Scalable Sparse Covariance Estimation via Self-Concordance

Anastasios Kyrillidis, Rabeeh Karimi Mahabadi,
Quoc Tran Dinh and **Volkan Cevher**[*]
École Polytechnique Fédérale de Lausanne
{anastasios.kyrillidis,rabeeh.karimimahabadi,quoc.trandinh,volkan.cevher}@epfl.ch

Abstract

We consider the class of convex minimization problems, composed of a self-concordant function, such as the $\log\det$ metric, a convex data fidelity term $h(\cdot)$ and, a regularizing – possibly non-smooth – function $g(\cdot)$. This type of problems have recently attracted a great deal of interest, mainly due to their omnipresence in top-notch applications. Under this *locally* Lipschitz continuous gradient setting, we analyze the convergence behavior of proximal Newton schemes with the added twist of a probable presence of inexact evaluations. We prove attractive convergence rate guarantees and enhance state-of-the-art optimization schemes to accommodate such developments. Experimental results on sparse covariance estimation show the merits of our algorithm, both in terms of recovery efficiency and complexity.

Introduction

Convex ℓ_1-regularized $\log\det$ divergence criteria have been proven to produce – both theoretically and empirically – consistent modeling in diverse top-notch applications. The literature on the setup and utilization of such criteria is expanding with applications in Gaussian graphical learning (Dahl, Vandenberghe, and Roychowdhury 2008; Banerjee, El Ghaoui, and d'Aspremont 2008; Hsieh et al. 2011), sparse covariance estimation (Rothman 2012), Poisson-based imaging (Harmany, Marcia, and Willett 2012), etc.

In this paper, we focus on the sparse covariance estimation problem. Particularly, let $\{\mathbf{x}_j\}_{j=1}^N$ be a collection of n-variate random vectors, i.e., $\mathbf{x}_j \in \mathbb{R}^n$, drawn from a joint probability distribution with covariance matrix $\mathbf{\Sigma}$. In this context, assume there may exist unknown marginal independences among the variables to discover; we note that $(\mathbf{\Sigma})_{kl} = 0$ when the k-th and l-th variables are independent. Here, we assume $\mathbf{\Sigma}$ is unknown and *sparse*, i.e., only a small number of entries are nonzero. Our goal is to recover the nonzero pattern of $\mathbf{\Sigma}$, as well as compute a good approximation, from a (possibly) limited sample corpus.

[*]This work is supported in part by the European Commission under Grant MIRG-268398, ERC Future Proof and SNF 200021-132548, SNF 200021-146750 and SNF CRSII2-147633.

Mathematically, one way to approximate $\mathbf{\Sigma}$ is by solving:

$$\mathbf{\Theta}^* = \arg\min_{\mathbf{\Theta}} \Big\{ -\log\det(\mathbf{\Theta}) + h(\mathbf{\Theta}) + g(\mathbf{\Theta}) \Big\}, \quad (1)$$

where $\mathbf{\Theta} \in \mathbb{R}^{n\times n}$ is the optimization variable, $h(\cdot) := 1/(2\rho)\cdot\|\mathbf{\Theta} - \widehat{\mathbf{\Sigma}}\|_{\text{F}}^2$ where $\widehat{\mathbf{\Sigma}}$ is the sample covariance and $g(\cdot) := \lambda/\rho \cdot \|\mathbf{\Theta}\|_1$ is a convex nonsmooth regularizer function, accompanied with an easily computable proximity operator (Combettes and Wajs 2005). and $\rho,\ \lambda > 0$.

Whereas there are several works (Becker and Fadili 2012; Lee, Sun, and Saunders 2012) that compute the minimizer of such composite objective functions, where the smooth term is generally a *Lipschitz* continuous gradient function, in (1) we consider a more tedious task: The objective function has only *locally Lipschitz* continuous gradient. However, one can easily observe that (1) is *self-concordant*; we refer to some notation and definitions in the Preliminaries section. Within this context, (Tran-Dinh, Kyrillidis, and Cevher 2013a) present a new convergence analysis and propose a series of proximal Newton schemes with provably quadratic convergence rate, under the assumption of *exact algorithmic calculations at each step of the method*.

Here, we extend the work of (Tran-Dinh, Kyrillidis, and Cevher 2013a) to include *inexact evaluations* and study how these errors propagate into the convergence rate. As a by-product, we apply these changes to propose the inexact **S**elf-**C**oncordant **OPT**imization (`iSCOPT`) framework. Finally, we consider the sparse covariance estimation problem as a running example for our discussions. The contributions are:

(i) We consider locally Lipschitz continuous gradient convex problems, similar to (1), where errors are introduced in the calculation of the descent direction step. Our analysis indicates that inexact strategies achieve similar convergence rates as the corresponding exact ones.

(ii) We present the inexact `SCOPT` solver (`iSCOPT`) for the sparse covariance estimation problem, with several variations that increase the convergence rate in practice.

Preliminaries

Notation: We reserve $\mathbf{vec}(\cdot)\colon \mathbb{R}^{n\times n} \rightarrow \mathbb{R}^{n^2\times 1}$ to denote the vectorization operator which maps a matrix into a vector, by stacking its columns and, let $\mathbf{mat}(\cdot)\colon \mathbb{R}^{n^2\times 1} \rightarrow \mathbb{R}^{n\times n}$ be the inverse operation. $\mathbf{I}$ denotes the identity matrix.

Definition 1 (Self-concordant functions (Nesterov and Nemirovskii 1994)). *A convex function* $\varphi(\cdot) : \mathrm{dom}(\varphi) \to \mathbb{R}$ *is self-concordant if* $|\varphi'''(x)| \leq 2\varphi''(x)^{3/2}, \forall x \in \mathrm{dom}(\varphi)$. *A function* $\psi(\cdot) : \mathrm{dom}(\psi) \to \mathbb{R}$ *is self-concordant if* $\varphi(t) \equiv \psi(\mathbf{x} + t\mathbf{v})$ *is self-concordant* $\forall \mathbf{x} \in \mathrm{dom}(\psi), \mathbf{v} \in \mathbb{R}^n$.

For $\mathbf{v} \in \mathbb{R}^n$, we define $\|\mathbf{v}\|_{\mathbf{x}} \equiv \left(\mathbf{v}^T \nabla^2 f(\mathbf{x}) \mathbf{v}\right)^{1/2}$ as the local norm around $\mathbf{x} \in \mathrm{dom}(f)$ with respect to $f(\cdot)$. The corresponding dual norm is $\|\mathbf{v}\|_{\mathbf{x}}^* \equiv \max_{\|\mathbf{u}\|_{\mathbf{x}} \leq 1} \mathbf{u}^T \mathbf{v} = \left(\mathbf{v}^T \nabla^2 f(\mathbf{x})^{-1} \mathbf{v}\right)^{1/2}$. We define $\omega(\cdot) : \mathbb{R} \to \mathbb{R}_+$ as $\omega(t) \equiv t - \ln(1+t)$, and $\omega_*(\cdot) : [0,1] \to \mathbb{R}_+$ as $\omega_*(t) \equiv -t - \ln(1-t)$. Note that $\omega(\cdot)$ and $\omega_*(\cdot)$ are both nonnegative, strictly convex, and increasing.

Problem reformulation: We can transform the matrix formulation of (1) in the following vectorized problem:

$$\min_{\mathbf{x} \in \mathbb{R}^p} \Big\{ F(\mathbf{x}) := \underbrace{-\log\det(\mathbf{mat}(\mathbf{x})) + h(\mathbf{x})}_{=f(\mathbf{x})} + g(\mathbf{x}) \Big\}, \quad (2)$$

for $\mathbf{mat}(\mathbf{x}) \equiv \mathbf{\Theta}$, $\mathbf{x} \in \mathbb{R}^p$, $p = n^2$, where $f(\mathbf{x})$ is a convex, self-condordant function and $g(\mathbf{x})$ is a proper, lower semi-continuous and non-smooth convex regularization term. For our discussions, we assume $g(\mathbf{x})$ is ℓ_1-norm-based.

The algorithm in a nutshell

For our convenience and without loss of generality, we use the vectorized reformulation in (2). Here, we describe the SCOPT optimization framework, proposed in (Tran-Dinh, Kyrillidis, and Cevher 2013a). SCOPT generates a sequence of putative solutions $\{\mathbf{x}_i\}_{i \geq 0}$, according to:

$$\mathbf{x}_{i+1} = (1 - \tau_i)\mathbf{x}_i + \tau_i \boldsymbol{\delta}_i^\star, \quad \tau_i = \frac{1}{\lambda_i + 1}, \quad (3)$$

where $\boldsymbol{\delta}_i^\star - \mathbf{x}_i \in \mathbb{R}^p$ is a descent direction, $\lambda_i := \|\boldsymbol{\delta}_i^\star - \mathbf{x}_i\|_{\mathbf{x}_i}$ and $\tau_i > 0$ is a step size along this direction. To compute $\boldsymbol{\delta}_i^\star$, we minimize the *non-smooth* convex surrogate of $F(\cdot)$ around $\mathbf{x}_i$; observe that λ_i assumes *exact* evaluations of $\boldsymbol{\delta}_i^\star$:

$$\boldsymbol{\delta}_i^\star = \arg\min_{\boldsymbol{\delta} \in \mathbb{R}^p} \{U(\boldsymbol{\delta}, \mathbf{x}_i) + g(\boldsymbol{\delta})\}; \quad (4)$$

$U(\boldsymbol{\delta}, \mathbf{x}_i)$ is a quadratic approximation of $f(\cdot)$ such that $U(\boldsymbol{\delta}, \mathbf{x}_i) := f(\mathbf{x}_i) + \nabla f(\mathbf{x}_i)^T(\boldsymbol{\delta} - \mathbf{x}_i) + \frac{1}{2}(\boldsymbol{\delta} - \mathbf{x}_i)^T \nabla^2 f(\mathbf{x}_i)(\boldsymbol{\delta} - \mathbf{x}_i)$, where $\nabla f(\mathbf{x}_i)$ and $\nabla^2 f(\mathbf{x}_i)$ denote the gradient (first-order) and Hessian (second-order) information of function $f(\cdot)$ around $\mathbf{x}_i \in \mathrm{dom}(F)$, respectively.

While quadratic approximations of smooth functions (of the form $U(\boldsymbol{\delta}, \mathbf{x}_i)$) have become *de facto* approaches for general convex *smooth* objective functions, to the best of our knowledge, there are not many works considering a composite *non-smooth and non-Lipschitz gradient* minimization case with provable convergence guarantees under the presence of errors in the descent direction evaluations.

Inexact solutions in (4)

An important ingredient for our scheme is the calculation of the descent direction through (4). For sparsity based applications, we use FISTA – a fast ℓ_1-norm regularized gradient method for solving (4) (Beck and Teboulle 2009) – and describe how to efficiently implement such solver for the case of sparse covariance estimation where $f(\mathbf{x}) = \frac{1}{2\rho}\|\mathbf{x} - \mathbf{vec}(\widehat{\mathbf{\Sigma}})\|_2^2 - \log\det(\mathbf{mat}(\mathbf{x}))$.

Given the current estimate $\mathbf{x}_i$, the gradient and the Hessian of $f(\cdot)$ around $\mathbf{x}_i$ can be computed respectively as: $\nabla f(\mathbf{x}_i) = \frac{1}{\rho}\left(\mathbf{x}_i - \mathbf{vec}(\widehat{\mathbf{\Sigma}})\right) - \mathbf{vec}\left(\mathbf{mat}(\mathbf{x}_i)^{-1}\right) \in \mathbb{R}^{p \times 1}$, $\nabla^2 f(\mathbf{x}_i) = \frac{\mathbf{I}}{\rho} + (\mathbf{mat}(\mathbf{x}_i)^{-1} \otimes \mathbf{mat}(\mathbf{x}_i)^{-1}) \in \mathbb{R}^{p \times p}$. Given the above, let $\mathbf{z} := \nabla f(\mathbf{x}_i) - \nabla^2 f(\mathbf{x}_i)\mathbf{x}_i$. After calculations on (4), we easily observe that (4) is equivalent to:

$$\boldsymbol{\delta}_i = \arg\min_{\boldsymbol{\delta}} \Big\{ \underbrace{\frac{1}{2}\boldsymbol{\delta}^T \nabla^2 f(\mathbf{x}_i)\boldsymbol{\delta} + \mathbf{z}^T\boldsymbol{\delta}}_{\varphi(\boldsymbol{\delta})} + g(\boldsymbol{\delta}) \Big\}, \quad (5)$$

where $\varphi(\cdot)$ is smooth and convex with Lipschitz constant L:

$$L = \frac{1}{\rho} + \frac{1}{\lambda_{\min}^2(\mathbf{mat}(\mathbf{x}_i))}, \quad (6)$$

where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of a matrix. Combining the above quantities in a ISTA-like procedure (Daubechies, Defrise, and De Mol 2004), we have:

$$\boldsymbol{\delta}^{k+1} = \mathcal{S}_{\frac{\lambda}{L\rho}}\left(\boldsymbol{\delta}^k - \frac{1}{L}\nabla\varphi(\boldsymbol{\delta}^k)\right), \quad (7)$$

where we use superscript k to denote the k-th iteration of the ISTA procedure (as opposed to the subscript i for the i-th iteration of (3)). Here, $\nabla\varphi(\boldsymbol{\delta}^k) = \nabla^2 f(\mathbf{x}_i)\boldsymbol{\delta}^k + \mathbf{z}$ and $\mathcal{S}_{\frac{\lambda}{L\rho}}(\mathbf{x}) := \mathrm{sign}(\mathbf{x})\max\{|\mathbf{x}| - \frac{\lambda}{L\rho}, 0\}$. Furthermore, to achieve an $\mathcal{O}(1/k^2)$ convergence rate, one can use acceleration techniques that lead to the FISTA algorithm, based on Nesterov's seminal work (Nesterov 1983). We repeat and extend FISTA's guarantees, as described in the next theorem; the proof is provided in the supplementary material.

Theorem 1. *Let* $\{\boldsymbol{\delta}^k\}_{k \geq 1}$ *be the sequence of estimates generated by FISTA. Moreover, define* $G(\boldsymbol{\delta}) := U(\boldsymbol{\delta}, \mathbf{x}_i) + g(\boldsymbol{\delta})$ *where* $\boldsymbol{\delta}^\star$ *is the minimizer with* $\|\boldsymbol{\delta}^\star\|_2^2 \leq c$ *for some global constant* $c > 0$. *Then, to achieve a solution* $\boldsymbol{\delta}^K$ *such that:*

$$G(\boldsymbol{\delta}^K) - G(\boldsymbol{\delta}^\star) \leq \epsilon, \ \epsilon > 0, \quad (8)$$

the FISTA algorithm requires at least $K := \left\lceil \sqrt{\frac{2Lc}{\epsilon}} - 1 \right\rceil$ *iterations. Moreover, it can be proved that:*

$$G(\boldsymbol{\delta}^K) - G(\boldsymbol{\delta}^\star) \geq \frac{1}{2}\|\boldsymbol{\delta}^\star - \boldsymbol{\delta}^K\|_{\mathbf{x}_i}^2$$

We note that, given accuracy ϵ, $\boldsymbol{\delta}^K$ satisfies (8) and $\boldsymbol{\delta}_i \leftarrow \boldsymbol{\delta}^K$ in the recursion (3). In general, c is not known apriori; in practice though, such a global constant can be found during execution, such that Theorem 1 is satisfied. A detailed description is given in the supplementary material.

For the sparse covariance problem, one can observe that L and $\mathbf{z}$ are precomputed once before applying FISTA iterations. Given $\mathbf{x}_i$, we compute $\lambda_{\min}(\mathbf{mat}(\mathbf{x}_i))$ in $\mathcal{O}(n^3)$ time complexity, while $\mathbf{z}$ can be computed with $\mathcal{O}(n^3)$ time cost using the Kronecker product property $\mathbf{vec}(\mathbf{AXB}) = (\mathbf{B}^T \otimes \mathbf{A})\mathbf{vec}(\mathbf{X})$. Similarly, $\nabla\varphi(\boldsymbol{\delta}^k)$ can be iteratively computed in $\mathcal{O}(n^3)$ time cost. Overall, the FISTA algorithm for this problem has $\mathcal{O}(K \cdot n^3)$ computational cost.

iSCOPT: Inexact SCOPT

Assembling the ingredients described above leads to Algorithm 1, which we call as the **I**next **S**elf-**C**oncordant **O**ptimization (iSCOPT) with the following convergence guarantees; our objective function satisfies the assumptions A.1, defined in (Tran-Dinh, Kyrillidis, and Cevher 2013a); the proof is provided in the supplementary material.

Theorem 2 (Global convergence guarantee). *Let* $\tau_i := \frac{\varepsilon_i - \sqrt{2\epsilon}}{\varepsilon_i(\varepsilon_i - \sqrt{2\epsilon}+1)} \in (0,1)$ *where* $\varepsilon_i := \|\boldsymbol{\delta}_i - \mathbf{x}_i\|_{\mathbf{x}_i}$ *is the* Newton decrement, $\boldsymbol{\delta}_i$ *is the solution of* (4) *and* ϵ *is the requested accuracy for solving* (4). *Assume* $\varepsilon_i \geq \sqrt{2\epsilon}$, $\forall i$, *and let the set* $\{\mathbf{x} \in \mathrm{dom}(F) : F(\mathbf{x}) \leq F(\mathbf{x}_0)\}$ *be bounded. Then,* iSCOPT *generates* $\{\mathbf{x}_i\}_{i\geq 0}$ *such that* $\mathbf{x}_{i+1}$ *satisfies:*

$$F(\mathbf{x}_{i+1}) \leq F(\mathbf{x}_i) - \xi(\tau_i), \quad \textit{where}$$

$$\xi(\tau_i) = -\omega_*(\tau_i\varepsilon_i) - \tau_i\left(\epsilon - \tfrac{1}{2}\left(\varepsilon_i - \sqrt{2\epsilon}\right)^2 - \tfrac{1}{2}\varepsilon_i^2\right) \geq 0,$$

$\forall i$, *i.e.,* $\{F(\mathbf{x}_i)\}_{i\geq 0}$ *is a* strictly non-increasing *sequence.*

Quadratic convergence rate of iSCOPT algorithm

For *strictly convex* criteria with unique solution $\mathbf{x}^\star$, the above proof guarantees convergence, i.e., $\{\mathbf{x}_i\}_{i\geq 0} \rightarrow \mathbf{x}^\star$ for sufficiently large i. Given this property, we prove the convergence rate towards the minimizer using *local information* in norm measures: as long as $\|\mathbf{x}_{i+1} - \mathbf{x}_i\|$ is away from 0, the algorithm has not yet converged to $\mathbf{x}^\star$. On the other hand, as $\|\mathbf{x}_{i+1} - \mathbf{x}_i\| \rightarrow 0$, the sequence $\{F(\mathbf{x}_i)\}_{i\geq 0}$ converges to its minimum and $\{\mathbf{x}_i\}_{i\geq 0} \rightarrow \mathbf{x}^\star$, as i increases.

Algorithm 1 Inexact SCOPT for sparse cov. estimation

Input: $\mathbf{x}_0, \rho, \lambda > 0, \sigma = \frac{3}{40}, \epsilon, \gamma > 0$.
while $\varepsilon_i \leq \gamma$ or $i \leq I^{\max}$ **do**
  Solve (4) for $\boldsymbol{\delta}_i$ with accuracy ϵ and parameters ρ, λ.
  Compute $\varepsilon_i = \|\boldsymbol{\delta}_i - \mathbf{x}_i\|_{\mathbf{x}_i}$
  if $(\varepsilon_i > \sigma)$
    $\mathbf{x}_{i+1} = (1-\tau_i)\mathbf{x}_i + \tau_i\boldsymbol{\delta}_i$ for $\tau_i = \frac{\varepsilon_i - \sqrt{2\epsilon}}{\varepsilon_i(\varepsilon_i - \sqrt{2\epsilon}+1)}$.
  else $\mathbf{x}_{i+1} = \boldsymbol{\delta}_i$
end while

In our analysis, we use the weighted distance $\|\mathbf{x}_{i+1} - \mathbf{x}_i\|_{\mathbf{x}_i}$ to characterize the *rate of convergence* of the putative solutions. By (3) and given $\boldsymbol{\delta}_i$ is a computable solution where $\|\boldsymbol{\delta}_i - \boldsymbol{\delta}_i^\star\|_{\mathbf{x}_i} \leq \sqrt{2\epsilon}$, we observe:

$$\|\mathbf{x}_{i+1} - \mathbf{x}_i\|_{\mathbf{x}_i} = \|\tau_i\left(\boldsymbol{\delta}_i - \mathbf{x}_i\right)\|_{\mathbf{x}_i} \propto \|\boldsymbol{\delta}_i - \mathbf{x}_i\|_{\mathbf{x}_i} := \varepsilon_i.$$

This setting is nearly algorithmic: given $\mathbf{x}_i$ and $\boldsymbol{\delta}_i$ at each iteration, we can observe the behavior of $\|\mathbf{x}_{i+1} - \mathbf{x}_i\|_{\mathbf{x}_i}$ through the *evolution* of $\{\varepsilon_i\}_{i\geq 0}$ and identify the region where this sequence decreases with a quadratic rate.

Definition 2. *We define the quadratic convergence region* $\mathcal{Q} = \{\mathbf{x}_i : \mathbf{x}_i \in dom(F)\}$ *as such where* $\{\mathbf{x}_i\}_{i\geq 0}$ *satisfies* $\varepsilon_{i+1} \leq \beta\varepsilon_i^2 + c$, *for some constant* $\beta > 0$, $\varepsilon_i < 1$ *and bounded and small constant* $c > 0$.

The following lemma provides a first step for a concrete characterization of β for the iSCOPT algorithm; the proof can be found in the supplementary material.

Lemma 1. *For any* τ_i *selection,* $\forall i$, *the* iSCOPT *algorithm generates the sequence* $\{\varepsilon_i\}_{i\geq 0}$ *such that* (9) *holds.*

We provide a series of corollaries and lemmata that justify the local quadratic convergence of our approach in theory.

Corollary 1. *In the ideal case where* $\boldsymbol{\delta}_i^\star$ *is computable exactly, i.e.,* $\epsilon = 0$, *the* iSCOPT *algorithm is identical to the* SCOPT *algorithm (Tran-Dinh, Kyrillidis, and Cevher 2013a).*

We apply the bound $\sqrt{2\epsilon} \leq \varepsilon_i$ to simplify (9) as:

$$\varepsilon_{i+1} \leq \frac{2}{1-2\tau_i\varepsilon_i} \cdot \frac{1-\tau_i+2\tau_i^2\varepsilon_i}{1-8\tau_i\varepsilon_i+8\tau_i^2\varepsilon_i^2} \cdot \varepsilon_i + \sqrt{2\epsilon} \quad (10)$$

Next, we describe the convergence rate of iSCOPT for the two distinct phases in our approach: full step size and damped step size; the proofs are provided in the supplementary material.

Theorem 3. *Assume* $\tau_i = 1$. *Then,* iSCOPT *satisfies:*

$$\varepsilon_{i+1} \leq \beta\varepsilon_i^2 + c,$$

where $\beta = \frac{4}{(1-2\varepsilon_i)(1-8\varepsilon_i+8\varepsilon_i^2)} = \mathcal{O}\left(\frac{1}{1-\varepsilon_i}\right)$, $c = \sqrt{2\epsilon}$ *and* ϵ *is user-defined. I.e.,* iSCOPT *has* locally quadratic convergence rate *where* $c > 0$ *is small-valued and bounded. Moreover, for* $\varepsilon_i \leq \frac{3}{40}$, $\forall i$, $\varepsilon_{i+1} \leq 14\varepsilon_i^2 + \sqrt{2\epsilon}$.

Theorem 4. *Assume the damped-step case where* $\tau_i = \frac{\varepsilon_i - \sqrt{2\epsilon}}{\varepsilon_i(\varepsilon_i - \sqrt{2\epsilon}+1)} \in (0,1)$. *Then,* iSCOPT *satisfies:*

$$\varepsilon_{i+1} \leq \beta\varepsilon_i^2 + c,$$

where $\beta = \frac{2\left(\varepsilon_i - \sqrt{2\epsilon}+1\right)}{1-2\varepsilon_i\left(\varepsilon_i - \sqrt{2\epsilon}\right)} \cdot \frac{\varepsilon_i^2 + 3\varepsilon_i + 2\epsilon}{\left(2+6\sqrt{2\epsilon}+2\epsilon\right) - \varepsilon_i\left(6+2\sqrt{2\epsilon}\right)} = \mathcal{O}\left(\frac{1}{1-\varepsilon_i}\right)$ *and* ϵ *is user-defined. I.e.,* iSCOPT *has* locally quadratic convergence rate *where* $c > 0$ *is small-valued and bounded. Moreover, for* $\varepsilon_i \leq \frac{3}{20}$, $\forall i$, $\varepsilon_{i+1} \leq 14\varepsilon_i^2 + \sqrt{2\epsilon}$.

An iSCOPT variant

Starting from a point far away from the true solution, Newton-like methods might not show the expected convergence behavior. To tackle this issue, we can further perform Forward Line Search (FLS) (Tran-Dinh, Kyrillidis, and Cevher 2013a): starting from the current estimate τ_i, one might perform a *forward* binary search in the range $[\tau_i, 1]$. The selection of the new step size $\widehat{\tau}_i$ is taken as the maximum-valued step size in $[\tau_i, 1]$, as long as $\widehat{\tau}_i$ decreases the objective function $F(\cdot)$, while satisfying any constraints in the optimization. The supplementary material contains illustrative examples which we omit due to lack of space.

Application to sparse covariance estimation

Covariance estimation is an important problem, found in diverse research areas. In classic portfolio optimization (Markowitz 1952), the covariance matrix over the asset returns is unknown and even the estimation of the most significant dependencies among assets might lead to meaningful decisions for portfolio optimization. Other applications of the sparse covariance estimation include inference in gene

$$\varepsilon_{i+1} \leq \frac{\left(1-\tau_i\varepsilon_i+\tau_i\sqrt{2\epsilon}\right)}{\left(1-\tau_i\varepsilon_i-\tau_i\sqrt{2\epsilon}\right)} \cdot \frac{\left(1-\tau_i\left(1-\sqrt{2\epsilon}\right)+\left(2\tau_i^2-\tau_i\right)\varepsilon_i\right)}{1-4\tau_i\left(\sqrt{2\epsilon}+\varepsilon_i\right)+2\tau_i^2\left(\sqrt{2\epsilon}+\varepsilon_i\right)^2}\left(\sqrt{2\epsilon}+\varepsilon_i\right)+\sqrt{2\epsilon} \tag{9}$$

Table 1: Summary of related work on sparse covariance estimation. Here, [1]: (Xue, Ma, and Zou 2012), [2]: (Bien and Tibshirani 2011), [3]: (Rothman 2012), [4]: (Wang 2012). All methods have the same $\mathcal{O}(n^3)$ time-complexity per iteration.

	[1]	[2]	[3]	[4]	This work
# of tuning parameters	2	1	2	1	2
Convergence guarantee	✓	✓	✓	–	✓
Convergence rate	Linear	–	–[†]	–[†]	Quadratic
Covariate distribution	Any	Gaussian	Any	Gaussian	Any

[†]To the best of our knowledge, block coordinate descent algorithms have known convergence *only* for the case of Lipschitz continuous gradient objective functions (Beck and Tetruashvili 2013).

dependency networks (Schäfer and Strimmer 2005), fMRI imaging (Varoquaux et al. 2010), data mining (Alqallaf et al. 2002), etc. Overall, sparse covariance matrices come with nice properties such as natural graphical interpretation, whereas are easy to be transfered and stored.

To this end, we consider the following problem: PROBLEM I: *Given n-dimensional samples $\{\mathbf{x}_j\}_{j=1}^N$, drawn from a joint probability density function with* unknown *sparse covariance $\mathbf{\Sigma} \succ 0$, we approximate $\mathbf{\Sigma}$ as the solution to the following optimization problem for some $\lambda, \rho > 0$:*

$$\mathbf{\Theta}^\star = \arg\min_{\mathbf{\Theta}} \left\{ \frac{1}{2\rho}\|\mathbf{\Theta}-\widehat{\mathbf{\Sigma}}\|_F^2 - \log\det(\mathbf{\Theta}) + \frac{\lambda}{\rho}\|\mathbf{\Theta}\|_1 \right\}$$

A summary of the related work on the sparse covariance problem is given in Table 1 and a more detailed discussion is provided in the supplementary material.

Experiments

All approaches are carefully implemented in MATLAB code with no C-coded parts. In all cases, we set $I^{\max} = 500$, $\gamma = 10^{-10}$ and $\epsilon = 10^{-8}$. A more extensive presentation of these results can be found in the supplementary material.

Benchmarking `iSCOPT`: time efficiency

To the best of our knowledge, only (Rothman 2012) considers the same objective function as in PROBLEM I. There, the proposed algorithm follows similar motions with the graphical Lasso method (Friedman, Hastie, and Tibshirani 2008).

To show the merits of our approach as compared with the state-of-the-art in (Rothman 2012), we generate $\mathbf{\Sigma} \equiv \mathbf{\Sigma}_3$ as a random positive definite covariance matrix with $\|\mathbf{\Sigma}_3\|_0 = k$. In our experiments, we test sparsity levels k such that $\frac{k}{n^2} = \{0.05, 0.1, 0.2\}$ and $n \in \{100, 1000, 2000\}$. Without loss of generality, we assume that the variables are drawn from a joint Gaussian probability distribution. Given $\mathbf{\Sigma}$, we generate $\{\mathbf{x}_j\}_{j=1}^N$ random n-variate vectors according to $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $N = \frac{n}{2}$. Then, the sample covariance matrix $\widehat{\mathbf{\Sigma}} = \frac{1}{N}\sum_{j=1}^N \mathbf{x}_j\mathbf{x}_j^T$ is ill-conditioned in all cases with $\text{rank}(\widehat{\mathbf{\Sigma}}) \leq \frac{n}{2}$. We observe that the number of unknowns is $\binom{n}{2} = \frac{n(n-1)}{2}$; in our testbed, this corresponds to estimation of 4950 up to $1,999,000$ variables. To compute L in (6), we use a power method scheme with $\texttt{Pw} = 20$ iterations. All algorithms under comparison are initialized with $\mathbf{x}_0 = \mathbf{vec}(\text{diag}(\widehat{\mathbf{\Sigma}}))$. As an execution wall time, we set $\texttt{T} = 3600$ seconds (1 hour). In all cases, we set $\rho = 0.1$.

Table 2 contains the summary of results. Overall, the proposed framework shows superior performance across diverse configuration settings, both in terms of time complexity and objective function minimization efficiency: both `iSCOPT` and `iSCOPT` FLS find solutions with lower objective function value, as compared to (Rothman 2012), within the same time frame. The regular `iSCOPT` algorithm performs relatively well in terms of computational time as compared to the rest of the methods. However, its convergence rate heavily depends on the *conservative* τ_i selection. We note that (4) benefits from warm-start strategies that result in convergence in Step 3 of Algorithm 1 within a few steps.

Benchmarking `iSCOPT`: reconstruction efficiency

We also measure the $\mathbf{\Sigma}$ reconstruction efficacy by solving PROBLEM I, as compared to other optimization formulations for sparse covariance estimation. We compare our $\mathbf{\Theta}^\star$ estimate with: (i) the Alternating Direction Method of Multipliers (ADMM) implementation (Xue, Ma, and Zou 2012), and (ii) the coordinate descent algorithm (Wang 2012).

Table 3 aggregates the experimental results in terms of the normalized distance $\frac{\|\mathbf{\Theta}^\star - \mathbf{\Sigma}\|_F}{\|\mathbf{\Sigma}\|_F}$ and the captured sparsity pattern in $\mathbf{\Sigma}$. Without loss of generality, we fix $\lambda = 0.5, \rho = 0.1$ for the case $n = 100$ and, $\lambda = 1.5, \rho = 0.1$ for the case $n = 2000$. `iSCOPT` framework is at least as competitive with the state-of-the-art implementations for sparse covariance estimation. It is evident that the proposed `iSCOPT` variant, based on self-concordant analysis, is at least one order of magnitude faster than the rest of algorithms under comparison. In terms of reconstruction efficacy, using our proposed scheme, we can achieve marginally better $\mathbf{\Sigma}$ reconstruction performance, as compared to (Xue, Ma, and

Table 2: Summary of comparison results for time efficiency.

Model				$F(\mathbf{\Theta}^\star)$ $(\times 10^2)$			Time (secs)		
	n		λ	[3]	iSCOPT	iSCOPT FLS	[3]	iSCOPT	iSCOPT FLS
$\mathbf{\Sigma}_3$	100	$\frac{k}{n^2} = 0.05$	1	32.013	**31.919**	**31.919**	8.288	9.996	**3.584**
		$\frac{k}{n^2} = 0.1$	0.5	36.190	**34.689**	**34.689**	10.470	12.761	**5.012**
		$\frac{k}{n^2} = 0.2$	0.5	62.143	**53.081**	**53.081**	18.446	14.720	**6.257**
	1000	$\frac{k}{n^2} = 0.05$	1	–	–	**2711.931**	$>$ T	$>$ T	**759.724**
		$\frac{k}{n^2} = 0.1$	1	–	–	**4734.251**	$>$ T	$>$ T	**875.344**
		$\frac{k}{n^2} = 0.2$	1	–	–	**5553.508**	$>$ T	$>$ T	**1059.709**

Table 3: Summary of comparison results for reconstruction of efficiency.

Model			$\|\mathbf{\Theta}^\star - \mathbf{\Sigma}\|_F / \|\mathbf{\Sigma}\|_F$			Time		
	n	N	[4]	[1]	iSCOPT FLS	[4]	[1]	iSCOPT FLS
$\mathbf{\Sigma}_3$	100	$n/2$	1.180	0.912	**0.908**	0.456	**0.252**	2.604
		n	0.920	0.554	**0.542**	0.494	**0.108**	0.155
		$10n$	0.396	0.192	**0.190**	0.451	0.108	**0.054**
	2000	$n/2$	–	**0.428**	**0.428**	$>$ T	350.145	**203.515**
		n	–	**0.352**	**0.352**	$>$ T	385.340	**167.688**
		$10n$	–	0.211	**0.209**	$>$ T	401.970	**122.535**

Zou 2012).

Sparse covariance estimates for portfolio optimization

Classical mean-variance optimization (MVO) (Markowitz 1952) corresponds to the following optimization problem:

$$\begin{aligned} &\underset{\mathbf{w}}{\text{minimize}} \quad \mathbf{w}^T \mathbf{\Sigma} \mathbf{w} \\ &\text{subject to} \quad \mathbf{w}^T \mathbf{r} = \mu, \ \ \sum_i w_i = C, \ \ w_i \geq 0, \ \forall i. \end{aligned} \tag{11}$$

Here, $\mathbf{\Sigma} \in \mathbb{S}_+^n$ is the *true* covariance matrix over a set of asset returns, $\mathbf{r} \in \mathbb{R}^n$ denotes the *true* asset returns of n stocks, $\mathbf{w}$ represents a weighted probability distribution over the set of assets such that $\sum_i w_i = C$ and C is the total capital to be invested. Without loss of generality, one can assume a normalized capital such that $\sum_i w_i = 1$. In such case, $\mathbf{w}^T \mathbf{\Sigma} \mathbf{w}$ is both the risk of the investment as well as a metric of *variance* of the portfolio selection.

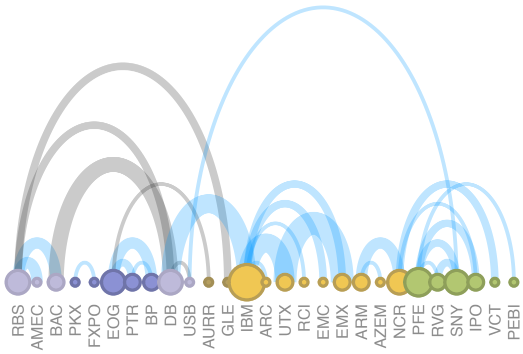
In practice, both $\mathbf{r}$ and $\mathbf{\Sigma}$ are unknown and MVO requires an estimation for both. Empirical estimates, such as $\widehat{\mathbf{\Sigma}}$, quickly become problematic in the large scale: the data amount required increases quadratically to be commensurate with the degree of dimensionality. Due to such difficulties, even a simple *equal weighted portfolio* $\mathbf{w}$ such that $w_i = 1/n, \forall i$, is often preferred in practice (DeMiguel, Garlappi, and Uppal 2009). Nevertheless, practitioners assume that many elements of the covariance matrix are zero, a property which is appealing due to its interpretability and ease of estimation. Moreover, there are cases in practice where most of the variables are correlated to only a few others.

Figure 1 shows some representative correlation estimates that we observed during the period 01.09.2009 and 31.08.2013. For this purpose, we use iSCOPT with $\lambda = 0.1$ and $\rho = 1$ to solve PROBLEM I and sort the non-diagonal elements of $\mathbf{\Theta}^\star$ to keep the most important correlations. We observe in practice some strong correlations between assets, while most of the rest entries in $\mathbf{\Theta}^\star$ have significantly small magnitude. This dataset contains 2833 stocks over a trading period of 1038 days, crawled from the Yahoo Finance website[1]. Stocks are retrieved from stock markets in the America (e.g., Dow Jones, NYSE, etc.), Europe (e.g., London Stock Exchange, etc.), Asia (e.g., Nikkei, etc) and Africa (e.g., South Africa's exchange).

Out-of-sample performance with synthetic data: It is apparent that both strong and weak correlations among stock assets are evident in practice. The behavior of non-diagonal entries in correlation matrix estimates is such that it is not easily distinguishable whether small values indicate weak dependence between variables or estimation fluctuations. Under these settings, (Hero and Rajaratnam 2011) propose that small values should be considered as zeros while only large values can be considered as good covariate estimates.

To measure the performance of using a sparse covariance estimate in MVO, we assume the following synthetic case: Let $\mathbf{\Sigma} \succ 0$ be a synthetically generated Gaussian covariance matrix to represent the correlations among assets. Further-

[1] http://finance.yahoo.com



Stock Abbr.	Company name	Stock Abbr.	Company name
RBS	Scotland Bank	ARC	Arc Document
AMEC	AMEC Group	UTX	United Tech.
BAC	Bank of America	RCI	Rogers Comm.
PKX	Posco	EMX	EMX Industries
FXPO	Ferrexpo	ARM	ARM Holdings
EOG	EOG Resources	AZEM	Azem Chemicals
PTR	PetroChina	NCR	NCR Electronics
BP	BP	PFE	Pfizer Inc.
DB	Deutsche Bank	RVG	Retro Virology
USB	U.S. Bank Corp.	SNY	Sanofi health
AURR	Aurora Russia	IPO	Intellectual Property
GLE	Glencore	VCT	Victrex Chemicals
IBM	IBM	PEBI	Port Erin BioFarma

Figure 1: We focus on three sectors: (i) bank industry (light purple), (ii) petroleum industry (dark purple), (iii) Computer science and microelectronics industry (light yellow), (iv) Pharmaceuticals/Chemistry industry (green). Any miscellaneous companies are denoted with dart yellow. Positive correlations are denoted with blue arcs; negative correlations with black arcs. The width of the arcs denotes the strength of the correlation - here, the maximum correlation (in magnitude) is 0.3934.

Table 4: Summary of portfolio optimization results– all strategies considered achieve the requested return μ.

Model			Risk $\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}$			Model			Risk $\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}$		
	λ	$\frac{k}{n^2}$ (%)	$\mathbf{w}(\widehat{\mathbf{\Sigma}})$	$\mathbf{w}_{\text{equal}}$	$\mathbf{w}(\mathbf{\Theta}^\star)$		λ	$\frac{k}{n^2}$ (%)	$\mathbf{w}(\widehat{\mathbf{\Sigma}})$	$\mathbf{w}_{\text{equal}}$	$\mathbf{w}(\mathbf{\Theta}^\star)$
$\mathbf{\Sigma}_3$ ($n=$ 1000, $N=90$)	1.4	0.5	0.0760	0.0065	**0.0053**	$\mathbf{\Sigma}_3$ ($n=$ 1000, $N=180$)	1.4	0.5	0.0223	0.0066	**0.0050**
	1.7	1	0.0810	0.0078	**0.0059**		1.7	1	0.0233	0.0076	**0.0072**
	2.3	5	0.0902	0.0158	**0.0129**		2.3	5	0.0513	0.0157	**0.0115**
	2.7	7	0.1968	0.0188	**0.0159**		2.7	7	0.0529	0.0183	**0.0139**
	3.0	10	0.2232	0.0223	**0.0196**		3.0	10	0.0706	0.0217	**0.0177**
	3.8	15	0.2463	0.0267	**0.0231**		3.8	15	0.0876	0.0264	**0.0202**
	4.5	20	0.2408	0.0307	**0.0257**		4.5	20	0.0872	0.0307	**0.0227**
	4.5	30	0.4925	0.0375	**0.0365**		4.5	30	0.1075	0.0373	**0.0291**

more, assume that only k entries of $\mathbf{\Sigma}$ are sufficiently larger than the rest of the entries. In our experiments below we set $n = 1000$ and consider a time window of $N = 90,\ 180$ days (i.e., a 3- and 6-month sampling period).

Given the above, both $\widehat{\mathbf{\Sigma}}$ and $\mathbf{\Theta}^\star$ are calculated – we use our algorithm for the latter. Using these two quantities, we then solve (11) for $\mathbf{\Sigma} \leftarrow \widehat{\mathbf{\Sigma}}$ and $\mathbf{\Sigma} \leftarrow \mathbf{\Theta}^\star$ for various expected returns μ and record the computed minimum risk portfolios $\mathbf{w}(\widehat{\mathbf{\Sigma}})$ and $\mathbf{w}(\mathbf{\Theta}^\star)$, respectively. Finally, given $\mathbf{w}(\widehat{\mathbf{\Sigma}})$ and $\mathbf{w}(\mathbf{\Theta}^\star)$, as well as the equal-weight portfolio $\mathbf{w}_{\text{equal}} := \frac{1}{n} \cdot \mathbb{1}_{n\times 1}$, we report the risk/variances achieved by the constructed portfolios, using the ground truth covariance $\mathbf{\Sigma}$. In Table 4, we report lower variances $\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}$ for portfolios $\mathbf{w}$ trained when $\mathbf{\Theta}^\star$ is used in (11), compared with the risk achieved by the equally-weighted portfolio or the sample covariance estimation where $\widehat{\mathbf{\Sigma}}$ is used. However, our approach comes with some cost to compute $\mathbf{\Theta}^\star$. The empirical strategy with $\mathbf{w}(\widehat{\mathbf{\Sigma}})$ has the worst performance in terms of minimum risk achieved for most of our testings; we point out that, in this case, $\widehat{\mathbf{\Sigma}}$ is a rank-deficient positive semidefinite matrix.

Discussion

A drawback of our approach is the combined setup of the parameters λ and ρ: one needs to identify selections that perform well on-the-fly, via a trial-and-error strategy. Unfortunately, such process might be inefficient in practice, especially in high dimensional cases. An interesting question to pursue is the *adaptive* setup of at least one of $\lambda,\ \rho$. Such adaptive strategies have attracted a great deal of interest ; c.f., (Hale, Yin, and Zhang 2008). One idea is to devise a path-following scheme with an adaptive ρ selection, where the resulting scheme solves approximately a series of problems as $\rho \to 0$ is adaptively updated (Tran-Dinh, Kyrillidis, and Cevher 2013b). We hope this paper triggers future efforts towards this research direction for further investigation.

References

Alqallaf, F. A.; Konis, K. P.; Martin, R. D.; and Zamar, R. H. 2002. Scalable robust covariance and correlation estimates for data mining. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 14–23. ACM.

Banerjee, O.; El Ghaoui, L.; and d'Aspremont, A. 2008. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research* 9:485–516.

Beck, A., and Teboulle, M. 2009. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *Image Processing, IEEE Transactions on* 18(11):2419–2434.

Beck, A., and Tetruashvili, L. 2013. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization* 23(4):2037–2060.

Becker, S., and Fadili, J. 2012. A quasi-newton proximal splitting method. In *Advances in Neural Information Processing Systems (NIPS)*, 2627–2635.

Bien, J., and Tibshirani, R. J. 2011. Sparse estimation of a covariance matrix. *Biometrika* 98(4):807–820.

Combettes, P. L., and Wajs, V. R. 2005. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation* 4(4):1168–1200.

Dahl, J.; Vandenberghe, L.; and Roychowdhury, V. 2008. Covariance selection for nonchordal graphs via chordal embedding. *Optimization Methods & Software* 23(4):501–520.

Daubechies, I.; Defrise, M.; and De Mol, C. 2004. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on pure and applied mathematics* 57(11):1413–1457.

DeMiguel, V.; Garlappi, L.; and Uppal, R. 2009. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *Review of Financial Studies* 22(5):1915–1953.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9(3):432–441.

Hale, E. T.; Yin, W.; and Zhang, Y. 2008. Fixed-point continuation for \ell_1-minimization: Methodology and convergence. *SIAM Journal on Optimization* 19(3):1107–1130.

Harmany, Z. T.; Marcia, R. F.; and Willett, R. M. 2012. This is spiral-tap: sparse poisson intensity reconstruction algorithmstheory and practice. *Image Processing, IEEE Transactions on* 21(3):1084–1096.

Hero, A., and Rajaratnam, B. 2011. Large-scale correlation screening. *Journal of the American Statistical Association* 106(496):1540–1552.

Hsieh, C.; Sustik, M.; Dhillon, I.; and Ravikumar, P. 2011. Sparse inverse covariance matrix estimation using quadratic approximation. *Advances in Neural Information Processing Systems (NIPS)* 24.

Lee, J.; Sun, Y.; and Saunders, M. 2012. Proximal newton-type methods for convex optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 836–844.

Markowitz, H. 1952. Portfolio selection. *The journal of finance* 7(1):77–91. Wiley Library.

Nesterov, Y., and Nemirovskii, A. S. 1994. *Interior-point polynomial algorithms in convex programming*, volume 13. SIAM.

Nesterov, Y. 1983. A method of solving a convex programming problem with convergence rate o (1/k2). In *Soviet Mathematics Doklady*, volume 27, 372–376.

Rothman, A. J. 2012. Positive definite estimators of large covariance matrices. *Biometrika* 99(3):733–740.

Schäfer, J., and Strimmer, K. 2005. An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics* 21(6):754–764.

Tran-Dinh, Q.; Kyrillidis, A.; and Cevher, V. 2013a. Composite self-concordant minimization. *arXiv preprint arXiv:1308.2867*.

Tran-Dinh, Q.; Kyrillidis, A.; and Cevher, V. 2013b. An inexact proximal path-following algorithm for constrained convex minimization. *arXiv preprint arXiv:1311.1756*.

Varoquaux, G.; Gramfort, A.; Poline, J.-B.; and Thirion, B. 2010. Brain covariance selection: better individual functional connectivity models using population prior. In *Advances in Neural Information Processing Systems (NIPS)*, volume 23. 2334–2342.

Wang, H. 2012. Two new algorithms for solving covariance graphical lasso based on coordinate descent and ecm. *arXiv preprint arXiv:1205.4120*.

Xue, L.; Ma, S.; and Zou, H. 2012. Positive definite ℓ_1 penalized estimation of large covariance matrices. *Journal of the American Statistical Association*.