Elementary Estimators for Sparse Covariance Matrices and other Structured Moments

Eunho Yang EUNHO@CS.UTEXAS.EDU
Department of Computer Science, The University of Texas, Austin, TX 78712, USA

Aurélie C. Lozano ACLOZANO@US.IBM.COM
IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA

Pradeep Ravikumar PRADEEPR@CS.UTEXAS.EDU
Department of Computer Science, The University of Texas, Austin, TX 78712, USA

Abstract

We consider the problem of estimating expectations of vector-valued feature functions; a special case of which includes estimating the covariance matrix of a random vector. We are interested in recovery under high-dimensional settings, where the number of features p is potentially larger than the number of samples n, and where we need to impose structural constraints. In a natural distributional setting for this problem, the feature functions comprise the sufficient statistics of an exponential family, so that the problem would entail estimating *structured moments* of exponential family distributions. For instance, in the special case of covariance estimation, the natural distributional setting would correspond to the multivariate Gaussian distribution. Unlike the inverse covariance estimation case, we show that the regularized MLEs for covariance estimation, as well as natural Dantzig variants, are *non-convex*, even when the regularization functions themselves are convex; with the same holding for the general structured moment case. We propose a class of elementary convex estimators, that in many cases are available in *closed-form*, for estimating general structured moments. We then provide a unified statistical analysis of our class of estimators. Finally, we demonstrate the applicability of our class of estimators via simulation and on real-world climatology and biology datasets.

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

1. Introduction

Covariance matrix estimation is an increasingly important problem with applications in varied multivariate settings. A motivating application for this paper is climate data analysis, specifically climate change detection (Ribes et al., 2009), where covariance estimation is used for the computation of so-called Empirical Orthogonal Functions (Wikle & Cressie, 1999), which in turn are used to determine climate variability indices such as the Arctic Oscillation. In classical statistical settings where the number of observations n is much larger than the number of dimensions p, a strong statistical estimator of the population covariance matrix is the sample covariance matrix itself. Its strong statistical guarantees however fail to hold in high-dimensional settings when $p > n$ (Johnstone, 2001; Johnstone & Lu, 2004). A natural distributional setting for the covariance estimation problem is when the random vector is multivariate Gaussian: in that case the sample covariance matrix serves as the maximum likelihood estimator (MLE). In high-dimensional regimes however, such MLEs are typically not consistent, and it is necessary to use structurally constrained estimation involving regularized MLEs. We show however that even with the use of convex regularization functions, regularized MLE estimators for the covariance matrix solve *non-convex* programs. We also show that natural Dantzig variants (the Dantzig estimator technically is defined for sparse linear regression) are also non-convex.

Practical high-dimensional covariance matrix estimators have thus typically focused not on likelihood-based regularized programs, but on thresholding and shrinkage. Ledoit & Wolf (2003) for instance proposed to shrink the sample covariance matrix to the identity matrix. Bickel & Levina (2008a); Rothman et al. (2009) proposed thresholding estimators for covariance matrices under the structural assumption that each row of the covariance matrix satis-

fies a weak sparsity assumption; El Karoui (2008) considered an alternative notion of sparsity based on the number of closed paths of any length in the associated graph. There has also been a line of work on banded covariance matrices, where the entries of the covariance matrix are assumed to decay based on their distance from the diagonal. For such banded covariance matrices, Furrer & Bengtsson (2007) proposed to shrink the sample covariance entries based on this distance from the diagonal; and Bickel & Levina (2008b); Cai et al. (2010) have analyzed the consistency of such banded estimators. In recent years, there have been considerable advances in estimation of high-dimensional parameters under varied structural constraints such as sparsity, group-sparsity, low-rank structure, etc. However, these are largely restricted to regularized MLE estimators, which for the covariance estimation case, as noted above, lead to non-convex programs. This leads to the following question:

"Can we provide tractable estimators with strong statistical guarantees for high-dimensional covariance matrices under general structural constraints?"

Note that under such general structural constraints (e.g. group sparsity), the specific thresholding and shrinkage based estimators discussed above, which are designed for their specific structure, would not be applicable. Recall that the population covariance matrix is the expectation of the outer-product of a centered random vector with itself. In this paper, we thus actually consider a generalization of the above question:

"Can we provide tractable estimators with strong statistical guarantees for expectations of general vector-valued feature functions (i.e. moments) under general structural constraints?"

Note that a natural distributional assumption for this general problem entails the random vector being drawn from an exponential family with sufficient statistics set to the feature functions, so that the task reduces to recovering the *moment parameters* of this exponential family. Even in this general structured moment setting, we show that the regularized MLE estimator, even with convex regularization functions, as well as a natural Dantzig variant, lead to *non-convex* programs. We note that this problem of recovering structured moments has been the subject of much less investigation when compared to the estimation of structured *canonical parameters* of such exponential family distributions (e.g. estimation of structured inverse covariance matrices). We conjecture this is in part because regularized MLEs for the estimation of such structured canonical parameters lead to convex programs, unlike the case with structured moments. The estimation of structured moments is nonetheless an important problem: it not only includes the important covariance matrix problem, which corresponds to the multivariate Gaussian exponential family case, but also the graphical model inference problem of estimating moment parameters for general positive graphical distributions such as Ising models.

Our approach to addressing the questions above involves an estimator that solves for a parameter with minimum structural complexity subject to certain very simple structural constraints. Our estimator is reminiscent of the form of the Dantzig estimator (Candès & Tao, 2007) for sparse linear regression, but it is actually available in *closed form* for the sparse covariance case, and corresponds to very simple operations in other structural constraint settings. Our class of algorithms are thus not only computationally practical, but also highly scalable. Interestingly, even though the class of estimators is elementary, in our unified statistical analysis of our class of algorithms for general structural constraints, we show that they come with strong statistical guarantees with near-optimal convergence rates. We illustrate the applicability of our framework via simulation and by applying it to two real-world problems, one on climate analysis, and the other on 3-D organization of chromosomes.

2. Setup

Let $X \in \mathbb{R}^p$ be a random vector with distribution $\mathbb{P}$, and let $\{X_i\}_{i=1}^n$ denote n i.i.d. observations drawn from $\mathbb{P}$. In this paper, we consider the task of estimating some moment parameter $\mu^* := \mathbb{E}[\phi(X)]$ of this distribution, where $\phi : \mathbb{R}^p \mapsto \mathbb{R}^m$ is some vector-valued feature function of interest. In the analysis to follow, it will be useful to consider the empirical expectation of the feature function, $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \phi(X_i)$.

2.1. Example: Estimating Covariance Matrices

A key example of the above problem is the estimation of the covariance matrix $\Sigma^* = \mathbb{E}[(X - \mathbb{E}(X))(X - \mathbb{E}(X))^\top]$. The empirical covariance matrix is then given by $\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n \big(X_i - \overline{X}\big)\big(X_i - \overline{X}\big)^\top$, where $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. A natural distributional setting for such covariance estimation is when the random vector X is multivariate Gaussian, with mean μ^*, and covariance matrix Σ^*:

$$\mathbb{P}(X; \mu^*, \Sigma^*) \propto \exp\Big(-1/2(X - \mu^*){\Sigma^*}^{-1}(X - \mu^*)^T\Big).$$

Under this distributional setting, a natural estimator of the covariance matrix is the regularized Gaussian maximum likelihood estimator:

$$\underset{\Sigma \succeq 0}{\text{minimize}} \Big\{ \langle\langle \Sigma^{-1}, \widehat{\Sigma}_n \rangle\rangle + \text{logdet}(\Sigma) + \lambda_n \mathcal{R}(\Sigma) \Big\}, \quad (1)$$

where $\mathcal{R}(\Sigma)$ is an arbitrary penalty function encouraging specific structure in the covariance matrix, and λ_n is the corresponding regularization parameter. For instance, a

structural assumption of sparsity of the underlying covariance matrix would suggest the use of the element-wise ℓ_1 norm regularization function.

Another natural estimator is based on the Dantzig estimator (Candès & Tao, 2007). The Dantzig estimator was developed for sparse linear regression, and estimates the parameter with the minimum ℓ_1 norm that at the same time satisfies a constraint entailed by the stationary condition of the ℓ_1-regularized least squares estimator. Following this resume, we first derive the stationary condition of (1) as

$$-\Sigma^{-1}\widehat{\Sigma}_n\Sigma^{-1} + \Sigma^{-1} + \lambda_n z = 0,$$

where z here is the subgradient of $\mathcal{R}(\Sigma)$. Assume the following dual function $\mathcal{R}^*(\cdot)$ is well-defined: $\mathcal{R}^*(A) = \sup_{\Sigma:\mathcal{R}(\Sigma)\neq 0} \frac{\langle A,\Sigma\rangle}{\mathcal{R}(\Sigma)}$. When $\mathcal{R}(\cdot)$ is a vector or matrix norm for instance, $\mathcal{R}^*(\cdot)$ is the corresponding dual norm. The subgradient z then has the following property that $\mathcal{R}^*(z) \leq 1$ (Watson, 1992), which in turn entails $\mathcal{R}^*\big(\Sigma^{-1}(\widehat{\Sigma}_n - \Sigma)\Sigma^{-1}\big) \leq \lambda_n$. Thus, the counterpart of the Dantzig estimator for estimating the structured covariance matrix of multivariate Gaussian can be written as

$$\begin{aligned} &\underset{\Sigma}{\text{minimize}}\ \mathcal{R}(\Sigma) \\ &\quad \text{s.t. } \mathcal{R}^*\Big(\Sigma^{-1}(\widehat{\Sigma}_n - \Sigma)\Sigma^{-1}\Big) \leq \lambda_n. \end{aligned} \tag{2}$$

Unfortunately, it can been seen both the estimators in (1) and (2) are *non-convex*, as stated in the following proposition.

Proposition 1. *The estimation problems in* (1) *and* (2) *are both non-convex, even when the regularization function* $\mathcal{R}(\cdot)$ *itself is a convex function.*

It thus remains to derive convex tractable estimators for structured covariance matrices, even under natural distributional settings.

2.2. Estimating General Moments

The development in the previous section for the specific example of structured covariance matrices extends to the general problem of estimating expected feature functions $\mu^* := \mathbb{E}[\phi(X)]$. As in the covariance estimation case, a natural distributional setting is when the random vector X is distributed as an exponential family, with sufficient statistics set $\phi(X)$:

$$\mathbb{P}(X;\theta) = \exp\Big\{\langle\theta, \phi(X)\rangle - A(\theta)\Big\}. \tag{3}$$

Suppose that this is a minimal exponential family, so that the parameter $\theta(\mu)$ that gives rise to expected sufficient statistics (hereafter, moments) μ is obtained as $\theta(\mu) = \nabla A^*(\mu)$ where $A^*(\cdot)$ is the conjugate dual function to $A(\cdot)$ (see Wainwright & Jordan (2008) for an expanded discussion of exponential families and moments). When the distribution (3) belongs to such a minimal exponential family, it can then be re-written using moment-based parameters as: $\mathbb{P}(X;\mu) = \exp\Big\{\langle\theta(\mu), \phi(X)\rangle - A(\theta(\mu))\Big\}$.

Remark. Note that the earlier multivariate Gaussian covariance estimation problem can be re-written in this setting. Specifically, the multivariate Gaussian distribution can be written in canonical parameterization (or Gaussian Markov random fields) as: $\mathbb{P}(X;\theta,\Theta) = \exp\Big\{\langle\theta, X\rangle + \langle\Theta, XX^T\rangle - A(\theta,\Theta)\Big\}$, where $A(\theta,\Theta) = 1/2\log(n\log(2\pi)) - \log\det(-2\Theta) - 1/4\,\theta^T\Theta^{-1}\theta$. The moments of this distribution are given as $\mu = \mathbb{E}[X] = -1/2\Theta^{-1}\theta$, and $\mathbb{E}[XX^T] = \mu\mu^T - 1/2\Theta^{-1}$; so that the centered second moment is given as $\Sigma = \mathbb{E}[(X-\mu)(X-\mu)^T] = -1/2\Theta^{-1}$. Note that the multivariate Gaussian can be equivalently parameterized (as in the previous section) in terms of these moments as:

$$\mathbb{P}(X;\mu,\Sigma) \propto \exp\Big\{-1/2(X-\mu)\Sigma^{-1}(X-\mu)^T\Big\}.$$

Given the empirical moments, $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \phi(X^{(i)})$, the negative log-likelihood can then be written as:

$$\mathcal{L}(\mu) := -\langle\theta(\mu), \widehat{\mu}_n\rangle + A\big(\theta(\mu)\big),$$

so that a regularized MLE with a regularization function $\mathcal{R}(\cdot)$ is given as:

$$\underset{\mu}{\text{minimize}}\Big\{-\langle\theta(\mu), \widehat{\mu}_n\rangle + A\big(\theta(\mu)\big) + \mathcal{R}(\mu)\Big\}, \tag{4}$$

which can be seen to be non-convex in general. Let us consider the Dantzig variant in this general setting. The gradient of the negative log-likelihood is given by

$$\begin{aligned} \nabla\mathcal{L}(\mu) &= -\nabla^2 A^*(\mu)\,\widehat{\mu}_n + \nabla^2 A^*(\mu)\,\nabla A\big(\theta(\mu)\big) \\ &= \nabla^2 A^*(\mu)\big(-\widehat{\mu}_n + \mu\big). \end{aligned}$$

Thus, the "Dantzig" variant of the structured moment estimator then takes the form:

$$\begin{aligned} &\underset{\mu}{\text{minimize}}\ \mathcal{R}(\mu) \\ &\quad \text{s.t. } \mathcal{R}^*\Big(\nabla^2 A^*(\mu)(\mu - \widehat{\mu}_n)\Big) \leq \lambda_n, \end{aligned} \tag{5}$$

which too can be seen to non-convex in general, as we already observed in (1) and (2) as a special case.

We thus get a counterpart of the proposition in the structured covariance case:

Proposition 2. *The estimation problems in* (4) *and* (5) *are both non-convex programs for general exponential families, even when the regularization function* $\mathcal{R}(\cdot)$ *itself is a convex function.*

2.3. Other Examples

Other important examples of multivariate exponential families where computing moments is of interest include the Ising model (see Ravikumar et al. (2010) and references therein) and the multivariate Bernoulli distribution (Dai et al., 2013). Indeed, probabilistic graphical model distributions in general, when positive, can be expressed as exponential family distributions, and computing moments of the resulting exponential families constitutes the important problem of graphical model inference; see Wainwright & Jordan (2008) for additional discussion and examples.

In the next section, we propose an elementary estimator, that is not only convex, but also has a closed form in many cases. In the section following that, we show that the estimator while elementary, nonetheless comes with strong statistical guarantees.

3. The Elem-Moment Estimator

The previous section showed that even under natural distributional settings, *natural* estimators such as the regularized MLE, as well as a Dantzig variant, yield non-convex optimization problems for recovering structured moments. In this paper, we thus consider the following elementary estimator, which we call the "Elem-Moment" estimator, that is specified by a regularization function $\mathcal{R}(\cdot)$:

$$\begin{aligned} &\underset{\mu}{\text{minimize}}\ \mathcal{R}(\mu) \\ &\quad \text{s.t. } \mathcal{R}^*(\widehat{\mu}_n - \mu) \leq \lambda_n. \end{aligned} \tag{6}$$

where $\mathcal{R}^*(\cdot)$ is the dual function.

It can be seen that the estimator in (6) solves a convex program, unlike the regularized MLE and Dantzig variants discussed in the previous section. Moreover, while it is reminiscent of the Dantzig estimator (Candès & Tao, 2007), for many typical settings of the regularization function $\mathcal{R}(\cdot)$, the elementary estimator is available in *closed-form*.

Suppose for instance the regularization function $\mathcal{R}(\cdot)$ is given by an "atomic" gauge function, as defined in Chandrasekaran et al. (2010). Specifically, suppose we are given a set $\mathcal{A} := \{\mathbf{a}_j\}_{j \in I}$ of very "simple" objects or "atoms", and that the regularization function $\mathcal{R}(\mu)$ can be written as

$$\mathcal{R}(\mu) = \inf_{\mathbf{c}} \Big\{ \sum_{j \in I} c_j : \mu = \sum_{j \in I} c_j \mathbf{a}_j, \mathbf{a}_j \in \mathcal{A}, c_j \geq 0 \Big\}. \tag{7}$$

There, they showed that this class includes many popular regularization functions including the ℓ_1 norm, ℓ_1/ℓ_q norms, the nuclear norm, and others.

The following proposition then specifies the solution of (6) with $\mathcal{R}(\cdot)$ set to such an atomic norm.

Proposition 3. *Suppose $\mathcal{R}(\cdot)$ is an "atomic" gauge function as specified by (7). Suppose also that the optimal coefficients solving (7) with μ set to the sample expected sufficient statistics $\widehat{\mu}_n$, are given as $\{\widehat{c}_i\}_{i \in I}$, so that $\widehat{\mu}_n = \sum_{i \in I} \widehat{c}_i \mathbf{a}_i$, and $\mathcal{R}(\widehat{\mu}_n) = \sum_{i \in I} \widehat{c}_i$. Then, the optimal solution $\widehat{\mu}$ of (6) is given by*

$$\widehat{\mu} = \sum_{i \in I} \max\{\widehat{c}_i - \lambda_n, 0\} \mathbf{a}_i.$$

Remark. As a special case of the above, consider the use of ℓ_1 regularization for off-diagonals (for recovering a sparse covariance matrix for instance). The regularization function $\mathcal{R}(\cdot)$ can then be written as $\|\Sigma\|_{1,\text{off}} := \sum_{i \neq j} |\Sigma_{ij}|$. The corresponding dual-norm can then be shown to be equal to: $\|\Sigma\|_{\infty,\text{off}} := \max_{i \neq j} |\Sigma_{ij}|$. Our elementary estimator (6) with this setting of the regularization function then takes the following form:

$$\begin{aligned} &\underset{\Sigma}{\text{minimize}}\ \|\Sigma\|_{1,\text{off}} \\ &\quad \text{s.t. } \|\widehat{\Sigma}_n - \Sigma\|_{\infty,\text{off}} \leq \lambda_n. \end{aligned} \tag{8}$$

It can be seen the solution is given by the element-wise soft-thresholding of $\widehat{\Sigma}_n$ (only for off-diagonal entries), so that $\widehat{\Sigma} = \mathcal{S}_{\lambda_n}(\widehat{\Sigma}_n)$, where $[\mathcal{S}_\lambda(u)]_i = \text{sign}(u_i)\ \max(|u_i| - \lambda, 0)$, is the soft-thresholding function.

4. Error Bounds

In this section, we show that the ease of computing our class of estimators does not come at the cost of strong statistical guarantees. We provide a general analytical framework for deriving error bounds for our class of estimators (6) in a high-dimensional setting, where the expected feature function $\mu^* = \mathbb{E}[\phi(X)]$ has some "structure".

For a formalization of the notion of structure, we follow the unified statistical framework of Negahban et al. (2012). There, they use subspace pairs $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$, where $\mathcal{M} \subseteq \overline{\mathcal{M}}$, to capture any structured parameter. $\mathcal{M}$ is the *model subspace* that captures the constraints imposed on the model parameter, which is typically *low-dimensional*. On the other hand, $\overline{\mathcal{M}}^\perp$ is the *perturbation subspace* of parameters that represents perturbations away from the model subspace. Following their terminology, we assume that the regularization function in (6) is *decomposable* with respect to a subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$:

(C1) $\mathcal{R}(u+v) = \mathcal{R}(u) + \mathcal{R}(v), \quad \forall u \in \mathcal{M},\ \forall v \in \overline{\mathcal{M}}^\perp.$

Such *decomposability* captures the suitability of a regularization function $\mathcal{R}(\cdot)$ to particular structure. As Negahban et al. (2012) showed, for standard structural constraints such as sparsity, low-rank, etc., we can define corresponding low-dimensional model subspaces, as well as regularization functions that are decomposable with respect to the corresponding subspace pairs.

Example 1. *Given any subset $S \subseteq \{1,\ldots,p\}$ of the coordinates, let $\mathcal{M}(S)$ be the* subspace *of vectors in $\mathbb{R}^p$ that have support contained in S. It can be seen that any parameter $\theta \in \mathcal{M}(S)$ would be atmost $|S|$-sparse. For this case, we use $\overline{\mathcal{M}}(S) = \mathcal{M}(S)$, so that $\overline{\mathcal{M}}^{\perp}(S) = \mathcal{M}^{\perp}(S)$. Negahban et al. (2012) show that the ℓ_1 norm $\mathcal{R}(\theta) = \|\theta\|_1$, commonly used as a sparsity-encouraging regularization function, is decomposable with respect to subspace pairs $(\mathcal{M}(S), \overline{\mathcal{M}}^{\perp}(S))$.*

We also use their definition of *subspace compatibility constant* that captures the relative value between the regularization function $\mathcal{R}(\cdot)$ and the error norm $\|\cdot\|$, over vectors in the subspace $\mathcal{M}$: $\Psi(\mathcal{M}, \|\cdot\|) := \sup_{u\in\mathcal{M}\setminus\{0\}} \frac{\mathcal{R}(u)}{\|u\|}$. We also define the projection operator $\Pi_{\bar{\mathcal{M}}}(u) := \operatorname{argmin}_{v\in\bar{\mathcal{M}}} \|u - v\|_2$.

For technical simplicity, we consider the case where μ^* is *exactly* structured with respect to some subspace pair:

(C2) There exists a structured subspace-pair $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$ such that the model parameter satisfies $\Pi_{\mathcal{M}^{\perp}}(\mu^*) = \mathbf{0}$.

Theorem 1. *Suppose we solve the estimation problem (6), such that true structured moment satisfies Condition (C2), the regularization function satisfies Condition (C1), and the constraint term λ_n is set as $\lambda_n \geq \mathcal{R}^*(\widehat{\mu}_n - \mu^*)$. Then, the optimal solution $\widehat{\mu}$ of (6) satisfies:*

$$\mathcal{R}^*(\widehat{\mu} - \mu^*) \leq 2\lambda_n\,, \tag{9}$$

$$\|\widehat{\mu} - \mu^*\|_2 \leq 4\lambda_n \Psi(\overline{\mathcal{M}})\,, \tag{10}$$

$$\mathcal{R}(\widehat{\mu} - \mu^*) \leq 8\lambda_n \Psi(\overline{\mathcal{M}})^2\,. \tag{11}$$

We note that Theorem 1 is a non-probabilistic result, and holds deterministically for any selection of λ_n or any distributional setting of the covariates X. While Theorem 1 builds on concepts and notations such as decomposable regularization functions from Negahban et al. (2012), it is worthwhile to note that their analysis does not apply to our class of estimators: there they consider regularized convex programs, whereas here, we consider an elementary class of constrained programs. Moreover, unlike their case, the form of our estimators also allows us to provide bounds in $\mathcal{R}^*(\cdot)$ norm, which guarantee the estimates are structured, under similar conditions to those imposed on convex regularized programs in Negahban et al. (2012).

While the result in Theorem 1 seems a bit abstract, in the sequel, we provide corollaries that obtain concrete instantiations of Theorem 1 for specific settings of the feature functions, structures and regularization functions $\mathcal{R}(\cdot)$, and distributional assumptions on X.

4.1. Bounds for Covariance Estimation

In what follows, we shall assume that the components of X are sub-Gaussian, that is, there exist some constants $c_0 \geq 0$ and $T > 0$, such that for any $|t| \leq T$,

$$\mathbb{E}\Big(e^{tX_i^2}\Big) \leq c_0, \quad i = 1,\ldots,p\,.$$

This condition is satisfied for instance if X follows a multivariate normal distribution or if the entries of X are bounded. For simplicity, we also assume that $\mathbb{E}(X) = \mathbf{0}$, so that $\mathbb{E}(XX^\top) = \Sigma^*$.

4.1.1. Covariance Matrices with Element-wise Sparsity of Off-Diagonals

As a concrete example, we first consider the case where the true covariance Σ^* has sparse off-diagonals: it has at most k non-zero off-diagonal elements. A natural variant of our elementary estimator in (6) under this assumption would be the one with $\mathcal{R}(\cdot) := \|\cdot\|_{1,\text{off}}$ as in (8).

Corollary 1. *Suppose that $\lambda_n = M\sqrt{\frac{\log p}{n}}$ for a sufficiently large constant M. Then, with probability at least $1 - C_1\exp(-C_2 n\lambda_n^2)$, we have*

$$\|\widehat{\Sigma} - \Sigma^*\|_{\infty,\text{off}} \leq 2M\sqrt{\frac{\log p}{n}}\,,$$

$$\|\widehat{\Sigma} - \Sigma^*\|_F \leq 4M\sqrt{\frac{k\log p}{n}}\,,$$

$$\|\widehat{\Sigma} - \Sigma^*\|_{1,\text{off}} \leq 8Mk\sqrt{\frac{\log p}{n}}\,,$$

for some constants $C_1, C_2 > 0$.

It is instructive to compare the results of this corollary to those in Bickel & Levina (2008a); Rothman et al. (2009) where authors consider estimating covariance matrices by thresholding, however they focus on the matrices that are invariant under permutations. We note that an application of our Theorem 1 is able to provide tighter bounds by decoupling the probabilistic component from the non-probabilistic bound; for instance, compare our *consistent* ℓ_2 (or Frobenius norm) error bound in (10) against that in Theorem 2 of Bickel & Levina (2008a) where authors provide the bound of $\|\widehat{\Sigma} - \Sigma^*\|_F = O\Big(\sqrt{\frac{kp\log p}{n}}\Big)$. Moreover, our results can be applied to any covariance matrix beyond subset of matrices discussed in Bickel & Levina (2008a); Rothman et al. (2009).

4.1.2. Covariance Matrices with Element-wise Group Sparsity

Suppose the indices $\{1,\ldots,p\} \times \{1,\ldots,p\}$ of the covariance matrix entries are partitioned into L disjoints groups $\mathcal{G} = \{G_1,\ldots,G_L\}$. Let d denote the maximum group cardinality $\max_{j=1}^{L} |G_j|$. Suppose that Σ^* is *group-sparse* with respect to this set of groups, so that $|\{j \in \{1,\ldots,L\} : \Sigma^*_{G_j} \neq 0| \leq k$, where $\Sigma^*_{G_j}$ is the vector comprising entries of Σ^* corresponding to the indices in G_j. Thus, the

element-wise support of Σ^* can be expressed as the union of at most k groups in $\mathcal{G}$. A natural regularization function $\mathcal{R}(\cdot)$ for this setting is the group-structured norm

$$\|\Sigma\|_{\mathcal{G},\nu} := \sum_{l=1}^{L} \|\Sigma_{G_l}\|_\nu, \quad \text{where } \nu \geq 2, \tag{12}$$

where $\|\cdot\|_\nu$ is the element-wise ℓ_ν vector norm.

Corollary 2. *Suppose that we solve the variant of the elementary estimator in (6), with the regularization function $\mathcal{R}(\cdot)$ set to the group-structured norm in (12), and with the constraint penalty λ_n set as $\lambda_n = Md^{1/\nu^\star}\sqrt{(\log d + \log L)/n}$ for a sufficiently large constant $M > 0$. Then, with probability at least $1 - C_1\exp(-C_2 n\lambda_n^2)$, we have*

$$\|\widehat{\Sigma} - \Sigma^*\|^*_{\mathcal{G},\nu} \leq 2Md^{1/\nu^\star}\sqrt{(\log d + \log L)/n}\,,$$
$$\|\widehat{\Sigma} - \Sigma^*\|_F \leq 4Md^{1/\nu^\star}\sqrt{k(\log d + \log L)/n}\,,$$
$$\|\widehat{\Sigma} - \Sigma^*\|_{\mathcal{G},\nu} \leq 8Mkd^{1/\nu^\star}\sqrt{(\log d + \log L)/n}$$

where $\|\Sigma\|^*_{\mathcal{G},\nu} := \max_g \|\Sigma_{G_g}\|_{\nu^*}$ for a constant ν^* satisfying $\frac{1}{\nu} + \frac{1}{\nu^*} = 1$.

5. Extension to Superposition Structures

While there have been considerable advances in structurally constrained high-dimensional estimation in recent years, there has been an increasing realization, especially in the context of matrix decomposition problems, that typical structural constraints such as sparsity, group-sparsity, etc. are too stringent, and may not be very realistic. Over the last few years, there has been an emerging line of work that addresses this issue by "mixing and matching" different structures (Chandrasekaran et al., 2011; Hsu et al., 2011; McCoy & Tropp, 2011; Xu et al., 2012; Jalali et al., 2010; Agarwal et al., 2012; Yang & Ravikumar, 2013). As an illuminating example, consider the principal component analysis (PCA) problem, where we are given i.i.d. random vectors $X_i \in \mathbb{R}^p$ where $X_i = U_i + v_i$. $U_i \sim N(0, \Theta^*)$, with a low-rank covariance matrix $\Theta^* = LL^T$, for some loading matrix $L \in \mathbb{R}^{p\times r}$, corresponds to the set of low-dimensional observations without noise; and $v_i \in \mathbb{R}^p$ is a noise/error vector that is typically assumed to be spherically Gaussian distributed, $v_i \sim N(0, \sigma^2 I_{p\times p})$, or in ideal settings $v_i = 0$. The goal in PCA is to then recover the covariance matrix Θ^* from samples $\{X_i\}_{i=1}^n$, with the "clean" structural constraint that Θ^* is low-rank. However, in realistic settings, with outliers, the noise vector may be distributed as $v_i \sim N(0, \Gamma^*)$, where Γ^* is elementwise sparse. In this case, the covariance matrix of X_i has the form $\Sigma^* = \Theta^* + \Gamma^*$, where Θ^* is low-rank, and Γ^* is sparse. Thus, Σ^* is neither low-rank nor sparse, but a superposition of two matrices, one of which is low-rank, and the other which is sparse.

The emerging line of work indicated above that address such superposition-structure is again based on regularized MLEs, which would have the same non-convexity caveats for our general structured moment problem, as detailed in the previous sections for even clean structural constraints. In this section, we thus extend our "Elem-Moment" estimators in (6) to cover such "superposition-structure" as well.

To set up our notation, we assume that the true moment is given as $\mu^* = \sum_{\alpha\in I}\mu^*_\alpha$, where μ^*_α is a "clean" structured parameter with respect to a subspace pair $(\mathcal{M}_\alpha, \overline{\mathcal{M}}^\perp_\alpha)$, for $\mathcal{M}_\alpha \subset \overline{\mathcal{M}}_\alpha$. For instance, the individual components μ^*_α could individually be sparse, low-rank, column-sparse, etc., while the overall moment parameter μ^* is none of these structures per se, just a superposition of these.

Our "Elem-Moment" class of estimators in (6), naturally extends to these superposition-structured problems as follows, in what we call "Elem-Super-Moment" estimators:

$$\underset{\mu_1,\mu_2,\ldots,\mu_{|I|}}{\text{minimize}} \sum_{\alpha\in I}\lambda_\alpha\mathcal{R}_\alpha(\mu_\alpha)$$
$$\text{s. t. } \mathcal{R}^*_\alpha\Big(\widehat{\mu}_n - \sum_{\alpha\in I}\mu_\alpha\Big) \leq \lambda_\alpha \quad \text{for } \forall\alpha \in I. \tag{13}$$

This class of problems can be solved via simple closed-form operations by employing the parallel proximal algorithm of Combettes & Pesquet (2008). The details of the algorithm are presented in the Supplementary Materials.

5.1. Error bounds

As a natural extension of (C2), we assume that each component exactly μ^*_α lies in its structured subspace:

(C3) $\Pi_{\mathcal{M}^\perp_\alpha}(\mu^*_\alpha) = \mathbf{0}, \quad \forall\alpha \in I\,.$

In recent work, Yang & Ravikumar (2013) extend the analysis of regularized convex programs of Negahban et al. (2012) from the vanilla structural constraint case to the superposition structural constraint case. Their analysis however is restricted to specialized regularized convex programs, and is not applicable to our class of elementary constrained "Elem-Super-Moment" estimators in (13). In the sequel, we thus derive an extension of our Theorem 1 to this superposition-structured setting.

We borrow the following condition from Yang & Ravikumar (2013), which is a structural incoherence condition ensuring that the non-interference of different structures:

(C4) (Structural Incoherence) Let $\Omega := \max_{\gamma_1,\gamma_2}\Big\{2 + \frac{3\lambda_{\gamma_1}\Psi_{\gamma_1}(\bar{\mathcal{M}}_{\gamma_1})}{\lambda_{\gamma_2}\Psi_{\gamma_2}(\bar{\mathcal{M}}_{\gamma_2})}\Big\}$. For any $\alpha, \beta \in I$, $\max\Big\{\sigma_{\max}\big(\mathcal{P}_{\bar{\mathcal{M}}_\alpha}\mathcal{P}_{\bar{\mathcal{M}}_\beta}\big), \sigma_{\max}\big(\mathcal{P}_{\bar{\mathcal{M}}_\alpha}\mathcal{P}_{\bar{\mathcal{M}}^\perp_\beta}\big), \sigma_{\max}\big(\mathcal{P}_{\bar{\mathcal{M}}^\perp_\alpha}\mathcal{P}_{\bar{\mathcal{M}}^\perp_\beta}\big)\Big\} \leq \frac{1}{16\Omega^2}$ where $\mathcal{P}_{\bar{\mathcal{M}}}$ denote the

Table 1. Average performance measures and standard errors for sparse plus low-rank covariance estimation.

	Method	Spectral	Frobenius	Nuclear	Matrix 1-norm
n=100,p=200	Elem-Super-Moment	**7.10 (0.15)**	**8.56(0.18)**	**35.87 (0.43)**	**11.65 (0.12)**
	Thresholding	8.30 (0.17)	10.43 (0.11)	45.84 (0.39)	19.85 (0.21)
	Well-conditioned	12.22 (0.12)	13.19 (0.17)	48.11 (0.45)	23.89(0.18)
n=100,p=400	Elem-Super-Moment	**25.63 (0.54)**	**26.67 (0.49)**	**198.76 (1.31)**	**50.77 (0.72)**
	Thresholding	33.55 (0.49)	41.91(0.60)	331.41 (2.05)	67.64 (0.73)
	Well-conditioned	35.71 (0.50)	34.83 (0.46)	207.97(2.27)	93.60 (0.91)

matrix corresponding to the projection operator for the subspace $\overline{\mathcal{M}}$.

Under these two mild conditions, we can provide the following statistical guarantees for our estimators:

Theorem 2. *Suppose that the true structured moment μ^* satisfies conditions (C3) and (C4). Furthermore, suppose we solve elementary estimators in (13) setting the constraint parameters λ_α such that $\lambda_\alpha \geq \mathcal{R}^*_\alpha(\widehat{\mu}_n - \mu^*)$. Then, the optimal solution $\{\widehat{\mu}_\alpha\}_{\alpha \in I}$ of (13) satisfies the following error bounds:*

$$\mathcal{R}^*_\alpha(\widehat{\mu} - \mu^*) \leq 2\lambda_\alpha \,, \tag{14}$$

$$\mathcal{R}_\alpha(\widehat{\mu}_\alpha - \mu^*_\alpha) \leq \frac{16|I|}{\lambda_\alpha}\Big(\max_{\alpha \in I} \lambda_\alpha \Psi(\overline{\mathcal{M}}_\alpha)\Big)^2 \,, \tag{15}$$

$$\|\widehat{\mu} - \mu^*\|_F \leq 4\sqrt{2|I|}\max_{\alpha \in I} \lambda_\alpha \Psi(\overline{\mathcal{M}}_\alpha) \,. \tag{16}$$

where $\widehat{\mu} = \sum_{\alpha \in I} \widehat{\mu}_\alpha$, and $\mu^* = \sum_{\alpha \in I} \mu^*_\alpha$.

5.2. Covariance Matrices with Low-rank plus Element-wise sparse Structures

As an illustration of superposition structured moments, we consider the case where $\Sigma^* = \Sigma^*_1 + \Sigma^*_2$, and where Σ^*_1 is a low-rank matrix, and Σ^*_2 is an element-wise sparse matrix. The natural selection from the estimation class (13) would be the following:

$$\begin{aligned} \underset{\Sigma_1, \Sigma_2}{\text{minimize}}\ & \lambda_1 \|\Sigma_1\|_* + \lambda_2 \|\Sigma_2\|_{1,\text{off}} \\ \text{s. t. } & \|\widehat{\Sigma}_n - (\Sigma_1 + \Sigma_2)\|_2 \leq \lambda_1 \\ & \|\widehat{\Sigma}_n - (\Sigma_1 + \Sigma_2)\|_{\infty,\text{off}} \leq \lambda_2, \end{aligned} \tag{17}$$

where $\|\cdot\|_*$ and $\|\cdot\|_2$ represent the nuclear norm and spectral norm of a matrix, respectively. Then, the consistency of this estimator can be easily derived as the following corollary of Theorem 2:

Corollary 3. *Suppose that the true structured covariance matrix is given as $\Sigma^* = \Sigma^*_1 + \Sigma^*_2$, where k_1 and k_2 denote the rank of Σ^*_1 and the number non-zero elements of Σ^*_2, respectively. Also suppose we solve the elementary estimator variant in (17) setting $\lambda_1 = M_1\sqrt{\frac{p}{n}}$ and $\lambda_2 = M_2\sqrt{\frac{\log p}{n}}$ for sufficiently large constants $M_1, M_2 > 0$. Then, with probability at least $1 - 2\exp(-Cp)$, the solution $\widehat{\Sigma}$ satisfies the following error bounds:*

$$\|\widehat{\Sigma} - \Sigma^*\|_2 \leq 2M_1\sqrt{\frac{p}{n}} \,, \|\widehat{\Sigma} - \Sigma^*\|_{\infty,\text{off}} \leq 2M_2\sqrt{\frac{\log p}{n}} \,,$$

$$\|\widehat{\Sigma} - \Sigma^*\|_F \leq 8\max\left\{M_1\sqrt{\frac{k_1 p}{n}}\,,\, M_2\sqrt{\frac{k_2 \log p}{n}}\right\},$$

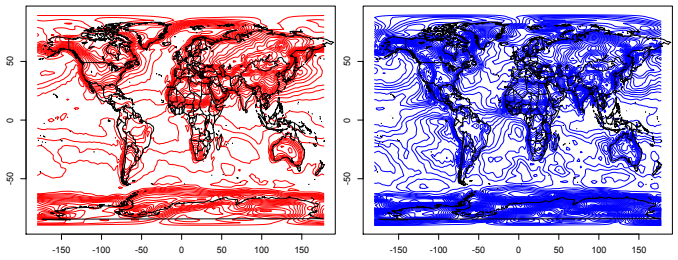
$$\|\widehat{\Sigma} - \Sigma^*\|_* \leq \tfrac{32}{M_1}\sqrt{\tfrac{n}{p}}\left[\max\left\{M_1\sqrt{\tfrac{k_1 p}{n}}\,,\, M_2\sqrt{\tfrac{k_2 \log p}{n}}\right\}\right]^2,$$

$$\|\widehat{\Sigma} - \Sigma^*\|_{1,\text{off}} \leq \tfrac{32}{M_2}\sqrt{\tfrac{n}{\log p}}\left[\max\left\{M_1\sqrt{\tfrac{k_1 p}{n}}\,,\, M_2\sqrt{\tfrac{k_2 \log p}{n}}\right\}\right]^2.$$

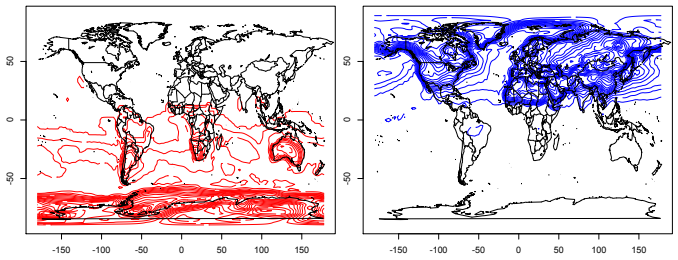
6. Experiments

Simulation We first confirm the usefulness of our framework in the presence of superposition structures. Specifically, we focus on covariance estimation where the true covariance has a sparse plus low-rank structure. We consider $\Sigma^* = \Sigma^*_1 + \Sigma^*_2$, where $\Sigma^*_1 = 0.5(1_p 1_p^T)$. and $\Sigma^*_2 = I_{p/5} \otimes (0.2(1_5 1_5^T) + 0.2I_5)$, where $\otimes$ denotes the Kronecker product. We perform 100 simulation runs. For each simulation run, we generate $n = 100$ observations from $N(0, \Sigma^*)$. We compare our "Elem-Super-Moment" estimator with the thresholding method of Bickel & Levina (2008a) and the well-conditioned estimator of Ledoit & Wolf (2003). For each method, the tuning parameters are set using 5-fold cross validation with Frobenius norm as described in Bickel & Levina (2008a). We consider $p = 200, 400$. As performance measures, we used the spectral, Frobenius, nuclear and matrix 1-norm of the difference between estimated and true covariance. The results presented in Table 1 show that "Elem-Super-Moment" clearly outperforms the other methods. In addition, our method is able to recover the sparsity pattern of the sparse component with True Positive and True Negative rates greater than 99.50% and 95.24% respectively.

Climate dataset We demonstrate the applicability of our class of estimators on a climatology dataset. We used 4-times daily surface temperature data from NCEP/NCAR Reanalysis 1. The data is for the year 2011 and uses a 2.5 degree latitude x 2.5 degree longitude global grid covering 90N - 90S, 0E - 357.5E, so that we have 144×73 locations. We considered each location as a feature, so that $p = 144 \times 73 = 10512$, and used observations across time



(a) PCA on the sample covariance matrix



(b) PCA on our Elem-Moment covariance estimate with ℓ_1 norm.

Figure 1. Contour plots of the first two principal components using PCA on (top) the sample covariance matrix, and (bottom) our Elem-Moment covariance estimate with ℓ_1 norm.

as samples, so that $n = 4 \times 365 = 1460$, and computed the $p \times p$ spatial sample covariance matrix. To evaluate the covariance matrix estimates, we used *empirical orthogonal functions* (EOFs), which actually correspond to the principal components of the covariance matrix, which are commonly employed in spatio-temporal statistics to investigate spatial patterns in data. By visualizing the spacial contour plots of a given EOF, one can get an idea of which geographical regions contribute greatly to that principle component. We depict these contour plots for the first two principal components using (a) PCA on the sample covariance matrix and (b) PCA on our Elem-Moment estimate with the regularization set to the ℓ_1 norm. As can be seen from the figures, our method clearly separates the Northern and Southern hemispheres, which is as expected by climate scientists. In contrast, PCA on the sample covariance itself is unable to make that distinction.

Hi-C dataset We also illustrate the usefulness of our class of estimators on data from Hi-C, a very recent methodology to study the 3-D architecture of genomes. Briefly, the data consists of the observed frequencies of chromatin interaction between any two genomic loci (out of a total of as many as $40,000$ loci). A comprehensive review of Hi-C is provided in Dekker et al. (2013). Note that these empirical interaction frequencies can be collated as a sample covariance matrix. The goal then is to estimate the true contact map comprising the expected interaction frequencies, which would correspond to the population covariance matrix. We consider two Hi-C datasets taken from Lieberman-Aiden et al. (2009) which are biological replicates assembled using different restriction enzymes (HindIII and NcoI). One approach to validate the contact map (or population covariance) estimators would be to estimate the contact map based on each dataset separately, and measure the "reproducibility" between both estimates, measured using Spearman correlation. We applied our "Elem-Super-Moment" estimator with two regularization functions set to ℓ_1 norm and the nuclear norm, to the normalized data provided by Hu et al. (2008). The raw data exhibits an average correlation across the 23 chromosomes of 0.7241, that of the normalized data 0.8041 (see Hu et al. (2008)). Our estimator improves the correlation further to 0.8355.

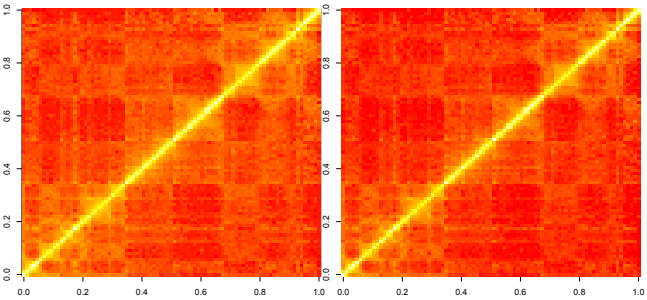


Figure 2. Our Elem-Super-Moment covariance estimates (with ℓ_1 and nuclear norms) for Hi-C data analysis on chromosome 14; (Left) HindIII dataset, and (Right) Ncol dataset

Visualizing the fine details of the contact map is challenging due to the high resolution of the map (2761×2761 matrix in this experiment). In Figure 2, we thus depict a portion of our estimated contact map for chromosomes 14 only, so the reader can get an idea of the data structure. As can be seen from Figure 2, the contact maps look quite similar for both biological replicates, which corroborates our high correlation value of 0.8355 noted earlier. (For comparison with the original data, a picture of the raw data is provided in Figure 1 B&D of Lieberman-Aiden et al. (2009).) The bright diagonals correspond to nearby interactions within the same chromosome. We can also distinguish some interaction blocks around the diagonals as well as more distal interactions. Given this encouraging preliminary analysis, we plan to perform an in-depth biological analysis using our Elem-Moment estimators in future work.

Acknowledgments

E.Y and P.R. acknowledge the support of ARO via W911NF-12-1-0390 and NSF via IIS-1149803, IIS-1320894, DMS-1264033.

References

Agarwal, A., Negahban, S., and Wainwright, M. J. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Annals of Statistics*, 40(2):1171–1197, 2012.

Bickel, P. J. and Levina, E. Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604, 2008a.

Bickel, P. J. and Levina, E. Regularized estimation of large covariance matrices. *Annals of Statistics*, 36(1):199–227, 2008b.

Cai, T. T., Zhang, C.-H., and Zhou, H. H. Optimal rates of convergence for covariance matrix estimation. *Annals of Statistics*, 2010. To appear.

Candès, E. and Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. The convex geometry of linear inverse problems. In *48th Annual Allerton Conference on Communication, Control and Computing*, 2010.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2), 2011.

Combettes, P. L. and Pesquet, J. C. A proximal decomposition method for solving convex variational inverse problems. *Inverse Problems*, 24(27), 2008.

Dai, B., Ding, S., and Wahba, G. Multivariate bernoulli distribution. *Bernoulli*, 19(4):1465–1483, 2013.

Dekker, J., Marti-Renom, M. A., and Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature reviews. Genetics*, 14(6):390–403, 2013.

El Karoui, N. Operator norm consistent estimation of large dimensional sparse covariance matrices. *Annals of Statistics*, 36 (6):2717–2756, 2008.

Furrer, R. and Bengtsson, T. Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255, 2007.

Hsu, D., Kakade, S. M., and Zhang, T. Robust matrix decomposition with sparse corruptions. *IEEE Trans. Inform. Theory*, 57: 7221–7234, 2011.

Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B., and Liu, J. S. Hicnorm: removing biases in hi-c data via poisson regression. *Bioinformatics*, 28(23):3131–3133, 2008.

Jalali, A., Ravikumar, P., Sanghavi, S., and Ruan, C. A dirty model for multi-task learning. In *Neur. Info. Proc. Sys. (NIPS)*, 23, 2010.

Johnstone, I. M. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29(2):295–327, 2001.

Johnstone, I. M. and Lu, A. Y. Sparse principal components analysis. *Unpublished Manuscript*, 2004.

Ledoit, O. and Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365411, 2003.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., and Ragoczy, T. [and others]. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, October 2009.

McCoy, M. and Tropp, J. A. Two proposals for robust pca using semidefinite programming. *Electron. J. Statist.*, 5:1123–1160, 2011.

Negahban, S., Ravikumar, P., Wainwright, M. J., and Yu, B. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.

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

Ribes, A., Azais, J. M., and Planton, S. Adaptation of the optimal fingerprint method for climate change detection using a well-conditioned covariance matrix estimate. *Journal of Climate Dynamics*, 33:707–722, 2009.

Rothman, A. J., Levina, E., and Zhu, J. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association (Theory and Methods)*, 104:177–186, 2009.

Wainwright, M. J. and Jordan, M. I. Graphical models, exponential families and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1—305, December 2008.

Watson, G. A. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33–45, 1992.

Wikle, C. K. and Cressie, N. A dimension reduced approach to space-time kalman filtering. *Biometrika*, 86:815–829, 1999.

Xu, H., Caramanis, C., and Sanghavi, S. Robust pca via outlier pursuit. *IEEE Transactions on Information Theory*, 58(5): 3047–3064, 2012.

Yang, E. and Ravikumar, P. Dirty statistical models. In *Neur. Info. Proc. Sys. (NIPS)*, 26, 2013.

Appendix

A. Proof of Proposition 3

Consider an arbitrary parameter $\tilde{\mu} = \sum_i \tilde{c}_i \tilde{\mathbf{a}}_i$ in the feasible set of parameters that satisfy the constraint of (6).

Let $\mathcal{P}_{\mathcal{M}}$ denote the matrix corresponding to the projection operator for the subspace $\mathcal{M}$.

Since $\tilde{\mu}$ should satisfy the constraint of $\mathcal{R}^*(\widehat{\mu}_n - \tilde{\mu}) \leq \lambda_n$, for any fixed index i in the atoms of $\widehat{\mu}_n$,

$$\begin{aligned}&\mathcal{R}\Big(\mathcal{P}_{\mathbf{a}_i^*}(\widehat{\mu}_n - \tilde{\mu})\Big)\\ &= \mathcal{R}\Big(\sum_j \tilde{c}_j \mathcal{P}_{\mathbf{a}_i^*}(\tilde{\mathbf{a}}_j) - c_i^* \mathbf{a}_i^*\Big) \leq \lambda_n,\end{aligned}$$

which implies $\max\{c_i^* - \lambda_n, 0\} \leq \mathcal{R}\Big(\sum_j \tilde{c}_j \mathcal{P}_{\mathbf{a}_i^*}(\tilde{\mathbf{a}}_j)\Big)$. By summing over all i, we obtain

$$\begin{aligned}\mathcal{R}(\widehat{\mu}) &= \sum_i \max\{c_i^* - \lambda_n, 0\} \leq \sum_i \mathcal{R}\Big(\sum_j \tilde{c}_j \mathcal{P}_{\mathbf{a}_i^*}(\tilde{\mathbf{a}}_j)\Big)\\ &\overset{(i)}{=} \mathcal{R}\Big(\sum_j \tilde{c}_j \sum_i \mathcal{P}_{\mathbf{a}_i^*}(\tilde{\mathbf{a}}_j)\Big) = \mathcal{R}\Big(\sum_j \tilde{c}_j \tilde{\mathbf{a}}_j\Big) = \mathcal{R}(\tilde{\mu}).\end{aligned}$$

Finally, since we can simply verify that $\widehat{\mu}$ is also feasible, we can conclude that $\widehat{\mu}$ is the optimal of (6).

B. Proof of Theorem 1

Let $\Delta := \widehat{\mu} - \mu^*$ be the error vector that we are interested in.

$$\begin{aligned}&\mathcal{R}^*(\widehat{\mu} - \mu^*) = \mathcal{R}^*(\widehat{\mu} - \widehat{\mu}_n + \widehat{\mu}_n - \mu^*)\\ &\leq \mathcal{R}^*(\widehat{\mu}_n - \widehat{\mu}) + \mathcal{R}^*(\widehat{\mu}_n - \mu^*) \leq 2\lambda_n\end{aligned}$$

By the fact that $\mu^*_{\mathcal{M}^\perp} = \mathbf{0}$, and the decomposability of $\mathcal{R}$ with respect to $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$,

$$\begin{aligned}&\mathcal{R}(\mu^*)\\ &= \mathcal{R}(\mu^*) + \mathcal{R}\big[\Pi_{\bar{\mathcal{M}}^\perp}(\Delta)\big] - \mathcal{R}\big[\Pi_{\bar{\mathcal{M}}^\perp}(\Delta)\big]\\ &= \mathcal{R}\big[\mu^* + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta)\big] - \mathcal{R}\big[\Pi_{\bar{\mathcal{M}}^\perp}(\Delta)\big]\\ &\overset{(i)}{\leq} \mathcal{R}\big[\mu^* + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta) + \Pi_{\bar{\mathcal{M}}}(\Delta)\big] + \mathcal{R}\big[\Pi_{\bar{\mathcal{M}}}(\Delta)\big]\\ &\quad - \mathcal{R}\big[\Pi_{\bar{\mathcal{M}}^\perp}(\Delta)\big]\\ &= \mathcal{R}\big[\mu^* + \Delta\big] + \mathcal{R}\big[\Pi_{\bar{\mathcal{M}}}(\Delta)\big] - \mathcal{R}\big[\Pi_{\bar{\mathcal{M}}^\perp}(\Delta)\big] \qquad (18)\end{aligned}$$

where the equality (i) holds by the triangle inequality of norm. Since (6) minimizes $\mathcal{R}(\widehat{\mu})$, we have $\mathcal{R}(\mu^* + \Delta) = \mathcal{R}(\widehat{\mu}) \leq \mathcal{R}(\mu^*)$. Combining this inequality with (18),

$$\mathcal{R}\big[\Pi_{\bar{\mathcal{M}}^\perp}(\Delta)\big] \leq \mathcal{R}\big[\Pi_{\bar{\mathcal{M}}}(\Delta)\big]. \qquad (19)$$

Moreover, by Hölder's inequality and the decomposability of $\mathcal{R}(\cdot)$,

$$\begin{aligned}&\|\Delta\|_2^2 = \langle\!\langle \Delta, \Delta \rangle\!\rangle \leq \mathcal{R}^*(\Delta)\,\mathcal{R}(\Delta) \leq 2\lambda_n \mathcal{R}(\Delta)\\ &= 2\lambda_n\big[\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta))\big] \leq 4\lambda_n \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta))\\ &\leq 4\lambda_n \Psi(\overline{\mathcal{M}}) \|\Pi_{\bar{\mathcal{M}}}(\Delta)\|_2 \qquad (20)\end{aligned}$$

where $\Psi(\overline{\mathcal{M}})$ is a simple notation for $\Psi(\overline{\mathcal{M}}, \|\cdot\|_2)$.

Since the projection operator is defined in terms of $\|\cdot\|_2$ norm, it is non-expansive: $\|\Pi_{\bar{\mathcal{M}}}(\Delta)\|_2 \leq \|\Delta\|_2$. Therefore, by (20), we have

$$\|\Pi_{\bar{\mathcal{M}}}(\Delta)\|_2 \leq 4\lambda_n \Psi(\overline{\mathcal{M}}), \qquad (21)$$

and plugging it back into (20) yields the error bound (10).

Finally, (11) is straightforward from (19) and (21)

$$\begin{aligned}&\mathcal{R}(\Delta) \leq 2\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta))\\ &\leq 2\Psi(\overline{\mathcal{M}}) \|\Pi_{\bar{\mathcal{M}}}(\Delta)\|_2 \leq 8\lambda_n \Psi(\overline{\mathcal{M}})^2.\end{aligned}$$

C. Proof of corollaries: Covariance Estimation in Section 4.1

In order to leverage Theorem 1, two ingredients need to be specified: (i) the convergence rate of $\mathcal{R}^*(\widehat{\mu}_n - \mu^*)$ for λ_n to satisfy $\lambda_n \geq \mathcal{R}^*(\widehat{\mu}_n - \mu^*)$, and (ii) the compatibility constant $\Psi(\overline{\mathcal{M}})$. In each corollary, we are going to show how these two components can be computed.

C.1. Proof of Corollary 1

For this case, we can directly appeal to the well known bound (e.g., the Lemma 1 of (Ravikumar et al., 2011)): Consider the following event:

$$\begin{aligned}&P(\|\widehat{\Sigma}_n - \Sigma\|_{\infty,\text{off}} > \delta)\\ &\leq 4\exp\big(-\frac{n\delta^2}{3200(\max_i \Sigma_{ii})^2} + \log p^2\big).\end{aligned}$$

By setting $\delta = 40(\max_i \Sigma_{ii})\sqrt{\frac{2\tau \log p}{n}}$, we see that the choice of λ_n is valid with probability at least $1 - C_1\exp(-C_2 n\lambda_n^2)$.

For the second ingredient, let $\mathcal{M} = \overline{\mathcal{M}}$ correspond to the support of Σ^*. We have $\psi(\overline{\mathcal{M}}, \|\cdot\|_2) = \sqrt{s}$, where s is the cardinality of the support of Σ^*.

C.2. Proof of Corollary 2

$\mathcal{R}^*(\widehat{\Sigma}_n - \Sigma^*) = \max_{l=1,\ldots,L} \Big\|[\widehat{\Sigma}_n]_{G_l} - [\Sigma^*]_{G_l}\Big\|_{\nu^\star}$. For a given entry (i,j) we have

$$P(|[\widehat{\Sigma}_n]_{ij} - \Sigma^*_{ij}| > t) \leq C_1\exp(-C_2 n t^2).$$

For a given group G_l, by union bound over the group elements, we have

$$P(|[\widehat{\Sigma}_n]_{ij} - \Sigma^*_{ij}| > t \text{ for all } (i,j) \in G_l) \leq C_1 \exp(-C_2 n t^2 + \log d).$$

This implies that

$$P(\|[\widehat{\Sigma}_n]_{G_l} - \Sigma^*_{G_l}\|_{\nu^\star} > t d^{1/\nu^\star}) \leq C_1 \exp(-C_2 n t^2 + \log d).$$

By a union bound over all groups we obtain

$$P\left(\max_{l=1,\ldots,L} \|[\widehat{\Sigma}_n]_{G_l} - \Sigma^*_{G_l}\|_2 > t d^{1/\nu^\star}\right) \leq C_1 \exp(-C_2 n t^2 + \log d + \log L).$$

This yields

$$P\left(\max_{l=1,\ldots,L} \|[\widehat{\Sigma}_n]_{G_l} - \Sigma^*_{G_l}\|_2 > \delta\right) \leq C_1 \exp(-C_2 n d^{-2/\nu^\star} \delta^2 + \log d + \log L).$$

We conclude by setting $\delta = d^{1/\nu^\star}\sqrt{(\log d + \log L)/n}$. Let $\mathcal{M} = \overline{\mathcal{M}}$ correspond to the support of Σ^*, which can be written as a union of groups in $\mathcal{G}$. Since $\nu \geq 2$, we have $\psi(\overline{\mathcal{M}}, \|\cdot\|_F) = \sqrt{k}$.

D. Proof of Theorem 2

In this proof, we consider the matrix parameter such as the covariance. Basically, the Frobenius norm can be simply replaced by ℓ_2 norm for the vector parameters. Let $\Delta_\alpha := \widehat{\mu}_\alpha - \mu^*_\alpha$, and $\Delta := \widehat{\mu} - \mu^* = \sum_{\alpha\in I} \Delta_\alpha$. The error bound (14) can be easily shown from the assumption in the statement with the constraint of (13). For every $\alpha \in I$,

$$\mathcal{R}^*_\alpha(\Delta) = \mathcal{R}^*_\alpha(\widehat{\mu} - \mu^*) = \mathcal{R}^*_\alpha(\widehat{\mu} - \widehat{\mu}_n + \widehat{\mu}_n - \mu^*) \leq \mathcal{R}^*_\alpha(\widehat{\mu}_n - \widehat{\mu}) + \mathcal{R}^*_\alpha(\widehat{\mu}_n - \mu^*) \leq 2\lambda_\alpha. \quad (22)$$

By the similar reasoning as in (18) with the fact that $\Pi_{\mathcal{M}^\perp_\alpha}(\mu^*_\alpha) = \mathbf{0}$ in (C3), and the decomposability of $\mathcal{R}_\alpha$ with respect to $(\mathcal{M}_\alpha, \overline{\mathcal{M}}^\perp_\alpha)$, we have

$$\mathcal{R}_\alpha(\mu^*_\alpha) \leq \mathcal{R}_\alpha\big[\mu^*_\alpha + \Delta_\alpha\big] + \mathcal{R}_\alpha\big[\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\big] - \mathcal{R}_\alpha\big[\Pi_{\bar{\mathcal{M}}^\perp_\alpha}(\Delta_\alpha)\big]. \quad (23)$$

Since $\{\widehat{\mu}_\alpha\}_{\alpha\in I}$ minimizes the objective function of (13),

$$\sum_{\alpha\in I} \lambda_\alpha \mathcal{R}_\alpha(\widehat{\mu}_\alpha) \leq \sum_{\alpha\in I} \lambda_\alpha \mathcal{R}_\alpha(\mu^*_\alpha).$$

Combining this inequality with (23), we have

$$\sum_{\alpha\in I} \lambda_\alpha \mathcal{R}_\alpha(\widehat{\mu}_\alpha) \leq \sum_{\alpha\in I} \lambda_\alpha \Big\{ \mathcal{R}_\alpha(\mu^*_\alpha + \Delta_\alpha) + \mathcal{R}_\alpha\big[\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\big] - \mathcal{R}_\alpha\big[\Pi_{\bar{\mathcal{M}}^\perp_\alpha}(\Delta_\alpha)\big] \Big\},$$

which implies

$$\sum_{\alpha\in I} \lambda_\alpha \mathcal{R}_\alpha\big[\Pi_{\bar{\mathcal{M}}^\perp_\alpha}(\Delta_\alpha)\big] \leq \sum_{\alpha\in I} \lambda_\alpha \mathcal{R}_\alpha\big[\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\big], \quad (24)$$

Now, for each structure $\alpha \in I$, we have an application of Hölder's inequality; $|\langle\!\langle \Delta, \Delta_\alpha \rangle\!\rangle| \leq \mathcal{R}^*_\alpha(\Delta)\,\mathcal{R}_\alpha(\Delta_\alpha) \leq 2\lambda_\alpha \mathcal{R}_\alpha(\Delta_\alpha)$ where the notation $\langle\!\langle A, B \rangle\!\rangle$ denotes the trace inner product, $\mathrm{trace}(A^\top B) = \sum_i \sum_j A_{ij} B_{ij}$, and we use the pre-computed bound in (22). Then, the Frobenius error $\|\Delta\|_F$ can be upper-bounded as follows:

$$\|\Delta\|^2_F = \langle\!\langle \Delta, \Delta \rangle\!\rangle = \sum_{\alpha\in I} \langle\!\langle \Delta, \Delta_\alpha \rangle\!\rangle \leq \sum_{\alpha\in I} |\langle\!\langle \Delta, \Delta_\alpha \rangle\!\rangle| \leq 2\sum_{\alpha\in I} \lambda_\alpha \mathcal{R}_\alpha(\Delta_\alpha) \leq 2\sum_{\alpha\in I} \Big\{ \lambda_\alpha \mathcal{R}_\alpha\big[\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\big] + \lambda_\alpha \mathcal{R}_\alpha\big[\Pi_{\bar{\mathcal{M}}^\perp_\alpha}(\Delta_\alpha)\big] \Big\} \leq 4\sum_{\alpha\in I} \lambda_\alpha \mathcal{R}_\alpha\big[\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\big] \leq 4\sum_{\alpha\in I} \lambda_\alpha \Psi(\overline{\mathcal{M}}_\alpha)\, \|\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\|_F \quad (25)$$

where $\Psi(\overline{\mathcal{M}}_\alpha)$ denotes the compatibility constant of space $\overline{\mathcal{M}}_\alpha$ with respect to the Frobenius norm: $\Psi(\overline{\mathcal{M}}_\alpha, \|\cdot\|_F)$.

Here, we define a key notation in the error bound:

$$\Phi := \max_{\alpha\in I} \lambda_\alpha \Psi(\overline{\mathcal{M}}_\alpha).$$

Armed with this notation, (25) can be rewritten as

$$\|\Delta\|^2_F \leq 4\Phi \sum_{\alpha\in I} \|\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\|_F \quad (26)$$

At this point, we directly appeal to the result in Proposition 2 of (Yang & Ravikumar, 2013) with a small modification:

Proposition 4. *Suppose that the structural incoherence condition (C4) as well as the condition (C3) hold. Then, we have*

$$2|\sum_{\alpha<\beta} \langle\!\langle \Delta_\alpha, \Delta_\beta \rangle\!\rangle| \leq \frac{1}{2}\sum_{\alpha\in I} \|\Delta_\alpha\|^2_F.$$

By this proposition, we have

$$\sum_{\alpha\in I} \|\Delta_\alpha\|^2_F \leq \|\Delta\|^2_F + 2|\sum_{\alpha<\beta} \langle\!\langle \Delta_\alpha, \Delta_\beta \rangle\!\rangle| \leq \|\Delta\|^2_F + \frac{1}{2}\sum_{\alpha\in I} \|\Delta_\alpha\|^2_F,$$

which implies $\sum_{\alpha\in I} \|\Delta_\alpha\|^2_F \leq 2\|\Delta\|^2_F$.

Moreover, since the projection operator is defined in terms of the Frobenius norm, it is non-expansive for all α:

$\|\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\|_F \le \|\Delta_\alpha\|_F$. Hence, we finally obtain

$$\Big(\sum_{\alpha\in I}\|\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\|_F\Big)^2 \le \Big(\sum_{\alpha\in I}\|\Delta_\alpha\|_F\Big)^2$$
$$\le |I|\sum_{\alpha\in I}\|\Delta_\alpha\|_F^2 \le 8|I|\Phi\sum_{\alpha\in I}\|\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\|_F,$$

and therefore,

$$\sum_{\alpha\in I}\|\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\|_F \le 8|I|\Phi \tag{27}$$

The Frobenius norm error bound (16) can be derived by plugging (27) back into (26):

$$\|\Delta\|_F \le 32|I|\Phi^2.$$

The proof of the final error bound (15) is straightforward from (24) and (27) as follows: for each fixed $\alpha \in I$,

$$\begin{aligned}
&\mathcal{R}_\alpha(\Delta_\alpha)\\
&\le \frac{1}{\lambda_\alpha}\Big\{\lambda_\alpha\mathcal{R}_\alpha\big[\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\big] + \lambda_\alpha\mathcal{R}_\alpha\big[\Pi_{\bar{\mathcal{M}}_\alpha^\perp}(\Delta_\alpha)\big]\Big\}\\
&\le \frac{1}{\lambda_\alpha}\Big\{\lambda_\alpha\mathcal{R}_\alpha\big[\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\big] + \sum_{\beta\in I}\lambda_\beta\mathcal{R}_\beta\big[\Pi_{\bar{\mathcal{M}}_\beta}(\Delta_\beta)\big]\Big\}\\
&\le \frac{2}{\lambda_\alpha}\sum_{\beta\in I}\lambda_\beta\mathcal{R}_\beta\big[\Pi_{\bar{\mathcal{M}}_\beta}(\Delta_\beta)\big]\\
&\le \frac{2}{\lambda_\alpha}\sum_{\beta\in I}\lambda_\beta\Psi(\overline{\mathcal{M}}_\beta)\|\Pi_{\bar{\mathcal{M}}_\beta}(\Delta_\beta)\|_F\\
&\le \frac{2\Phi}{\lambda_\alpha}\sum_{\beta\in I}\|\Pi_{\bar{\mathcal{M}}_\beta}(\Delta_\beta)\|_F \le \frac{16|I|\Phi^2}{\lambda_\alpha},
\end{aligned}$$

which completes the proof.

D.1. Proof of Corollary 3

The proof for an element-wise sparse component is already proven in Section C.1. At the same time, for a low-rank component, we can directly appeal to the results for clean models (Agarwal et al., 2012):

$$\lambda_1 = 4\sqrt{\|\Sigma^*\|_2}\sqrt{\frac{p}{n}} \ge \|\Sigma^* - \widehat{\Sigma}_n\|_2$$

with probability at least $1 - 2\exp(-C_1 p)$. The subspace compatibility of any matrix A with rank k can be easily derived as

$$\sup_{A\neq 0}\frac{\|A\|_*}{\|A\|_F} \le \sqrt{k_1}.$$

E. A Parallel Proximal algorithm for "Elem-Super-Moment" Estimation

The class of "Elem-Super-Moment" estimators solves

$$\begin{aligned}
\underset{\mu_1,\mu_2,\ldots,\mu_{|I|}}{\text{minimize}} \ &\sum_{\alpha\in I}\lambda_\alpha\mathcal{R}_\alpha(\mu_\alpha)\\
\text{s.t. } &\mathcal{R}^*_\alpha(\widehat{\mu}_n - \mu) \le \lambda_\alpha \quad \text{for } \forall\alpha\in I\\
&\mu = \sum_{\alpha\in I}\mu_\alpha.
\end{aligned} \tag{28}$$

Let $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_{|I|})$. Consider the operators $L_\alpha(\boldsymbol{\mu}) = \mu_\alpha$, for $\alpha \in I$ and $L_{\text{tot}}(\boldsymbol{\mu}) = \sum_{\alpha\in I}\mu_\alpha$. Then the problem can be rewritten as

$$\begin{aligned}
\underset{\boldsymbol{\mu}}{\text{minimize}} \ &\sum_{\alpha\in I}\lambda_\alpha\mathcal{R}_\alpha(L_\alpha(\boldsymbol{\mu}))\\
\text{s.t. } &\mathcal{R}^*_\alpha(\widehat{\mu}_n - L_{\text{tot}}(\boldsymbol{\mu})) \le \lambda_\alpha \quad \text{for } \forall\alpha\in I.
\end{aligned} \tag{29}$$

For all $\alpha \in I$ let

$$f_\alpha(\cdot) = \lambda_\alpha\mathcal{R}_\alpha(L_\alpha(\cdot)).$$

Define the indicator function of a set C as

$$\imath_C : x \mapsto \begin{cases} 0, & \text{if } x \in C\\ +\infty, & \text{if } x \notin C.\end{cases}$$

and let

$$g_\alpha(\cdot) = \imath_{(\mathcal{R}^*_\alpha(\widehat{\mu}_n - L_{\text{tot}}(\cdot))\le\lambda_\alpha)}.$$

Then observe that (29) can be rewritten as

$$\begin{aligned}
\underset{\bar{\boldsymbol{\mu}}_1,\ldots\bar{\boldsymbol{\mu}}_{|I|},\tilde{\boldsymbol{\mu}}_1,\ldots\tilde{\boldsymbol{\mu}}_{|I|}}{\text{minimize}} \ &\sum_{\alpha\in I}f_\alpha(\bar{\boldsymbol{\mu}}_\alpha) + \sum_{\alpha\in I}g_\alpha(\tilde{\boldsymbol{\mu}}_\alpha)\\
\text{s.t. } &\bar{\boldsymbol{\mu}}_1 = \ldots = \bar{\boldsymbol{\mu}}_{|I|} = \tilde{\boldsymbol{\mu}}_1 = \ldots = \tilde{\boldsymbol{\mu}}_{|I|}.
\end{aligned} \tag{30}$$

We can then apply the parallel proximal method (Algorithm 3.1 of Combettes & Pesquet (2008)), which is derived from the classical Douglas-Rachford algorithm (Combettes & Pesquet, 2008), and obtain Algorithm 1. In this splitting algorithm, each function f_α is used separately via its own proximal operator. The same holds for each function g_α. Note that

$$\text{prox}_{2|I|\gamma f_\alpha} = \text{prox}_{2|I|\gamma\lambda_\alpha\mathcal{R}_\alpha\circ L_\alpha}$$

and

$$\text{prox}_{2|I|\gamma g_\alpha} = \text{prox}_{2|I|\gamma\imath_{(\mathcal{R}^*_\alpha(\widehat{\mu}_n - L_{\text{tot}}(\cdot))\le\lambda_\alpha)}}$$

For various popular choices of regularization R_α these proximal operators have simple closed-form formulas.

This can be seen by applying Lemma 2.4 of Combettes & Pesquet (2008) which states that if L is a bounded linear operator such that $L \circ L^* = \kappa \mathrm{Id}$ for some finite $\kappa > 0$ then

$$\mathrm{prox}_{h \circ L} = \mathrm{Id} + \frac{1}{\kappa} L^* \circ (\mathrm{prox}_{\kappa h} - \mathrm{Id}) \circ L.$$

and by noting that L_α and L_{tot} are such bounded linear operators.

Algorithm 1 Parallel proximal algorithm

Initialization: $\gamma > 0$, $(\bar{\boldsymbol{\mu}}_\alpha^0)_{\alpha \in I}$ and $(\tilde{\boldsymbol{\mu}}_\alpha^0)_{\alpha \in I}$
Set $\boldsymbol{\mu}^0 = \frac{1}{2|I|} \sum_{\alpha \in I} (\bar{\boldsymbol{\mu}}_\alpha^0 + \tilde{\boldsymbol{\mu}}_\alpha^0)$.
for $i = 0, 1, \ldots$ **do**
  for $\alpha \in I$ **do**
    $\bar{\boldsymbol{p}}_\alpha^i = \mathrm{prox}_{2|I|\gamma f_\alpha} \bar{\boldsymbol{\mu}}_\alpha^i$ and $\tilde{\boldsymbol{p}}_\alpha^i = \mathrm{prox}_{2|I|\gamma g_\alpha} \tilde{\boldsymbol{\mu}}_\alpha^i$.
  end for
  $\boldsymbol{p}^i = \frac{1}{2|I|} \sum_{\alpha \in I} (\bar{\boldsymbol{p}}_\alpha^i + \tilde{\boldsymbol{p}}_\alpha^i)$.
  $0 < \rho_i < 2$
  for $\alpha \in I$ **do**
    $\bar{\boldsymbol{\mu}}_\alpha^{i+1} = \bar{\boldsymbol{\mu}}_\alpha^i + \rho_i (2\boldsymbol{p}^i - \boldsymbol{\mu}^i - \bar{\boldsymbol{p}}_\alpha^i)$.
    $\tilde{\boldsymbol{\mu}}_\alpha^{i+1} = \tilde{\boldsymbol{\mu}}_\alpha^i + \rho_i (2\boldsymbol{p}^i - \boldsymbol{\mu}^i - \tilde{\boldsymbol{p}}_\alpha^i)$.
  end for
  $\boldsymbol{\mu}^{i+1} = \boldsymbol{\mu}^i + \rho_i (\boldsymbol{p}^i - \boldsymbol{\mu}^i)$.
end for