n w_i (y_i - \langle x_i, \theta \rangle)^2 + \lambda \|\theta\|_1 \tag{3}$$

where $\Delta_h := \{w : w \in [0,1]^n, \ 1^T w = h\}$ is the h-scaled capped unit simplex, $\mathbb{B}_1$ is the ℓ_1-norm ball, and the constraint $\theta \in \rho \mathbb{B}_1$ (or equivalently $\|\theta\|_1 \leq \rho$) ensures that the optimum of non-convex problem (3) exists as discussed in Loh and Wainwright [25]. This constraint on θ is a theoretical safeguard, since problem (3) is equivalent to the problem (2) when ρ is large enough.

A family of trimmed estimators Based on the reformulation (3), we propose the family of trimmed estimators for general high-dimensional problems: given a collection of arbitrary corrupted samples $Z_1^n = \{Z_1, \ldots, Z_n\}$, and a differentiable (possibly non-convex) loss function $\bar{\mathcal{L}}$, we solve

$$\underset{w \in \Delta_h, \theta \in \rho \mathbb{B}_\mathcal{R}}{\text{minimize}} \ f(w, \theta) := \frac{1}{h} \sum_{i=1}^n w_i \bar{\mathcal{L}}(\theta; Z_i) + \lambda \mathcal{R}(\theta) \tag{4}$$

where $\mathcal{R}(\cdot)$ is a decomposable norm used as a regularizer [29] to encourage particular low-dimensional structure of the estimator, and $\mathbb{B}_\mathcal{R}$ is the unit ball for $\mathcal{R}(\cdot)$ (in other words, the constraint $\theta \in \rho \mathbb{B}_\mathcal{R}$ ensures $\mathcal{R}(\theta) \leq \rho$). The parameter h selects the number of samples (or sum of weights) used in the training. Ideally, h is the number of uncorrupted samples in G, but we can tune h by cross-validation along with parameter λ. Cross-validation is a popular choice in practice and has been recently analyzed for the vanilla Lasso estimator [13]. Though beyond the scope of this paper, it would be interesting to explore the

applicability of other selection strategies including self-tuning approaches which are pivotal with respect to the penalty parameters [5, 4] and approaches estimating h by treating it as concomitant similarly to [22]

Having a general choice of loss function $\mathcal{L}$ in (4) is particularly important, given the wide variety of losses used by practitioners in statistics and machine learning.

Motivating example 2: Graphical models Gaussian graphical models (GGMs) form a powerful class of statistical models for representing distributions over a set of variables [23]. These models use undirected graphs to encode conditional independence assumptions among the variables, which is particularly convenient for exploring network structures. GGMs are widely used in variety of domains, including computational biology [32], natural language processing [27], image processing [48, 19, 14], statistical physics [20], and spatial statistics [37].

In such high-dimensional settings, sparsity constraints are particularly pertinent for estimating GGMs, as they encourage only a few parameters to be non-zero and induce graphs with few edges. The most widely used estimator, the Graphical Lasso minimizes the negative Gaussian log-likelihood regularized by the ℓ_1 norm of the entries (or the off-diagonal entries) of the precision matrix (see [52, 17, 3]). This estimator has strong statistical guarantees (see e.g. [35]). The corresponding optimization problem is a log-determinant program that can be solved with interior point methods [7] or by co-ordinate descent algorithms [17, 3]. Alternatively neighborhood selection [28, 50] can be used to estimate conditional independence relationships separately for each node in the graph, via Lasso linear regression [43]. Under certain assumptions, the sparse GGM structure can still be recovered even under high-dimensional settings.

The aforementioned approaches rest on a fundamental assumption: the multivariate normality of the observations. However, outliers and corruption are frequently encountered in high-dimensional data (see e.g. [15] for gene expression data). Contamination of a few observations can drastically affect the quality of model estimation. It is therefore imperative to devise procedures that can cope with observations deviating from the model assumptions. Despite this fact, little attention has been paid to robust estimation of high-dimensional graphical models. Partially Relevant work includes [16], which leverages multivariate t-distributions for robustified inference and the EM algorithm. They also propose an alternative t-model which adds flexibility to the classical t but requires the use of Monte Carlo EM or variational approximation as the likelihood function is not available explicitly. Another pertinent work is that of [41] which introduces a robustified likelihood function. A two-stage procedure is proposed for model estimation, where the graphical structure is first obtained via coordinate gradient descent and the concentration matrix coefficients are subsequently re-estimated using iterative proportional fitting so as to guarantee positive definiteness of the final estimate.

A special case of the proposed family is that of the *Trimmed Graphical Lasso* for robust estimation of sparse GGMs:

$$\underset{\Theta \in \Omega \cap R\mathbb{B}_1, w \in \Delta_h}{\text{minimize}} \; \left\langle\!\left\langle \Theta, \frac{1}{h} \sum_{i=1}^{n} w_i X^{(i)} (X^{(i)})^\top \right\rangle\!\right\rangle - \log \det(\Theta) + \lambda \|\Theta\|_{1,\text{off}} . \qquad (5)$$

Here for matrices $U \in \mathbb{R}^{p \times p}$ and $V \in \mathbb{R}^{p \times p}$, $\langle\!\langle U, V \rangle\!\rangle$ denotes the trace inner product $\text{tr}(A B^T)$. For a matrix $U \in \mathbb{R}^{p \times p}$ and parameter $a \in [1, \infty]$, $\|U\|_a$ denotes the element-wise ℓ_a norm, and $\|U\|_{a,\text{off}}$ does the element-wise ℓ_a norm only for off-diagonal entries. For example, $\|U\|_{1,\text{off}} := \sum_{i \neq j} |U_{ij}|$.

We provide statistical guarantees on the consistency of this estimator. To the best of our knowledge, this is in stark contrast with prior work on robust sparse GGM estimation (e.g. [16, 41]) which are not statistically guaranteed in theory.

3. Statistical guarantees of trimmed estimators

In this section, we provide a statistical analysis of the family of structurally regularized estimators (4). In order to simplify the notation in our theorem and its corollaries, we assume without loss of generality that the number of good samples is known a priori and the tuning parameter h in (4) is exactly set as the genuine samples size, $|G|$. This is an unrealistic assumption, but as long as we set h smaller than $|G|$, the statements in the main theorem and its corollaries apply.

Noting that the optimization problem (4) is non-convex, estimators returned by iterative methods for (4) will be stationary points. We call $(\widetilde{\theta}, \widetilde{w})$ a *local minimum* of (4) when

1. $\widetilde{\theta}$ is a local minimum of $g_1(\theta) := f(\theta, \widetilde{w})$ and
2. $\widetilde{w}$ is a global minimum of $g_2(w) := f(\widetilde{\theta}, w)$.

These are precisely the points that are found by the algorithms developed in Section 4. In this section, we give statistical error bounds for **any** such points.

Consider any such local minimum $(\widetilde{\theta}, \widetilde{w})$. While we are mainly interested in the error bounds of our estimator for target parameter θ^* (that is, $\widetilde{\theta} - \theta^*$), we first define w^* as follows: for the index $i \in G$, w_i^* is simply set to $\widetilde{w}_i$ so that $w_i^* - \widetilde{w}_i = 0$. Otherwise for the index $i \in B$, we set $w_i^* = 0$. Note that while θ^* is fixed unconditionally, w^* is dependent on $\widetilde{w}$. However, w^* is fixed given $\widetilde{w}$.

In order to guarantee bounded errors, we first assume that given $(\widetilde{\theta}, \widetilde{w})$, the following restricted strong convexity condition for $(\widetilde{\theta}, \widetilde{w})$ holds:

(C-1) (Restricted strong convexity (RSC) on θ) We overload notation and use $\mathcal{L}(\theta, w)$ to denote $\frac{1}{h} \sum_{i=1}^{n} w_i \bar{\mathcal{L}}(\theta; Z_i)$. Then, for any possible $\Delta := \theta - \theta^*$, the differentiable loss function $\bar{\mathcal{L}}$ satisfies

$$\langle \nabla_\theta \mathcal{L}(\theta^* + \Delta, w^*) - \nabla_\theta \mathcal{L}(\theta^*, w^*), \Delta \rangle \geq \kappa_l \|\Delta\|_2^2 - \tau_1(n, p) \mathcal{R}(\Delta)^2,$$

where κ_l is a curvature parameter, and $\tau_1(n, p)$ is a tolerance function on n and p.

Note that this condition is slightly different from the standard restricted strong convexity condition because of the dependency on w^* and therefore on $\widetilde{w}$. Each local optimum has its own restricted strong convexity condition. In case of no corruption with $w_i^* = 1$ for all i, this condition will be trivially reduced to the standard RSC condition, under which the standard general M-estimator has been analyzed (see Negahban et al. [29] for details).

We additionally require the following condition for a successful estimation with (4) on corrupted samples:

(C-2) Consider arbitrary local optimum $(\widetilde{\theta}, \widetilde{w})$. Letting $\widetilde{\Delta} := \widetilde{\theta} - \theta^*$ and $\widetilde{\Gamma} := \widetilde{w} - w^* \in [0, 1]^n$,

$$\big\langle \nabla_\theta \mathcal{L}\big(\theta^* + \widetilde{\Delta}, w^* + \widetilde{\Gamma}\big) - \nabla_\theta \mathcal{L}\big(\theta^* + \widetilde{\Delta}, w^*\big), \, \widetilde{\Delta} \big\rangle$$
$$\geq -\tau_2(n,p)\|\widetilde{\Delta}\|_2 - \tau_3(n,p)\mathcal{R}(\widetilde{\Delta})\,.$$

(C-2) can be understood as a structural incoherence condition between θ and w. This type of condition is also needed for the guarantees of extended LASSO [31] and other dirty statistical models with more than a single parameter [49]. Due to the dependence on $\widetilde{w}$, each local optimum will have its *own* conditions (C-1) and (C-2). We will see later in this section that these two conditions are mild enough for the popular estimators (such as linear models and GGMs) to satisfy.

Armed with these conditions, we state the main theorem on the error bounds of (4):

Theorem 1. *Consider an M-estimator from* (4) *with* **any** *local minimum* $(\widetilde{\theta}, \widetilde{w})$, *and suppose that it satisfies the conditions* (C-1) *and* (C-2). *Suppose also that the regularization parameter λ in* (4) *is set as*

$$\lambda \geq 4 \max \left\{ \mathcal{R}^*\Big(\nabla_\theta \mathcal{L}\big(\theta^*, w^*\big)\Big), \, 2\rho\tau_1(n,p) + \tau_3(n,p) \right\} \tag{6}$$

where $\mathcal{R}^*(v)$ is the dual norm of $\mathcal{R}(\cdot)$: $\sup_{u \in \mathbb{R}^p \setminus \{0\}} \frac{\langle u, v \rangle}{\mathcal{R}(u)}$. Then the following error bounds for $\widetilde{\theta}$ are guaranteed for a given model space $\mathcal{M}$:

$$\|\widetilde{\theta} - \theta^*\|_2 \leq \frac{1}{\kappa_l}\Big(\frac{3\lambda\Psi}{2} + \tau_2(n,p)\Big) \quad and \quad \mathcal{R}\big(\widetilde{\theta} - \theta^*\big) \leq \frac{2}{\lambda\,\kappa_l}\Big(2\lambda\Psi + \tau_2(n,p)\Big)^2,$$

where

$$\Psi := \sup_{u \in \mathcal{M} \setminus \{0\}} R(u)/\|u\|_2$$

measures the compatibility between $\mathcal{R}(\cdot)$ and ℓ_2 norms.

For sparse vectors, $\Psi := \sup_{u \in \mathcal{M} \setminus \{0\}} \|u\|_1/\|u\|_2 = \sqrt{k}$ where k is the sparsity of true parameter θ^*, and $\mathcal{M}$ is the space of vectors with the correct support set [29].

The statement in Theorem 1 is applicable to any local minimum of (4), and it holds deterministically. Probabilistic statements come in when the condition on λ_n specified in Theorem 1 is satisfied. In (6), λ is chosen based

on $\mathcal{R}^*\big(\nabla_\theta \mathcal{L}(\theta^*, w^*)\big)$ similarly to Negahban et al. [29]. We shall see that the remaining terms with tolerance functions τ in (6) have the same order as $\mathcal{R}^*\big(\nabla_\theta \mathcal{L}(\theta^*, w^*)\big)$ for the specific cases of linear models and GGMs developed in the next sections.

3.1. Statistical guarantees of high-dimensional least trimmed squares

We now focus on the special case of high-dimensional linear regression, and apply Theorem 1 to problem (3). In particular, if $i \in G$, $y_i = \langle x_i, \theta^* \rangle + \epsilon_i$ where the observation noise ϵ_i follows zero mean and has sub-Gaussian tails. Otherwise, for $i \in B$, $y_i = \langle x_i, \theta^* \rangle + \delta_i$ where δ_i is the amount of arbitrary corruption.

While Theorem 1 holds for any setting under (C-1) and (C-2), we consider the following natural setting with dependent design matrices. This setting has been widely studied in past work on conventional (without outliers) high dimensional linear models (e.g. [29]):

(LTS1) (Σ-Gaussian ensemble) Each sample x_i is i.i.d. sampled from $N(0, \Sigma)$.

(LTS2) (Sub-Gaussian noise) The noise vector $\epsilon \in \mathbb{R}^n$ is zero-mean and has sub-Gaussian tails, which means that for any fixed vector v such that $\|v\|_2 = 1$, $\mathbb{P}\left[|\langle v, \epsilon \rangle| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$ for all $t > 0$. The sub-Gaussian is quite a wide class of distributions, and contains the Gaussian family as well as and all bounded random variables.

(LTS3) (Column normalization) Let $X \in \mathbb{R}^{n \times p}$ be the design matrix whose i-th row is the covariate i-th sample: $x_i^\top$, and $X^j \in \mathbb{R}^n$ be the j-th column vector of X. Then, $\frac{\|X^j\|_2}{\sqrt{h}} \leq 1$. As pointed out in Negahban et al. [29], we can always rescale linear models with out loss of generality to satisfy this condition.

In addition to the standard conditions above, we further assume the following conditions, which will guarantee structural incoherence (C-2):

(C-h) Let h be the number of good samples: $|G| = h$ and hence $|B| = n - h$. Then, we assume that at least half of samples are genuine and uncorrupted so that $\frac{|G| - |B|}{|G|} \geq \alpha$ where $0 < \alpha \leq 1$. If we assume that 40% of samples are corrupted, then $\alpha = 1/3$.

(LTS4) We set the tuning parameter ρ in (3) as $\rho \leq \frac{C_1}{2}\sqrt{\frac{h}{\log p}}$ for some constant C_1. This setting requires that the number of good samples h is larger than or equal to $\left(\frac{2k\|\theta^*\|_\infty}{C_1}\right)^2 \log p$ so that the true regression parameter θ^* is feasible for the objective.

Under these conditions, we can recover the following error bounds of high-dimensional LTS (3), as a corollary of Theorem 1:

Corollary 1. *Consider corrupted linear models* (1) *when* $\|\theta^*\|_0 \leq k$. *Suppose that conditions* (C-h), (LTS1), (LTS2), (LTS3), *and* (LTS4) *hold. Also suppose that we find a local minimum* $(\widetilde{\theta}, \widetilde{w})$ *of* (3), *choosing*

$$\lambda = c\sqrt{\frac{\log p}{h}}$$

where c is some constant dependent on Σ, σ and the upper bound of $\frac{(\max_i \delta_i^2)|B|}{h}$ [1]. *Then, $(\widetilde{\theta}, \widetilde{w})$ is guaranteed to satisfy* (C-1) *and* (C-2) *for the specific case of* (3), *and have the following error bounds: for some constant c' depending on c, Σ and the portion of genuine samples α in* (C-h),

$$\|\widetilde{\theta} - \theta^*\|_2 \leq c'\left(\sqrt{\frac{k \log p}{h}} + \sqrt{\frac{|B| \log p}{h}}\right),$$

$$\|\widetilde{\theta} - \theta^*\|_1 \leq \frac{4c'}{\lambda}\left(\sqrt{\frac{k \log p}{h}} + \sqrt{\frac{|B| \log p}{h}}\right)^2$$

with probability at least $1 - c_1 \exp(-c_1' h\lambda^2)$ for some universal positive constants c_1 and c_1'.

Note that Corollary 1 concerns any *single* local minimum. For the guarantees of *multiple* local optima simultaneously, we may use a union bound from the corollary.

Remarks While the error bounds in Corollary 1 hold under the condition $|B| < |G|$ ((C-h)), in order to have meaningful and consistent bounds, α in (C-h) should be properly lower bounded so that $\frac{|B| \log p}{h}$ converges to zero as n and p increase. When the fraction of corruptions is linearly proportional to the number of overall samples n, then the parameters are estimated within some bounded error involving the *logarithm* of n. It might imply that our method is more resilient to massive outliers than other existing works such as [12] having up to an additive error bound of $\sqrt{p}$ under linear fraction of corruptions. Also note that this is also the case in the ℓ_2 error bound of extended Lasso analyzed in Nguyen and Tran [31], which has a similar term with asymptotically the same rate. Specifically, the extended Lasso estimator solves:

$$\underset{\theta, e}{\text{minimize}} \ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \langle x_i, \theta \rangle - e_i\right)^2 + \lambda_\theta \|\theta\|_1 + \lambda_e \|e\|_1$$

---

[1] From a theoretical standpoint it is impossible to recover the true signal by trimming when the amount of corruptions exceed that of normal observation errors. An adversarial 'corruption' could be constructed to point to the wrong estimator, and be a local minimum of the formulation. The required assumption for success is that the squared loss for corrupted samples should be less than or equal to squared loss for good samples: $(\max_i \delta_i^2)|B| \leq C_2^2 h$ for some constant C_2. In practice, large deviations that are not adversarial (e.g. random) are easily detected by the approach, even when their magnitude far exceeds normal observations errors.

where λ_e is the regularization parameter for parameter e capturing corruptions. e is encouraged to be sparse to reflect the fact that only a fraction of samples is corrupted. The ℓ_2 norm-based error rate in Corollary 1 is almost the same as that of extended Lasso: $\|\widehat{\theta}_{\text{E\_Lasso}} - \theta^*\|_2 = O\left(\sqrt{\frac{k \log p}{n}} + \sqrt{\frac{|B| \log n}{n}}\right)$ under the standard Gaussian design setting (LTS1), which is asymptotically the same as that in Corollary 1 (since $1/h \leq 2/n$ under (C-h)).

However, it is important to revisit the conditions required for the statistical guarantees of extended Lasso. Besides an extended version of the restricted eigenvalue condition, Nguyen and Tran [31] assumes a mutual incoherence condition, which in turn requires $c\sqrt{\|\Sigma\|_2} \max\left\{\frac{k}{|B|}, \frac{|B|}{k}\right\}\left(\sqrt{\frac{k}{n}} + \sqrt{\frac{|B|}{n}} + \sqrt{\frac{\log p}{n}}\right) \leq \frac{1}{16}$ for some large and fixed constant c. Provided that k and $|B|$ are fixed, the inequality can hold for a large enough sample size n. However, when $|B|$ grows with n, this condition will be violated; for example if (i) a square root fraction of samples is corrupted ($|B| = \alpha\sqrt{n}$) for a fixed k or (ii) a linear fraction of n is corrupted ($|B| = \alpha n$), then $c'\sqrt{\|\Sigma\|_2}$ can easily exceed $1/16$. Our experimental results of Section 5 will confirm this observation: as the fraction of corruptions increases, the performance of extended Lasso deteriorates compared to that of our estimator (3).

Statistical guarantees when covariates are corrupted   In the linear model (1), corruption is considered in the space of the response variable $y_i \in \mathbb{R}$: namely an additional random variable $\delta_i \in \mathbb{R}$ is used to model corruption in the response space. Even in the case where we have outliers with corrupted covariates $x_i + \delta' \in \mathbb{R}^p$, δ_i can be understood as the mean-shift variable to model $\langle \delta', \theta^* \rangle$. For linear models, modeling outliers in the parameter space or modeling them in the output space is thus equivalent (In constrast, for more general GLM settings, the link function is not the identity function and both approaches are distinct, see e.g. [51]). Nevertheless, when outliers stem from corrupted covariates, condition (LTS1) might be violated. For this setting, we introduce the following alternative condition:

(LTS5) (Σ-Gaussian ensemble) Each sample x_i in G is i.i.d. sampled from $N(0, \Sigma_G)$. Let X^B be the sub-design matrix in $\mathbb{R}^{|B| \times p}$ corresponding to outliers. Then, we define $f(X^B)$ such that $\|X^B\|_2 \leq f(X^B)\sqrt{|B| \log p}$.

Under condition (LTS5) we recover results similar to Corollary 1:

Corollary 2. *Consider linear models in* (1) *where* $\|\theta^*\|_0 \leq k$. *Suppose that all the conditions* (C-h), (LTS2), (LTS3), (LTS4) *and* (LTS5) *hold. Also suppose that we choose the regularization parameter*

$$\lambda = c\sqrt{\frac{\log p}{h}}$$

where c is some constant dependent on Σ_G, $f(X^B)$, σ and the upper bound of $\frac{(\max_i \delta_i^2)|B|}{h}$. *Then, $(\widetilde{\theta}, \widetilde{w})$ is guaranteed to have the following error bounds as*

before: for some constant c' depending on c, Σ_G and the portion of genuine samples α in (C-h),

$$\|\widetilde{\theta} - \theta^*\|_2 \leq c'\left(\sqrt{\frac{k\log p}{h}} + \sqrt{\frac{|B|\log p}{h}}\right),$$

$$\|\widetilde{\theta} - \theta^*\|_1 \leq \frac{4c'}{\lambda}\left(\sqrt{\frac{k\log p}{h}} + \sqrt{\frac{|B|\log p}{h}}\right)^2$$

with probability at least $1 - c_1\exp(-c_1' h\lambda^2)$ for some universal positive constants c_1 and c_1'.

3.2. Statistical guarantees of trimmed graphical Lasso

We now focus on Gaussian graphical models and provide the statistical guarantees of our Trimmed Graphical Lasso estimator as presented in Section 2 (Motivating Example 2). Our theory in this section provides the statistical error bounds on **any** local minimum of (5). We use $\|U\|_{\mathrm{F}}$ and $\|U\|_2$ to denote the Frobenius and spectral norms, respectively.

Let $X = (X_1, X_2, \ldots, X_p)$ be a zero-mean Gaussian random field parameterized by $p \times p$ concentration matrix Θ^*:

$$\mathbb{P}(X; \Theta^*) = \exp\left(-\frac{1}{2}\langle\!\langle\Theta^*, XX^\top\rangle\!\rangle - A(\Theta^*)\right) \tag{7}$$

where $A(\Theta^*)$ is the log-partition function of Gaussian random field. Here, the probability density function in (7) is associated with p-variate Gaussian distribution, $N(0, \Sigma^*)$ where $\Sigma^* = (\Theta^*)^{-1}$.

We consider the case where the number of random variables p may be substantially larger than the number of sample size n, however, the concentration parameter of the underlying distribution is sparse so that the number of non-zero off-diagonal entries of θ^* is at most k: $|\{\Theta_{ij}^* : \Theta_{ij}^* \neq 0 \text{ for } i \neq j\}| \leq k$.

We now investigate how easily we can satisfy the conditions in Theorem 1. Intuitively it is impossible to recover true parameter by weighting approach as in (5) when the amount of corruptions exceeds that of normal observation errors.

To this end, suppose that we have some upper bound on the corruptions:

(TGL1) For some function $f(\cdot)$, we have $\left(\|X^B\|_2\right)^2 \leq f(X^B)\sqrt{h\log p}$

where X^B denotes the sub-design matrix in $\mathbb{R}^{|B|\times p}$ corresponding to outliers. Under this assumption, we can recover the following error bounds of Trimmed Graphical Lasso (5), as a new corollary of Theorem 1:

Corollary 3. *Consider corrupted Gaussian graphical models with conditions* (C-h) *and* (TGL1). *Suppose that we compute the local optimum $(\widetilde{\Theta}, \widetilde{w})$ of (5) choosing*

$$\lambda = 4\max\left\{8(\max_i \Sigma_{ii}^*)\sqrt{\frac{30\log p}{h - |B|}} + \frac{|B|}{h}\|\Sigma^*\|_\infty\,,\ f(X^B)\sqrt{\frac{\log p}{h}}\right\}$$

$$\leq \frac{c_1 - f(X^B)\sqrt{\frac{|B|\log p}{h}}}{3R}\,.$$

Then, $(\widetilde{\theta}, \widetilde{w})$ is guaranteed to satisfy (C-1) *and* (C-2) *for the specific case of* (5) *and have the error bounds of*

$$\|\widetilde{\Theta} - \Theta^*\|_{\mathrm{F}} \leq \frac{1}{\kappa_l}\left(\frac{3\lambda\sqrt{k+p}}{2} + f(X^B)\sqrt{\frac{|B|\log p}{h}}\right) \quad \textit{and}$$

$$\|\widetilde{\Theta} - \Theta^*\|_{1,\mathrm{off}} \leq \frac{2}{\lambda\,\kappa_l}\left(3\lambda\sqrt{k+p} + f(X^B)\sqrt{\frac{2|B|\log p}{n}}\right)^2 \qquad (8)$$

with probability at least $1 - c_2\exp(-c_2'h\lambda^2)$ for some universal positive constants c_2 and c_2'.

In Corollary 3, the term $\sqrt{k+p}$ captures the relation between element-wise ℓ_1 norm and the error norm $\|\cdot\|_{\mathrm{F}}$ including *diagonal entries.*

If we further assume that the number of corrupted samples scales with $\sqrt{n}$ at most:

(TGL2) $|B| \leq a\sqrt{n}$ for some constant $a \geq 0$,

then we can derive the following result as another corollary of Theorem 1:

Corollary 4. *Consider corrupted Gaussian graphical models, and compute the local minimum $(\widetilde{\Theta}, \widetilde{w})$ of* (5)*, setting*

$$\lambda = c\sqrt{\frac{\log p}{n}}, \quad c := 4\max\left\{16(\max_i \Sigma_{ii}^*)\sqrt{15} + \frac{2a\|\Sigma^*\|_\infty}{\sqrt{\log p}}\,,\ \sqrt{2}f(X^B)\right\}.$$

Suppose that the conditions (C-h)*,* (TGL1) *and* (TGL2) *hold. Then, if the sample size n is lower bounded as*

$$n \geq \max\left\{16a^2\,,\ \left(\|\Theta^*\|_2 + 1\right)^4\left(3Rc + f(X^B)\sqrt{2|B|}\right)^2(\log p)\right\}\,,$$

then $(\widetilde{\Theta}, \widetilde{w})$ is guaranteed to satisfy (C-1) *and* (C-2) *for the specific case of* (5) *and have the following error bound:*

$$\|\widetilde{\Theta} - \Theta^*\|_{\mathrm{F}} \leq \frac{1}{\kappa_l}\left(\frac{3c}{2}\sqrt{\frac{(k+p)\log p}{n}} + f(X^B)\sqrt{\frac{2|B|\log p}{n}}\right) \qquad (9)$$

with probability at least $1 - c_1\exp(-c_1'h\lambda^2)$ for some universal positive constants c_1 and c_1'.

Note that an $\|\cdot\|_{1,\mathrm{off}}$-norm error bound can also be easily derived using the selection of λ from (8).

Remarks Corollary 4 reveals an interesting result: even when $O(\sqrt{n})$ samples out of total n samples are corrupted, our estimator (5) can successfully recover the true parameter with guaranteed error in (9). The first term in this bound is $O\left(\sqrt{\frac{(k+p)\log p}{n}}\right)$ which exactly matches the Frobenius error bound for the case without outliers (see Ravikumar et al. [35], Loh and Wainwright [26] for example). Due to the outliers, the performance degrades with the second term, which is $O\left(\sqrt{\frac{|B|\log p}{n}}\right)$. To the best of our knowledge, our results are the first statistical error bounds available in the litterature on parameter estimation for Gaussian graphical models with outliers.

When outliers follow a Gaussian graphical model Now let us provide a concrete example and show how $f(X^B)$ in (TGL1) is precisely specified in this case:

(TGL3) Outliers in the set B are drawn from another Gaussian graphical model (7) with a parameter $(\Sigma_B)^{-1}$.

This can be understood as a Gaussian mixture model where most of the samples are drawn from $(\Theta^*)^{-1}$ which we want to estimate, and a small portion of samples are drawn from Σ_B. In this case, Corollary 4 can be further shaped as follows:

Corollary 5. *Suppose that the conditions* (C-h), (TGL2) *and* (TGL3) *hold. Then the statement in Corollary 4 holds with* $f(X^B) := \frac{4\sqrt{2}a\left(1+\sqrt{\log p}\right)^2 \|\Sigma_B\|_2}{\sqrt{\log p}}$.

4. Optimization for trimmed estimators

While the objective function $f(w, \theta)$ in (4) is non-convex in (w, θ), it simplifies for block w or θ held fixed. Perhaps for this reason, prior algorithms for trimmed approaches [39, 1] alternated between solving for θ and w. These approaches are inefficient, since each solve in θ is as expensive as finding the original (untrimmed) estimator.

We take advantage of the computational complexity gap between subproblems in θ and in w. With w fixed, the problem in θ is equivalent to classic high-dimensional problems, e.g. Lasso. In contrast, the problem in w for fixed θ is the simple linear program

$$\operatorname*{minimize}_{w \in \Delta_h} \sum_{i=1}^{n} w_i \bar{\mathcal{L}}(\theta; Z_i) \tag{10}$$

with all dependence on the predictors captured by the current losses $\bar{\mathcal{L}}(\theta; Z_i)$. The solution is obtained setting $w_i = 1$ for the h smallest values of $\bar{\mathcal{L}}(\theta; Z_i)$, and setting remaining w_i to 0.

We exploit structure using partial minimization. Similar ideas have been used for optimizing a range of nonlinear least squares problems [18] as well as more

---

Algorithm 1 Partial Minimization using Proximal Gradient Descent for (11)

Initialize $\theta^{(0)}$, $t = 0$
repeat
    Compute $w^{(t)}$ given $\theta^{(t)}$ as the *global* minimum of (10)
    Given $w^{(t)}$, compute the direction $\mathcal{G}^{(t+1)} \leftarrow \frac{1}{h} \sum_{i=1}^{n} w_i^{(t)} \nabla_\theta \bar{\mathcal{L}}(\theta^{(t)}; y_i, x_i)$
    Update $\theta^{(t+1)} \leftarrow \mathcal{S}_{\eta^{(t+1)}\lambda}(\theta^{(t)} - \eta^{(t+1)}\mathcal{G}^{(t+1)})$, with $\eta^{(t)}$ selected using line search.
until stopping criterion is satisfied

---

general problems involving nuisance parameters [2]. Rather than an alternating scheme (similar to that of [1] for least squares) where we solve *multiple* 'weighted' regularized problems to completion, we can rewrite the problem as follows:

$$\underset{\theta \in \rho \mathbb{B}_\mathcal{R}}{\text{minimize}}\, \widetilde{\mathcal{L}}(\theta) + \lambda \mathcal{R}(\theta), \quad \widetilde{\mathcal{L}}(\theta) := \min_{w \in \Delta_h} \frac{1}{h} \sum_{i=1}^{n} w_i \bar{\mathcal{L}}(\theta; Z_i) = \frac{1}{h} \sum_{i=1}^{n} w_i(\theta) \bar{\mathcal{L}}(\theta; Z_i).$$
(11)

Problem (11) is equivalent to (4). The reader can verify that $\widetilde{\mathcal{L}}(\theta)$ is non-smooth and non-convex[2]. However, partial minimization provides a way to modify any descent method for fitting an M-estimator to find the corresponding trimmed estimator (11). Algorithm 1 gives a description of the steps involved for the specific case of proximal gradient descent. The algorithm uses the proximal mapping, which for the case of ℓ_1 regularization is the soft-thresholding operator defined as $[S_\nu(u)]_i = \text{sign}(u_i) \max(|u_i| - \nu, 0)$. We assume that we pick ρ sufficiently large, so one does not need to enforce the constraint $\mathcal{R}(\theta) \leq \rho$ explicitly.

When the loss $\mathcal{L}$ is convex and smooth with Lipschitz continuous gradient, the proximal gradient has a global convergence theory (see e.g. [30]). Convergence of the extended Algorithm 1 is analyzed in the following proposition.

Proposition 1. *Consider any monotonic algorithm $\mathcal{A}$ for solving $\widehat{\theta} \in \text{argmin}_\theta F(\theta) := \mathcal{L}(\theta) + \lambda \mathcal{R}(\theta)$, i.e. (i) $\mathcal{A}$ guarantees that $F(\theta^{k+1}) \leq F(\theta^k)$ and (ii) for any fixed $w \in \Delta_h$, $\mathcal{A}$ produces converging sequence of $\{\theta^{(t)}\}$ when solving $\text{argmin}_\theta F(\theta; w) := f(w, \theta)$. If $\mathcal{A}$ is extended to solve (11) using partial minimization (10), the monotonic property is preserved, at least one limit point exists, and every limit point of the sequence $\{(\theta^{(t)}, w^{(t)})\}$ is a stationary point of (4). Moreover, if F is convex, and estimators over each feasible data selection have different optimal values, then $w^{(t)}$ converge in finitely many steps, and the extended algorithm converges to a local minimum[3] of (4).*

Finite convergence of the weights $w^{(t)}$ is an important point for practical implementation, since once the weights converge, one is essentially solving a single estimation problem, rather than a sequence of such problems. In particular, after finitely many steps, the extended algorithm inherits all properties of the original algorithm $\mathcal{A}$ for the M-estimator over the selected data.

---

[2]When $h = 1$, trimming equates to minimizing the *minimum* of $\mathcal{L}_i$, a problem which is nonsmooth and nonconvex.

[3]$\widetilde{\theta}$ is a local minimum of $g_1(\theta) := f(\theta, \widetilde{w})$ and $\widetilde{w}$ is a global minimum of $g_2(w) := f(\widetilde{\theta}, w)$.

5. Simulated data experiments

We illustrate the generality of our approach by considering sparse logistic regression, trace-norm regularized multi-reponse regression and sparse GGMs (For experiments with sparse linear models, see Alfons et al. [1]).

5.1. Simulations for sparse logistic regression

We begin with sparse logistic regression. We adopt an experimental protocol similar to [51]. We consider $p = 200$ features. The parameter vectors have $k = \sqrt{p}$ non-zero entries sampled i.i.d. from $N(0,1)$. The data matrix X is such that each of its n observations is sampled from a standard Normal distribution $N(0, I_p)$. Given each observation, we draw a true class label from $\{0,1\}$ following the logistic regression model. We show two scenarios, selecting either $\sqrt{n}$ or $0.1n$ samples with the highest amplitude of $\langle \theta^*, x_i \rangle$ and flipping their labels. We compare the ℓ_2 errors over 100 simulation runs of the new estimator with those of vanilla Lasso for logistic regression, and with two extended Lasso methods for logistic regression of [51] (with "error in parameter" and in "error in output") as the sample size n increases. The tuning parameters for each method are set via holdout where we use a separate dataset of size n with the same type and amount of corruption. This tuning also applies to the trimming parameter h. Specifically, we set a grid for the trimming parameter ranging from 5% to 45% of the sample size with an increment of 5%.

Figure 1 shows that the trimmed approach has both better performance (achieves lower errors), and is faster, matching the computational efficiency of the vanilla Lasso method. This result is anticipated by Proposition 1: the weights $w^{(t)}$ converge in finitely many steps, and then we are essentially solving the Lasso with a fixed weight set thereafter.

5.2. Simulations for trace-norm regularized regression

Beyond the ℓ_1 penalty, we consider trace-norm regularized multi response regression. We set $\mathcal{R}(\Theta) = \|\Theta\|_*$, for $\Theta \in \mathbb{R}^{p \times q}$. We consider $n = 50$ samples, $p = 300$ covariates, and $q = 10$ responses. Each entry of X is generated independently from $N(0,1)$. To generate the true low rank weights, we first sample a $p \times q$ matrix of coefficients, with each coefficient sampled independently from $N(0,1)$. We then set the true parameter matrix to the best rank 3 approximation of the sample, obtained using an SVD. For clean samples in G, we then set the error term as $\epsilon_i \sim N(0, 0.01)$. The contaminated terms are generated with an error term as $\delta_i \sim N(2,1)$. We consider varying corruption levels ranging from 5% to 30%. The parameters are tuned via holdout validation as in the previous section and we present the average ℓ_2 error based on 100 simulation runs. Figure 2 further illustrates the computational advantage of the partial minimization scheme described in Section 4 for general structures.
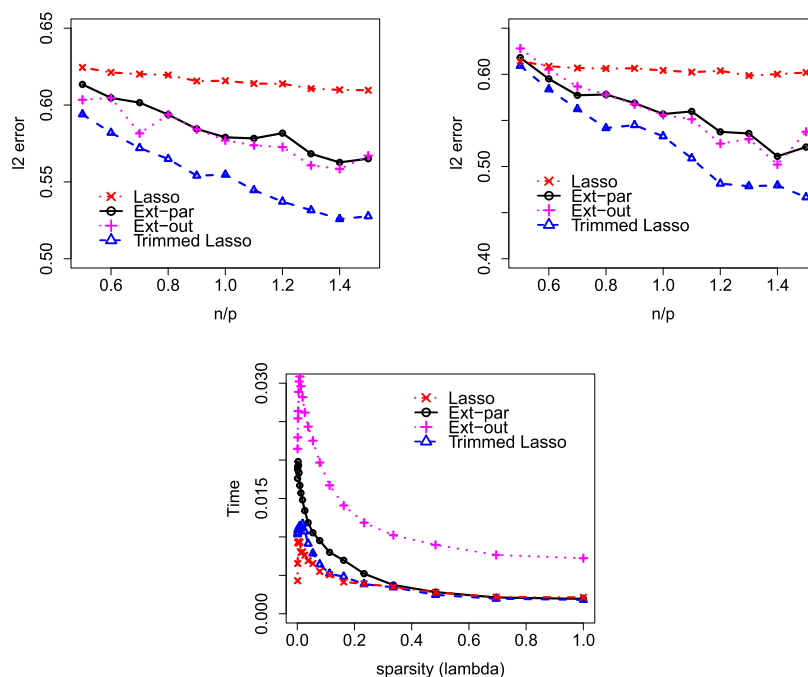
FIG 1. ℓ_2 error vs.sample size n under logistic regression model (a) $\sqrt{n}$ corruptions (b) $0.1n$ corruptions. (c) Timing comparison for $0.1n$ corruptions and $n = p$.

TABLE 1
Average ℓ_2 error for comparison methods on simulated data under low-rank multi response linear models with contaminated data.

Contamination %	No trimming	Low-Rank LTS
5%	20.43	**19.20**
10%	33.49	**25.10**
20%	33.70	**26.05**
30%	40.78	**30.10**

5.3. Simulations for Gaussian graphical models

We compare the Trimmed Graphical Lasso (trim-glasso) algorithm against the vanilla Graphical Lasso(glasso) [17]; the t-lasso and t*-lasso methods [16], and robust-LL: the robustified-likelihood approach of [41].

Our simulation setup is similar to [41] and is a akin to gene regulatory networks. Namely we consider four different scenarios where the outliers are generated from models with different graphical structures. Specifically, each sample is generated from the following mixture distribution:

$$y_k \sim (1 - p_0)N_p(0, \theta^{-1}) + \frac{p_0}{2} N_p(-\mu, \theta_o^{-1}) + \frac{p_0}{2} N_p(\mu, \theta_o^{-1}), \quad k = 1, \ldots, n,$$
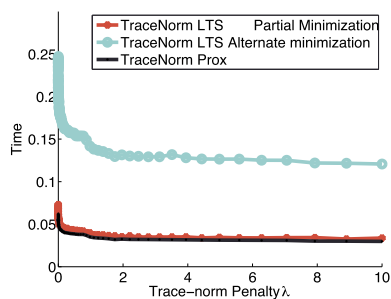
FIG 2. *Average timing of TraceNorm-LTS with partial minimization,TraceNorm-LTS with full alternate minimization, and TraceNorm-Prox under 20% of contaminated data.*
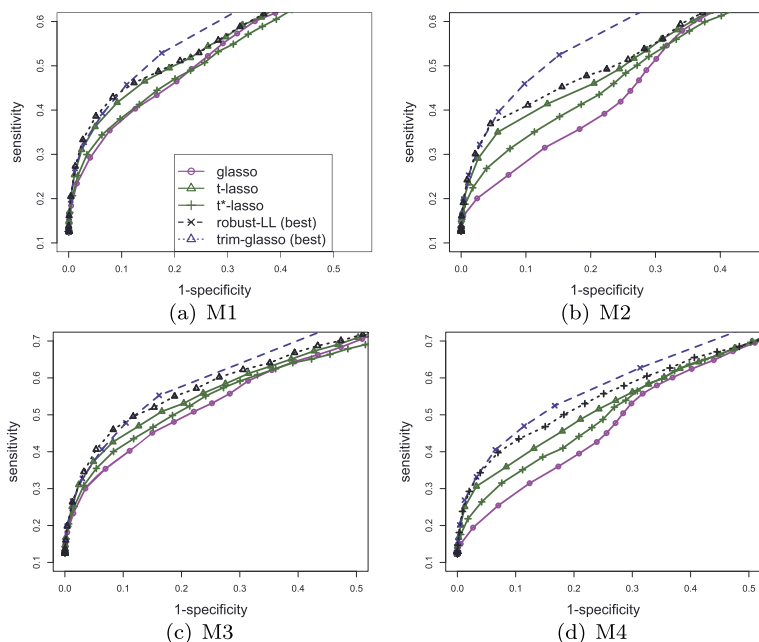

FIG 3. *Average ROC curves for the comparison methods for contamination scenarios M1–M4.*

where $p_o = 0.1, n = 100$, and $p = 150$. Four different outlier distributions are considered:

M1: $\mu = (1, \ldots, 1)^T, \theta_o = \tilde{\theta}$,   M2: $\mu = (1.5, \ldots, 1.5)^T, \theta_o = \tilde{\theta}$,
M3: $\mu = (1, \ldots, 1)^T, \theta_o = I_p$,   M4: $\mu = (1.5, \ldots, 1.5)^T, \theta_o = I_p$.

For each simulation run, θ is a randomly generated precision matrix corresponding to a network with 9 hub nodes simulated as follows. Let A be the adjacency of the network. For all $i < j$ we set $A_{ij} = 1$ with probability 0.03, and zero otherwise. We set $A_{ji} = A_{ij}$. We then randomly select 9 hub nodes and set
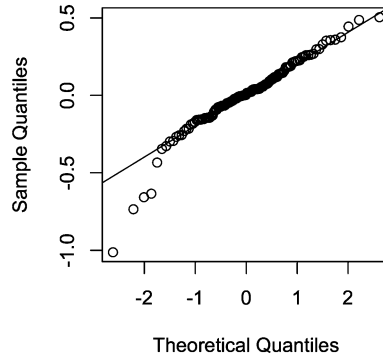
Fig 4. *QQ-plots of fitted residuals for the Sparse-LTS method in the genomic study.*

the elements of the corresponding rows and columns of A to one with probability 0.4 and zero otherwise. Using A, the simulated nonzero coefficients of the precision matrix are sampled as follows. First we create a matrix E so that $E_{i,j} = 0$ if $A_{i,j} = 0$, and $E_{i,j}$ is sampled uniformly from $[-0.75, -0.23] \cup [0.25, 0.75]$ if $A_{i,j} \neq 0$. Then we set $E = \frac{E + E^T}{2}$. Finally we set $\theta = E + (0.1 - \Lambda_{\min}(E))I_p$, where $\Lambda_{\min}(E)$ is the smallest eigenvalue of E. $\tilde{\theta}$ is a randomly generated precision matrix in the same way θ is generated.

For the robustness parameter β of the robust-LL method, we consider $\beta \in \{0.005, 0.01, 0.02, 0.03\}$ as recommended in [41]. For the trim-glasso method we consider $\frac{100h}{n} \in \{90, 85, 80\}$. Since all the robust comparison methods converge to a stationary point, we tested various initialization strategies for the concentration matrix, including I_p, $(S + \lambda I_p)^{-1}$ and the estimate from glasso. We did not observe any noticeable impact on the results.

Figure 3 presents the average ROC curves of the comparison methods over 100 simulation data sets for scenarios M1–M4 as the tuning parameter λ varies. In the figure, for robust-LL and trim-glasso methods, we depict the best curves with respect to parameter β and h respectively. The detailed results for all the values of β and h considered are provided in the appendix.

From the ROC curves we can see that our proposed approach is competitive compared the alternative robust approaches t-lasso, t*-lasso and robust-LL. The edge over glasso is even more pronounced for scenarios M2, M4. Surprisingly, trim-glasso with $h/n = 80\%$ achieves superior sensitivity for nearly any specificity.

Computationally the trim-glasso method is also competitive compared to alternatives. The average run-time over the path of tuning parameters λ is 45.78s for t-lasso, 22.14s for t*-lasso, 11.06s for robust-LL, 1.58s for trimmed lasso, 1.04s for glasso. Experiments were run on R in a single computing node with a Intel Core i5 2.5GHz CPU and 8G memory. For t-lasso, t*-lasso and robust-LL we used the R implementations provided by the methods' authors. For glasso we used the glassopath package.

TABLE 2
Average Trimmed Mean Square Error from 10-fold cross validation for comparison methods on the Yeast dataset.

Method	T-MSE
Lasso	0.137
LAD-Lasso	0.132
Extended Lasso	0.093
ROMP	0.135
Sparse-LTS	**0.081**

TABLE 3
Marker position of SNPs selected on chromosome 8 by comparison methods for the Yeast dataset.

LAD-Lasso	Sparse-LTS
111682	46007
213237	46055
	111682
	111683
	111686
	111687
	111690

6. Application genomic analysis

6.1. Analysis of Yeast genotype and expression data

We apply Sparse-LTS, Extended Lasso [31], LAD Lasso [47], standard Least Squares Lasso estimator [43], and ROMP [12] to the analysis of yeast genotype and gene expression data. We employ the "yeast" dataset from [9]. The data set concerns $n = 112$ *F1 segregants* from a yeast genetic cross between two strains: *BY* and *RM*. For each of these 112 samples, we observe $p = 3244$ SNPs (These genotype data are our predictors x) and focus on the gene expression of gene *GPA1* (our response y), which is involved in pheromone response [9]. For both Sparse-LTS-Ada and Sparse LTS considering a total of $|B| = 11$ contaminated observations lead to the best predictive performance on the uncontaminated data. In addition, the QQ-plots of the fitted residuals from the various comparison methods indicated heavy left tails (see Figure 4). This suggests that it might be advisable to use robust methods.

We compare the trimmed mean square error (T-MSE) computed from 10-folds cross validation for each method, where for each method we exclude the 11 observations with largest residual absolute error. From Table 2 we can see thatSparse-LTS exhibit the smallest T-MSE.

We conclude by examining the SNPs selected by the methods achieving the lowest T-MSE: Sparse-LTS and LAD Lasso. Out of $p = 3244$ SNPs, Sparse-LTS selected 30 SNPs, and LAD Lasso chose 61 SNPs. Table 3 provides a list of the SNPs selected on chromosome 8, which is where gene *GPA1* resides. In the dataset, there is a total of 166 SNPs on chromosome 8. From the table
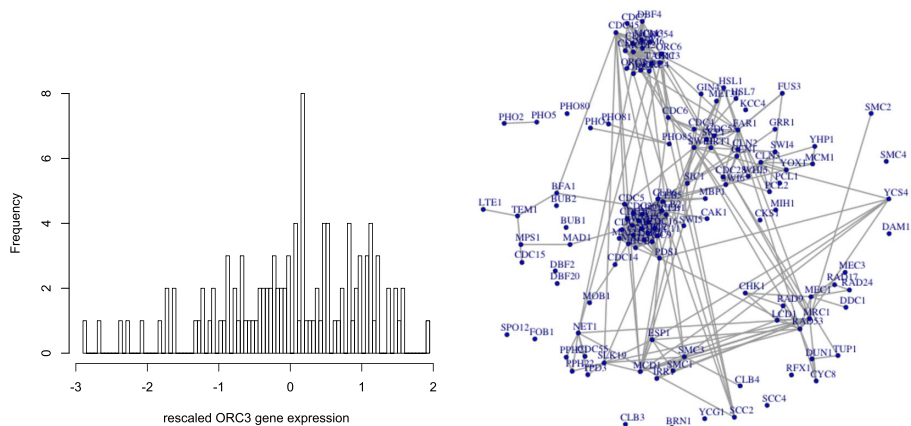
FIG 5. *(a) Histogram of standardized gene expression levels for gene* ORC3. *(b) Network estimated by trim-glasso.*

we can see that there is some overlap in terms of the selected SNPs across the various methods. Sparse-LTS tends to select a larger number of SNPs on chromosome 8 even though it selects fewer SNPs in total (namely within and beyond chromosome 8). Five of these are very close to *GPA1* which is consistent with the fact that *GPA1* can directly inhibit the mating signal by binding to its own subunit [40].

6.2. *Application to the analysis of Yeast gene expression data*

We analyze a yeast microarray dataset generated by [8]. The dataset concerns $n = 112$ yeast segregants (instances). We focused on $p = 126$ genes (variables) belonging to cell-cycle pathway as provided by the KEGG database [21]. For each of these genes we standardize the gene expression data to zero-mean and unit standard deviation. We observed that the expression levels of some genes are clearly not symmetric about their means and might include outliers. For example the histogram of gene *ORC3* is presented in Figure 5(a). For the robust-LL method we set $\beta = 0.05$ and for trim-glasso we use $h/n = 80\%$. We use 5-fold-CV to choose the tuning parameters for each method. After λ is chosen for each method, we rerun the methods using the full dataset to obtain the final precision matrix estimates.

Figure 5(b) shows the cell-cycle pathway estimated by our proposed method. For comparison the cell-cycle pathway from the KEGG [21] is provided in Figure 6. It is important to note that the KEGG graph corresponds to what is currently known about the pathway. It should *not* be treated as the ground truth. Certain discrepancies between KEGG and estimated graphs may also be caused by inherent limitations in the dataset used for modeling. For instance, some edges in cell-cycle pathway may not be observable from gene expression
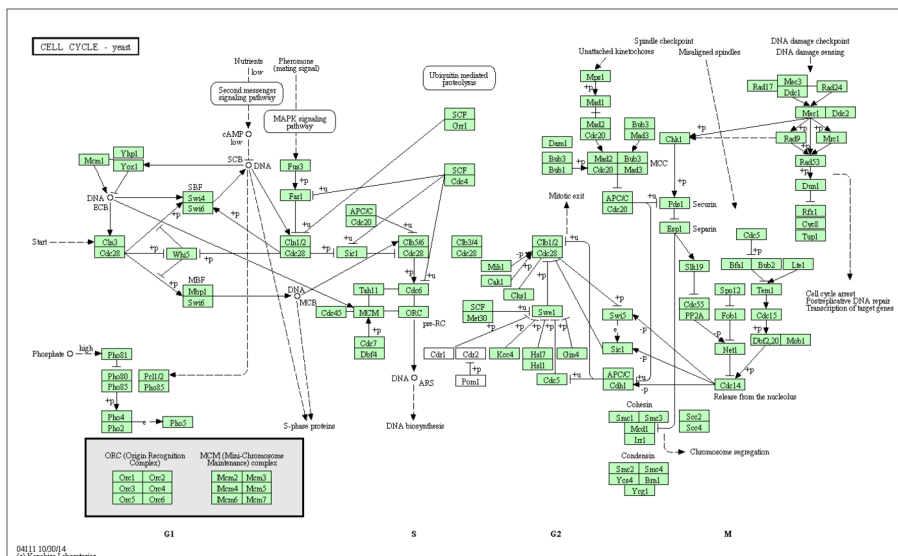
FIG 6. *Reference Yeast Cell Signaling Network from the KEGG database [21].*

data. Additionally, the perturbation of cellular systems might not be strong enough to enable accurate inference of some of the links.

glasso tends to estimate more links than the robust methods. We postulate that the lack of robustness might result in inaccurate network reconstruction and the identification of spurious links. Robust methods tend to estimate networks that are more consistent with that from the KEGG (F_1-score of 0.23 for glasso, 0.37 for t*-lasso, 0.39 for robust-NLL and 0.41 for trim-glasso, where the F_1 score is the harmonic mean between precision and recall). For instance our approach recovers several characteristics of the KEGG pathway. For instance, genes *CDC6* (a key regulator of DNA replication playing important roles in the activation and maintenance of the checkpoint mechanisms coordinating S phase and mitosis) and *PDS1* (essential gene for meiotic progression and mitotic cell cycle arrest) are identified as a hub genes, while genes *CLB3, BRN1, YCG1* are unconnected to any other genes.

7. Concluding remarks

We presented a family of trimmed estimators for a wide class of structured high-dimensional problems. We provided general results on their statistical convergence rates and consistency. In particular our results for sparse linear regression and gaussian graphical models allow to precisely characterize the impact of corruptions on the statistical performance of the resulting estimators, while recovering the rates of their 'untrimmed' counterparts under clean data. We showed how to efficiently adapt existing optimization algorithms to solve the

modified trimmed problems. Relevant directions for future work include specializing our theoretical analysis to generalized linear models, applying and analyzing trimmed approaches for more general structural regularizations, and the study of concomitant selection of the amount of trimming.

Appendix A: Proof of Theorem 1

We use the shorthand for local optimal error vector: $\widetilde{\Delta} := \widetilde{\theta} - \theta^*$ and $\widetilde{\Gamma} := \widetilde{w} - w^*$ where $(\widetilde{\theta}, \widetilde{w})$ is an *arbitrary* local optimum of M-estimator of (4). Our proof mainly uses the fact that $(\widetilde{\theta}, \widetilde{w})$ is a local minimum of (4) satisfying

$$\big\langle \nabla_\theta \mathcal{L}\big(\theta^* + \widetilde{\Delta}, w^* + \widetilde{\Gamma}\big), \widetilde{\theta} - \theta \big\rangle \leq -\big\langle \partial \lambda \|\theta^* + \widetilde{\Delta}\|_1, \widetilde{\theta} - \theta \big\rangle \quad \text{for any feasible } \theta.$$

This inequality comes from the first order stationary condition (see Loh and Wainwright [25] for details) in terms of only θ fixing w at $\widetilde{w}$. In order to provide the complete proof of the theorem, we need to define the set of notations on the model space, perturbation space and corresponding projections following Negahban et al. [29]. The sparse LTS (3) is a typical example of (4), and such notations can be naturally defined based on the true support set S. In this proof, we specifically focus on the case with $\mathcal{R}(\cdot) := \|\cdot\|_1$ for notational simplicity, but statements here can be seamlessly extendible for the general regularizer $\mathcal{R}(\cdot)$ and the appropriately defined model/perturbation spaces.

If we take $\theta = \theta^*$ above, we have

$$\big\langle \nabla_\theta \mathcal{L}\big(\theta^* + \widetilde{\Delta}, w^* + \widetilde{\Gamma}\big), \widetilde{\Delta} \big\rangle \leq -\big\langle \partial \lambda \|\theta^* + \widetilde{\Delta}\|_1, \widetilde{\Delta} \big\rangle \overset{(i)}{\leq} \lambda(\|\theta^*\|_1 - \|\widetilde{\theta}\|_1)$$

$$\leq \lambda(\|\theta^*\|_1 + \|\widetilde{\Delta}_{S^c}\|_1 - \|\widetilde{\Delta}_{S^c}\|_1 - \|\widetilde{\theta}\|_1) = \lambda(\|\theta^* + \widetilde{\Delta}_{S^c}\|_1 - \|\widetilde{\Delta}_{S^c}\|_1 - \|\widetilde{\theta}\|_1)$$

$$\overset{(ii)}{\leq} \lambda\big(\|\theta^* + \widetilde{\Delta}_{S^c} + \widetilde{\Delta}_S\|_1 + \|\widetilde{\Delta}_S\|_1 - \|\widetilde{\Delta}_{S^c}\|_1 - \|\widetilde{\theta}\|_1\big) = \lambda(\|\widetilde{\Delta}_S\|_1 - \|\widetilde{\Delta}_{S^c}\|_1),\tag{12}$$

where S is true support set of θ^*, the inequalities (i) and (ii) hold by respectively the convexity and the triangular inequality of ℓ_1 norm.

Now, by the RSC condition in (C-1), we obtain

$$\kappa_l \|\widetilde{\Delta}\|_2^2 - \tau_1(n,p)\|\widetilde{\Delta}\|_1^2$$
$$\leq \big\langle \nabla_\theta \mathcal{L}\big(\theta^* + \widetilde{\Delta}, w^*\big) - \nabla_\theta \mathcal{L}\big(\theta^*, w^*\big), \widetilde{\Delta} \big\rangle$$
$$= \Big\langle \nabla_\theta \mathcal{L}\big(\theta^* + \widetilde{\Delta}, w^* + \widetilde{\Gamma}\big) - \nabla_\theta \mathcal{L}\big(\theta^* + \widetilde{\Delta}, w^* + \widetilde{\Gamma}\big)$$
$$+ \nabla_\theta \mathcal{L}\big(\theta^* + \widetilde{\Delta}, w^*\big) - \nabla_\theta \mathcal{L}\big(\theta^*, w^*\big), \widetilde{\Delta} \Big\rangle,$$

which is equivalent with

$$\kappa_l \|\widetilde{\Delta}\|_2^2 - \tau_1(n,p)\|\widetilde{\Delta}\|_1^2 + \big\langle \nabla_\theta \mathcal{L}\big(\theta^* + \widetilde{\Delta}, w^* + \widetilde{\Gamma}\big) - \nabla_\theta \mathcal{L}\big(\theta^* + \widetilde{\Delta}, w^*\big), \widetilde{\Delta} \big\rangle$$
$$\leq \big\langle \nabla_\theta \mathcal{L}\big(\theta^* + \widetilde{\Delta}, w^* + \widetilde{\Gamma}\big) - \nabla_\theta \mathcal{L}\big(\theta^*, w^*\big), \widetilde{\Delta} \big\rangle.\tag{13}$$

Combining (12), (13) and (C-2) yields

$$\kappa_l\|\widetilde{\Delta}\|_2^2 - \tau_1(n,p)\|\widetilde{\Delta}\|_1^2 - \tau_2(n,p)\|\widetilde{\Delta}\|_2 - \tau_3(n,p)\|\widetilde{\Delta}\|_1$$
$$\leq -\left\langle \nabla_\theta \mathcal{L}\big(\theta^*, w^*\big),\, \widetilde{\Delta} \right\rangle + \lambda \left(\|\widetilde{\Delta}_S\|_1 - \|\widetilde{\Delta}_{S^c}\|_1\right)$$
$$\leq \left\|\nabla_\theta \mathcal{L}\big(\theta^*, w^*\big)\right\|_\infty \|\widetilde{\Delta}\|_1 + \lambda \left(\|\widetilde{\Delta}_S\|_1 - \|\widetilde{\Delta}_{S^c}\|_1\right).$$

Since the theorem assumes $\max\left\{\|\nabla_\theta\mathcal{L}\big(\theta^*, w^*\big)\|_\infty, 2\rho\tau_1(n,p) + \tau_3(n,p)\right\} \leq \frac{\lambda}{4}$, we can conclude that

$$0 \leq \kappa_l\|\widetilde{\Delta}\|_2^2$$
$$\leq \left\|\nabla_\theta\mathcal{L}\big(\theta^*, w^*\big)\right\|_\infty\|\widetilde{\Delta}\|_1 + \lambda\left(\|\widetilde{\Delta}_S\|_1 - \|\widetilde{\Delta}_{S^c}\|_1\right)$$
$$\qquad + \left(2\rho\tau_1(n,p) + \tau_3(n,p)\right)\|\widetilde{\Delta}\|_1 + \tau_2(n,p)\|\widetilde{\Delta}\|_2$$
$$\leq \frac{3\lambda}{2}\|\widetilde{\Delta}_S\|_1 - \frac{\lambda}{2}\|\widetilde{\Delta}_{S^c}\|_1 + \tau_2(n,p)\|\widetilde{\Delta}\|_2. \tag{14}$$

As a result, we can finally have an ℓ_2 error bound as follows:

$$\kappa_l\|\widetilde{\Delta}\|_2^2 \leq \frac{3\lambda}{2}\|\widetilde{\Delta}_S\|_1 + \tau_2(n,p)\|\widetilde{\Delta}\|_2$$
$$\leq \frac{3\lambda\sqrt{k}}{2}\|\widetilde{\Delta}_S\|_2 + \tau_2(n,p)\|\widetilde{\Delta}\|_2 \leq \left(\frac{3\lambda\sqrt{k}}{2} + \tau_2(n,p)\right)\|\widetilde{\Delta}\|_2$$

implying that

$$\|\widetilde{\Delta}\|_2 \leq \frac{1}{\kappa_l}\left(\frac{3\lambda\sqrt{k}}{2} + \tau_2(n,p)\right).$$

At the same time in order to derive ℓ_1 error bound, we again use the inequality by (14):

$$\|\widetilde{\Delta}_{S^c}\|_1 \leq 3\|\widetilde{\Delta}_S\|_1 + \frac{2}{\lambda}\tau_2(n,p)\|\widetilde{\Delta}\|_2\,.$$

Hence,

$$\|\widetilde{\Delta}\|_1 \leq \|\widetilde{\Delta}_S\|_1 + \|\widetilde{\Delta}_{S^c}\|_1 \leq 4\|\widetilde{\Delta}_S\|_1 + \frac{2}{\lambda}\tau_2(n,p)\|\widetilde{\Delta}\|_2$$
$$\leq 4\sqrt{k}\|\widetilde{\Delta}_S\|_2 + \frac{2}{\lambda}\tau_2(n,p)\|\widetilde{\Delta}\|_2$$
$$\leq \left(4\sqrt{k} + \frac{2}{\lambda}\tau_2(n,p)\right)\|\widetilde{\Delta}\|_2 \leq \frac{2}{\lambda\kappa_l}\left(2\lambda\sqrt{k} + \tau_2(n,p)\right)^2,$$

which completes the proof.

Appendix B: Proof of Corollary 1 and Corollary 2 (results for LTS)

We begin with specifying (C-1) and (C-2) for the showcasing example of (3):

$$\frac{1}{h}\sum_{i=1}^n w_i^*\langle x_i,\, \widetilde{\Delta}\rangle^2 \geq \kappa_l\|\widetilde{\Delta}\|_2^2 - \tau_1(n,p)\mathcal{R}(\widetilde{\Delta})^2, \quad \text{and} \tag{15}$$

$$\frac{1}{h} \sum_{i=1}^{n} \widetilde{\Gamma}_i \big(\langle x_i, \, \theta^* + \widetilde{\Delta} \rangle - y_i \big) \langle x_i, \, \widetilde{\Delta} \rangle \geq -\tau_2(n,p) \|\widetilde{\Delta}\|_2 - \tau_3(n,p) \mathcal{R}(\widetilde{\Delta}) \,. \quad (16)$$

In order to directly utilize Theorem 1 for linear models, we only need to show that (15) (for the condition (C-1)) and (16) (for (C-2)) hold. Throughout the proof, we use the fact that all elements in $\widetilde{\Gamma}$ corresponding to G (set of good examples) are all zeros: $\widetilde{\Gamma}_G = \mathbf{0}$ by construction.

First, consider the condition (C-1) in (15): $\frac{1}{h} \sum_{i=1}^{n} w_i^* \langle x_i, \Delta \rangle^2$. Recall that we constructed w^* as follows: w_i^* is simply set to $\widetilde{w}_i$ if $i \in G$, and $w_i^* = 0$ for $i \in B$. Hence, by construction, $\sum_{i \in G} w_i^* \geq h - (n-h)$ (since $\mathbf{1}^\top w = h$), and at least $\frac{h-(n-h)}{2}$ samples in G have $\widetilde{w}_i$ (therefore w_i^*) larger than $\frac{h-(n-h)}{2h}$. Let $\overline{G}$, which is the subset of G, be the set of such samples.

Then, $\frac{1}{h} \sum_{i=1}^{n} w_i^* \langle x_i, \Delta \rangle^2$ can be lower bounded as follows:

$$\frac{1}{h} \sum_{i=1}^{n} w_i^* \langle x_i, \, \Delta \rangle^2$$
$$= \frac{1}{h} \sum_{i \in G} w_i^* \langle x_i, \, \Delta \rangle^2 \geq \frac{1}{h} \sum_{i \in \overline{G}} w_i^* \langle x_i, \, \Delta \rangle^2 \geq \frac{h - (n-h)}{2h^2} \sum_{i \in \overline{G}} \langle x_i, \, \Delta \rangle^2 \,.$$

Noting that all $x_i \in \overline{G}$ are uncorrupted and iid sampled from $N(0, \Sigma_G)$, we can appeal to the result in Raskutti et al. [34]: with probability at least $1 - c_1 \exp\big(- c_2 |\overline{G}| \big)$,

$$\frac{1}{|\overline{G}|} \sum_{i \in \overline{G}} \langle x_i, \, \Delta \rangle^2 \geq \kappa_1 \|\Delta\|_2^2 - \kappa_2 \frac{\log p}{|\overline{G}|} \|\Delta\|_1^2 \quad \text{for all } \Delta \in \mathbb{R}^p \quad (17)$$

where κ_1 and κ_2 are strictly positive constants depending only on Σ_G. Therefore,

$$\frac{1}{h} \sum_{i=1}^{n} w_i^* \langle x_i, \, \Delta \rangle^2 \geq \frac{\big(h - (n-h) \big) |\overline{G}|}{2h^2} \left(\kappa_1 \|\Delta\|_2^2 - \kappa_2 \frac{\log p}{|\overline{G}|} \|\Delta\|_1^2 \right),$$

hence, (15) holds with

$$\kappa_l = \frac{\kappa_1 (2h - n)^2}{4h^2} \,, \, \tau_1(n,p) = \frac{\kappa_2 (2h - n) \log p}{2h^2} \quad (18)$$

since $|\overline{G}| \geq h - (n-h)$ as discussed.

Now, we consider the condition (C-2) in (16).

$$\frac{1}{h} \sum_{i=1}^{n} \Gamma_i \big(\langle x_i, \, \theta^* + \Delta \rangle - y_i \big) \langle x_i, \, \Delta \rangle$$
$$= \frac{1}{h} \sum_{i=1}^{n} \Gamma_i \langle x_i, \, \Delta \rangle^2 - \frac{1}{h} \sum_{i=1}^{n} \Gamma_i \big(\epsilon_i + \delta_i \big) \langle x_i, \, \Delta \rangle$$

$$\geq -\frac{1}{h} \sum_{i=1}^{n} \Gamma_i (\epsilon_i + \delta_i) \langle x_i, \Delta \rangle$$

where the inequality comes from (1) and from the fact that Γ_i is always greater than 0: if $i \in G$, $\Gamma_i = 0$, and if $i \in B$, $\Gamma_i := \widetilde{w}_i - w_i^* \geq 0$ since $\widetilde{w}_i \geq 0$ and $w_i^* = 0$.

Now, we follow similar strategy as in Nguyen and Tran [31]: given Δ, we divide the index of Δ into the disjoint exhaustive subsets $S_1, S_2, \ldots, S_q$ of size $|B|$ such that S_1 contains $|B|$ largest absolute elements in Δ, and so on. Then, we have

$$\left| \sum_{i=1}^{n} \Gamma_i (\epsilon_i + \delta_i) \langle x_i, \Delta \rangle \right| = \left| \sum_{i \in B} \Gamma_i \delta_i \langle x_i, \Delta \rangle \right| = \left| \sum_{i \in B} \Gamma_i \delta_i \sum_{j=1}^{q} \langle [x_i]_{S_j}, [\Delta]_{S_j} \rangle \right|$$

$$\leq \sum_j \left| \sum_{i \in B} \Gamma_i \delta_i \langle [x_i]_{S_j}, [\Delta]_{S_j} \rangle \right| \leq \sum_j \sqrt{\sum_{i \in B} \Gamma_i^2 \delta_i^2} \sqrt{\sum_{i \in B} \langle [x_i]_{S_j}, [\Delta]_{S_j} \rangle^2}$$

$$\leq \sqrt{\sum_{i \in B} \Gamma_i^2 \delta_i^2} \left(\max_j \|X_{S_j}^B\|_2 \right) \sum_j \|[\Delta]_{S_j}\|_2$$

$$\leq \sqrt{|B|} \left(\max_{i \in B} |\Gamma_i \delta_i| \right) \left(\max_j \|X_{S_j}^B\|_2 \right) \sum_j \|[\Delta]_{S_j}\|_2$$

$$\leq \sqrt{|B|} \max_{i \in B} |\delta_i| \underbrace{\left(\max_j \|X_{S_j}^B\|_2 \right)}_{\text{(I)}} \underbrace{\left(\sum_j \|[\Delta]_{S_j}\|_2 \right)}_{\text{(II)}}$$

where we use the fact that $\Gamma_i = 0$ if $i \in G$ and the Cauchy-Schwarz inequalities, and $X_{S_j}^B$ denotes $|B| \times |S_j|$ sub-matrix of $X^B \in \mathbb{R}^{|B| \times p}$ corresponding only to indices S_j.

(I): Provided $|B| \geq \exp(1)$, $\binom{p}{|B|} \leq \left(\frac{\exp(1)p}{|B|} \right)^{|B|} \leq p^{|B|}$. As discussed in Vershynin [45], Nguyen and Tran [31], for every $t > 0$,

$$\frac{1}{\sqrt{|B|}} \max_j \|X_{S_j}^B\|_2 \leq \sqrt{\|\Sigma_B\|_2} (2 + t)$$

with probability at least $1 - 2\binom{p}{|B|} \exp(-t^2 |B|/2) \geq 1 - 2\exp(-t^2 |B|/2 + |B| \log p)$. Setting $t = 2\sqrt{\log p}$, we have

$$\max_j \|X_{S_j}^B\|_2 \leq 2(1 + \sqrt{\log p}) \sqrt{\|\Sigma_B\|_2} \sqrt{|B|}$$

with probability $1 - p^{-|B|}$. In the proof of Corollary 2, $\max_j \|X_{S_j}^B\|_2 \leq f(X^B)$ $\sqrt{|B| \log p}$ by assumption (LTS5), and the remaining proof would be exactly the same.

(II): by the standard bound in Candès et al. [11], we obtain

$$\sum_j \|[\Delta]_{S_j}\|_2 = \|[\Delta]_{S_1}\|_2 + \sum_{j=2}^{q} \|[\Delta]_{S_j}\|_2 \leq \|[\Delta]_{S_1}\|_2 + \frac{1}{\sqrt{|B|}} \sum_{j=2}^{q} \|[\Delta]_{S_j}\|_1$$

$$\leq \|\Delta\|_2 + \frac{1}{\sqrt{|B|}}\|\Delta\|_1 \,.$$

Combining all pieces together yields

$$\frac{1}{h}\sum_{i=1}^{n}\Gamma_i\big(\langle x_i,\,\theta^*+\Delta\rangle - y_i\big)\langle x_i,\,\Delta\rangle \geq -\frac{1}{h}\sum_{i=1}^{n}\Gamma_i\big(\epsilon_i+\delta_i\big)\langle x_i,\,\Delta\rangle$$

$$\geq -4\sqrt{\log p}\sqrt{\|\Sigma_B\|_2}\,\max_{i\in B}|\delta_i|\,\frac{|B|}{h}\left(\|\Delta\|_2 + \frac{1}{\sqrt{|B|}}\|\Delta\|_1\right),$$

hence, we can guarantee (16) with functions

$$\tau_2(n,p) = 4\sqrt{\|\Sigma_B\|_2}\,\max_{i\in B}|\delta_i|\sqrt{\log p}\,\frac{|B|}{h}\,,$$

$$\tau_3(n,p) = 4\sqrt{\|\Sigma_B\|_2}\,\max_{i\in B}|\delta_i|\sqrt{\log p}\,\frac{\sqrt{|B|}}{h}\,.$$

To complete the proof, we need to specify the quantity $\big\|\frac{1}{h}\sum_{i=1}^{n}w_i^*\big(\langle x_i,\,\theta^*\rangle - y_i\big)x_i\big\|_\infty = \big\|\frac{1}{h}\sum_{i\in G}w_i^*\epsilon_i x_i\big\|_\infty$ for the appropriate choice of λ as stated in Theorem 1. By the sub-Gaussian property of noise vector ϵ in (LTS2): for any fixed vector v such that $\|v\|_2 = 1$,

$$\mathbb{P}\Big[|\langle v,\,\epsilon\rangle| \geq t\Big] \leq 2\exp\Big(-\frac{t^2}{2\sigma^2}\Big) \quad \text{for all } t > 0\,.$$

Given vector x_i, let x_i^j be the j-th element of vector x_i. Using the column normalization condition (LTS3) with $0 \leq w^* \leq 1$, we have for all $j = 1, \ldots, p$

$$\mathbb{P}\bigg[\Big|\frac{1}{h}\sum_{i\in G}w_i^*x_i^j\epsilon_i\Big| \geq t\bigg] \leq 2\exp\Big(-\frac{ht^2}{2\sigma^2}\Big) \quad \text{for all } t > 0\,,$$

and consequently by the union bound over,

$$\mathbb{P}\bigg[\Big\|\frac{1}{h}\sum_{i\in G}w_i^*x_i^j\epsilon_i\Big\|_\infty \geq t\bigg] \leq 2\exp\Big(-\frac{ht^2}{2\sigma^2} + \log p\Big) \quad \text{for all } t > 0\,.$$

Setting $t^2 = \frac{4\sigma^2\log p}{h}$, we obtain $\big\|\frac{1}{h}\sum_{i\in G}w_i^*\epsilon_i x_i\big\|_\infty \leq \sqrt{\frac{4\sigma^2\log p}{h}}$ with probability at least $1 - c\exp(-c'h\lambda^2)$.

Now, we have all pieces to utilize Theorem 1. The assumption on choosing λ in the statement is satisfied as follows:

$$2\rho\tau_1(n,p) + \tau_3(n,p) = 2\rho\frac{\kappa_2(2h-n)\log p}{2h^2} + 4\sqrt{\|\Sigma_B\|_2}\,\max_{i\in B}|\delta_i|\sqrt{\log p}\frac{\sqrt{|B|}}{h}$$

$$\leq C_1\sqrt{\frac{h}{\log p}}\frac{\kappa_2(2h-n)\log p}{2h^2} + 4\sqrt{\|\Sigma_B\|_2}C_2\sqrt{\frac{h}{|B|}}\sqrt{\log p}\frac{\sqrt{|B|}}{h}$$

$$\leq \Big(\frac{1}{2} C_1 \kappa_2 + 4 C_2 \sqrt{\| \Sigma_B \|_2} \Big) \sqrt{\frac{\log p}{h}}$$

where C_2 is some constant satisfying $C_2^2 \geq \frac{(\max_i \delta_i^2)|B|}{h}$, and we use the condition (LTS4). Finally, the RSC constant in (18) can be simply lower bounded with the assumption (C-h):

$$\frac{\kappa_1 (2h - n)^2}{4h^2} \geq \kappa_1 \frac{\alpha^2}{4} \,,$$

hence we can have the bounds as stated.

Appendix C: Results for trimmed graphical Lasso

C.1.   Useful lemma(s)

Lemma 1 (Lemma 1 of Ravikumar et al. [35]). *Suppose that $\{X^{(i)}\}_{i=1}^n$ are iid samples from $N(0, \Sigma)$ with $n \geq 40 \max_i \Sigma_{ii}$. Let $\mathcal{A}$ be the event that*

$$\left\| \frac{1}{n} \sum_{i=1}^n X^{(i)} (X^{(i)})^\top - \Sigma \right\|_\infty \leq 8 (\max_i \Sigma_{ii}) \sqrt{\frac{10 \tau \log p}{n}}$$

where τ is any constant greater than 2. Then, the probability of event $\mathcal{A}$ occurring is at least $1 - 4/p^{\tau - 2}$.

Lemma 2 (Section B.4 of Loh and Wainwright [26]). *For any $\Delta \in \mathbb{R}^{p \times p}$ such that $\| \Delta \|_F \leq 1$,*

$$\left\langle\!\!\left\langle \left(\Theta^* \right)^{-1} - \left(\Theta^* + \Delta \right)^{-1}, \Delta \right\rangle\!\!\right\rangle \geq \left(\| \Theta^* \|_2 + 1 \right)^{-2} \| \Delta \|_F^2 \,.$$

C.2.   Proof of Corollary 3

Although Theorem 1 can be seamlessly applied for the Trimmed Graphical Lasso as well, we need to restrict our attention to the case of $\| \Delta \|_F \leq 1$ in order to guarantee the (vanilla) restricted strong convex in Lemma 2 (which is the standard technique even for the case without outliers as developed in Loh and Wainwright [26]). Toward this, we first show that $\| \Delta \|_F \leq 1$ actually holds under the conditions:

Lemma 3. *Suppose that the condition (C-2) holds. Moreover, $4 \max \big\{ \| \frac{1}{h} \sum_{i=1}^n w_i^* X^{(i)} (X^{(i)})^\top - (\Theta^*)^{-1} \|_\infty, \tau_3(n, p) \big\} \leq \lambda \leq \frac{\kappa_l - \tau_2(n,p)}{3R}$. Then, for $(\widetilde{\Theta}, \widetilde{w})$, $\| \widetilde{\Delta} \|_F \leq 1$.*

Proof. The Lemma 3 can be proved by the fact $-\log \det \Theta$ is a convex function. Hence, the function $f : [0, 1] \to \mathbb{R}$ given by $f(t; \Theta^*, \widetilde{\Delta}) := -\log \det \left(\Theta^* + t \widetilde{\Delta} \right)$ is also convex in t, and $\left\langle\!\!\left\langle -(\Theta^* + \widetilde{\Delta})^{-1}, \widetilde{\Delta} \right\rangle\!\!\right\rangle \geq \left\langle\!\!\left\langle -(\Theta^* + t\widetilde{\Delta})^{-1}, \widetilde{\Delta} \right\rangle\!\!\right\rangle$ for $t \in [0, 1]$ (see Loh and Wainwright [26] for details).

Now, suppose that $\|\widetilde{\Delta}\|_F \geq 1$. Then, we have

$$\left\langle\!\!\left\langle\left(\Theta^*\right)^{-1} - \left(\Theta^* + \widetilde{\Delta}\right)^{-1}, \widetilde{\Delta}\right\rangle\!\!\right\rangle \geq \left\langle\!\!\left\langle\left(\Theta^*\right)^{-1} - \left(\Theta^* + t\widetilde{\Delta}\right)^{-1}, \widetilde{\Delta}\right\rangle\!\!\right\rangle$$
$$= \frac{1}{t}\left\langle\!\!\left\langle\left(\Theta^*\right)^{-1} - \left(\Theta^* + t\widetilde{\Delta}\right)^{-1}, t\widetilde{\Delta}\right\rangle\!\!\right\rangle. \tag{19}$$

Since $\|\widetilde{\Delta}\|_F \geq 1$, we can set $t = \frac{1}{\|\widetilde{\Delta}\|_F} \leq 1$ so that $\|t\widetilde{\Delta}\|_F = 1$. Hence, by applying Lemma 2 for $t\widetilde{\Delta}$, we obtain

$$\left\langle\!\!\left\langle\left(\Theta^*\right)^{-1} - \left(\Theta^* + t\widetilde{\Delta}\right)^{-1}, t\widetilde{\Delta}\right\rangle\!\!\right\rangle \geq \kappa_l\|t\widetilde{\Delta}\|_F^2 = \kappa_l.$$

Combining with (19) yields

$$\left\langle\!\!\left\langle\left(\Theta^*\right)^{-1} - \left(\Theta^* + \widetilde{\Delta}\right)^{-1}, \widetilde{\Delta}\right\rangle\!\!\right\rangle \geq \kappa_l\|\widetilde{\Delta}\|_F. \tag{20}$$

Now, from (12) and (20) followed by the condition (C-2) and Hölder's inequity, we can obtain

$$\kappa_l\|\widetilde{\Delta}\|_F \leq \left\langle\!\!\left\langle(\Theta^*)^{-1} - \frac{1}{h}\sum_{i=1}^n \widetilde{w}_i X^{(i)}(X^{(i)})^\top, \widetilde{\Delta}\right\rangle\!\!\right\rangle + \lambda(\|\widetilde{\Delta}_S\|_{1,\text{off}} - \|\widetilde{\Delta}_{S^c}\|_{1,\text{off}})$$

$$\leq \left\langle\!\!\left\langle(\Theta^*)^{-1} - \frac{1}{h}\sum_{i=1}^n \widetilde{w}_i X^{(i)}(X^{(i)})^\top, \widetilde{\Delta}\right\rangle\!\!\right\rangle + \lambda\|\widetilde{\Delta}\|_{1,\text{off}}$$

$$\leq \left\langle\!\!\left\langle(\Theta^*)^{-1} - \frac{1}{h}\sum_{i=1}^n w_i^* X^{(i)}(X^{(i)})^\top + \frac{1}{h}\sum_{i=1}^n w_i^* X^{(i)}(X^{(i)})^\top\right.\right.$$
$$\left.\left. - \frac{1}{h}\sum_{i=1}^n \widetilde{w}_i X^{(i)}(X^{(i)})^\top, \widetilde{\Delta}\right\rangle\!\!\right\rangle + \lambda\|\widetilde{\Delta}\|_{1,\text{off}}$$

$$\leq \left\|\frac{1}{h}\sum_{i=1}^n w_i^* X^{(i)}(X^{(i)})^\top - (\Theta^*)^{-1}\right\|_\infty \cdot \|\widetilde{\Delta}\|_1 + \tau_2(n,p)\|\widetilde{\Delta}\|_F$$
$$+ \tau_3(n,p)\|\widetilde{\Delta}\|_1 + \lambda\|\widetilde{\Delta}\|_{1,\text{off}}.$$

By the choice of λ in the assumption of the statement and by the fact that $\|\widetilde{\Delta}\|_{1,\text{off}} \leq \|\widetilde{\Delta}\|_1$ and $\|\widetilde{\Delta}\|_1 \leq \|\widetilde{\Theta}\|_1 + \|\Theta^*\|_1 \leq 2R$, we can rearrange the above inequality into

$$\|\widetilde{\Delta}\|_F \leq \frac{3\lambda}{2\left(\kappa_l - \tau_2(n,p)\right)}\|\widetilde{\Delta}\|_1 \leq \frac{3\lambda R}{\left(\kappa_l - \tau_2(n,p)\right)} \leq 1,$$

which conflicts with the assumption in the beginning of this proof. Hence, by contradiction, we can conclude $\|\widetilde{\Delta}\|_F \leq 1$ under conditions in the statement. □

Since for this particular example, the modified restricted strong convexity condition in (C-1) is identical as the vanilla case (which is already proved in Lemma 2), the only remaining to utilize Theorem 1 is to specify the quantity

$\tau_2(n,p)$ and $\tau_3(n,p)$ in (C-2). Toward this, we follow similar strategy as in Nguyen and Tran [31]: given Δ, we divide the index of Δ into the disjoint exhaustive subsets $S_1, S_2, \ldots, S_q$ of size $|B|$ such that S_1 contains $|B|$ largest absolute elements in Δ, and so on. Then, we have

$$\left| \left\langle\!\!\left\langle \sum_{i=1}^{n} \widetilde{\Gamma}_i X^{(i)}(X^{(i)})^\top, \widetilde{\Delta} \right\rangle\!\!\right\rangle \right| = \left| \left\langle\!\!\left\langle \sum_{i \in B} \widetilde{\Gamma}_i X^{(i)}(X^{(i)})^\top, \widetilde{\Delta} \right\rangle\!\!\right\rangle \right|$$

$$= \left| \sum_{j=1}^{q} \left\langle\!\!\left\langle \sum_{i \in B} \widetilde{\Gamma}_i \big[X^{(i)}(X^{(i)})^\top\big]_{S_j}, \big[\widetilde{\Delta}\big]_{S_j} \right\rangle\!\!\right\rangle \right|$$

$$\leq \sum_{j=1}^{q} \left| \left\langle\!\!\left\langle \sum_{i \in B} \widetilde{\Gamma}_i \big[X^{(i)}(X^{(i)})^\top\big]_{S_j}, \big[\widetilde{\Delta}\big]_{S_j} \right\rangle\!\!\right\rangle \right|.$$

Let $D_{\widetilde{\Gamma}}$ be a $|B| \times |B|$ diagonal matrix whose i-th diagonal entry is $[D_{\widetilde{\Gamma}}]_{ii} := \widetilde{\Gamma}_i$. Let also X^B is a $|B| \times p$ design matrix for samples in the set B. Finally $X^B_{S_j}$ denotes a $|B| \times |S_j|$ sub-matrix of X^B whose columns are indexed by S_j. Then,

$$\sum_{j=1}^{q} \left| \left\langle\!\!\left\langle \sum_{i \in B} \widetilde{\Gamma}_i \big[X^{(i)}(X^{(i)})^\top\big]_{S_j}, \big[\widetilde{\Delta}\big]_{S_j} \right\rangle\!\!\right\rangle \right| = \sum_{j=1}^{q} \left| \text{Trace}\Big(\big[\widetilde{\Delta}\big]_{S_j}^\top [X^B_{S_j}]^\top D_{\widetilde{\Gamma}} X^B_{S_j} \Big) \right|$$

$$= \sum_{j=1}^{q} \left| \left\langle\!\!\left\langle X^B_{S_j} \big[\widetilde{\Delta}\big]_{S_j}, D_{\widetilde{\Gamma}} X^B_{S_j} \right\rangle\!\!\right\rangle \right| \leq \sum_{j=1}^{q} \big\| X^B_{S_j} \big[\widetilde{\Delta}\big]_{S_j} \big\|_{\mathrm{F}} \big\| D_{\widetilde{\Gamma}} X^B_{S_j} \big\|_{\mathrm{F}}$$

$$\leq \sqrt{|B|}\Big(\max_j \big\| X^B_{S_j} \big\|_2 \Big)^2 \underbrace{\Big(\sum_j \big\| [\widetilde{\Delta}]_{S_j} \big\|_{\mathrm{F}} \Big)}_{(\mathrm{I})}.$$

(I): by the standard bound in Candès et al. [11], we obtain

$$\sum_{j} \big\| [\widetilde{\Delta}]_{S_j} \big\|_{\mathrm{F}} = \big\| [\widetilde{\Delta}]_{S_1} \big\|_{\mathrm{F}} + \sum_{j=2}^{q} \big\| [\widetilde{\Delta}]_{S_j} \big\|_1 \leq \big\| [\widetilde{\Delta}]_{S_1} \big\|_{\mathrm{F}} + \frac{1}{\sqrt{|B|}} \sum_{j=2}^{q} \big\| [\widetilde{\Delta}]_{S_j} \big\|_1$$

$$\leq \big\| \widetilde{\Delta} \big\|_{\mathrm{F}} + \frac{1}{\sqrt{|B|}} \big\| \widetilde{\Delta} \big\|_1.$$

Combining all pieces together yields

$$\left| \left\langle\!\!\left\langle \frac{1}{h} \sum_{i=1}^{n} \widetilde{\Gamma}_i X^{(i)}(X^{(i)})^\top, \widetilde{\Delta} \right\rangle\!\!\right\rangle \right| \leq \frac{\sqrt{|B|}}{h} \Big(\max_j \big\| X^B_{S_j} \big\|_2 \Big)^2 \left(\big\| \widetilde{\Delta} \big\|_{\mathrm{F}} + \frac{1}{\sqrt{|B|}} \big\| \widetilde{\Delta} \big\|_1 \right),$$

hence, we can guarantee the condition (C-2) with functions

$$\tau_2(n,p) = f(X^B) \sqrt{\frac{|B| \log p}{h}} \quad \text{and}$$

$$\tau_3(n,p) = f(X^B) \sqrt{\frac{\log p}{h}}.$$

To complete the proof, we also need to specify the quantity $\left\| \frac{1}{h} \sum_{i=1}^{n} w_i^* \times X^{(i)} (X^{(i)})^\top - (\Theta^*)^{-1} \right\|_\infty = \left\| \frac{1}{h} \sum_{i \in G} w_i^* X^{(i)} (X^{(i)})^\top - (\Theta^*)^{-1} \right\|_\infty$ for the appropriate choice of λ as stated in Theorem 1. Recall that we constructed w^* as follows: w_i^* is simply set to $\widetilde{w}_i$ if $i \in G$, and $w_i^* = 0$ for $i \in B$. Let $\overline{G}$ be the subset of G such that $w_i^* = 1$ and $\overline{G}^c$ be the subset such that $w_i^* = 0$. Then, we have $h \geq |\overline{G}| \geq h - |B|$, and hence $h - |\overline{G}| \leq |B|$. Now, by Lemma 1, we can obtain the following bound:

$$\left\| \frac{1}{h} \sum_{i=1}^{n} w_i^* X^{(i)} (X^{(i)})^\top - (\Theta^*)^{-1} \right\|_\infty = \left\| \frac{|\overline{G}|}{h} \frac{1}{|\overline{G}|} \sum_{i \in G} X^{(i)} (X^{(i)})^\top - (\Theta^*)^{-1} \right\|_\infty$$

$$= \left\| \frac{|\overline{G}|}{h} \left(\frac{1}{|\overline{G}|} \sum_{i \in \overline{G}} X^{(i)} (X^{(i)})^\top - (\Theta^*)^{-1} \right) - \left(\frac{h - |\overline{G}|}{h} \right) (\Theta^*)^{-1} \right\|_\infty$$

$$\leq \left\| \frac{|\overline{G}|}{h} \left(\frac{1}{|\overline{G}|} \sum_{i \in \overline{G}} X^{(i)} (X^{(i)})^\top - (\Theta^*)^{-1} \right) \right\|_\infty + \left\| \left(\frac{h - |\overline{G}|}{h} \right) (\Theta^*)^{-1} \right\|_\infty$$

$$\leq \left\| \frac{1}{|\overline{G}|} \sum_{i \in \overline{G}} X^{(i)} (X^{(i)})^\top - (\Theta^*)^{-1} \right\|_\infty + \frac{|B|}{h} \|\Sigma^*\|_\infty$$

$$\leq 8 (\max_i \Sigma_{ii}^*) \sqrt{\frac{10\tau \log p}{\overline{G}}} + \frac{|B|}{h} \|\Sigma^*\|_\infty \leq 8 (\max_i \Sigma_{ii}^*) \sqrt{\frac{10\tau \log p}{h - |B|}} + \frac{|B|}{h} \|\Sigma^*\|_\infty$$

with probability at least $1 - 4/p^{\tau - 2}$ for any $\tau > 2$.

C.3. Proof of Corollary 4

Under (TGL2), $h - |B| \geq (n - a\sqrt{n}) - a\sqrt{n}$. Hence, if $n \geq 16a^2$, then $h - |B| \geq (n - a\sqrt{n}) - a\sqrt{n} \geq \frac{n}{2}$. Moreover, $\frac{|B|}{h} \leq \frac{a\sqrt{n}}{n/2} \leq \frac{2a}{\sqrt{n}}$. Therefore, from the Corollary 3, the selection of λ in the statement satisfies $\lambda \geq 4 \max \left\{ \|\frac{1}{h} \sum_{i=1}^{n} w_i^* \times X^{(i)} (X^{(i)})^\top - (\Theta^*)^{-1} \|_\infty, \tau_3(n, p) \right\}$.

Furthermore, as long as $n \geq \left(\|\Theta^*\|_2 + 1 \right)^4 \left(3Rc + f(X^B) \sqrt{2|B|} \right)^2 (\log p)$,

$$\lambda = c \sqrt{\frac{\log p}{n}} \leq \frac{\left(\|\Theta^*\|_2 + 1 \right)^{-2} - f(X^B) \sqrt{\frac{2|B| \log p}{n}}}{3R},$$

and therefore we have $\lambda \leq \frac{\kappa_l - f(X^B) \sqrt{\frac{|B| \log p}{h}}}{3R}$ where c is defined as $4 \max \{ 16 (\max_i \Sigma_{ii}^*) \sqrt{5\tau} + \frac{2a \|\Sigma^*\|_\infty}{\sqrt{\log p}}, \sqrt{2} f(X^B) \}$, as stated.

C.4. Proof of Corollary 5

In this proof, we simply need to specify the quantity $f(X^B)$ under the condition (TGL2) and (TGL3), and then we can appeal to the result in Corollary 4.

Provided $|B| \geq \exp(1)$, $\binom{p}{|B|} \leq \left(\frac{\exp(1)p}{|B|} \right)^{|B|} \leq p^{|B|}$. As discussed in Vershynin [45], Nguyen and Tran [31], if (TGL3) holds, for every $t > 0$, we have

$$\frac{1}{\sqrt{|B|}} \max_j \|X_{S_j}^B\|_2 \leq \sqrt{\|\Sigma_B\|_2} \, (2+t)$$

with probability at least $1-2\binom{p}{|B|}\exp(-t^2|B|/2) \geq 1-2\exp(-t^2|B|/2+|B|\log p)$. Setting $t = 2\sqrt{\log p}$, we obtain

$$\max_j \|X_{S_j}^B\|_2 \leq 2(1+\sqrt{\log p})\sqrt{\|\Sigma_B\|_2}\sqrt{|B|}$$

with probability $1 - p^{-|B|}$. Therefore, under (TGL2),

$$\left(\max_j \|X_{S_j}^B\|_2\right)^2 \leq 4\big(1+\sqrt{\log p}\big)^2\|\Sigma_B\|_2|B| \leq 4a\big(1+\sqrt{\log p}\big)^2\|\Sigma_B\|_2\sqrt{n}$$

$$= \frac{4a\big(1+\sqrt{\log p}\big)^2\|\Sigma_B\|_2\sqrt{n}}{\sqrt{h\log p}}\sqrt{h\log p} \leq \frac{4\sqrt{2}a\big(1+\sqrt{\log p}\big)^2\|\Sigma_B\|_2}{\sqrt{\log p}}\sqrt{h\log p},$$

as specified in the statement.

Appendix D: Proof of Proposition 1

We assume that the initial algorithm $\mathcal{A}$ is monotone, so the objective is non-increasing in each iteration. Therefore for any w we know

$$f(\theta^{(t+1)}; w) - f(\theta^{(t)}; w) \leq 0\,.$$

Setting $w = w^{(t)}$ above, we have

$$f(\theta^{(t+1)}; w^{(t)}) \leq f(\theta^{(t)}; w^{(t)})\,. \tag{21}$$

Since $w^{(t+1)}$ is computed to minimize $\min_w f(\theta^{(t+1)}; w)$, we also know

$$f(\theta^{(t+1)}; w^{(t+1)}) \leq f(\theta^{(t+1)}; w^{(t)})\,. \tag{22}$$

By combining (21) and (22), we obtain

$$f(\theta^{(t+1)}; w^{(t+1)}) \leq f(\theta^{(t)}; w^{(t)}) \quad \text{for all} \quad t, \tag{23}$$

establishing monotonic decrease of function values. Since the domain of F is compact, we know a limit point exists.

Next, we can take each $w^{(t)}$ to be a vertex of the capped simplex, since the subproblem for w is a linear program (indeed, our implementation only chooses vertex solutions $w^{(t)}$). Therefore, along a subsequence t_k that converges to any limit point $(v, \overline{\theta})$, the weights w^{t_k} converge to v after finitely many steps (since all vertices are separated by some positive distance). Once $w^{(t_k)}$ have converged v, iterates in the extended framework are identical to those generated by the original algorithm $\mathcal{A}$ for the associated data selection, and therefore $\overline{\theta}$ is a stationary point for the associated M-estimator. Then $(v, \overline{\theta})$ is a stationary point for the overall problem.

Suppose now that two limit points correspond to two different vertices v_1 and v_2. Each vertex of the capped simplex corresponds to a selection of data points, which we call $\mathcal{D}_1$ and $\mathcal{D}_2$. Consider subsequences t_{k_1} and t_{k_2} which converge to v_1 and v_2, respectively. Along each subsequence, $w^{(t_{k_i})}$ converge to v_i after finitely many steps as discussed above, and again the iterates of the extended algorithm are identical to those generated by $\mathcal{A}$ for the M-estimators defined over $\mathcal{D}_1$ and $\mathcal{D}_2$.

In order to make a stronger statement, we need to make stronger assumptions. Suppose that

1. the original M-estimator is convex, and
2. the optimization problems over each vertex v_k (corresponding to data selection $\mathcal{D}_k$) have different optimal values.

Then there exists an $\epsilon > 0$ so that without loss of generality, $f(\theta_1^*; v_1) + \epsilon \leq f(\theta_1^*; v_2)$. Now, since each problem is convex over its respective dataset, we can guarantee that after $k \geq T$ steps of $\mathcal{A}$ along the subsequence t_{k_1}, we have $f(\theta^{t_{k_1}}; v_1) < f(\theta_1^*; v_1) + \frac{\epsilon}{2} < f(\theta_1^*; v_2)$, and it is impossible for the algorithm to return to v_2 by the already established descent property (23). The number of iterations can be precisely quantified, see e.g. Nesterov [30].

The contradiction ensures that the weights converge after finitely many steps to a single vertex v. Once the weights converge to v, we know that *all iterates* of the extended algorithm are identical to those of $\mathcal{A}$ for the convex problem defined over selection $\mathcal{D}$ associated to v, and the extended algorithm converges to a stationary point of the problem.

Acknowledgements

References

[1] Alfons, A., Croux, C., and Gelper, S. (2013), "Sparse least trimmed squares regression for analyzing high-dimensional large data sets," *Ann. Appl. Stat.*, 7, 226–248. MR3086417
[2] Aravkin, A. Y. and Van Leeuwen, T. (2012), "Estimating nuisance parameters in inverse problems," *Inverse Problems*, 28, 115016. MR2997233
[3] Bannerjee, O., Ghaoui, L. E., and d'Aspremont, A. (2008), "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data," *Jour. Mach. Lear. Res.*, 9, 485–516. MR2417243

[4] Belloni, A., Chernozhukov, V., Kaul, A., Rosenbaum, M., and Tsybakov, A. B. (2017), "Pivotal Estimation via Self-Normalization for High-Dimensional Linear Models with Error in Variables," *arXiv preprint arXiv:1708.08353*.

[5] Belloni, A., Chernozhukov, V., and Wang, L. (2011), "Square-root lasso: pivotal recovery of sparse signals via conic programming," *Biometrika*, 98, 791–806. MR2860324

[6] Bhatia, K., Jain, P., and Kar, P. (2015), "Robust Regression via Hard Thresholding," in *Neur. Info. Proc. Sys. (NIPS)*.

[7] Boyd, S. and Vandenberghe, L. (2004), *Convex optimization*, Cambridge, UK: Cambridge University Press. MR2061575

[8] Brem, R. B. and Kruglyak, L. (2005), "The landscape of genetic complexity across 5,700 gene expression traits in yeast," *Proceedings of the National Academy of Sciences of the United States of America*, 102, 1572–1577.

[9] Brem, R. B., Storey, J. D., Whittle, J., and Kruglyak, L. (2005), "Genetic interactions between polymorphisms that affect gene expression in yeast." *Nature*, 436, 701–703.

[10] Bunea, F. (2008), "Honest variable selection in linear and logistic regression models via l1 and l1 + l2 penalization," *Electron. J. Stat.*, 2, 1153–1194. MR2461898

[11] Candès, E., Romberg, J., and Tao, T. (2006), "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, 59, 1207–1223. MR2230846

[12] Chen, Y., Caramanis, C., and Mannor, S. (2013), "Robust High Dimensional Sparse Regression and Matching Pursuit," *The Proceedings of the International Conference on Machine Learning (ICML)*.

[13] Chetverikov, D., Liao, Z., and Chernozhukov, V. (2017), "On cross-validated Lasso," *Arxiv preprint arXiv:1605.02214*.

[14] Cross, G. and Jain, A. (1983), "Markov Random Field Texture Models," *IEEE Trans. PAMI*, 5, 25–39.

[15] Daye, Z., Chen, J., and H., L. (2012), "High-Dimensional Heteroscedastic Regression with an Application to eQTL Data Analysis," *Biometrics*, 68, 316–326. MR2909888

[16] Finegold, M. and Drton, M. (2011), "Robust graphical modeling of gene networks using classical and alternative T-distributions," *The Annals of Applied Statistics*, 5, 1057–1080. MR2840186

[17] Friedman, J., Hastie, T., and Tibshirani, R. (2007), "Sparse inverse covariance estimation with the graphical Lasso," *Biostatistics*.

[18] Golub, G. and Pereyra, V. (2003), "Separable nonlinear least squares: the variable projection method and its applications," *Inverse Problems*, 19, R1–R26. MR1991786

[19] Hassner, M. and Sklansky, J. (1978), "Markov Random Field Models of Digitized Image Texture," in *ICPR78*, pp. 538–540.

[20] Ising, E. (1925), "Beitrag zur Theorie der Ferromagnetismus," *Zeitschrift für Physik*, 31, 253–258.

[21] Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014), "Data, information, knowledge and principle: back to metabolism in KEGG," *Nucleic Acids Res.*, 42, D199–D205.

[22] Lambert-Lacroix, S., Zwald, L., et al. (2011), "Robust regression through the Huber's criterion and adaptive lasso penalty," *Electronic Journal of Statistics*, 5, 1015–1053. MR2836768

[23] Lauritzen, S. (1996), *Graphical models*, Oxford University Press, USA. MR1419991

[24] Liu, L., Shen, Y., Li, T., and Caramanis, C. (2018), "High dimensional robust sparse regression," *Arxiv preprint arXiv:1805.11643*.

[25] Loh, P. and Wainwright, M. J. (2015), "Regularized M-estimators with Nonconvexity: Statistical and Algorithmic Theory for Local Optima," *Journal of Machine Learning Research (JMLR)*, 16, 559–616. MR3335800

[26] Loh, P.-L. and Wainwright, M. J. (2013), "Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima," in *Neur. Info. Proc. Sys. (NIPS)*, 26.

[27] Manning, C. D. and Schutze, H. (1999), *Foundations of Statistical Natural Language Processing*, MIT Press. MR1722790

[28] Meinshausen, N. and Bühlmann, P. (2006), "High-dimensional graphs and variable selection with the Lasso," *Annals of Statistics*, 34, 1436–1462. MR2278363

[29] Negahban, S., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012), "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statistical Science*, 27, 538–557. MR3025133

[30] Nesterov, Y. (2004), *Introductory lectures on convex optimization*, vol. 87 of *Applied Optimization*, Kluwer Academic Publishers, Boston, MA, a basic course. MR2142598

[31] Nguyen, N. H. and Tran, T. D. (2013), "Robust Lasso with missing and grossly corrupted observations," *IEEE Trans. Info. Theory*, 59, 2036–2058. MR3043781

[32] Oh, J. H. and Deasy, J. O. (2014), "Inference of radio-responsive gene regulatory networks using the graphical lasso algorithm," *BMC Bioinformatics*, 15, S5.

[33] Prasad, A., Suggala, A. S., Balakrishnan, S., and Ravikumar, P. (2018), "Robust Estimation via Robust Gradient Estimation," *Arxiv preprint arXiv:1802.06485*.

[34] Raskutti, G., Wainwright, M. J., and Yu, B. (2010), "Restricted Eigenvalue Properties for Correlated Gaussian Designs," *Journal of Machine Learning Research (JMLR)*, 99, 2241–2259. MR2719855

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

[36] Recht, B., Fazel, M., and Parrilo, P. A. (2010), "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, 52, 471–501. MR2680543

[37] Ripley, B. D. (1981), *Spatial statistics*, New York: Wiley. MR0624436

[38] Rosenbaum, M. and Tsybakov, A. B. (2010), "Sparse recovery under matrix uncertainty," *The Annals of Statistics*, 2620–2651. MR2722451

[39] Rousseeuw, P. J. (1984), "Least median of squares regression," *J. Amer. Statist. Assoc.*, 79, 871–880. MR0770281

[40] Stratton, H., Zhou, J., Reed, S., and Stone, D. (1996), "The Mating-Specific Galpha Protein of Saccharomyces cerevisiae Downregulates the Mating Signal by a Mechanism That Is Dependent on Pheromone and Independent of Gbetagamma Sequestration," *Molecular and Cellular Biology*.

[41] Sun, H. and Li, H. (2012), "Robust Gaussian graphical modeling via l1 penalization," *Biometrics*, 68, 1197–206. MR3040026

[42] Tibshirani, J. and Manning, C. D. (2014), "Robust Logistic Regression using Shift Parameters." in *ACL (2)*, pp. 124–129.

[43] Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288. MR1379242

[44] van de Geer, S. and Buhlmann, P. (2009), "On the conditions used to prove oracle results for the Lasso," *Electronic Journal of Statistics*, 3, 1360–1392. MR2576316

[45] Vershynin, R. (2012), "Introduction to the non-asymptotic analysis of random matrices," in *Compressed Sensing: Theory and Applications*, eds. Eldar, Y. and Kutyniok, G., Cambridge University Press, pp. 210–268, forthcoming. MR2963170

[46] Wainwright, M. J. (2009), "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)," *IEEE Trans. Information Theory*, 55, 2183–2202. MR2729873

[47] Wang, H., Li, G., and Jiang, G. (2007), "Robust regression shrinkage and consistent variable selection through the LAD-lasso," *Journal of Business and Economics Statistics*, 25, 347–355. MR2380753

[48] Woods, J. (1978), "Markov Image Modeling," *IEEE Transactions on Automatic Control*, 23, 846–850.

[49] Yang, E. and Ravikumar, P. (2013), "Dirty Statistical Models," in *Neur. Info. Proc. Sys. (NIPS)*, 26.

[50] Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. (2012), "Graphical Models via Generalized Linear Models," in *Neur. Info. Proc. Sys. (NIPS)*, 25.

[51] Yang, E., Tewari, A., and Ravikumar, P. (2013), "On Robust Estimation of High Dimensional Generalized Linear Models," in *Inter. Joint Conf. on Artificial Intelligence*, 13.

[52] Yuan, M. and Lin, Y. (2007), "Model selection and estimation in the Gaussian graphical model," *Biometrika*, 94, 19–35. MR2367824

[53] Zhang, X., Xu, C., Zhang, Y., Zhu, T., and Cheng, L. (2017), "Multivariate Regression with Grossly Corrupted Observations: A Robust Approach and its Applications," *Arxiv preprint arXiv:1701.02892*.

[54] Zhang, X., Zhao, L., Boedihardjo, A. P., and Lu, C.-T. (2017), "Robust Regression via Heuristic Hard Thresholding," in *International Joint Conference on Artificial Intelligence (IJCAI)*.