Biometrika (2017), **xx**, x, *pp*. 1–14
Printed in Great Britain

A proximal distance algorithm for sparse covariance estimation

BY JASON XU[†], KENNETH LANGE[†,*]

Departments of Biomathematics[†], Statistics[*], and Human Genetics[*], University of California Los Angeles, 621 Charles E. Young Drive, Los Angeles, California, 90095-1766, U.S.A.

jqxu@ucla.edu klange@ucla.edu

SUMMARY

We consider the problem of estimating a covariance matrix under a patternless sparsity assumption by minimizing a penalized multivariate Gaussian likelihood. In contrast to the majority of approaches which rely on shrinkage penalties, our formulation regularizes the distance from the estimate to a symmetric sparsity set. To solve the resulting non-convex optimization problem, we propose a proximal distance algorithm and establish its global convergence properties. This algorithm is efficient in practice, amenable to the high-dimensional setting where the dimension p exceeds the number of observations n, and produces a positive definite solution. We demonstrate empirically that these merits lead to favorable performance in terms of both accuracy and runtime compared to existing methods over a suite of simulation studies. In addition, we consider two case studies, beginning with estimating forecast correlations from international migration data. Our analysis of a flow cytometry dataset suggest that the marginal and conditional dependency networks are more similar than previously concluded.

Some key words: Covariance; Sparse estimation; Majorization-minimization; Proximal algorithm; Penalized likelihood, Distance penalty; Sequential unconstrained minimization; Cell signaling.

We consider estimating a covariance matrix from a sample of vectors drawn from the normal distribution. The task is central to the analysis of multivariate data, but poses several statistical challenges. It is generally recognized that two of the chief problems among these concern high-dimensionality and the positive-definiteness constraint (Pourahmadi, 2011). The number of parameters to be estimated grows quadratically in the dimension p, quickly exceeding the number of available samples n. It was noted as early as 1956 (Stein, 1956) that the sample covariance degrades into a poor estimator when p/n is large, and is well known to be singular in the high dimensional setting where $p > n$. Since then, covariance estimation has remained an active area of research, and many approaches have sought to mitigate the curse of dimensionality by imposing parsimony through assumptions on the size or structure of the effective parameters. In this paper, we focus on the setting where the covariance matrix is only presumed to be *sparse*. Such an assumption is essential in reducing the number of free parameters, and has proven to be reasonable in many applications. Further, the sparsity structure has an important interpretation: zero entries in the covariance matrix encode marginal independencies between variables.

Under the sparsity assumption, many regularized estimators have been proposed; Pourahmadi (2011) and Fan et al. (2016) provide an excellent overview. Some assume a known ordering of variables: tapering, banding, and solutions based on the Cholesky decomposition generally are sensitive to permutations of the variables (Wu & Pourahmadi, 2003; Bickel & Levina, 2008b;

© 2017 Biometrika Trust

Levina et al., 2008; Cai et al., 2010; Bien et al., 2016). When no natural ordering is available, a simple approach is available by *thresholding* the sample covariance, or setting small estimated elements to zero (Karoui, 2008; Bickel & Levina, 2008a; Rothman et al., 2009; Cai & Liu, 2011). Though such elementwise operations induce sparsity straightforwardly, it is well-documented that the resulting estimator is not always positive definite under finite samples, as the property applies to all elements simultaneously. Adding a log-determinant barrier term (Rothman, 2012) or appealing to alternating directions methods (Xue et al., 2012) have been proposed to address this issue. In general, great care must be taken to ensure positive definiteness by selecting the thresholding constant from an appropriate range, which may be too narrow to induce an effective amount of shrinkage in many cases (Azose & Raftery, 2016).

Penalized likelihood techniques offer an alternative, and are arguably the preferred class of methods for estimating a sparse inverse covariance or *precision* matrix. In this case, the negative Gaussian log-likelihood is convex, and fast algorithms such as the graphical lasso make estimation easy under an ℓ_1 penalty. Folded concave penalties are often desirable instead, and can be considered via a majorization-minimization (MM) algorithm, which entails solving a sequence of reweighed lasso-type problems iteratively. The problem is decidedly more difficult for covariance matrices, whose negative log-likelihood is non-convex, and this approach has been much less widely used. Lam & Fan (2009) study the properties of such an ℓ_1 penalized formulation, and Bien & Tibshirani (2011) propose a majorization-minimization algorithm that makes use of generalized gradient descent. Convergence relies on a Lipschitz continuity assumption, ensured by restricting the solution space to a subset of the positive definite cone. Enforcing the constraint entails another optimization subproblem within an iterative minimization scheme, which can be cumbersome and leads to instability in the high-dimensional setting. Step-size selection has a considerable effect on performance, and the choice often incurs a tradeoff between stability and practical rate of convergence.

Extending this work, our method revisits the penalized likelihood framework for sparse covariance estimation. We propose a novel way to impose sparsity in this setting by way of a *distance penalty* that keeps the estimate close to its projection onto a set constraint. Such a penalty marks a departure from shrinkage-based rules dominating both likelihood and thresholding techniques alike, and a similar approach has proven effective for generalized linear model regression under rank or sparsity constraints (Xu et al., 2017). We do not assume any ordering nor knowledge of the location of zero entries; the method thus performs model selection along with producing a positive definite estimate of the covariance matrix. In contrast with existing approaches, our method does not rely on a global shrinkage penalty. Increasing the penalty parameter corresponds to the exact constrained formulation in the limit, rather than producing the trivial solution. Further, our formulation offers several computational advantages. In particular, both positive definiteness and the descent property of MM algorithms can be practically enforced via simple backtracking. As a result, the algorithm enjoys global convergence to a stationary point of the objective, and the estimate is guaranteed to be positive definite. These merits are demonstrated through extensive simulation studies, and our method is applied to cell signaling and international migration data.

1. BACKGROUND AND PENALIZED FORMULATION

We consider estimating the covariance matrix Σ given n independent, identically distributed random vectors $X_1, \ldots, X_n \sim N_p(0, \Sigma)$. As we focus on estimation of Σ, we assume zero mean

without loss of generality. In this scenario, the negative log-likelihood of the data is

$$\begin{aligned}\mathcal{L}(\Sigma) &= \frac{np}{2}\ln 2\pi + \frac{n}{2}\ln\det\Sigma + \frac{n}{2}\text{tr}(\Sigma^{-1}S) \\ &= \frac{n}{2}\ln\det\Sigma + \frac{n}{2}\text{tr}(\Sigma^{-1}S) + K \end{aligned} \tag{1}$$

where $S = \frac{1}{n}\sum_{i=1}^{N} X_i X_i^T$ denotes the sample covariance and K is a constant term with respect to Σ. When the data are weakly dependent or do not follow a normal distribution, estimation may still proceed on the basis of (1), which becomes correctly interpreted as the quasi-likelihood. We seek to maximize the likelihood subject to a sparse, positive definite solution. Formally, this can be cast as the optimization problem

$$\hat{\Sigma} = \underset{\Sigma\in\mathcal{C},\Sigma\succ 0}{\text{argmin}}\left[\ln\det\Sigma + \text{tr}(\Sigma^{-1}S)\right], \text{ where } \mathcal{C} = \left\{\Sigma\in\mathbb{R}^{p\times p} : \Sigma = \Sigma^T, \|\Sigma\|_0 \leq 2k+p\right\}. \tag{2}$$

The constraint set is written such that k denotes the number of nonzero entries in the upper triangle; that is, the number of free parameters in Σ. Directly solving (2) is difficult: even without considering the positive definiteness constraint $\Sigma \succ 0$ for now, optimizing over $\binom{\binom{p}{2}}{k}$ sparsity patterns for a model with k effective parameters quickly become combinatorially intractable. Instead, it is common to add a lasso penalty to the negative log-likelihood $\mathcal{L}(\Sigma)$, well-known to promote sparse solutions in a wide range of problems. For covariance estimation, Bien & Tibshirani (2011) consider such a penalty applied to $A * \Sigma$, where A has non-negative entries and $*$ denotes the Hadamard or element-wise product. The resulting optimization problem

$$\text{Minimize}\left\{\ln\det\Sigma + \text{tr}(\Sigma^{-1}S) + \lambda\|P*\Sigma\|_1\right\} \text{ subject to } \Sigma\succ 0 \tag{3}$$

remains nontrivial. Though the objective is non-convex, it takes a special form as a difference of convex functions. Making use of this structure, Bien & Tibshirani (2011) propose a majorization-minimization (MM) approach described in the following section. Including a lasso penalty as proxy for the sparsity constraint entails shrinking the solution *globally* toward zero, known to bias the parameters and often introduce spurious covariates. Nonetheless, its nice properties have made the approach a popular one, producing fast algorithms by preserving nice properties such as convexity and continuity in the objective when available. For instance, this is the case for the graphical lasso toward estimating the precision matrix. The ℓ_1-penalized problem (3) remains difficult, and lasso-type regularization still suffers from bias and false positives without benefiting from the full slate of its usual computational advantages. In particular, enforcing the positive definiteness condition is difficult, giving rise to sub-problems that must be solved with an additional optimization routine such as alternating direction methods embedded within the iterative scheme. While our focus is on penalized likelihood estimation, thresholding approaches (Rothman et al., 2009; Rothman, 2012) rely on the same global shrinkage principle, and alternating direction methods have also been considered in this context to enforce positive definiteness (Xue et al., 2012).

We suggest an alternate approach to (2) by instead minimizing the penalized objective

$$f(\Sigma) \equiv \mathcal{L}(\Sigma) + \frac{\rho}{2}\text{dist}(\Sigma,\mathcal{C})^2. \tag{4}$$

The distance penalty term encourages the solution Σ to be close to a constraint set $\mathcal{C}$, and is precisely zero whenever $\Sigma\in\mathcal{C}$. This formulation approaches equivalence to the original constrained problem (2) in the limit as ρ tends to ∞. Such a penalty not only allows straightforward incorporation of constraints onto a loss function while avoiding global shrinkage, but lends itself naturally to practical algorithms constructed within the majorization-minimization framework.

1·1. *Majorization-minimization*

Majorization-minimization (MM) algorithms are gaining traction for solving large-scale optimization problems arising in statistics and machine learning (Mairal, 2015). An MM algorithm successively minimizes a sequence of surrogate functions $q(\Sigma \mid \Sigma_k)$ which lie above the objective function $f(\Sigma)$ about the estimate Σ_k at current iteration k. Decreasing $q(\Sigma \mid \Sigma_k)$ automatically engenders a descent in the objective $f(\Sigma)$ as well (Lange, 2016). The celebrated expectation-maximization (EM) algorithm (Dempster et al., 1977) for maximum likelihood estimation under missing data provides a prominent example of an MM algorithm, where the surrogate q is the expected complete data log-likelihood. Majorization holds whenever two conditions are satisfied: tangency at the current point $q(\Sigma_k \mid \Sigma_k) = f(\Sigma_k)$, and domination $q(\Sigma \mid \Sigma_k) \geq f(\Sigma)$ for every $\Sigma \in \mathbb{R}^{p \times p}$. The algorithm is then defined by the following iteration:

$$\Sigma_{k+1} = \arg\min_{\Sigma} q(\Sigma \mid \Sigma_k).$$

That the descent property holds is readily apparent via two inequalities

$$f(\Sigma_{k+1}) \leq q(\Sigma_{k+1} \mid \Sigma_k) \leq q(\Sigma_k \mid \Sigma_k) = f(\Sigma_k). \tag{5}$$

In particular, because the inequalities (5) are sufficient, we see that domination for all Σ or exact minimization of q at each step are not always strictly necessary.

The MM principle thus offers a general recipe for converting a hard optimization problem into a sequence of manageable subproblems. The penalty $\operatorname{dist}(\Sigma, \mathcal{C})^2$ can be brought into this framework by way of distance majorization (Chi et al., 2014). Begin by rewriting the penalty equivalently in terms of the Euclidean or Frobenius norm: $\operatorname{dist}(\Sigma, \mathcal{C}) = \|\Sigma - P_{\mathcal{C}}(\Sigma)\|_F$, where $P_{\mathcal{C}}(\Sigma)$ denotes the projection of Σ onto $\mathcal{C}$. The inequality $\operatorname{dist}(\Sigma, \mathcal{C})^2 \leq \|\Sigma - P_{\mathcal{C}}(\Sigma_k)\|_F^2$ now follows from the definitions of the set distance and the projection operator $P_{\mathcal{C}}(\Sigma)$. This majorization is useful in practice as $P_{\mathcal{C}}(\Sigma_k)$ on the right hand side is now constant with respect to Σ. Making use of the squared distance renders the term differentiable: $\nabla \frac{1}{2}\|\Sigma - P_{\mathcal{C}}(\Sigma_k)\|_F^2 = \Sigma - P_{\mathcal{C}}(\Sigma_k)$. The following section describes how these distance majorizations give way to an MM algorithm that converts our task (2) into a sequence of unconstrained, smooth problems.

2. SEQUENTIAL UNCONSTRAINED MINIMIZATION

In this section, we derive quadratic surrogates q which more closely approximate f than commonly used tangent plane majorizations. The algorithm is an instance of the recently introduced proximal distance principle (Lange & Keys, 2015), where surrogates are constructed in the form $q(x \mid x_k) = g(x) + \frac{\rho}{2}\|x - P_{\mathcal{C}}(x_k)\|^2$ for minimizing a loss function g under constraint set $\mathcal{C}$. Its namesake refers to the minimizer of $q(x \mid x_k)$, given by $\operatorname{prox}_{\rho^{-1}g}[P_{\mathcal{C}}(x_k)]$, where the proximal mapping with parameter λ is defined

$$\operatorname{prox}_{\lambda g}(y) \equiv \operatorname*{argmin}_x g(x) + \frac{1}{2\lambda}\|x - y\|_2^2.$$

We see that $\operatorname{prox}_{\lambda g}(y)$ compromises between minimizing f and being close to y, with parameter λ modulating the tradeoff. Following the solution as $\rho \to \infty$ is the philosophy behind the proximal distance principle, respecting the constraint $\mathcal{C}$ in the limit. Proximal operators are closely related to gradient methods. Indeed, in generalized gradient algorithms with a lasso penalty $\frac{\lambda}{2}\|\cdot\|_1$, the proximal mapping produces the soft-thresholding step $S_{\lambda t}[\nabla g(x)]$. Polson et al. (2015) provide an excellent overview of proximal methods in statistics.

2·1. *Constructing and minimizing the surrogates*

Recall $\mathcal{L}(\Sigma) = \ln\det\Sigma + \operatorname{tr}(\Sigma^{-1}S)$. We define a sequence of quadratic approximations q to the objective by expanding: taking matrix derivatives of $\mathcal{L}(\Sigma)$ in directions U and V,

$$\begin{aligned} d_V\mathcal{L} &= \operatorname{tr}(\Sigma^{-1}V) - \operatorname{tr}(\Sigma^{-1}V\Sigma^{-1}S) \\ d_U d_V\mathcal{L} &= -\operatorname{tr}(\Sigma^{-1}U\Sigma^{-1}V) + \operatorname{tr}(\Sigma^{-1}U\Sigma^{-1}V\Sigma^{-1}S) + \operatorname{tr}(\Sigma^{-1}V\Sigma^{-1}U\Sigma^{-1}S) \\ V^T d^2\mathcal{L}V &= -\operatorname{tr}(\Sigma^{-1}V\Sigma^{-1}V) + 2\operatorname{tr}(\Sigma^{-1}V\Sigma^{-1}V\Sigma^{-1}S), \end{aligned}$$

where the quadratic form in the last line is obtained by setting $U = V$. The second differential simplifies considerably by approximating $S \approx \Sigma$, a maneuver is akin to scoring as $\mathbb{E}(S) = \Sigma$:

$$V^T d^2\mathcal{L}V \approx \operatorname{tr}(\Sigma^{-1}V\Sigma^{-1}V).$$

We may now define q by taking the second-order Taylor expansion of the log-likelihood $\mathcal{L}$ about the current estimate

$$\begin{aligned} q(\Sigma \mid \Sigma_k) = \mathcal{L}(\Sigma_k) &+ \operatorname{tr}[\Sigma_k^{-1}(\Sigma - \Sigma_k)] - \operatorname{tr}[\Sigma_k^{-1}S\Sigma_k^{-1}(\Sigma - \Sigma_k)] \\ &+ \frac{1}{2}\operatorname{tr}[\Sigma_k^{-1}(\Sigma - \Sigma_k)\Sigma_k^{-1}(\Sigma - \Sigma_k)] + \frac{\rho}{2}\|\Sigma - P_{\mathcal{C}}(\Sigma_k)\|_F^2. \end{aligned}$$

where our distance penalty appears as the final term. In contrast to ℓ_1-penalized functions, this surrogate is differentiable:

$$dq(\Sigma \mid \Sigma_k)/d\Sigma = \rho[\Sigma - P_{\mathcal{C}}(\Sigma_k)] + \Sigma_k^{-1} - \Sigma_k^{-1}S\Sigma_k^{-1} + \Sigma_k^{-1}(\Sigma - \Sigma_k)\Sigma_k^{-1}.$$

Setting to zero and rearranging reveals that minimizing $q(\Sigma \mid \Sigma_k)$ amounts to solving

$$\rho P_{\mathcal{C}}(\Sigma_k) + \Sigma_k^{-1}S\Sigma_k^{-1} = \rho\Sigma + \Sigma_k^{-1}\Sigma\Sigma_k^{-1}. \tag{6}$$

We will abbreviate the left-hand side

$$\rho P_{\mathcal{C}}(\Sigma_k) + \Sigma_k^{-1}S\Sigma_k^{-1} \equiv D.$$

This matrix equation can be solved directly upon vectorization: denoting the Kronecker product $\otimes$, (6) becomes

$$\operatorname{vec}(D) = \rho\operatorname{vec}(\Sigma) + (\Sigma_k^{-1} \otimes \Sigma_k^{-1})\operatorname{vec}(\Sigma).$$

We may now immediately invert to obtain the solution

$$\operatorname{vec}(\widetilde{\Sigma}) = \left[\rho I + (\Sigma_k^{-1} \otimes \Sigma_k^{-1})\right]^{-1}\operatorname{vec}(D)$$

and recover the minimizer $\widetilde{\Sigma}$ by reshaping. Since $(\Sigma_k^{-1} \otimes \Sigma_k^{-1})^{-1} = \Sigma_k \otimes \Sigma_k$, the Sherman-Woodbury-Morrison lemma shows that matrix inversion is required only once by expressing the solution

$$\operatorname{vec}(\widetilde{\Sigma}) = \left[\frac{1}{\rho}I - \frac{1}{\rho^2}\left(\Sigma_k \otimes \Sigma_k + \frac{1}{\rho}I\right)^{-1}\right]\operatorname{vec}(D). \tag{7}$$

We briefly remark that estimation based on the sample correlation using the same algorithm is straightforward, analogous to what has been suggested in previous studies and resulting in penalization on the correlation scale, detailed in the Appendix.

Due to the Kronecker product, computing the analytic solution (7) becomes expensive for large N. Upon inspection, (6) takes the general form $AX + XB = D$, known as a Sylvester

equation in X. Like the closely related and better-known Lyapunov equations central in dynamical systems, Sylvester equations frequently arise in control theory and eigenvalue problems. As such, they are well-studied objects. It is known that the equation has a unique solution if and only if eigenvalues of A and $-B$ are distinct. This is true in our case by positive definiteness of Σ_0^{-1}. More saliently, we may borrow numerical methods from the control theory literature. An algorithm due to Bartels & Stewart (1972) provides a more efficient solution than evaluating (7). The first step and crux of this procedure lies in transforming the problem into Schur form by computing decompositions $A = URU^T, B = VSV^T$ via the QR algorithm. Because R, S are upper triangular, the equivalent triangular system $R^TY + YS = U^TDU$ becomes soluble by simple back-substitution. Multiplication then recovers the original solution $X = UYV^T$. The computational complexity is thereby reduced from $\mathcal{O}(p^6)$ operations required to compute (7) to $\mathcal{O}(p^3)$. Current state-of-the-art implementations are variations on this theme and possess the same overall complexity (Simoncini, 2016).

2·2. *Backtracking and gradient interpretation*

Because D may not be positive definite, neither is the minimizer $\widetilde{\Sigma}$ of (6). Further, the surrogate q does not strictly majorize f, and as such it is possible that minimizing q does not yield a descent in f. Both of these issues are handled via backtracking: for positive definite Σ_k,

$$\Sigma_{k+1} = \Sigma_k + (1/2^s)(\widetilde{\Sigma} - \Sigma_k) \tag{8}$$

is bound to produce a descent in f with $\Sigma_k \succ 0$ as we have expanded about $\ln\det\Sigma$. This is made rigorous in proving Theorem 1; to see this, consider minimization of a quadratic equivalently as one step of Newton's algorithm. We may express

$$\widetilde{\Sigma} = \Sigma_k + v; \qquad v = -H(\Sigma_k)^{-1}\nabla q(\Sigma_k \mid \Sigma_k), \tag{9}$$

where H is the scoring approximation to the second differential. We avoid explicitly writing the unwieldy tensor, instead specifying that H generates the quadratic form $\mathrm{tr}(\Sigma^{-1}V\Sigma^{-1}S)$. The update is implicitly a gradient step, and v is a descent direction. A decrease in f along with positive definiteness is thus guaranteed with step-halving

$$\Sigma_{k+1} = \Sigma_k + \eta^s v$$

where s is the number of backtracking steps. Choosing $\eta = 1/2$ recovers (8).

Before proceeding, we pause to compare to the surrogates proposed by Bien & Tibshirani (2011) for intuition. Based on the concave-convex procedure of Yuille & Rangarajan (2003), the approach makes use of the tangent plane majorization

$$g(\Sigma \mid \Sigma_0) = \ln\det\Sigma_0 + \mathrm{tr}(\Sigma_0^{-1}\Sigma) - p + \mathrm{tr}(\Sigma^{-1}S) + \lambda\|A * \Sigma\|_1$$

to the ℓ_1-penalized objective (3). The resulting MM iteration

$$\hat{\Sigma}_t = \operatorname*{argmin}_{\Sigma\succ 0}\left[\mathrm{tr}(\hat{\Sigma}_{t-1}^{-1}\Sigma) + \mathrm{tr}(\Sigma^{-1}S) + \lambda\|A * \Sigma\|_1\right]$$

is carried out by proximal gradient descent. The choice of step-size parameter is crucial to the practical rate of convergence. Because the linear approximation loosely approximates f, a given step-size may be well-suited at some points but drastically overshoot the minimum or leave the positive definite cone at others. Whenever the latter occurs, an additional subproblem arises and is solved iteratively using alternating directions methods. In our experience with the method, decreasing the step-size trades off speed for stability and vice versa. The quadratic expansion of the log-likelihood produces a surrogate q that more closely hugs f, but sacrifices strict majorization

at all Σ. In practice, this trade-off is well worth as illustrated in Section 3·1. Making use of a much tighter quadratic approximation more efficiently decrements f, and the minimizer $\widetilde{\Sigma}$ very rarely fails to cause a decrease in f. Working backwards from the minimum of q avoids tuning a step-size, and simple backtracking bypasses the need for another inner optimization problem within an iterative scheme.

2·3. *Convergence*

Because the symmetric sparsity set $\mathcal{C}$ is non-convex, the projection operator may be multi-valued at unusual points Σ_k in theory. The penalty $\text{dist}(\Sigma, \mathcal{C})^2$ is only guaranteed to be differentiable everywhere when $\mathcal{C}$ is convex. Surrogates $q(\Sigma \mid \Sigma_k)$ are differentiable, but the objective $f(\Sigma)$ becomes only semi-differentiable, and standard analyses of the convergence of gradient methods or proximal distance algorithms do not apply. Nevertheless, continuity of f is enough to appeal to theory by Zangwill (1969) that applies when the algorithm entails a point-to-set mapping. The following global convergence result is proven in the Appendix.

THEOREM 1. *Consider the algorithm A mapping $\Sigma_{k+1} = \Sigma_k + \eta v \in A(\Sigma_k)$ where v is as in* (9) *and $\eta \in (0, 1)$. Let the initial point Σ_0 be positive definite. Then the sequence of iterates $\{\Sigma_k\}$ satisfying $\Sigma_{k+1} = A(\Sigma_k)$ lies within the positive definite cone, and the limit points of any convergent subsequence of $\{\Sigma_k\}$ are critical points of $f(\Sigma)$.*

We clarify that global convergence refers to convergence to a critical point from any initial Σ_0, rather than to a global minimizer. The result is stated in terms of subsequences in case the set of stationary points is infinite, corresponding to the unidentifiable setting. The solution set is connected due to monotonicity, and the entire sequence Σ_k converges to a local minimum whenever Γ is finite by Theorem 3.1 in (Meyer, 1976).

3. EMPIRICAL RESULTS

We illustrate the practical merits of our algorithm on a suite of simulated examples and two real data analyses.

3·1. *Simulation study*

An open-source Julia implementation of our algorithm is available at the first author's website. (To do: **Add cross-validation results when competing method finishes on simulation suite**). Figure 1 considers synthetic data examples studied in (Bien & Tibshirani, 2011). This includes a model in which nonzero entries are randomly chosen at eight percent sparsity, a first-order moving average model, and a cliques model where $p > n$. The distance penalty has similar area under the ROC curve, but is clearly preferable in terms of entropy loss especially in the high-dimensional setting. Upon reproducing their results, we confirm that calling ADMM to enforce the positive definiteness constraint is rather rare on the settings considered with $n > p$. However, this is not the case in the high-dimensional setting, where the sample covariance cannot be full rank, and in our experience, numerical errors arise when alternating between generalized gradient steps and ADMM subroutines. We see in Figure 1 that some instability is present, more notably for the adaptive version of the generalized gradient where the reciprocal of the sample covariance entries are used as weights in the penalty, akin to adaptive lasso. At best, it is necessary to significantly reduce step size, resulting in very small steps toward convergence.

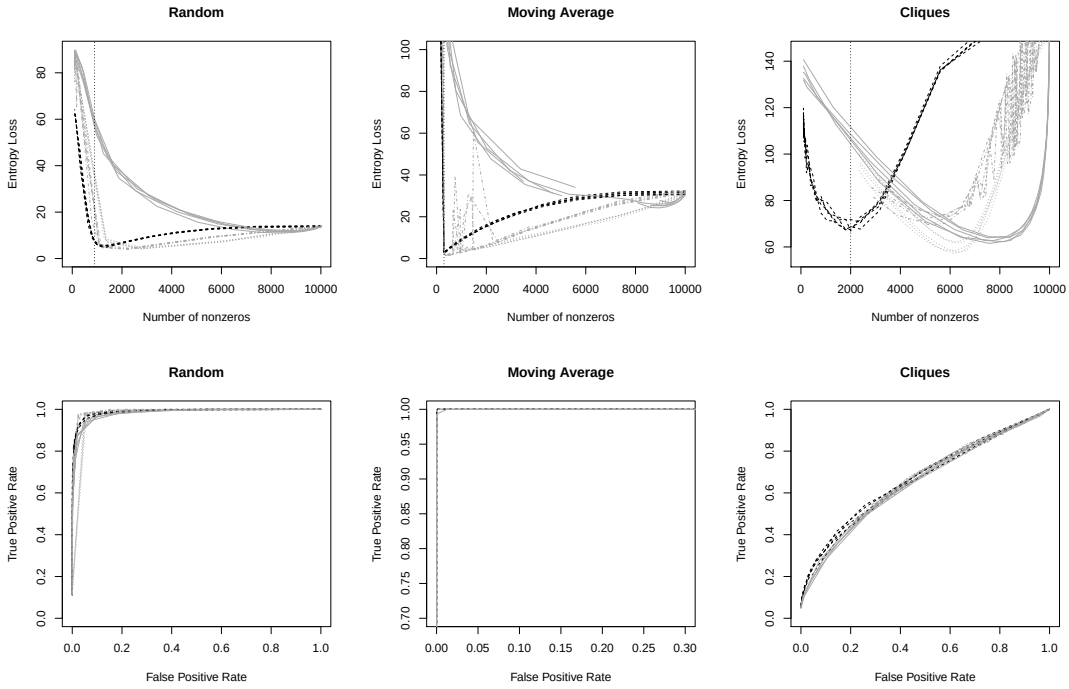

Fig. 1. Black dashed line represents proximal distance results. Gray solid, dotted, and dot-dash lines represent thresholding, generalized gradient, and "adaptive" generalized gradient respectively. The vertical line marks the number of nonzeros in the true covariance matrix.

3·2. *International migration data*

Projecting international migration at the country-specific scale is important toward shaping policy such as social welfare program planning. Probabilistic projections are desirable in providing uncertainty toward this "barely predictable" goal, as described by Bijak & Wiśniowski (2010), but existing global models typically assume that forecast errors are uncorrelated across countries. In this setting, it becomes important to estimate the marginal dependency structure: though modeling with the independence assumption may be well-calibrated for individual countries, ignoring correlations will yield under- or over-estimates of projected migration.

We consider international migration forecast data from the United Nations World Population Prospects (UN WPP) division. The dataset consist of net migration estimates every five years in each country from 1950 to 2010. Following Azose & Raftery (2016), our goal is to estimate the correlation structure among forecast errors. The observations ϵ_t $(t = 1, \ldots, 11)$ are residual vectors from an AR(1) model for net migration in all countries; ϵ_t are assumed to be independent and identically distributed according to a multivariate normal distribution. We are basing inference on a small sample, seeking to estimate the correlation matrix R with roughly $18,000$ effective parameters corresponding to country pairs on the basis of 11 measurements. The Pearson sample correlation is known to degrade in such settings, revealing spurious correlations. Azose & Raftery (2016) consider a Bayesian approach, shrinking *a priori* untrustworthy elements toward zero via penalizing country pairs that are far apart, do not share a colonial relationship, and are in the same region according to the WPP partition into 22 regions based on geographical and cultural affinity. The authors consider a slight modification of the approach in (Bien & Tibshirani,

2011) to perform maximum *a posteriori* estimation. They note that the method is slow on a problem of this size. In particular, Azose & Raftery (2016) remark that cross-validation is infeasible, and instead choose λ via a heuristic criterion.

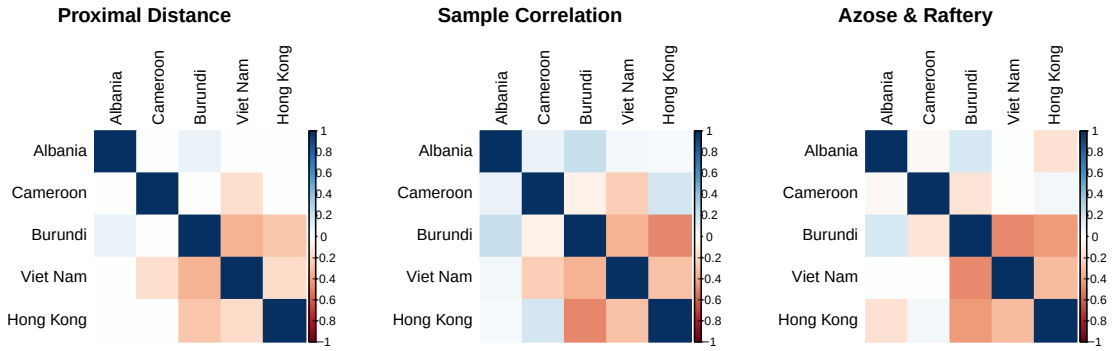

Fig. 2. Estimated correlations on a random subset of countries. The proximal distance algorithm results in a sensible sparsity pattern based on criteria suggested by Azose & Raftery (2016). Though not visibly discernable, their maximum *a posteriori* estimate produces no sparse entries on this subset.

Our algorithm converges in under a minute on a standard laptop computer applied to these data, making cross-validation practical. Our estimate suggests 6145 zeroes, considerably more sparse than the estimate in (Azose & Raftery, 2016) consisting of 323 zero entries.

To illustrate the difference in estimates, we consider a random subset of five countries. Further details of the data analysis and additional random subsets are included in the Appendix. Despite the absence of prior knowledge, our method reveals relationships that are qualitatively consistent with the criteria used by Azose & Raftery (2016) in the design of the prior. As apparent in Figure 2, zero entries estimated under our method correspond to country pairs that we may not expect to be correlated under such criteria, while the method of Azose & Raftery (2016) fails to produce sparse entries.

To compare quality of estimates quantitatively, we advocate the extended Bayesian information criterion (EBIC). EBIC is shown to be more suitable in high dimension settings, though we report standard AIC and BIC for completeness in Table 1. Unsurprisingly, the denser estimate by Azose & Raftery (2016) achieves a lower negative log-likelihood on the data, but the other measures we consider favor our sparser solution. We hesitate to conclude that our estimate $\hat{R}$ is preferable due to the very limited amount of data. Indeed, the use of sensible prior knowledge in such a setting is prudent. At the very least, the analysis shows that without assuming any information *a priori*, our method is competitive with a principled recent approach, while its computational advantages newly make feasible cross-validated estimation on a high-dimensional problem.

3·3. *Flow cytometry data*

Next we revisit data from a cell signaling study by Sachs et al. (2005), comprised of flow cytometry measurements on $p = 11$ proteins and $n = 7466$ cells. Previously analyzed in the original graphical lasso paper (Friedman et al., 2007) as well as the ℓ_1-penalized covariance

Table 1. *Model Selection Criteria*

	$\hat{\mathcal{L}}$	EBIC	BIC	AIC	Mean $\hat{\mathcal{L}}_{CV}$ (error)
Azose & Raftery	$\mathbf{-572.1}$	43208.1	41591.1	34499.8	1254.1 (363.6)
Proximal Distance	-341.7	**39701.3**	**28091.3**	**23216.6**	**434.2 (140.8)**

Comparison of the maximum a posteriori estimate by Azose & Raftery (2016) and our estimate in terms of negative log-likelihood, extended BIC, BIC, AIC, and cross-validated negative log-likelihood. Standard error of cross-validated estimates are in parentheses.

estimation study (Bien & Tibshirani, 2011), the dataset has some historical significance, and provides a small enough example to run existing algorithms and easily visualize the complete solutions. We produce two estimates of the Markov graph using the graphical lasso whose solutions contain $k = 9$ and $k = 16$ edges. Bien & Tibshirani (2011) apply their method to estimate the covariance graph, which does not coincide with estimates of the Markov graph at matched sparsity levels. This is no surprise since the covariance graph offers a distinct interpretation: a missing edge tells us that the concentration of one protein does not inform us about another, whereas a missing edge in the Markov graph indicates that the concentration of one protein gives no information about the other *conditional* on all other concentrations. While this difference is indeed crucial, our results suggest that the covariance graph may be more similar to the Markov graph than solutions using an ℓ_1 penalized approach may suggest.

Figure 3 displays covariance graphs obtained via the generalized gradient descent of Bien & Tibshirani (2011) and under our proximal distance algorithm at sparsity levels matched to the Markov graphs. It is visually clear that our estimate of $\hat{\Sigma}$ has more common edges with the Markov graph. While there is no known ground truth dependency network as a baseline and the covariance and Markov graphs do not always coincide, these results suggest that their discrepancy may be overstated based on generalized gradient descent. This may be either due to unwanted shrinkage or convergence to a poor local minimum. It is again difficult to produce a complete range of sparsity levels, and the generalized gradient estimate on the bottom row features one fewer edge than desired, though it yields the closest sparsity estimate before transitioning to $k = 11$ over a grid search of mesh size 1e-7. Even in the extreme case (not pictured) where penalty parameters are chosen so that only $k = 1$ edge is present, the proximal distance algorithm and graphical lasso agree in producing only the edge Mek—Raf, while generalized gradient estimates the edge Erk—Akt. We briefly remark that when a specific sparsity level k is desired as is the case here, our problem formulation allows us to directly specify k, while ℓ_1 approaches require tedious calibration due to the lack of an explicit relationship between k and penalty parameter λ.

4. DISCUSSION

We propose a novel approach to estimating a sparse covariance matrix that does not appeal to global shrinkage. Our technique makes use of a natural and interpretable penalty based on the distance to the constraint represented as a set, and its merits are justified theoretically and showcased empirically. Building on groundwork laid by Bien & Tibshirani (2011), the majorization-minorization once again provides a framework from which to build algorithms toward penalized likelihood estimation in a difficult, non-convex setting. Our methodological contributions mark a substantial improvement upon existing work, owing to the efficiency of the surrogate functions as well as a surprising connection to control theory that arises from our distance penalty formulation.

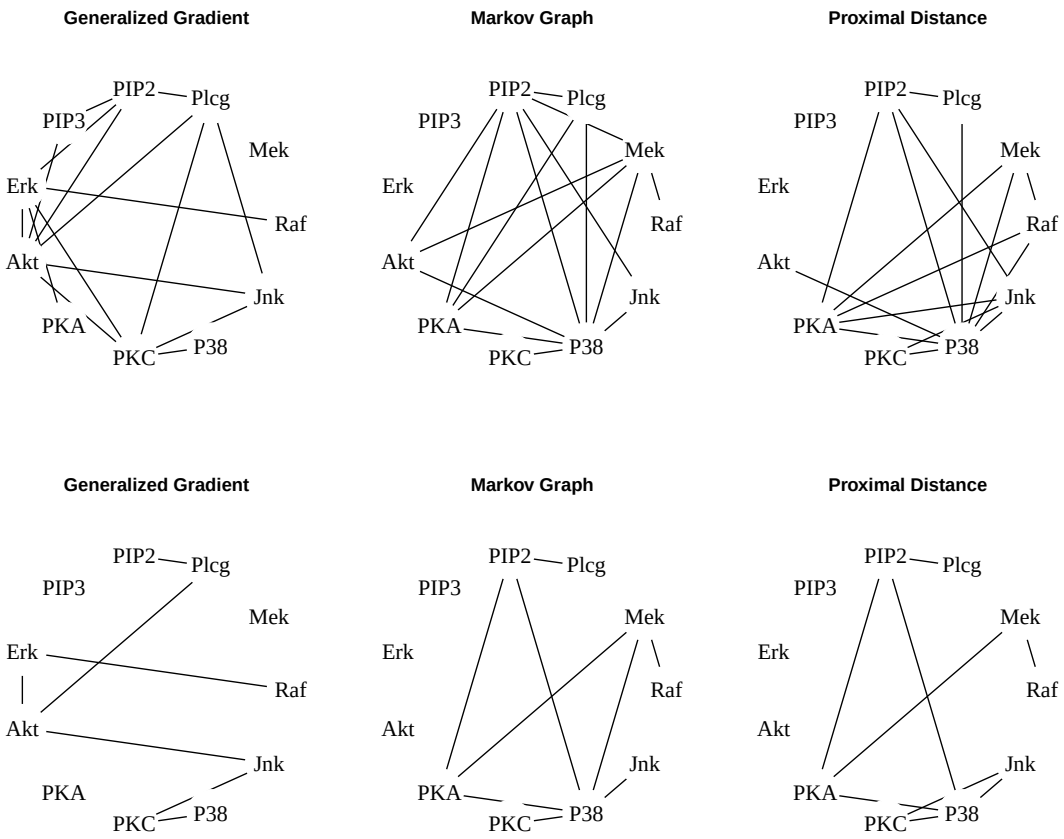

Fig. 3. Estimated covariance graphs under generalized gradient descent (left) and our proximal distance algorithm (right) compared to the graphical lasso estimate of the Markov graph (center). Top and bottom panels display two settings in which sparsity levels are matched between methods. In each case, we see that the proposed method produces more edges in common with the Markov graph.

Future work may pursue extensions for large-scale applications with lower than cubic complexity. Iterative approaches offer a practical alternative for large-scale A. Note an explicit sequential approach to solving the subproblem (6) is given by taking $\Sigma_{k+1} = \frac{1}{\rho}D - \frac{1}{\rho}\Sigma_0^{-1}\Sigma_k\Sigma_0^{-1}$ for iterations k. As long as $\rho > \|\Sigma_0^{-2}\|$, we have

$$\frac{1}{\rho}\|\Sigma_0^{-1}\Sigma\Sigma_0^{-1}\| = \frac{1}{\rho}\|\Sigma_0^{-2}\| \cdot \|\Sigma\| < 1.$$

More generally, rational Krylov subspace methods are well-suited to solve very large systems (6), and work especially well when A is not dense. For further details and current implementations, we refer the interested reader to Simoncini (2016).

We note that previous convergence results for proximal distance algorithms do not consider non-convex sets, and our proof nowhere assumes $\mathcal{C}$ must be the sparsity set. Our convergence analysis considers a fixed penalty parameter ρ, and is justified as we may consider all previous

values part of initialization and still retain the local optimality guarantees. A joint analysis of the algorithm maps across iterations that considers the changing objective functions is beyond the scope of this paper, but will be elucidating in fully understanding our proposed method.

We have rather naively set an increasing step size scheme for *all* examples, and that doing so naively works well empirically across the board should be considered an advantage. Nevertheless, a closer look at different step size schedules will be fruitful, and potentially crucial in some applications.

(To do:)

-Link to Julia implementation

5. ACKNOWLEDGEMENTS

We would like to thank Jon Azose and Adrian Raftery for sharing the UN migration forecasting correlations, and Jacob Bien for providing the cell signaling data.

APPENDIX 1

Proof of Theorem 1

To establish global convergence, we show that the conditions of Zangwill's Global Convergence Theorem on page 91 of Zangwill (1969) are satisfied. In analyzing our algorithm, we make use of a point-to-set map A because the projection $P_{\mathcal{C}}(\Sigma_k)$ appearing in q may be multi-valued at unusual points Σ_k. We denote the set of stationary points of A by Γ. The theorem statement is reproduced in our notation below for convenience:

THEOREM A1. *(Global Convergence Theorem) Let algorithm $A : X \to \mathcal{P}(X)$ be a point-to-set map and suppose that given Σ_0, the sequence $\{\Sigma_k\}$ is generated such that $\Sigma_{k+1} \in A(\Sigma_k)$. Let $\Gamma \subset X$ be the solution set, and assume*

i. all iterates Σ_k are contained in a compact set $S \subset X$

ii. there is a continuous function f such that

a) if $\Sigma \notin \Gamma$, then $f(\Omega) < f(\Sigma)$ for all $\Omega \in A(\Sigma)$

b) if $\Sigma \in \Gamma$, then $f(\Omega) \leq f(\Sigma)$ for all $\Omega \in A(\Sigma)$

iii. the mapping A is closed at points outside of Γ

Then the limit of any convergent subsequence of $\{\Sigma_k\}$ is a solution.

We begin by proving coercivity of $\mathcal{L}(\Sigma)$ toward showing that $\{\Sigma_k\}$ is contained in a compact set.

LEMMA A1. *The negative log-likelihood $\mathcal{L}(\Sigma)$ is a coercive function.*

Proof. It suffices to show that $\ln\det\Sigma + \text{tr}(\Sigma^{-1}S)$ is coercive. Let the singular values of Σ be denoted $\sigma_1 \geq \sigma_2 \geq \ldots > 0$, and let $s_1 \geq s_2 \geq \ldots > 0$ be the singular values of S. We have $\|\Sigma\| \to \infty$ if and only if at least one $\sigma_i \to \infty$. Similarly, $\|\Sigma^{-1}\| \to \infty$ is equivalent to some $\sigma_i \to 0$. The matrix analogue of the Cauchy-Schwarz inequality due to Von Neumann and Fan tells us that $\text{tr}(\Sigma^{-1}S) \geq \sum_i \frac{s_i}{\sigma_i}$. We also have $\ln\det\Sigma \geq \sum_i \ln\sigma_i$. Let us consider their sum $\sum_i \ln\sigma_i + \frac{s_i}{\sigma_i}$ which bounds $\mathcal{L}$ below. Since

$$\min_i \ln\sigma_i + \frac{s_i}{\sigma_i} \geq \ln s_i + 1,$$

the summands are lower bounded as $\sigma_i \to \infty$, while the first term $\sum_i \ln\sigma_i \to \infty$. On the other hand, if $\sigma_i \to 0$, the growth of s_i/σ_i dominates the $\ln\sigma_i$ term, so that $\mathcal{L}(\Sigma) \to \infty$ in either case. That is, $\mathcal{L}(\Sigma)$ is a coercive function. □

Before proving the next lemma, recall that the minimizer of (6) may be written $\widetilde{\Sigma} = \Sigma_k + v$. Here $v = -H(\Sigma_k)^{-1}\nabla q(\Sigma_k \mid \Sigma_k)$ and H is the approximate second differential that generates quadratic form $\mathrm{tr}(\Sigma^{-1}V\Sigma^{-1}S)$. The algorithm map is defined $\Sigma_{k+1} = \Sigma_k + \eta v \in A(\Sigma_k)$ where $\eta \in (0, 1)$.

LEMMA A2. *The algorithm $A(\Sigma_k)$ is a closed map. Further, any point $\Sigma_{k+1} \in A(\Sigma_k)$ decreases f and lies within the positive definite cone for all $\Sigma_k \notin \Gamma$.*

Proof. By Lemma A1, $\mathcal{L}$ is coercive, and since $\mathrm{dist}(\Sigma, \mathcal{C})^2$ is non-negative, their sum f is coercive. This implies compactness of the sub-level set $\mathcal{S}_f(\Sigma_k) = \{\Sigma : f(\Sigma) \leq f(\Sigma_k)\}$ for any point Σ_k. Let $\Sigma_k \notin \Gamma$. Because Σ_k is not stationary, there exists a point $z_k \in P_{\mathcal{C}}(\Sigma_k)$ such that

$$\nabla q(\Sigma_k \mid \Sigma_k) = \nabla \mathcal{L}(\Sigma_k) + \rho(\Sigma_k - z_k) \neq 0.$$

This implies $v \neq 0$, and together with continuity of f on $\mathcal{S}_f(\Sigma_k)$, A is a closed mapping by Theorem 8.4 in Luenberger & Ye (1984). Since $q(\Sigma_k \mid \Sigma_k) = f(\Sigma_k)$, $\nabla q(\Sigma_k \mid \Sigma_k)$ is a semi-derivative of f and v is therefore a descent direction in f. Thus, we can choose $\eta \in (0, 1)$ such that $\Sigma_{k+1} = \Sigma_k + \eta v \in A(\Sigma_k)$ satisfies $f(\Sigma_{k+1}) < f(\Sigma_k)$. Finally, because $\Sigma_{k+1} \in \mathcal{S}_f(\Sigma_k)$, its singular values are bounded away from zero as $f(\Sigma_{k+1})$ must be finite, so that Σ_{k+1} remains positive definite. □

Now we are ready to prove Theorem 1 by a direct application of Zangwill's theorem.

Proof. The sub-level set $\mathcal{S}_f(\Sigma_0)$ is compact, and by Lemmas A1 and A2, all points $\{\Sigma_k\}$ satisfying $\Sigma_{k+1} \in A(\Sigma_k)$ lie in $\mathcal{S}_f(\Sigma_0)$. Lemma A2 further shows that $\Sigma_k \succ 0$ for every k, and its proof shows that f is a continuous function such that if $\Sigma \notin \Gamma$, then $f(\Omega) < f(\Sigma)$ for all $\Omega \in A(\Sigma)$. Further, if $\Sigma \in \Gamma$, then $\nabla \mathcal{L}(\Sigma) + \rho(\Sigma - z) = 0$ for all $z \in P_{\mathcal{C}}(\Sigma)$, so that the gradient is well-defined and $\nabla f(\Sigma) = 0$. That is, whenever $\Sigma \in \Gamma$, the condition $f(\Omega) \leq f(\Sigma)$ holds for all $\Omega \in A(\Sigma)$. Finally, A is closed outside of Γ directly as a result of Lemma A2. Therefore, Theorem A1 applies, and for any convergent subsequence of $\Sigma_k \to \Sigma_*$, the limit $\Sigma_* \in \Gamma$ is a critical point. □

APPENDIX 2

Estimation using the sample correlation

As Bien & Tibshirani (2011) note, let $R = V^{-1/2}SV^{-1/2}$ denote the sample correlation, where $D = \mathrm{diag}(S_{11}, \ldots, S_{pp})$ contains the observed variances. When estimating with R, we may consider minimizing

$$\ln\det\Theta + \mathrm{tr}(\Theta^{-1}R) + \frac{\rho}{2}\|\Theta - P_{\mathcal{C}}(\Theta)\|_F^2$$

over $\Theta \succ 0$. Upon rewriting in terms of $\Sigma = V^{1/2}\Theta V^{1/2}$ equivalently as

$$\sum_{i=1}^{p} -\ln S_{ii} + \ln\det\Sigma + \mathrm{tr}(\Sigma^{-1}S) + \frac{\rho}{2}\mathrm{dist}(V^{-1/2}\Sigma V^{-1/2}, \mathcal{C})^2,$$

we see the task is equivalent to solving (4) with penalization on the correlation scale.

REFERENCES

AZOSE, J. J. & RAFTERY, A. E. (2016). Estimating large correlation matrices for international migration. *arXiv preprint arXiv:1605.08759* .

BARTELS, R. H. & STEWART, G. W. (1972). Solution of the matrix equation ax+ xb= c [f4]. *Communications of the ACM* **15**, 820–826.

BICKEL, P. J. & LEVINA, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics* , 2577–2604.

BICKEL, P. J. & LEVINA, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics* , 199–227.

BIEN, J., BUNEA, F. & XIAO, L. (2016). Convex banding of the covariance matrix. *Journal of the American Statistical Association* **111**, 834–845.

BIEN, J. & TIBSHIRANI, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika* **98**, 807–820.

BIJAK, J. & WIŚNIOWSKI, A. (2010). Bayesian forecasting of immigration to selected european countries by using expert knowledge. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **173**, 775–796.

CAI, T. & LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* **106**, 672–684.

CAI, T. T., ZHANG, C.-H., ZHOU, H. H. et al. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics* **38**, 2118–2144.

CHI, E. C., ZHOU, H. & LANGE, K. (2014). Distance majorization and its applications. *Mathematical programming* **146**, 409–436.

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)* , 1–38.

FAN, J., LIAO, Y. & LIU, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal* **19**.

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441.

KAROUI, N. E. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics* , 2717–2756.

LAM, C. & FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics* **37**, 4254.

LANGE, K. (2016). *MM Optimization Algorithms*. SIAM.

LANGE, K. & KEYS, K. L. (2015). The proximal distance algorithm. *arXiv preprint arXiv:1507.07598* .

LEVINA, E., ROTHMAN, A., ZHU, J. et al. (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics* **2**, 245–263.

LUENBERGER, D. G. & YE, Y. (1984). *Linear and nonlinear programming*, vol. 2. Springer.

MAIRAL, J. (2015). Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization* **25**, 829–855.

MEYER, R. R. (1976). Sufficient conditions for the convergence of monotonic mathematicalprogramming algorithms. *Journal of computer and system sciences* **12**, 108–121.

POLSON, N. G., SCOTT, J. G. & WILLARD, B. T. (2015). Proximal algorithms in statistics and machine learning. *Statistical Science* **30**, 559–581.

POURAHMADI, M. (2011). Covariance estimation: The glm and regularization perspectives. *Statistical Science* , 369–387.

ROTHMAN, A. J. (2012). Positive definite estimators of large covariance matrices. *Biometrika* **99**, 733–740.

ROTHMAN, A. J., LEVINA, E. & ZHU, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association* **104**, 177–186.

SACHS, K., PEREZ, O., PE'ER, D., LAUFFENBURGER, D. A. & NOLAN, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**, 523–529.

SIMONCINI, V. (2016). Computational methods for linear matrix equations. *SIAM Review* **58**, 377–441.

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, vol. 1.

WU, W. B. & POURAHMADI, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90**, 831–844.

XU, J., CHI, E. C. & LANGE, K. (2017). Generalized linear model regression under distance-to-set penalties. *arXiv preprint arXiv:1711.01341* .

XUE, L., MA, S. & ZOU, H. (2012). Positive-definite 1-penalized estimation of large covariance matrices. *Journal of the American Statistical Association* **107**, 1480–1491.

YUILLE, A. L. & RANGARAJAN, A. (2003). The concave-convex procedure. *Neural computation* **15**, 915–936.

ZANGWILL, W. I. (1969). *Nonlinear programming: a unified approach*. Prentice-Hall.