UC Santa Barbara

Title

A Convex Pseudolikelihood Framework for High Dimensional Partial Correlation Estimation with Convergence Guarantees

Permalink

Journal

ISSN

Authors

Khare, Kshitij
Oh, Sang-Yun
Rajaratnam, Bala

Publication Date

DOI

Peer reviewed

A convex pseudo-likelihood framework for high dimensional partial correlation estimation with convergence guarantees

Kshitij Khare, *University of Florida, USA*
Sang-Yun Oh, *Stanford University, USA*
Bala Rajaratnam, *Stanford University, USA*

Abstract

Sparse high dimensional graphical model selection is a topic of much interest in modern day statistics. A popular approach is to apply ℓ_1-penalties to either (1) parametric likelihoods, or, (2) regularized regression/pseudo-likelihoods, with the latter having the distinct advantage that they do not explicitly assume Gaussianity. As none of the popular methods proposed for solving pseudo-likelihood based objective functions have provable convergence guarantees, it is not clear if corresponding estimators exist or are even computable, or if they actually yield correct partial correlation graphs. This paper proposes a new pseudo-likelihood based graphical model selection method that aims to overcome some of the shortcomings of current methods, but at the same time retain all their respective strengths. In particular, we introduce a novel framework that leads to a convex formulation of the partial covariance regression graph problem, resulting in an objective function comprised of quadratic forms. The objective is then optimized via a coordinatewise approach. The specific functional form of the objective function facilitates rigorous convergence analysis leading to convergence guarantees; an important property that cannot be established using standard results, when the dimension is larger than the sample size, as is often the case in high dimensional applications. These convergence guarantees ensure that estimators are well-defined under very general conditions, and are always computable. In addition, the approach yields estimators that have good large sample properties and also respect symmetry. Furthermore, application to simulated/real data, timing comparisons and numerical convergence is demonstrated. We also present a novel unifying framework that places all graphical pseudo-likelihood methods as special cases of a more general formulation, leading to important insights.

Keywords: Sparse inverse covariance estimation, Graphical model selection, Soft thresholding, Partial correlation graph, Convergence guarantee, Generalized pseudo-likelihood, Gene regulatory network

1 Introduction

One of the hallmarks of modern day statistics is the advent of high-dimensional datasets arising particularly from applications in the biological sciences, environmental sciences and finance. A central quantity of interest in such applications is the covariance matrix Σ of high dimensional random vectors. It is well known that the sample covariance matrix $\mathbf{S}$ can be a poor estimator of Σ, especially when p/n is large, where n is the sample size and p is the number of variables in the dataset. Hence $\mathbf{S}$ is not a useful estimator for Σ for high-dimensional datasets, where often either $p \gg n$ ("large p, small n") or when p is comparable to n and both are large ("large p, large n"). The basic problem here is that the number of parameters in Σ is of the order p^2. Hence in the settings mentioned above, the sample size is often not large enough to obtain a good estimator.

For many real life applications, the quantity of interest is the inverse covariance/partial covariance matrix $\Omega = \Sigma^{-1}$. In such situations, it is often reasonable to assume that there are only a few significant partial correlations and the other partial correlations are negligible in comparison. In mathematical terms, this amounts to making the assumption that the inverse covariance matrix $\Omega = \Sigma^{-1} = ((\omega_{ij}))_{1 \leq i,j \leq p}$ is sparse, i.e., many entries in Ω are zero. Note that $\omega_{ij} = 0$ is equivalent to saying that the partial correlation between the ith and j^{th} variables is zero (under Gaussianity, this reduces to the statement that the ith and j^{th} variables are conditionally independent given the other variables). The zeros in Ω can be conveniently represented by partial correlation graphs. The assumption of a sparse graph is often deemed very reasonable in applications. For example, as Peng et al. (2009) point out, among 26 examples of published networks compiled by Newman (2003), 24 networks had edge density less than 4%.

A number of methods have been proposed for identifying sparse partial correlation graphs in the penalized likelihood and penalized regression based framework (Meinshausen and Bühlmann, 2006, Friedman et al., 2008, Peng et al., 2009, Friedman et al., 2010). The main focus here is estimation of the sparsity pattern. Many of these methods do not necessarily yield positive definite estimates of Ω. However, once a sparsity pattern is established, a positive definite estimate can be easily obtained using efficient methods (see Hastie et al. (2009), Speed and Kiiveri (1986)).

The penalized likelihood approach induces sparsity by minimizing the (negative) log-likelihood function with an ℓ_1 penalty on the elements of Ω. In the Gaussian setup, this approach was pursued by Banerjee et al. (2008) and others. Friedman et al. (2008) proposed the *graphical lasso* ("Glasso") algorithm for the above minimization problem, and is substantially faster than earlier methods. In recent years, many interesting and useful methods

have been proposed for speeding up the performance of the graphical lasso algorithm (see Mazumder and Hastie (2012) for instance). It is worth noting that for these methods to provide substantial improvements over the graphical lasso, certain assumptions are required on the number and size of the connected components of the graph implied by the zeros in $\hat{\Omega}$ (the minimizer).

Another useful approach introduced by Meinshausen and Bühlmann (2006) estimates the zeros in Ω by fitting separate lasso regressions for each variable given the other variables. These individual lasso fits give neighborhoods that link each variable to others. Peng et al. (2009) improve this neighborhood selection (NS) method by taking the natural symmetry in the problem into account (i.e., $\Omega_{ij} = \Omega_{ji}$), as not doing so could result in less efficiency and contradictory neighborhoods.

In particular, the SPACE (Sparse PArtial Correlation Estimation) method was proposed by Peng et al. (2009) as an effective alternative to existing methods for sparse estimation of Ω. The SPACE procedure iterates between (1) updating partial correlations by a joint lasso regression and (2) separately updating the partial variances. As indicated above, it also accounts for the symmetry in Ω and is computationally efficient. Peng et al. (2009) show that under suitable regularity conditions, SPACE yields consistent estimators in high dimensional settings. All the above properties make SPACE an attractive regression based approach for estimating sparse partial correlation graphs. In the examples presented in Peng et al. (2009), the authors find that empirically the SPACE algorithm seems to converge really fast. It is however not clear if SPACE will converge in general. Convergence is of course critical so that the corresponding estimator is always guaranteed to exist and is therefore meaningful, both computationally and statistically. In fact, as we illustrate in Section 2, the SPACE algorithm might fail to converge in simple cases, for both the standard choices of weights suggested in Peng et al. (2009). Motivated by SPACE, Friedman et al. (2010) present a coordinate-wise descent approach (the "Symmetric lasso"), which may be considered as a symmetrized version of the approach in Meinshausen and Bühlmann (2006). As we show in Section 2.3, it is also not clear if the Symmetric lasso will converge.

In this paper, we present a new method called the CONvex CORrelation selection methoD (CONCORD) algorithm for sparse estimation of Ω. The algorithm obtains estimates of Ω by minimizing an objective function, which is jointly convex, but more importantly comprised of quadratic forms in the entries of Ω. The subsequent minimization is performed via coordinate-wise descent. The convexity is strict if $n \geq p$, in which case standard results guarantee the convergence of the coordinate-wise descent algorithm to the unique global minimum. If $n < p$, the objective function may not be strictly convex. As a result, a unique global minimum may not exist, and existing theory does not guarantee convergence of the

Property	METHOD				
	NS	SPACE	SYMLASSO	SPLICE	CONCORD
Symmetry		+	+	+	+
Convergence guarantee (fixed n)	N/A				+
Asymptotic consistency $(n, p \to \infty)$	+	+			+

Table 1: Comparison of regression based graphical model selection methods. A "+" indicates that a specified method has the given property. A blank space indicates the absence of a property. "N/A" stands for not applicable.

sequence of iterates of the coordinate-wise descent algorithm to a global minimum. In Section 4, by exploiting the quadratic forms present in the objective, it is rigorously demonstrated that the sequence of iterates does indeed converge to a global minimum of the objective function regardless of the dimension of the problem. Furthermore, it is shown in Section 6 that the CONCORD estimators are asymptotically consistent in high dimensional settings under regularity assumptions identical to Peng et al. (2009). Hence, our method preserves all the attractive properties of SPACE, while also providing a theoretical guarantee of convergence to a global minimum. In the process CONCORD yields an estimator $\hat{\Omega}$ that is well-defined and is always computable. The strengths of CONCORD are further illustrated in the simulations and real data analysis presented in Section 5. A comparison of the relevant properties of different estimators proposed in the literature is provided in Table 1 (Neighborhood selection (NS) by Meinshausen and Bühlmann (2006), SPACE by Peng et al. (2009), Symmetric lasso (SYMLASSO) by Friedman et al. (2010), SPLICE by Rocha et al. (2008) and CONCORD). The table shows that the CONCORD algorithm preserves all the attractive properties of existing algorithms, while also providing rigorous convergence guarantees. Another major contribution of the paper is the development of a unifying framework that renders the different pseudo-likelihood based graphical model selection procedures as special cases. This general formulation facilitates a direct comparison between the above pseudo-likelihood based methods and gives deep insights into their respective strengths and weaknesses.

The remainder of the paper is organized as follows. Section 2 briefly describes the SPACE algorithm and presents examples where it fails to converge. This section motivates our work and also analyzes other regression-based or pseudo-likelihood methods that have been proposed. Section 3 introduces the CONCORD method and presents a general framework that unifies recently proposed pseudo-likelihood methods. Section 4 establishes convergence

of CONCORD to a global minimum, even if $n < p$. Section 5 illustrates the performance of the CONCORD procedure on simulated and real data. Comparisons to SPACE and Glasso are provided. When applied to gene expression data, the results given by CONCORD are validated in a significant way by a recent extensive breast cancer study. Section 6 establishes large sample properties of the CONCORD approach. Concluding remarks are given in Section 7. The supplemental document contains proofs of some of the results in the paper.

2   The SPACE algorithm and convergence properties

Let the random vector $\mathbf{Y}^k = \left(y_1^k, y_2^k, \cdots, y_p^k\right)'$, $k = 1, 2, \cdots, n$ denote *i.i.d.* observations from a multivariate distribution with mean vector $\mathbf{0}$ and covariance matrix Σ. Let $\Omega = \Sigma^{-1} = ((\omega_{ij}))_{1 \leq i,j \leq p}$ denote the inverse covariance matrix, and let $\boldsymbol{\rho} = (\rho^{ij})_{1 \leq i < j \leq p}$ where $\rho^{ij} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}$ denotes the partial correlation between the ith and j^{th} variable for $1 \leq i \neq j \leq p$. Note that $\rho^{ij} = \rho^{ji}$ for $i \neq j$. Denote the sample covariance matrix by $\mathbf{S}$, and the sample corresponding to the ith variable by $\mathbf{Y}_i = (y_i^1, y_i^2, \cdots, y_i^n)'$.

2.1   The SPACE algorithm

Peng et al. (2009) propose the following novel iterative algorithm to estimate the partial correlations $\{\rho^{ij}\}_{1 \leq i < j \leq p}$ and the partial covariances $\{\omega_{ii}\}_{1 \leq i \leq p}$ corresponding to Ω (see Algorithm 1).

2.2   Convergence Properties of SPACE

From empirical studies, Peng et al. (2009) find that the SPACE algorithm converges quickly. As mentioned in the introduction, it is not immediately clear if convergence can be established theoretically. In an effort to understand such properties, we now place the SPACE algorithm in a useful optimization framework.

Lemma 1. *For the choice of weights, $w_i = \omega_{ii}$, the SPACE algorithm corresponds to an iterative partial minimization procedure (IPM) for the following objective function:*

$$
\begin{aligned}
Q_{spc}(\Omega) &= \frac{1}{2} \sum_{i=1}^p \left(-n \log \omega_{ii} + \omega_{ii} \| \mathbf{Y}_i - \sum_{j \neq i} \rho^{ij} \sqrt{\frac{\omega_{jj}}{\omega_{ii}}} \mathbf{Y}_j \|^2 \right) + \lambda \sum_{1 \leq i < j \leq p} \left| \rho^{ij} \right| \\
&= \frac{1}{2} \sum_{i=1}^p -n \log \omega_{ii} + \frac{1}{2} \omega_{ii} \| \mathbf{Y}_i + \sum_{j \neq i} \frac{\omega_{ij}}{\omega_{ii}} \mathbf{Y}_j \|^2 + \lambda \sum_{1 \leq i < j \leq p} \left| \rho^{ij} \right|.
\end{aligned} \tag{1}
$$

4

Algorithm 1 (SPACE pseudocode)

Input: Standardize data to have mean zero and standard deviation one

Input: Fix maximum number of iterations: r_{max}

Input: Fix initial estimate: ($\hat{\omega}_{ii}^{(0)} = 1/s_{ii}$ as suggested)

Input: Choose weights[a]: w_i ($w_i = \omega_{ii}$ or $w_i = 1$)

Set $r \leftarrow 1$

repeat

    `## Update partial correlations`

    Update $\hat{\boldsymbol{\rho}}^{(r)}$ by minimizing (with current estimates $\{\hat{\omega}_{ii}^{(r-1)}\}_{i=1}^{p}$ as fixed)

$$\frac{1}{2}\sum_{i=1}^{p}\left(w_i\|\mathbf{Y}_i - \sum_{j\neq i}\rho^{ij}\sqrt{\frac{\hat{\omega}_{jj}^{(r-1)}}{\hat{\omega}_{ii}^{(r-1)}}}\mathbf{Y}_j\|_2^2\right) + \lambda\sum_{1\leq i<j\leq p}|\rho^{ij}| \qquad (2)$$

    `## Update conditional variances`

    Update $\{\omega_{ii}^{(r)}\}_{i=1}^{p}$ by computing (with fixed $\hat{\rho}_{ij}^{(r-1)}$ and $\hat{\omega}_{ii}^{(r-1)}$ for all i and j)

$$\frac{1}{\hat{\omega}_{ii}^{(r)}} = \frac{1}{n}\|\mathbf{Y}_i - \sum_{j\neq i}(\hat{\rho}^{ij})^{(r-1)}\sqrt{\frac{\hat{\omega}_{jj}^{(r-1)}}{\hat{\omega}_{ii}^{(r-1)}}}\mathbf{Y}_j\|_2^2 \qquad (3)$$

  for $i = 1, \ldots, p$.

  $r \leftarrow r + 1$

  Update weights: w_i

until $r == r_{max}$

Return $(\hat{\boldsymbol{\rho}}^{(r_{max})}, \{\hat{\omega}_{ii}^{(r_{max})}\}_{i=1}^{p})$

---

[a]Peng et al. (2009) suggest two natural choices of weights w_i: (1) uniform weights $w_i = 1$ for all $i = 1, 2, \ldots, p$ (ii) partial variance weights $w_i = \omega_{ii}$.

Proof: Note that when fixing the diagonals $\{\omega_{ii}\}_{i=1}^{p}$, the minimization in (2) in the SPACE algorithm (with weights $w_i = \omega_{ii}$), corresponds to minimizing Q_{spc} with respect to $\boldsymbol{\rho}$. Now, let $\hat{\omega}_{ii}$ be the minimizer of Q_{spc} with respect to ω_{ii}, fixing $\{\beta_{ij}\}_{1\leq i\neq j\leq p}$ (where $\beta_{ij} = \rho^{ij}\sqrt{\frac{\omega_{jj}}{\omega_{ii}}} = -\frac{\omega_{ij}}{\omega_{ii}}$). Then, it follows that

$$\hat{\omega}_{ii} = \left(\frac{1}{n}\|\mathbf{Y}_i - \sum_{j\neq i}\beta_{ij}\mathbf{Y}_j\|_2^2\right)^{-1} \qquad (4)$$

The result follows by comparing (4) with the updates in (3). □

Although Lemma 1 identifies SPACE as an IPM, existing theory for iterative partial mini-

mization (see for example Zangwill (1969), Jensen et al. (1991), Lauritzen (1996), etc) only guarantees that every accumulation point of the sequence of iterates is a stationary point of the objective function Q_{spc}. To establish convergence, one needs to prove that every contour of the function Q_{spc} contains only finitely many stationary points. It is not clear if this latter condition holds for the function Q_{spc}. Moreover, for choice of weights $w_i = 1$, the SPACE algorithm does not appear to have an iterative partial minimization interpretation.

To improve our understanding of the convergence properties of SPACE, we started by testing the algorithm on simple examples. On some examples, SPACE converges very quickly; however, examples can be found where SPACE does not converge when using the two possible choices for weights: partial variance weights ($w_i = \omega_{ii}$) and uniform weights ($w_i = 1$). We now give an example of the lack of convergence.

Example 1: Consider the following population covariance and inverse covariance matrices:

$$\Omega = \begin{pmatrix} 3.0 & 2.1 & 0.0 \\ 2.1 & 3.0 & 2.1 \\ 0.0 & 2.1 & 3.0 \end{pmatrix}, \ \Sigma = \Omega^{-1} = \begin{pmatrix} 8.500 & -11.667 & 8.167 \\ -11.667 & 16.667 & -11.667 \\ 8.167 & -11.667 & 8.500 \end{pmatrix} \tag{5}$$

A sample of $n = 100$ *i.i.d.* vectors was generated from the corresponding $\mathcal{N}(\mathbf{0}, \Sigma)$ distribution. The data was standardized and the SPACE algorithm was run with choice of weights $w_i = \omega_{ii}$ and $\lambda = 160$. After the first few iterations successive SPACE iterates alternate between the following two matrices:

$$\begin{pmatrix} 29.009570 & 27.266460 & 0.000000 \\ 27.266460 & 51.863320 & 24.680140 \\ 0.000000 & 24.680140 & 26.359350 \end{pmatrix} \text{ and } \begin{pmatrix} 28.340040 & 27.221520 & -0.705390 \\ 27.221520 & 54.255190 & 24.569900 \\ -0.705390 & 24.569900 & 25.753040 \end{pmatrix}, \tag{6}$$

thereby establishing non-convergence of the SPACE algorithm in this example (see also Figure 1(a)). Note that the two matrices in (6) have different sparsity patterns. A similar example of non-convergence of SPACE with uniform weights is provided in Supplemental Section N.

A natural question to ask is whether the non-convergence of SPACE is pathological or whether is it widespread in settings of interest. To this end, the following simulation study was undertaken.

Example 2: We created a sparse 100×100 matrix Ω with edge density 4% and a condition number of 100. A total of 100 multivariate Gaussian datasets (with $n = 100$) having mean vector zero and covariance matrix $\Sigma = \Omega^{-1}$ were generated. Table 2 summarizes the number of times (out of 100) SPACE1 (SPACE with uniform weights) and SPACE2 (SPACE with partial variance weights) do not converge within 1500 iterations. When they do converge,

6

the mean number of iterations are 22.3 for SPACE1 and 14.1 for SPACE2 (note that since the original implementation of SPACE by Peng et al. (2009) was programmed to stop after 3 iterations, we modified the implementation to allow for more iterations in order to check for convergence of parameter estimates). It is clear from Table 2 that both variations of SPACE, using unit weights as well as ω_{ii} weights, exhibit extensive non-convergence behavior. Our simulations suggest that the convergence problem is exacerbated as the condition number of Ω increases.

SPACE1 ($w_i = 1$)			SPACE2 ($w_i = \omega_{ii}$)		
λ^*	NZ	NC	λ^*	NZ	NC
0.026	60.9%	92	0.085	79.8%	100
0.099	19.7%	100	0.160	28.3%	0
0.163	7.6%	100	0.220	10.7%	0
0.228	2.9%	100	0.280	4.8%	0
0.614	0.4%	0	0.730	0.5%	97

Table 2: Number of simulations (out of 100) that do not converge within 1500 iterations (NC) for select values of penalty parameter ($\lambda^* = \lambda/n$). Average percentage of non-zeros (NZ) in $\hat{\Omega}$ are also shown.

2.3 Symmetric lasso

The Symmetric lasso algorithm was proposed as a useful alternative to SPACE in a recent work by Friedman et al. (2010). Symmetric lasso minimizes the following (negative) pseudo-likelihood:

$$Q_{\mathrm{sym}}(\boldsymbol{\alpha}, \breve{\Omega}) = \frac{1}{2} \sum_{i=1}^{p} \left[n \log \alpha_{ii} + \frac{1}{\alpha_{ii}} \|\mathbf{Y}_i + \sum_{j \neq i} \omega_{ij} \alpha_{ii} \mathbf{Y}_j\|^2 \right] + \lambda \sum_{1 \leq i < j \leq p} |\omega_{ij}|. \qquad (7)$$

where $\alpha_{ii} = 1/\omega_{ii}$. Here $\boldsymbol{\alpha}$ denotes the vector with entries α_{ii} for $i = 1, \ldots, p$ and $\breve{\Omega}$ denotes the matrix Ω with diagonal entries set to zero. A comparison of (1) and (7) shows a deep connection between SPACE (with $w_i = \omega_{ii}$) and Symmetric lasso objective functions. In particular, the $Q_{\mathrm{sym}}(\boldsymbol{\alpha}, \breve{\Omega})$ objective function in (7) is a reparameterization of (1): the only difference is that the ℓ_1 penalty on the elements of $\boldsymbol{\rho}$ is replaced by a penalty on the elements of Ω in (7). The minimization of the objective function in (7) is performed by coordinate-wise descent on $(\boldsymbol{\alpha}, \breve{\Omega})$. Symmetric lasso is indeed a useful and computationally efficient procedure. However, theoretical properties such as convergence or asymptotic consistency have not yet been established. The following lemma investigates the properties of the objective function used in Symmetric lasso.

7

Lemma 2. *The Symmetric Lasso objective function in* (7) *is a non-convex function of* $(\boldsymbol{\alpha}, \breve{\Omega})$.

The proof of Lemma 2 is given in Supplemental Section A. The arguments in the proof of Lemma 2 demonstrate that the objective function used in Symmetric lasso is not convex, or even bi-convex in the parameterization used above. However, it can be shown that the SYMLASSO objective function is jointly convex in the elements of Ω (see Lee and Hastie (2014) and Supplemental section L). It is straightforward to check that the coordinatewise descent algorithms for both parameterizations are exactly the same. However, unless a function is strictly convex, there are no general theoretical guarantes of convergence for the corresponding coordinatewise descent algorithm. Indeed, when $n < p$, the SYMLASSO objective function is not strictly convex. Therefore, it is not clear if the coordinate descent algorithm converges in general. We conclude this section by remarking that both SPACE and symmetric lasso are useful additions to the graphical model selection literature, especially because they both respect symmetry and give computationally fast procedures.

2.4 The SPLICE algorithm

The SPLICE algorithm (Sparse Pseudo-Likelihood Inverse Covariance Estimates) was proposed by Rocha et al. (2008) as an alternative means to estimate Ω. In particular, the SPLICE formulation uses an ℓ_1-penalized regression based pseudo-likelihood objective function parameterized by matrices $\mathbf{D}$ and $\mathbf{B}$ where $\Omega = \mathbf{D}^{-2}(\mathbf{I} - \mathbf{B})$. The diagonal matrix $\mathbf{D}$ has elements $d_{jj} = 1/\sqrt{\omega_{jj}}$, $j = 1, \ldots, p$. The (asymmetric) matrix $\mathbf{B}$ has as columns the vectors of regression coefficients, $\beta_j \in \mathbb{R}^p$. These coefficients, β_j, arise when regressing $\mathbf{Y}_j$ on the remaining variables. A constraint on each β_j is imposed so that regression of $\mathbf{Y}_j$ is performed without including itself as a predictor variable: i.e., $\beta_{jj} = 0$. Based on the above properties, the ℓ_1-penalized pseudo-likelihood objective function of SPLICE algorithm (without the constant term) is given by

$$Q_{\mathrm{spl}}(\mathbf{B}, \mathbf{D}) = \frac{n}{2} \sum_{i=1}^{p} \log(d_{ii}^2) + \frac{1}{2} \sum_{i=1}^{p} \frac{1}{d_{ii}^2} \| \mathbf{Y}_i - \sum_{j \neq i} \beta_{ij} \mathbf{Y}_j \|^2 + \lambda \sum_{i<j} |\beta_{ij}|. \tag{8}$$

In order to optimize (8) with respect to $\mathbf{B}$ and $\mathbf{D}$, Rocha et al. (2008) also propose an iterative algorithm that alternates between maximizing $\mathbf{B}$ fixing $\mathbf{D}$, followed by maximizing $\mathbf{D}$ fixing $\mathbf{B}$. As with other regression-based graphical model selection algorithms, a proof of convergence of SPLICE is not available. The following lemma gives the convexity properties of the SPLICE objective function.

Lemma 3. *i) The SPLICE objective function* $Q_{spl}(\mathbf{B}, \mathbf{D})$ *is not jointly convex in* $(\mathbf{B}, \mathbf{D})$.

ii) Under the transformation $\mathbf{C} = \mathbf{D}^{-1}$, $Q_{spl}(\mathbf{B}, \mathbf{C})$ *is bi-convex.*

The proof of Lemma 3 is given in Supplemental Section B. The convergence properties of the SPLICE algorithm is not immediately clear since its objective function is non-convex. Furthermore, it is not clear whether the SPLICE solution yields a global optimum.

3   CONCORD: A convex pseudo-likelihood framework for sparse partial covariance estimation

The two pseudo-likelihood based approaches, SPACE and Symmetric lasso, have several attractive properties such as computational efficiency, simplicity and use of symmetry. They also do not directly depend on the more restrictive Gaussian assumption. Additionally, Peng et al. (2009) also establish (under suitable regularity assumptions) consistency of SPACE estimators for distributions with sub-Gaussian tails. However, none of the existing pseudo-likelihood based approaches yield a method that is provably convergent. In Section 2.2, we showed that there are instances where SPACE does not converge. As explained earlier, convergence is critical as this property guarantees well defined estimators which always exist, and are computable regardless of the data at hand. An important research objective therefore is the development of a pseudo-likelihood framework which preserves all the attractive properties of SPACE and SYMLASSO, and at the same time, leads to theoretical guarantees of convergence. It is however not clear immediately how to achieve this goal. A natural approach to take is to develop a convex formulation of the problem. Such an approach can yield many advantages, including 1) Guarantee of existence of a global minimum, 2) Better chance of convergence using convex optimization algorithms, 3) Deeper theoretical analysis of the properties of the solution and corresponding algorithm. As we have shown, the SPACE objective function is not jointly convex in the elements of Ω (or any natural reparameterization). Hence, one is not in a position to leverage tools from convex optimization theory for understanding its behavior. The SYMLASSO objective function is jointly convex in the elements of Ω. However, unless a function is strictly convex, there are no general guarantees of convergence for the corresponding coordinatewise descent algorithm. Indeed, when $n < p$, the SYMLASSO objective function is not strictly convex, and it is not clear if the corresponding coordinatewise descent algorithm converges.

In this section, we introduce a new approach for estimating Ω, called the CONvex CORrelation selection methoD (CONCORD) that aims to achieve the above objective. The CONCORD algorithm constructs sparse estimators of Ω by minimizing an objective function that is jointly convex in the entries of Ω. We start by introducing the objective function for the

CONCORD method and then proceed to derive the details of the corresponding coordinate-wise descent updates. Convergence is not obvious, as the function may not be strictly convex if $n < p$. It is proved in Section 4 that the corresponding coordinate-wise descent algorithm does indeed converge to a global minimum. Computational complexity and running time comparisons for CONCORD are given in Sections 3.3 and 5.1, respectively. Subsequently, large sample properties of the resulting estimator are established in Section 6 in order to provide asymptotic guarantees in the regime when both the dimension p and the sample size n tend to infinity. Thereafter, the performance of CONCORD on simulated data, and real data from biomedical and financial applications is demonstrated. Such analysis serves to establish that CONCORD preserves all the attractive properties of existing pseudo-likelihood methods and additionally provides the crucial theoretical guarantee of convergence and existence of a well-defined solution.

3.1 The CONCORD objective function

In order to develop a convex formulation of the pseudo-likelihood graphical model selection problem let us first revisit the formulation of the SPACE objective function in (1) with arbitrary weights w_i instead of ω_{ii}.

$$Q_{\text{spc}}(\Omega) = \frac{1}{2} \sum_{i=1}^{p} \left(-n \log \omega_{ii} + w_i \| \mathbf{Y}_i - \sum_{j \neq i} \rho^{ij} \sqrt{\frac{\omega_{jj}}{\omega_{ii}}} \mathbf{Y}_j \|_2^2 \right) + \lambda \sum_{1 \leq i < j \leq p} \left| \omega^{ij} \right| \qquad (9)$$

Now note that the above objective is not jointly convex in the elements of Ω since, 1) The middle term for the regression with the choices $w_i = 1$ or $w_i = \omega_{ii}$ is not a jointly convex function of the elements of Ω. 2) The penalty term is on the partial correlations $\rho^{ij} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}$ and is hence not a jointly convex function of the elements of Ω.

Now note the following for the regression term:

$$w_i \| \mathbf{Y}_i - \sum_{j \neq i} \rho^{ij} \sqrt{\frac{\omega_{jj}}{\omega_{ii}}} \mathbf{Y}_j \|_2^2 = w_i \| \mathbf{Y}_i + \sum_{j \neq i} \frac{\omega_{ij}}{\omega_{ii}} \mathbf{Y}_j \|_2^2 \qquad \left(\because \rho^{ij} = \frac{-\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}} \right)$$

$$= w_i \| \frac{1}{\omega_{ii}} (\omega_{ii} \mathbf{Y}_i + \sum_{j \neq i} \omega_{ij} \mathbf{Y}_j) \|_2^2$$

$$= \frac{w_i}{\omega_{ii}^2} \| \sum_{j=1}^{p} \omega_{ij} \mathbf{Y}_j \|_2^2$$

$$= \frac{w_i}{\omega_{ii}^2} \left(\omega_{\bullet i}' \mathbf{Y}' \mathbf{Y} \omega_{\bullet i} \right)$$

The choice of weights $w_i = \omega_{ii}^2$ yields

$$w_i \|\mathbf{Y}_i - \sum_{j \neq i} \rho^{ij} \sqrt{\frac{\omega_{jj}}{\omega_{ii}}} \mathbf{Y}_j\|_2^2 = \omega_{\bullet i}' \mathbf{Y}' \mathbf{Y} \omega_{\bullet i} \geq 0 \qquad (10)$$

The above expression in (10) is a quadratic form (and hence jointly convex) in the elements of Ω. Putting the ℓ_1-penalty term on the partial covariances ω_{ij} instead of on the partial correlations ρ^{ij} yields the following jointly convex objective function:

$$Q_{\text{con}}(\Omega) =: \mathcal{L}_{\text{con}}(\Omega) + \lambda \sum_{1 \leq i < j \leq p} |\omega_{ij}|$$

$$=: -\sum_{i=1}^{p} n \log \omega_{ii} + \frac{1}{2} \sum_{i=1}^{p} \|\omega_{ii} \mathbf{Y}_i + \sum_{j \neq i} \omega_{ij} \mathbf{Y}_j\|_2^2 + \lambda \sum_{1 \leq i < j \leq p} |\omega_{ij}|. \qquad (11)$$

The function $\mathcal{L}_{\text{con}}(\Omega)$ can be regarded as a pseudo-likelihood function in the spirit of Besag (1975). Since $-\log x$ and $|x|$ are convex functions, and $\sum_{i=1}^{p} \|\omega_{ii} \mathbf{Y}_i + \sum_{j \neq i} \omega_{ij} \mathbf{Y}_j\|^2$ is a positive semi-definite quadratic form in Ω, it follows that $Q_{\text{con}}(\Omega)$ is a jointly convex function of Ω (but not necessarily strictly convex). As we shall see later, this particular formulation above helps us establish theoretical guarantees of convergence (see Section 4), and, consequently, yields a regression based graphical model estimator that is well defined and is always computable. Note that the $n/2$ in (9) has been replaced by n in (11). The point is elaborated further in Remark 4. We now proceed to derive the details of the coordinate-wise descent algorithm for minimizing $Q_{\text{con}}(\Omega)$.

3.2 A coordinatewise minimization algorithm for minimizing $Q_{\text{con}}(\Omega)$

Let $\mathcal{A}_p$ denote the set of $p \times p$ real symmetric matrices. Let the parameter space $\mathcal{M}$ be defined as

$$\mathcal{M} := \{\Omega \in \mathcal{A}_p : \ \omega_{ii} > 0, \text{ for every } 1 \leq i \leq p\}.$$

Note that as in other regression based approaches (see Peng et al. (2009)), we have deliberately not restricted Ω to be positive definite as the main goal is to estimate the sparsity pattern in Ω. As mentioned in the introduction, a positive definite estimator can be obtained by using standard methods (Hastie et al. (2009), Xu et al. (2011)) once a partial correlation graph has been determined.

Let us now proceed to optimizing $Q_{\text{con}}(\Omega)$. For $1 \leq i \leq j \leq p$, define the function

$T_{ij} : \mathcal{M} \to \mathcal{M}$ by

$$T_{ij}(\Omega) = \underset{\{\tilde{\Omega}:(\tilde{\Omega})_{kl}=\omega_{kl} \ \forall(k,l)\neq(i,j)\}}{\arg\min} Q_{\text{con}}(\tilde{\Omega}). \tag{12}$$

For each (i,j), $T_{ij}(\Omega)$ gives the matrix where all the elements of Ω are left as is except the $(i,j)^{th}$ element. The $(i,j)^{th}$ element is replaced by the value that minimizes $Q_{\text{con}}(\Omega)$ with respect to ω_{ij} holding all other variables ω_{kl}, $(k,l) \neq (i,j)$ constant. We now proceed to evaluate $T_{ij}(\Omega)$ explicitly.

Lemma 4. *The function $T_{ij}(\Omega)$ defined in (12) can be computed in closed form. In particular, for $1 \leq i \leq p$,*

$$(T_{ii}(\Omega))_{ii} = \frac{-\sum_{j\neq i}\omega_{ij}s_{ij} + \sqrt{\left(\sum_{j\neq i}\omega_{ij}s_{ij}\right)^2 + 4s_{ii}}}{2s_{ii}}. \tag{13}$$

For $1 \leq i < j \leq p$,

$$(T_{ij}(\Omega))_{ij} = \frac{S_{\frac{\lambda}{n}}\left(-\left(\sum_{j'\neq j}\omega_{ij'}s_{jj'} + \sum_{i'\neq i}\omega_{i'j}s_{ii'}\right)\right)}{s_{ii} + s_{jj}}, \tag{14}$$

where s_{ij} is the $(i,j)^{th}$ entry of $\frac{1}{n}\mathbf{Y}^T\mathbf{Y}$, and $S_\lambda(x) := \text{sign}(x)(|x| - \lambda)_+$.

The proof is given in Supplemental Section C. An important contribution of Lemma 4 is that it gives the necessary ingredients for designing a coordinate descent approach to minimizing the CONCORD objective function. More specifically, (13) can be used to update the partial variance terms, and (14) can be used to update the partial covariance terms. The coordinate-wise descent algorithm for CONCORD is summarized in Algorithm 2. The zeros in the estimated partial covariance matrix can then subsequently be used to construct a partial covariance or partial correlation graph.

The following procedure can be used to select the penalty parameter λ. Define the residual sum of squares (RSS) for $i = 1, \ldots, p$ as

$$RSS_i(\lambda) = \sum_{k=1}^{n}\left(y_i^k - \sum_{j\neq i}\frac{\omega_{ij}}{\omega_{ii}}y_j^k\right)^2.$$

Further, the i-th component of BIC type score can be defined as

$$BIC_i(\lambda) = n\log(RSS_i(\lambda)) + \log n \cdot |\{j : j \neq i, \omega_{ij,\lambda} \neq 0\}|.$$

12

Algorithm 2 (CONCORD pseudocode)

Input: standardize data to have mean zero and standard deviation one
Input: Fix maximum number of iterations: r_{max}
Input: Fix initial estimate: $\hat{\Omega}^{(0)}$
Input: Fix convergence threshold: ϵ
Set $r \leftarrow 1$
converged = FALSE
Set $\hat{\Omega}^{\text{current}} \leftarrow \hat{\Omega}^{(0)}$
repeat
    $\hat{\Omega}^{\text{old}} \leftarrow \hat{\Omega}^{\text{current}}$

    `## Updates to partial covariances` ω_{ij}
    for $i \leftarrow 1, 2, \cdots, p-1$ **do**
        for $j \leftarrow i+1, \cdots, p$ **do**

$$\hat{\omega}_{ij}^{\text{current}} \leftarrow \left(T_{ij}(\Omega^{\text{current}})\right)_{ij} \tag{15}$$

        end for
    end for

    `## Updates to partial variances` ω_{ii}
    for $i \leftarrow 1, 2, \cdots, p$ **do**

$$\hat{\omega}_{ii}^{\text{current}} \leftarrow \left(T_{ii}(\Omega^{\text{current}})\right)_{ii} \tag{16}$$

    end for

    $\hat{\Omega}^{(r)} \leftarrow \hat{\Omega}^{\text{current}}$
    `## Convergence checking`
    if $\|\hat{\Omega}^{\text{current}} - \hat{\Omega}^{\text{old}}\|_{\max} < \epsilon$ **then**
        converged = TRUE
    else
        $r \leftarrow r + 1$
    end if

until converged = TRUE or $r > r_{\max}$
Return final estimate: $\hat{\Omega}^{(r)}$

The penalty parameter λ can be chosen to minimize the sum $BIC(\lambda) = \sum_{i=1}^{p} BIC_i(\lambda)$.

3.3   Computational complexity

We now proceed to show that the computational cost of each iteration of CONCORD is $min\left(O(np^2), O(p^3)\right)$, that is, the CONCORD algorithm is competitive with other proposed methods. The updates in Equations in (15) and (16) are implemented differently depending on whether $n \geq p$ or $n < p$.

Case 1 ($n \geq p$): Let us first consider the case when $n \geq p$. Note that both sums in (14) are inner products between a row in $\hat{\Omega}$ and a row in $\mathbf{S}$. Clearly, computing these sums require $O(p)$ operations each. Similarly, the update in (13) requires $O(p)$ operations. Since there are $O(p^2)$ entries in Ω, one complete sweep of updates over all entries in $\hat{\Omega}$ would require $O(p^3)$ operations.

Case 2 ($n < p$): Let us now consider the case when $n < p$. We show below that the updates can be performed in $O(np^2)$ operations. The main idea here is that the coordinate-wise calculations at each iteration, which involves an inner product of two $p \times 1$ vectors, can be reduced to an inner product calculation involving auxiliary variables (residual variables to be more specific) of dimension $n \times 1$. The following lemmas are essential ingredients in calculating the computational complexity in this setting. In particular, Lemma 5 expresses the inner product calculations in (13) and (14) in terms of residual vectors.

Lemma 5. *For $1 \leq i, j \leq p$,*

$$\sum_{k \neq j} \omega_{ik} s_{jk} = -\omega_{ij} s_{jj} + \omega_{ii} \mathbf{Y}_j' \mathbf{r}_i,$$

where $\mathbf{Y}_j$ is the jth column of the data matrix $\mathbf{Y}$, and $\mathbf{r}_i = \mathbf{Y}_i + \sum_{k \neq i} \frac{\omega_{ik}}{\omega_{ii}} \mathbf{Y}_k$ is an n-vector of residuals of regressing $\mathbf{Y}_i$ on the rest.

The following lemma now quantifies the computational cost of updating the residual vectors during each iteration of the CONCORD algorithm.

Lemma 6. *Define the residual vector $\mathbf{r}_m$ for $m = 1, 2, \ldots, p$ as follows:*

$$\mathbf{r}_m = \mathbf{r}_m(\Omega) = \mathbf{Y}_m + \sum_{k \neq m} \frac{\omega_{mk}}{\omega_{mm}} \mathbf{Y}_k \tag{17}$$

where $\Omega = ((\Omega_{ij}))_{1 \leq i, j \leq p}$. Then,

1. *For $m \neq k, l$, the residual vector $\mathbf{r}_m$ is functionally independent of ω_{kl}. (The term ω_{kl} appears only in the expressions for the residual vectors $\mathbf{r}_k$ and $\mathbf{r}_l$.)*

2. *Fix all the elements of Ω except ω_{kl}. Suppose ω_{kl} is changed to ω_{kl}^*. Then, updating the residual vectors $\mathbf{r}_k$ and $\mathbf{r}_l$ requires $O(n)$ operations. (Hence, updating $\mathbf{r}_k$ and $\mathbf{r}_l$ after each update in (15) requires $O(n)$ operations.)*

3. *For $m \neq k$, the residual vector $\mathbf{r}_m$ is functionally independent of ω_{kk}. (The term ω_{kk} appears only in the expression for the residual vector $\mathbf{r}_k$.)*

4. *Fix all elements of Ω except ω_{kk}. Suppose ω_{kk} is changed to ω_{kk}^*. Then, updating the residual vector $\mathbf{r}_k$ requires $O(n)$ operations. (Hence, updating $\mathbf{r}_k$ after each update in (16) requires $O(n)$ operations).*

The proofs of Lemmas 5 and 6 are straightforward and are given in Supplemental Sections D and E. Note that the inner product between $\mathbf{y}_j$ and $\mathbf{r}_i$ takes $O(n)$ operations. Hence, by Lemma 5 the updates in (15) and (16) require $O(n)$ operations. Also, after each update in (15) and (16) the residual vectors need to be appropriately modified. By Lemma 6, this modification can also be achieved in $O(n)$ operations. As a result, one complete sweep of updates over all entries in $\hat{\Omega}$ can be performed in $O(np^2)$ operations.

Hence, we conclude that the computational complexity of the CONCORD algorithm is competitive with the SPACE and Symmetric lasso algorithms, which are also $\min\left(O(np^2), O(p^3)\right)$.

3.4 A unifying framework for pseudo-likelihood based graphical model selection

In this section, we provide a unifying framework which formally connects the five pseudo-likelihood formulations considered in this paper, namely, SPACE1, SPACE2, SYMLASSO, SPLICE and CONCORD (counting two choices for weights in the SPACE algorithm as two different formulations). Recall that the random vectors $\mathbf{Y}^k = \left(y_1^k, y_2^k, \cdots, y_p^k\right)'$, $k = 1, 2, \cdots, n$ denote *i.i.d.* observations from a multivariate distribution with mean vector $\mathbf{0}$ and covariance matrix Σ, the precision matrix is given by $\Omega = \Sigma^{-1} = ((\omega_{ij}))_{1 \leq i,j \leq p}$, and $\mathbf{S}$ denotes the sample covariance matrix. Let Ω_D denote the diagonal matrix with i^{th} diagonal entry given by ω_{ii}. Lemma 7 below formally identifies the relationship between all five of the regression-based pseudo-likelihood methods.

Lemma 7. *i) The (negative) pseudo-likelihood functions of CONCORD, SPACE1, SPACE2, SYMLASSO and SPLICE formulations can be expressed in matrix form as follows (up to reparameterization):*

ii) All five pseudo-likelihoods above correspond to a unified or generalized form of the

	Regression form	Matrix form	
$\mathcal{L}_{\text{con}}(\Omega)$	$\frac{1}{2}\sum_{i=1}^{p}\left[-n\log\omega_{ii}^2 + \|\omega_{ii}\mathbf{Y}_i + \sum_{j\neq i}\omega_{ij}\mathbf{Y}_j\|_2^2\right]$	$\frac{n}{2}\left[-\log|\Omega_D^2| + \text{tr}(\mathbf{S}\Omega^2)\right]$	(18)
$\mathcal{L}_{\text{spc},1}(\Omega_D,\boldsymbol{\rho})$	$\frac{1}{2}\sum_{i=1}^{p}\left[-n\log\omega_{ii} + \|\mathbf{Y}_i - \sum_{j\neq i}\rho^{ij}\sqrt{\frac{\omega_{jj}}{\omega_{ii}}}\mathbf{Y}_j\|_2^2\right]$	$\frac{n}{2}\left[-\log|\Omega_D| + \text{tr}(\mathbf{S}\Omega\Omega_D^{-2}\Omega)\right]$	(19)
$\mathcal{L}_{\text{spc},2}(\Omega_D,\boldsymbol{\rho})$	$\frac{1}{2}\sum_{i=1}^{p}\left[-n\log\omega_{ii} + \omega_{ii}\|\mathbf{Y}_i - \sum_{j\neq i}\rho^{ij}\sqrt{\frac{\omega_{jj}}{\omega_{ii}}}\mathbf{Y}_j\|_2^2\right]$	$\frac{n}{2}\left[-\log|\Omega_D| + \text{tr}(\mathbf{S}\Omega\Omega_D^{-1}\Omega)\right]$	(20)
$\mathcal{L}_{\text{sym}}(\boldsymbol{\alpha},\Omega_F)$	$\frac{1}{2}\sum_{i=1}^{p}\left[n\log\alpha_{ii} + \frac{1}{\alpha_{ii}}\|\mathbf{Y}_i + \sum_{j\neq i}\omega_{ij}\alpha_{ii}\mathbf{Y}_j\|^2\right]$	$\frac{n}{2}\left[-\log|\Omega_D| + \text{tr}(\mathbf{S}\Omega\Omega_D^{-1}\Omega)\right]$	(21)
$\mathcal{L}_{\text{spl}}(\mathbf{B},\mathbf{D})$	$\frac{1}{2}\sum_{i=1}^{p}\left[n\log(d_{ii}^2) + \frac{1}{d_{ii}^2}\|\mathbf{Y}_i - \sum_{j\neq i}\beta_{ij}\mathbf{Y}_j\|_2^2\right]$	$\frac{n}{2}\left[-\log|\Omega_D| + \text{tr}(\mathbf{S}\Omega\Omega_D^{-1}\Omega)\right]$	(22)

Gaussian log-likelihood function

$$\mathcal{L}_{\text{uni}}(G(\Omega), H(\Omega)) = \frac{n}{2}\left[-\log\det G(\Omega) + \text{tr}(\mathbf{S}H(\Omega))\right],$$

where $G(\Omega)$ and $H(\Omega)$ are functions of Ω. The functions G and H which characterize the pseudo-likelihood formulations corresponding to CONCORD, SPACE1, SPACE2, SYM-LASSO and SPLICE are given as follows:

$$G_{con}(\Omega) = \Omega_D^2, \qquad H_{con}(\Omega) = \Omega^2$$

$$G_{spc,1}(\Omega) = \Omega_D, \qquad H_{spc,1}(\Omega) = \Omega\Omega_D^{-2}\Omega$$

$$G_{spc,2}(\Omega) = G_{sym}(\Omega) = G_{spl}(\Omega) = \Omega_D, \qquad H_{spc,2}(\Omega) = H_{sym}(\Omega) = H_{spl}(\Omega) = \Omega\Omega_D^{-1}\Omega$$

The proof of Lemma 7 is given in Supplemental Section F. The above lemma gives various useful insights into the different pseudo-likelihoods that have been proposed for the inverse covariance estimation problem. The following remarks discuss these insights.

Remark 1. Note that when $G(\Omega) = H(\Omega) = \Omega$, $\mathcal{L}(G(\Omega), H(\Omega))$ corresponds to the standard (negative) Gaussian log-likelihood function.

Remark 2. Note that $\Omega_D^{-1}\Omega$ is a re-scaling of Ω so as to make all the diagonal elements one (hence sparsity between Ω and $\Omega_D^{-1}\Omega$ are the same). In this sense, the SPACE2, SYMLASSO and SPLICE algorithms make the same approximation to the Gaussian likelihood with the log determinant term, $\log|\Omega|$, replaced by $\log|\Omega_D|$. The trace term $\text{tr}(\mathbf{S}\Omega)$ is approximated by $\text{tr}(\mathbf{S}\Omega\Omega_D^{-1}\Omega)$. Moreover, if Ω is sparse, then $\Omega_D^{-1}\Omega$ is close to the identity matrix, i.e., $\Omega_D^{-1}\Omega \approx I + \mathbf{C}$ for some $\mathbf{C}$. In this case, the term in the Gaussian likelihood $\text{tr}(\mathbf{S}\Omega)$ is perturbed by an off-diagonal matrix $\mathbf{C}$ resulting in an expression of the form $\text{tr}(\mathbf{S}\Omega(I + \mathbf{C}))$.

Remark 3. Conceptually, the sole source of difference between the three regularized versions of the objective functions of SPACE2, SYMLASSO and SPLICE algorithms is in the way in which the ℓ_1-penalties are specified. SPACE2 applies the penalty to the partial correlations, SYMLASSO to the partial covariances and SPLICE to the symmetrized regression coefficients.

Remark 4. Note that the CONCORD method approximates the Normal likelihood by approximating the $\log|\Omega|$ term by $\log|\Omega_D^2|$, and $\text{tr}(\mathbf{S}\Omega)$ by $\text{tr}(\mathbf{S}\Omega^2)$. Hence, the CONCORD algorithm can be considered as a reparameterization of the Gaussian likelihood with the concentration matrix Ω^2 (together with an approximation to the log determinant term). More specifically,

$$\mathcal{L}_{\text{con}}(\Omega) = \mathcal{L}_{\text{uni}}(\Omega_D^2, \Omega^2) = \frac{n}{2}\left(-\log\det\Omega_D^2 + \text{tr}(\mathbf{S}\Omega^2)\right) = n\left(-\log\det\Omega_D + \frac{1}{2}\text{tr}(\mathbf{S}\Omega^2)\right),$$

and justifies the appearance of "n" as compared to "$n/2$" in the CONCORD objective in (11). In Supplemental Section G, we illustrate the usefulness of this correction based on the insight from our unification framework, and show that it leads to better estimates of Ω.

4   Convergence of CONCORD

We now proceed to consider the convergence properties of the CONCORD algorithm. Note that $Q_{\text{con}}(\Omega)$ is not differentiable. Also, if $n < p$, then $Q_{\text{con}}(\Omega)$ is not necessarily strictly convex. Hence, the global minimum may not be unique, and as discussed below, the convergence of the coordinatewise minimization algorithm to a global minimum does not follow from existing theory. Note that although $Q_{\text{con}}(\Omega)$ is not differentiable, it can be expressed as a sum of a smooth function of Ω and a separable function of Ω (namely $\lambda\sum_{1\leq i<j\leq p}|\omega_{ij}|$). Tseng (1988, 2001) proves that under certain conditions, every cluster point of the sequence of iterates of the coordinatewise minimization algorithm for such an objective function is a stationary point of the objective function. However, if the function is not strictly convex, there is no general guarantee that the sequence of iterates has a unique cluster point, i.e., there is no theoretical guarantee that the sequence of iterates converges. The following theorem shows that the cyclic coordinatewise minimization algorithm applied to the CONCORD objective function converges to a global minimum. A proof of this result can be found in Supplemental Section H.

Theorem 1. *If $S_{ii} > \mathbf{0}$ for every $1 \leq i \leq p$, the sequence of iterates $\left\{\hat{\Omega}^{(r)}\right\}_{r\geq 0}$ obtained by Algorithm 2 converges to a global minimum of $Q_{con}(\Omega)$. More specifically, $\hat{\Omega}^{(r)} \to \hat{\Omega} \in \mathcal{M}$ as $r \to \infty$ for some $\hat{\Omega}$, and furthermore $Q_{con}(\hat{\Omega}) \leq Q_{con}(\Omega)$ for all $\Omega \in \mathcal{M}$.*

Remark 5. If $n \geq 2$, and none of the underlying p marginal distributions (corresponding to the p-variate distribution for the data vectors) is degenerate, it follows that the diagonal entries of the data covariance matrix S are strictly positive with probability 1.
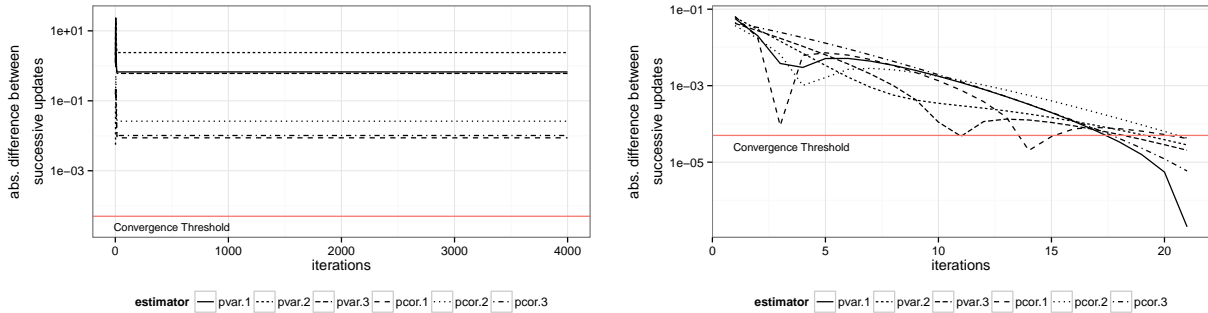
(a) SPACE algorithm (partial variance weights) applied to dataset in Example 1.

(b) CONCORD algorithm applied to dataset in Example 1.

Figure 1: Illustrations of the non-convergence of SPACE and convergence of CONCORD. The y-axes are log scaled. For SPACE, log absolute difference between entries of successive estimates becomes constant (thus indicating non-convergence).

With theory in hand, we now proceed to numerically illustrate the convergence properties established above. When CONCORD is applied to the dataset in Example 1, convergence is achieved (see Figure 1(b)), whereas SPACE does not converge (see Figure 1(a)).

5   Applications

5.1   Simulated Data

5.1.1   Timing Comparison

We now proceed to compare the timing performance of CONCORD with Glasso and the two different versions of SPACE. The acronyms SPACE1 and SPACE2 denote SPACE estimates using uniform weights and partial variance weights, respectively. We first consider the setting $p = 1000, n = 200$. For the purposes of this simulation study, a $p \times p$ positive definite matrix Ω (with $p = 1000$) with condition number 10 was used. Thereafter, 50 independent datasets were generated, each consisting of $n = 200$ *i.i.d.* samples from a $\mathcal{N}_p(0, \Sigma = \Omega^{-1})$ distribution. For each dataset, the four algorithms were run until convergence for a range of penalty parameter values. We note that the default number of iterations for SPACE in the R function by Peng et al. (2009) is 3. However, given the convergence issues for SPACE, we ran SPACE until convergence or until 50 iterations (whichever is smaller). The timing results (averaged over the 100 datasets) in the top part of Table 3 below show wall clock times until convergence (in seconds) for Glasso, CONCORD, SPACE1 and SPACE2.

One can see that in the $p = 1000, n = 200$ setting, CONCORD is uniformly faster than its

competitors. Note the low penalty parameter cases correspond to high dimensional settings where the estimated covariance matrix is typically poorly conditioned and the log-likelihood surface is very flat. The results in Table 3 indicate that in such settings CONCORD is faster than its competitors by orders of magnitude (even though Glasso is implemented in Fortran). Both SPACE1 and SPACE2 are much slower than CONCORD and Glasso in this setting. The wall clock time for an iterative algorithm can be thought of as a function of the number of iterations until convergence, the order of computations for a single iteration, and also the implementation details (such as choice of software, efficiency of the code etc.). Note that the order of computations for a single iteration is same for SPACE and CONCORD, and lower than that of Glasso when $n < p$. It is likely that the significant increase in the wall clock time for SPACE is due to implementation details and the larger number of iterations required for convergence (or non-convergence, since we are stopping SPACE if the algorithm does not satisfy the convergence criterion by 50 iterations).

We further compare the timing performance of CONCORD and Glasso for $p = 3000$ with $n = 600$ and $n = 900$ (SPACE is not considered here because of the timing issues mentioned above. These issues are amplified in this more demanding setting). A $p \times p$ positive definite matrix Ω (with $p = 3000$) with 3% sparsity is used. Thereafter, 50 independent datasets were generated, each consisting of $n = 600$ *i.i.d.* samples from a $\mathcal{N}_p(0, \Sigma = \Omega^{-1})$ distribution. The same exercise was repeated with $n = 900$. The timing results (averaged over the 100 datasets) in the bottom part of Table 3 below show wall clock times until convergence (in seconds) for Glasso, CONCORD, SPACE1 and SPACE2 for various penalty parameter values. It can be seen that in both the $n = 600$ and $n = 900$ cases, CONCORD was around ten times faster than Glasso.

In conclusion, these simulation results in this subsection illustrate that CONCORD is much faster as compared to SPACE and Glasso, especially in very high dimensional settings. We also note that a downloadable version of the CONCORD algorithm has been developed in R, and is freely available at http://cran.r-project.org/web/packages/gconcord.

5.1.2 Model selection comparison

In this section, we perform a simulation study in which we compare the model selection performance of CONCORD and Glasso when the underlying data is drawn from a multivariate-t distribution (the reasons for not considering SPACE are provided in a remark at the end of this section). The data is drawn from a multivariate-t distribution to illustrate the potential benefit of using penalized regression methods (CONCORD) outside the Gaussian setting.

For the purposes of this study, using a similar approach as in Peng et al. (2009), a $p \times p$ sparse positive definite matrix Ω (with $p = 1000$) with condition number 13.6 is chosen. Using

p = 1000, n = 200											
	Glasso			CONCORD			SPACE1 ($w_i = 1$)			SPACE2 ($w_i = \omega_{ii}$)	
λ	NZ	Time	λ^*	NZ	Time	λ^*	NZ	Time	λ^*	NZ	Time
0.14	4.77%	87.60	0.12	4.23%	6.12	0.10	4.49%	101.78	0.16	100.00%	19206.55
0.19	0.87%	71.47	0.17	0.98%	5.10	0.17	0.64%	99.20	0.21	1.76%	222.00
0.28	0.17%	5.41	0.28	0.15%	5.37	0.28	0.14%	138.01	0.30	0.17%	94.59
0.39	0.08%	5.30	0.39	0.07%	4.00	0.39	0.07%	75.55	0.40	0.08%	108.61
0.51	0.04%	6.38	0.51	0.04%	4.76	0.51	0.04%	49.59	0.51	0.04%	132.34

p = 3000, n = 600						p = 3000, n = 900					
	Glasso			CONCORD			Glasso			CONCORD	
λ	NZ	Time	λ^*	NZ	Time	λ	NZ	Time	λ^*	NZ	Time
0.09	2.71%	1842.74	0.09	2.10%	266.69	0.09	0.70%	1389.96	0.09	0.64%	298.21
0.10	1.97%	1835.32	0.10	1.59%	235.49	0.10	0.44%	1395.42	0.10	0.41%	298.00
0.10	1.43%	1419.41	0.10	1.19%	232.67	0.10	0.27%	1334.78	0.10	0.26%	302.15

Table 3: Timing comparison (in seconds) for $p = 1000$, 3000 and varying n. SPACE is run until convergence or 50 iterations (whichever is smaller). Note that SPACE1 and SPACE2 are much slower compared than CONCORD and Glasso in wall time, for the $p = 1000$ simulation. Hence, for $p = 3000$, only Glasso and CONCORD are compared. Here, λ denotes the value of the penalty parameter for the respective algorithms, with $\lambda^* = \lambda/n$ for CONCORD and SPACE. NZ denotes the percentage of non-zero entries in the corresponding estimator.

this Ω for each sample size $n = 200$, $n = 400$ and $n = 800$, 50 datasets, each having *i.i.d.* multivariate-t distribution with mean zero and covariance matrix $\Sigma = \Omega^{-1}$, are generated. We compare the model selection performance of Glasso and CONCORD in this heavy tailed setting with receiver operating characteristic (ROC) curves, which compare false positive rates (FPR) and true positive rates (TPR). Each ROC curve is traced out by varying the penalty parameter λ over 50 possible values.

We use the Area-under-the-curve (AUC) as a means to compare model selection performance. This measure is frequently used to compare ROC curves (Fawcett, 2006, Friedman et al., 2010). The AUC of a full ROC curve resulting from perfect recovery of zero/non-zero structure in Ω would be 1. In typical real applications, FPR is controlled to be sufficiently low. We therefore compare model selection performance when FPR is less than 15% (or 0.15). When controlling FPR to be less than 0.15, a perfect method will yield AUC of 0.15. Table 4 provides the median of the AUCs (divided by 0.15 to normalize to 1), as well as the interquartile ranges (IQR) over the 50 datasets for $n = 200$, $n = 400$ and $n = 800$.

	n = 200		n = 400		n = 800	
Solver	Median	IQR	Median	IQR	Median	IQR
Glasso	0.745	0.032	0.819	0.030	0.885	0.029
CONCORD	0.811	0.011	0.887	0.012	0.933	0.013

Table 4: Median and IQR of area-under-the-curve (AUC) for 50 simulations. Each simulation yields a ROC curve from which the AUC is computed for FPR in the interval [0, 0.15] and normalized to 1.

Table 4 above shows that CONCORD has a much better model selection performance as compared to Glasso. Moreover, it turns out that CONCORD has a higher AUC than Glasso for every single one of the 150 datasets (50 each for $n = 200, 400$ and 800). We note that CONCORD not only recovers the sparsity structure more accurately in general, it also has much less variation.

Remark: Note that we need to simulate 50 datasets for each of the above three sample sizes. For each of these datasets, an algorithm has to be run for 50 different penalty parameter values. In totality, this amounts to running the algorithm 7500 times. As we demonstrated in the simulations in Section 5.1.1, when SPACE is run until convergence (or terminated after the number of iterations is 50), then SPACE's intractability makes it infeasible to run it 7500 times. As an alternative, one could follow the approach of Peng et al. (2009) and stop SPACE after running 3 iterations. However, given the possible non-convergence issues associated with SPACE, it is not clear if the resulting estimate is meaningful. Even so, if we follow this approach of stopping SPACE after three iterations, we find that CONCORD outperforms SPACE1 and SPACE2. For example, if we consider the $n = 200$ case, then the median AUC value for SPACE1 is 0.779 (with $IQR = 0.054$) and the median AUC value for SPACE2 is 0.802 (with $IQR = 0.013$).

5.2 Application to breast cancer data

We now illustrate the performance of the CONCORD method on a real dataset. To facilitate comparison, we consider data from a breast cancer study (Chang et al., 2005) on which SPACE was illustrated. This dataset contains expression levels of 24481 genes on 248 patients with breast cancer. The dataset also contains extensive clinical data including survival times.

Following the approach in Peng et al. (2009) we focus on a smaller subset of genes. This reduction can be achieved by utilizing clinical information that is provided together with the microarray expression dataset. In particular, survival analysis via univariate Cox regression with patient survival times is used to select a subset of genes closely associated with breast cancer. A choice of p-value < 0.0003 yields a reduced dataset with 1107 genes. This subset

of the data is then mean centered and scaled so that the median absolute deviation is 1 (as outliers seem to be present). Following a similar approach to that in Peng et al. (2009), penalty parameters for each partial correlation graph estimation method were chosen so that each partial correlation graph yields 200 edges.

Partial correlation graphs can be used to identify genes that are biologically meaningful and can lead to gene therapeutic targets. In particular, there is compelling evidence from the biomedical literature that highly connected nodes are central to biological networks (Carter et al., 2004, Jeong et al., 2001, Han et al., 2004). To this end, we focus on identifying the 10 most highly connected genes ("hub" genes) identified by each partial correlation graph estimation method. Table 6 in Supplemental Section I summarizes the top 10 hub genes obtained by CONCORD, SYMLASSO, SPACE1 and SPACE2. The table also gives references from the biomedical literature that places these genes in the context of breast cancer. These references illustrate that most of the identified genes are indeed quite relevant in the study of breast cancer. It can also be seen that there is a large level of overlap in the top 10 genes identified by the four methods. There are also however some notable differences. For example, *TPX2* has been identified only by CONCORD. Bibby et al. (2009) suggests that mutation of Aurora A - a known general cancer related gene - reduces cellular activity and mislocalization due to loss of interaction with *TPX2*. Moreover, a recent extensive study by Maxwell et al. (2011)[1] identifies a gene regulatory mechanism in which *TPX2*, Aurora A, *RHAMM* and *BRCA1* play a key role. This finding is especially significant given that *BRCA1* (breast cancer type 1 susceptibility protein) is one of the most well known genes linked to breast cancer. We also remark that if a higher number of hub genes are targeted (like the top 20 or top 100 vs. the top 10), CONCORD identifies additional genes not discovered by existing methods. However, identification of even a single important gene can lead to significant findings and novel gene therapeutic targets, since many gene silencing experiments often focus on one or two genes at a time.

We conclude this section by remarking that CONCORD is a useful addition to the graphical models literature as it is competitive with other methods in terms of model selection accuracy, timing, relevance for applications, and also gives provable convergence guarantees.

5.3 Application to portfolio optimization

We now consider the efficacy of using CONCORD in a financial portfolio optimization setting where a stable estimate of the covariance matrix is often required. We follow closely the exposition to the problem as given in Won et al. (2012). A portfolio of financial instruments

---

[1] http://www.ncbi.nlm.nih.gov/pubmed/22110403

constitutes a collection of both risky and risk-free assets held by a legal entity. The return on the overall portfolio over a given holding period is defined as the weighted average of the returns on the individual assets, where the weights for each asset corresponds to its proportion in monetary terms. The primary objective of the portfolio optimization problem is to determine the weights that maximize the overall return on the portfolio subject to a certain level of risk (or vice versa). In Markowitz mean-variance portfolio (MVP) theory, this risk is taken to be the the standard deviation of the portfolio (Markowitz, 1952). As noted in Luenberger (1997) & Merton (1980), the optimal portfolio weights or the optimal allocation depends critically on the mean and covariance matrix of the individual asset returns, and hence estimation of these quantities is central to MVP. As one of the goals in this paper is to illustrate the efficacy of using CONCORD to obtain a stable covariance matrix estimate, we shall consider the *minimum variance* portfolio problem, as compared to the *mean-variance* portfolio optimization problem. The former requires estimating only the covariance matrix and thus presents an ideal setting for comparing covariance estimation methods in the portfolio optimization context (see Chan et al. (1999) for more details). In particular, we aim to compare the performance of CONCORD with other covariance estimation methods, for the purposes of constructing a minimum variance portfolio. The performance of each of the different methods and the associated strategies will be compared over a sustained period of time in order to assess their respective merits.

5.3.1 Minimum variance portfolio rebalancing

The minimum variance portfolio selection problem is defined as follows. Given p risky assets, let r_{it} denote the return of asset i over period t; which in turn is defined as the change in its price over time period t, divided by the price at the beginning of the period. As usual, let Σ_t denote the covariance matrix of the daily returns, $r_t^T = (r_{1t}, r_{2t}, \ldots, r_{pt})$. The portfolio weights $w_k^T = (w_{1k}, w_{2k}, \ldots, w_{pk})$ denote the weight of asset $i = 1, \ldots, p$ in the portfolio for the k-th time period. A long position or a short position for asset i during period k is given by the sign of w_{ik}, i.e., $w_{ik} > 0$ for long, and $w_{ik} < 0$ for short positions respectively. The budget constraint can be written as $\mathbf{1}^T w_k = 1$, where $\mathbf{1}$ denotes the vector of all ones. Note that the risk of a given portfolio as measured by the standard deviation of its return is simply $(w_k^T \Sigma w_k)^{1/2}$.

The minimum variance portfolio selection problem for investment period k can now be formally defined as follows:

$$\text{minimize } w_k^T \Sigma w_k \quad \text{subject to } \mathbf{1}^T w_k = 1. \tag{23}$$

As (23) above is a simple quadratic program, it has an analytic solution given by $w_k^\star = (\mathbf{1}^T \Sigma^{-1} \mathbf{1})^{-1} \Sigma^{-1} \mathbf{1}$. Note that the solution depends on the theoretical covariance matrix Σ. In practice, the parameter Σ has to be estimated.

The most basic approach to the portfolio selection problem often makes the unrealistic assumption that returns are stationary in time. A standard approach to dealing with the non-stationarity in such financial time series is to use a periodic rebalancing strategy. In particular, at the beginning of each investment period $k = 1, 2, \ldots, K$, portfolio weights $w_k = (w_{1k}, \ldots, w_{pk})'$ are computed from the previous N_{est} days of observed returns (N_{est} is called the "estimation horizon"). These portfolio weights are then held constant for the duration of each investment period. The process is repeated at the start of the next investment period and is often referred to as "rebalancing." More details of the rebalancing strategy are provided in Supplemental section J.3.

5.3.2 Application to the Dow Jones Industrial Average

We now consider the problem of investing in the stocks that feature in the Dow Jones Industrial Average (DJIA) index. The DJIA is a composite blue chip index consisting of 30 stocks (note that Kraft Foods (KFT) data was removed in our analysis due to its limited data span[2]. Table 7 in Supplemental Section J.1 lists the 29 component stocks used in our analysis.

Rebalancing time points were chosen to be every four weeks starting from 1995/02/18 to 2012/10/26 (approximately 17 years), and are shown in Table 8 in Supplemental Section J.2. Start and end dates of each period are selected to be calendar weeks, and need not coincide with a trading day. The total number of investment periods is 231, and the number of trading days in each investment period varies between 15 and 20 days. We shall compare the following five methods for estimating the covariance matrix: sample covariance, graphical lasso (Glasso) of Friedman et al. (2008), CONCORD, condition number regularized estimator (CondReg) of Won et al. (2012), and the Ledoit-Wolf estimator of Ledoit and Wolf (2004). We consider various choices of N_{est}, in particular, $N_{\text{est}} \in \{35, 40, 45, 50, 75, 150, 225, 300\}$ in our analysis. Note that once a choice for N_{est} is made, it is kept constant for all the 231 investment periods.

Note that for ℓ_1-penalized regression methods such as the Glasso and CONCORD methods, a value for the penalty parameter has to be chosen. For the purposes of this study, cross-validation was performed within each estimation horizon so as to minimize the residual sum of squares from out-of-sample prediction averaged over all stocks. Further details

---

[2]KFT was a component stock of the DJIA form 9/22/2008 to 9/13/2012. From 9/14/2012, KFT was replaced with United Health Group (UNH).

N_{est}	Sample	Glasso	CONCORD	CondReg	Ledoit-Wolf	DJIA
35	0.357	**0.489**	**0.487**	**0.486**	0.470	0.185
40	0.440	**0.491**	**0.490**	0.473	0.439	0.185
45	0.265	0.468	**0.473**	0.453	0.388	0.185
50	0.234	**0.481**	**0.482**	0.458	0.407	0.185
75	0.379	0.403	**0.475**	0.453	0.368	0.185
150	0.286	0.353	**0.480**	0.476	0.384	0.185
225	0.367	0.361	**0.502**	0.494	0.416	0.185
300	0.362	0.359	**0.505**	0.488	0.409	0.185

Table 5: Realized Sharpe ratio of different investment strategies corresponding to different estimators with various N_{est}. The maximum annualized Sharpe ratios for each row, and others within 1% of this maximum, are highlighted in bold.

are given in Supplemental Section J.4. The condition number regularized (CondReg) and Ledoit-Wolf estimators each use different criteria to perform cross-validation. The readers is referred to Won et al. (2012) and Ledoit and Wolf (2004) for details on the cross-validation procedure for these methods.

For comparison purposes with Won et al. (2012), we use the following quantities to assess the performance of the five MVR strategies: *Realized return, Realized risk, Realized Sharpe ratio (SR), Turnover, Size of the short side* and *Normalized wealth growth.* Precise definitions of these quantities are given in Supplemental Section J.5.

Table 5 gives the realized Sharpe ratios of all MVR strategies for the different choices of estimation horizon N_{est}. The column DJIA stands for the passive index tracking strategy that tracks the Dow Jones industrial average index. It is clear from Table 5 that the CONCORD method performs uniformly well across different choices of estimation horizons.

Figure 2 shows normalized wealth growth over the trading horizon for the choice $N_{\text{est}} = 225$. Normalized wealth growth curve for another choice $N_{\text{est}} = 75$ is provided in Supplemental section J.5. These plots demonstrate that CONCORD is either very competitive or better than leading covariance estimation methods.

We also note that trading costs associated with CONCORD are the lowest for most choices of estimation horizons, and are very comparable with CondReg for $N_{\text{est}} = \{35, 40\}$ (See Table 12 in Supplemental Section J.5). Moreover, CONCORD also has by far the lowest short side for most choices of estimation horizons. This property reduces the dependence on borrowed capital for shorting stocks and is also reflected in the higher normalized wealth growth.
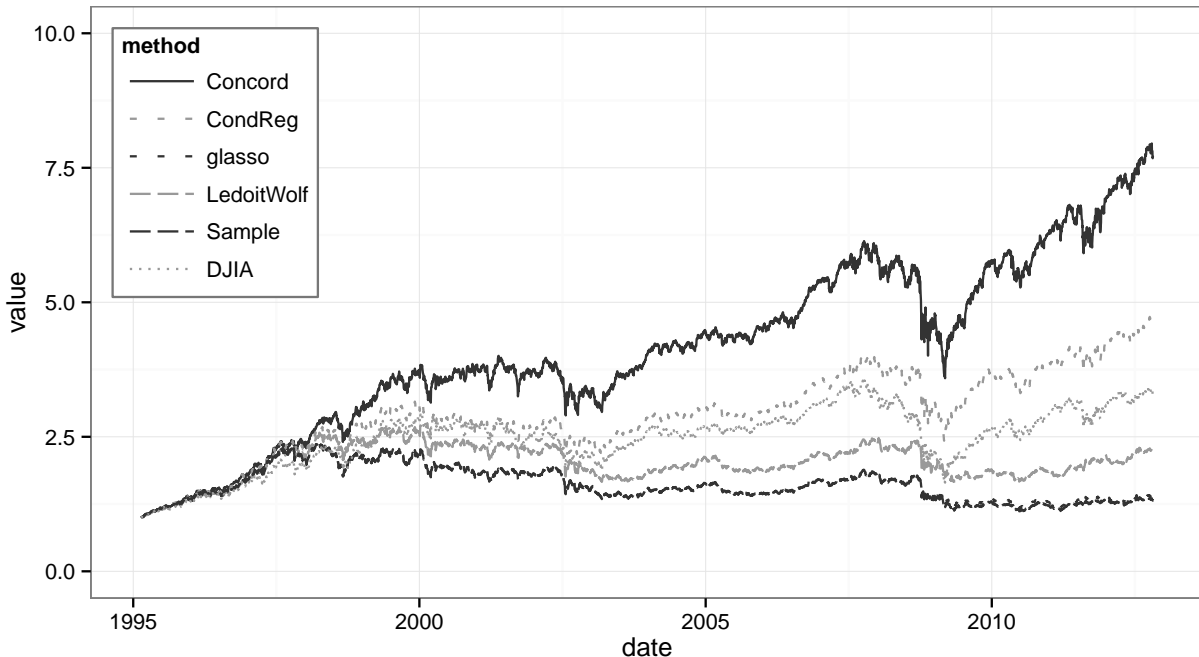
Figure 2: Normalized wealth growth after adjusting for transaction costs (0.5% of principal) and borrowing costs (interest rate of 7% APR) with $N_{\text{est}} = 225$.

6 Large sample properties

In this section, large sample properties of the CONCORD algorithm, estimation consistency and oracle properties under suitable regularity conditions are investigated. We adapt the approach in Peng et al. (2009) with suitable modifications. Now let the dimension $p = p_n$ vary with n so that our treatment is relevant to high dimensional settings. Let $\{\bar{\Omega}_n\}_{n \geq 1}$ denote the sequence of true inverse covariance matrices. As in Peng et al. (2009), for consistency purposes, we assume the existence of suitably accurate estimates of the diagonal entries, and consider the accuracy of the estimates of the off-diagonal entries obtained after running the CONCORD algorithm with diagonal entries fixed. In particular, the following assumption is made:

- (A0 - Accurate diagonal estimates) There exist estimates $\{\widehat{\alpha}_{n,ii}\}_{1 \leq i \leq p_n}$ such that for any $\eta > 0$, there exists a constant $C > 0$ such that

$$\max_{1 \leq i \leq p_n} |\widehat{\alpha}_{n,ii} - \bar{\omega}_{ii}| \leq C \left(\sqrt{\frac{\log n}{n}} \right),$$

holds with probability larger than $1 - O(n^{-\eta})$.

26

Note that the theory that follows is valid when the estimates $\{\widehat{\alpha}_{n,ii}\}_{1\leq i\leq p_n}$ and the estimates of the off-diagonal entries are obtained from the same dataset. When $\limsup_{n\to\infty}\frac{p_n}{n} < 1$, Peng et al. (2009) show that the diagonal entries of S^{-1} can be used as estimates of the diagonal entries of Ω. However, no such general recipe is provided in Peng et al. (2009) for the case $p_n > n$. Nevertheless, establishing consistency in the above framework is useful, as it indicates that the estimators obtained are statistically well-behaved when n and p both increase to infinity.

For vectors $\omega^o \in \mathbb{R}^{\frac{p_n(p_n-1)}{2}}$ and $\omega^d \in \mathbb{R}^{p_n}_+$, the notation $\mathcal{L}_n(\omega^o, \omega^d)$ stands for $\frac{\mathcal{L}_{\mathrm{con}}}{n}$ ($\mathcal{L}_{\mathrm{con}}$ is defined in (11)) evaluated at a matrix with off-diagonal entries ω^o and diagonal entries ω^d. Let $\bar{\omega}^o_n = ((\bar{\omega}_{n,ij}))_{1\leq i<j\leq p_n}$ denote the vector of off-diagonal entries of $\bar{\Omega}_n$, and $\widehat{\boldsymbol{\alpha}}_{p_n} \in \mathbb{R}^{p_n}_+$ denotes the vector with entries $\{\widehat{\alpha}_{n,ii}\}_{1\leq i\leq p_n}$. Let $\mathcal{A}_n$ denote the set of non-zero entries in the vector $\bar{\omega}^o_n$, and let $q_n = |\mathcal{A}_n|$. Let $\bar{\Sigma}_n = \bar{\Omega}^{-1}_n$ denote the true covariance matrix for every $n \geq 1$. The following standard assumptions are required.

- (A1 - Bounded eigenvalues) The eigenvalues of $\bar{\Omega}_n$ are bounded below by $\lambda_{min} > 0$, and bounded above by $\lambda_{max} < \infty$ uniformly for all n.

- (A2 - Sub Gaussianity) The random vectors $\mathbf{Y}^1, \ldots, \mathbf{Y}^n$ are *i.i.d.* sub-Gaussian for every $n \geq 1$, i.e., there exists a constant $c > 0$ such that for every $\mathbf{x} \in \mathbb{R}^{p_n}$, $E\left[e^{\mathbf{x}'\mathbf{Y}^i}\right] \leq e^{c\mathbf{x}'\bar{\Sigma}_n\mathbf{x}}$, and for every $i,j > 0$, there exists $\eta_j > 0$ such that $E\left[e^{t(Y^i_j)^2}\right] < K$ whenever $|t| < \eta_j$. Here K is independent of i and j.

- (A3 - Incoherence condition) There exists $\delta < 1$ such that for all $(i,j) \notin \mathcal{A}_n$,

$$\left| \bar{\mathcal{L}}''_{ij,\mathcal{A}_n}(\bar{\Omega}_n) \left[\bar{\mathcal{L}}''_{\mathcal{A}_n,\mathcal{A}_n}(\bar{\Omega}_n) \right]^{-1} \mathrm{sign}(\bar{\omega}^o_{\mathcal{A}_n}) \right| \leq \delta,$$

where for $1 \leq i,j,t,s \leq p_n$ satisfying $i < j$ and $t < s$,

$$\bar{\mathcal{L}}''_{ij,ts}(\bar{\Omega}_n) := E_{\bar{\Omega}_n}\left((\mathcal{L}''_n(\bar{\Omega}_n))_{ij,ts} \right) = \bar{\Sigma}_{n,js}1_{\{i=t\}} + \bar{\Sigma}_{n,it}1_{\{j=s\}} + \bar{\Sigma}_{n,is}1_{\{j=t\}} + \bar{\Sigma}_{n,jt}1_{\{i=s\}}.$$

Conditions analogous to (A3) have been used in Zhao and Yu (2006), Peng et al. (2009), Meinshausen and Bühlmann (2006) to establish high-dimensional model selection consistency. In the context of lasso regression, Zhao and Yu (2006) show that such a condition (which they refer to as an irrepresentable condition) is almost necessary and sufficient for model selection consistency, and provide some examples when this condition is satisfied. We provide some examples of situations where the condition (A3) is satisfied, along the lines of Zhao and Yu (2006), in Supplemental section M.

27

Define $\bar{\theta}_n^o = ((\bar{\theta}_{n,ij}))_{1 \leq i < j \leq p_n} \in \mathbb{R}^{p_n(p_n-1)/2}$ by $\bar{\theta}_{n,ij} = \frac{\bar{\omega}_{n,ij}}{\sqrt{\hat{\alpha}_{n,ii}\hat{\alpha}_{n,jj}}}$ for $1 \leq i < j \leq p_n$. Let $s_n = \min_{(i,j)\in\mathcal{A}_n} \bar{\omega}_{n,ij}$. The assumptions above can be used to establish the following theorem.

Theorem 2. *Suppose that assumptions (A0)-(A3) are satisfied. Suppose $p_n = O(n^\kappa)$ for some $\kappa > 0$, $q_n = o\left(\sqrt{n/\log n}\right)$, $\sqrt{\frac{q_n \log n}{n}} = o(\lambda_n)$, $\lambda_n\sqrt{n/\log n} \to \infty$, $\frac{s_n}{\sqrt{q_n}\lambda_n} \to \infty$ and $\sqrt{q_n}\lambda_n \to 0$, as $n \to \infty$. Then there exists a constant C such that for any $\eta > 0$, the following events hold with probability at least $1 - O(n^{-\eta})$.*

- *There exists a minimizer $\hat{\omega}_n^o = ((\hat{\omega}_{n,ij}))_{1 \leq i < j \leq p_n}$ of $Q_{con}(\omega^o, \hat{\boldsymbol{\alpha}}_n)$.*

- *Any minimizer $\hat{\omega}_n^o$ of $Q_{con}(\omega^o, \hat{\boldsymbol{\alpha}}_n)$ satisfies $\|\hat{\omega}_n^o - \bar{\omega}_n^o\|_2 \leq C\sqrt{q_n}\lambda_n$ and $\mathrm{sign}(\hat{\omega}_{n,ij}) = \mathrm{sign}(\bar{\omega}_{n,ij})$, $\forall\ 1 \leq i < j \leq p_n$.*

The proof of the above theorem is provided in Supplemental section K.

7 Conclusion

This paper proposes a novel regression based graphical model selection method that aims to overcome some of the shortcomings of current methods, but at the same time retain their respective strengths. We first place the highly useful SPACE method in an optimization framework, which in turn allows us to identify SPACE with a specific objective function. These and other insights lead to the formulation of the CONCORD objective function. It is then shown that the CONCORD objective function is comprised of quadratic forms, is convex, and can be regarded as a penalized pseudo-likelihood. A coordinate-wise descent algorithm that minimizes this objective, via closed form iterates, is proposed, and subsequently analyzed. The convergence of this coordinate-wise descent algorithm is established rigorously, thus ensuring that CONCORD leads to well defined symmetric partial correlation estimates that are always computable - a guarantee that is not available with popular regression based methods. Large sample properties of CONCORD establish consistency of the method as both the sample size and dimension tend to infinity. The performance of CONCORD is also illustrated via simulations and is shown to be competitive in terms of graphical model selection accuracy and timing. CONCORD is then applied to a biomedical dataset and to a finance dataset, leading to novel findings. Last but not least, a framework that unifies all pseudo-likelihood methods is established, yielding important insights.

Given the attractive properties of CONCORD, a natural question that arises is whether one should move away from penalized likelihood estimation (such as Glasso) and rather use only pseudo-likelihood methods. We note that CONCORD is attractive over Glasso for several reasons: Firstly, it does not assume Gaussianity and is hence more flexible. Secondly,

the computational complexity per iteration of CONCORD is lower than that of Glasso. Thirdly, CONCORD is faster (in terms of wall clock time) than Glasso by an entire order of magnitude in higher dimensions. Fourthly, CONCORD delivers better model selection performance. It is however important to note that if there is a compelling reason to assume multivariate Gaussianity (which some applications may warrant), then using both Glasso and CONCORD can potentially be useful for affirming multivariate associations of interest. In this sense, the two classes of methods could be complementary in many practical applications.

References

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.

Besag, J. (1975). Statistical Analysis of Non-Lattice Data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 24(3):179–195.

Bibby, R. A., Tang, C., Faisal, A., Drosopoulos, K., Lubbe, S., Houlston, R., Bayliss, R., and Linardopoulos, S. (2009). A cancer-associated aurora A mutant is mislocalized and misregulated due to loss of interaction with TPX2. *The Journal of Biological Chemistry*, 284(48):33177–84.

Carter, S. L., Brechbühler, C. M., Griffin, M., and Bond, A. T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics (Oxford, England)*, 20(14):2242–50.

Chan, L. K., Karceski, J., and Lakonishok, J. (1999). On portfolio optimization: Forecasting covariances and choosing the risk model. Working Paper 7039, National Bureau of Economic Research.

Chang, H. Y. et al. (2005). Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proceedings of the National Academy of Sciences of the United States of America*, 102(10):3738–3743.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

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

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Applications of the lasso and grouped lasso to the estimation of sparse graphical models. Technical report, Stanford University.

Han, J.-D. J. et al. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):88–93.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning.* Springer.

Jensen, S. r. T., Johansen, S. r., and Lauritzen, S. L. (1991). Globally Convergent Algorithms for Maximizing Likelihood Function. *Biometrika*, 78(4):867–877.

Jeong, H., Mason, S. P., Barabasi, A.-L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41–42.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press, USA.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Lee, J. D. and Hastie, T. J. (2014). Learning the structure of mixed graphical models. *to appear in Journal of Computational and Graphical Statistics*.

Luenberger, D. G. (1997). *Investment Science*. Oxford University Press, USA.

Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance*, 7(1):77–91.

Maxwell, C. A., Bentez, J., Gmez-Bald, L., Osorio, A., Bonifaci, N., Fernndez-Ramires, R., Costes, S. V., Guin, E., Chen, H., Evans, G. J. R., Mohan, P., Catal, I., Petit, A., Aguilar, H., Villanueva, A., Aytes, A., Serra-Musach, J., Rennert, G., Lejbkowicz, F., Peterlongo, P., Manoukian, S., Peissel, B., Ripamonti, C. B., Bonanni, B., Viel, A., Allavena, A., Bernard, L., Radice, P., Friedman, E., Kaufman, B., Laitman, Y., Dubrovsky, M., Milgrom, R., Jakubowska, A., Cybulski, C., Gorski, B., Jaworska, K., Durda, K., Sukiennicki, G., Lubiski, J., Shugart, Y. Y., Domchek, S. M., Letrero, R., Weber, B. L., Hogervorst, F. B. L., Rookus, M. A., Collee, J. M., Devilee, P., Ligtenberg, M. J., van der Luijt, R. B., Aalfs, C. M., Waisfisz, Q., Wijnen, J., van Roozendaal, C. E. P., Easton, D. F., Peock, S., Cook, M., Oliver, C., Frost, D., Harrington, P., Evans, D. G., Lalloo, F., Eeles, R., Izatt, L., Chu, C., Eccles, D., Douglas, F., Brewer, C., Nevanlinna, H., Heikkinen, T., Couch, F. J., Lindor, N. M., Wang, X., Godwin, A. K., Caligo, M. A., Lombardi, G., Loman, N., Karlsson, P., Ehrencrona, H., von Wachenfeldt, A., Bjork Barkardottir, R., Hamann, U., Rashid, M. U., Lasa, A., Calds, T., Andrs, R., Schmitt, M., Assmann, V., Stevens, K., Offit, K., Curado, J., Tilgner, H., Guig, R., Aiza, G., Brunet, J., Castellsagu, J., Martrat, G., Urruticoechea, A., Blanco, I., Tihomirova, L., Goldgar, D. E., Buys, S., John, E. M., Miron, A., Southey, M., Daly, M. B., Schmutzler, R. K., Wappenschmidt, B., Meindl, A., Arnold, N., Deissler, H., Varon-Mateeva, R., Sutter, C., Niederacher, D., Imyamitov, E., Sinilnikova, O. M., Stoppa-Lyonne, D., Mazoyer, S., Verny-Pierre, C., Castera, L., de Pauw, A., Bignon, Y.-J., Uhrhammer, N., Peyrat, J.-P., Vennin, P., Fert Ferrer, S., Collonge-Rame, M.-A., Mortemousque, I., Spurdle, A. B., Beesley, J., Chen, X., Healey,

S., Barcellos-Hoff, M. H., Vidal, M., Gruber, S. B., Lzaro, C., Capell, G., McGuffog, L., Nathanson, K. L., Antoniou, A. C., Chenevix-Trench, G., Fleisch, M. C., Moreno, V., Pujana, M. A., HEBON, EMBRACE, SWE-BRCA, BCFR, GEMO Study Collaborators, and kConFab (2011). Interplay between brca1 and rhamm regulates epithelial apicobasal polarization and may influence risk of breast cancer. *PLoS Biol*, 9(11):e1001199.

Mazumder, R. and Hastie, T. (2012). Exact Covariance Thresholding into Connected Components for Large-Scale Graphical Lasso. *The Journal of Machine Learning Research*, 13:781–794.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462.

Merton, R. C. (1980). On estimating the expected return on the market: An exploratory investigation. Working Paper 444, National Bureau of Economic Research.

Newman, M. (2003). The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial Correlation Estimation by Joint Sparse Regression Models. *Journal of the American Statistical Association*, 104(486):735–746.

Rocha, G., Zhao, P., and Yu, B. (2008). A path following algorithm for Sparse Pseudo-Likelihood Inverse Covariance Estimation (SPLICE). Technical report, Statistics Department, UC Berkeley, Berkeley, CA.

Speed, T. P. and Kiiveri, H. T. (1986). Gaussian Markov Distributions over Finite Graphs. *The Annals of Statistics*, 14(1):138–150.

Tseng, P. (1988). Coordinate ascent for maximizing nondifferentiable concave functions. Technical report, Massachusetts Institute of Technology.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494.

Won, J.-H., Lim, J., Kim, S.-J., and Rajaratnam, B. (2012). Condition Number Regularized Covariance Estimation. *Journal of the Royal Statistical Society: Series B*.

Xu, P.-F., Guo, J., and He, X. (2011). An Improved Iterative Proportional Scaling Procedure for Gaussian Graphical Models. *Journal of Computational and Graphical Statistics*, 20(2):417–431.

Zangwill, W. (1969). *Nonlinear programming: a unified approach.* Prentice-Hall international series in management. Prentice-Hall, Englewood Cliffs, NJ.

Zhao, P. and Yu, B. (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Supplemental Section

A   Proof of Lemma 2

Let $\mathbf{Y}$ denote the $n \times p$ matrix with j^{th} column given by $\mathbf{Y}_j$ for $j = 1, 2, \ldots, p$. Define $Q_{\text{sym}}(\boldsymbol{\alpha}, \breve{\Omega}) = \frac{1}{2} \left(\sum_{j=1}^{p} \mathcal{L}_{\text{sym},j}(\alpha_{jj}, \breve{\Omega}_j) \right) + \lambda \left(\sum_{1 \leq i < j \leq p} |\omega_{ij}| \right)$ so that

$$\mathcal{L}_{\text{sym},j}(\alpha_{jj}, \breve{\Omega}_j) = n \log \alpha_{jj} + \frac{1}{\alpha_{jj}} \|\mathbf{Y}_j + \mathbf{Y}\breve{\Omega}_j \alpha_{jj}\|_2^2 \tag{24}$$

where $\boldsymbol{\alpha} = (\alpha_{11} \ \alpha_{22} \ \cdots \ \alpha_{pp})'$, $\alpha_{jj} = 1/\omega_{jj}$ and $\breve{\Omega}_j$ is the jth column of $\breve{\Omega}$. Recall that $\breve{\Omega}$ is the matrix Ω with zeros in place of the diagonal entries. If follows that

$$\frac{\partial Q_{\text{sym}}(\boldsymbol{\alpha}, \breve{\Omega})}{\partial \alpha_{jj}} = \frac{n}{\alpha_{jj}} - \frac{\mathbf{Y}_j'\mathbf{Y}_j}{\alpha_{jj}^2} + \breve{\Omega}_j'\mathbf{Y}'\mathbf{Y}\breve{\Omega}_j, \quad \text{and} \quad \frac{\partial^2 Q_{\text{sym}}(\boldsymbol{\alpha}, \breve{\Omega})}{\partial \alpha_{jj}^2} = -\frac{n}{\alpha_{jj}^2} + 2\frac{\mathbf{Y}_j'\mathbf{Y}_j}{\alpha_{jj}^3} \tag{25}$$

It is clear that in general $\partial^2 Q_{\text{sym}}(\boldsymbol{\alpha}, \breve{\Omega})/\partial \alpha_{jj}^2 \not\geq 0$. Hence, $Q_{\text{sym}}(\boldsymbol{\alpha}, \breve{\Omega})$ is not convex.

B   Proof of Lemma 3

Proof. i) Rewrite the SPLICE objective function $Q_{\text{spl}}(\mathbf{B}, \mathbf{D}) = \mathcal{L}_{\text{spl}}(\mathbf{B}, \mathbf{D}) + \lambda \sum_{i<j} |\beta_{ij}|$ where

$$\mathcal{L}_{\text{spl}}(\mathbf{B}, \mathbf{D}) = \frac{1}{2} \left[n \log \det(\mathbf{D}^2) + \text{tr}(\mathbf{D}^{-2}\mathbf{A}) \right],$$

and $\mathbf{A} = [a_{ij}] = (\mathbf{I} - \mathbf{B})\mathbf{Y}'\mathbf{Y}(\mathbf{I} - \mathbf{B}')$. The function $\mathcal{L}_{\text{spl}}(\mathbf{B}, \mathbf{D})$ with all variables fixed except d_{jj} is given by

$$\mathcal{L}_{\text{spl},j}(\mathbf{B}, d_{jj}) = \frac{1}{2} \left[n \log d_{jj}^2 + \frac{a_{jj}}{d_{jj}^2} \right] + \text{constants}.$$

Now,

$$\frac{\partial Q_{\text{spl}}(\mathbf{B}, \mathbf{D})}{\partial d_{jj}} = \frac{n}{d_{jj}} - \frac{a_{jj}}{d_{jj}^3}$$

$$\frac{\partial^2 Q_{\text{spl}}(\mathbf{B}, \mathbf{D})}{\partial d_{jj}^2} = -\frac{n}{d_{jj}^2} + 3\frac{a_{jj}}{d_{jj}^4}$$

It is clear in general $\partial Q_{\text{spl}}^2(\mathbf{B}, \mathbf{D})/\partial d_{jj}^2 \not\geq 0$. Hence $Q_{\text{spl}}(\mathbf{B}, \mathbf{D})$ is not convex.

ii) Similarly, define $Q_{\mathrm{spl}}^*(\mathbf{B}, \mathbf{C}) = \mathcal{L}_{\mathrm{spl}}^*(\mathbf{B}, \mathbf{C}) + \lambda \sum_{i<j} |\beta_{ij}|$ where

$$\mathcal{L}_{\mathrm{spl}}^*(\mathbf{B}, \mathbf{C}) = \frac{1}{2} \left[n \log \mathbf{C}^{-2} + \mathrm{tr}(\mathbf{C}^2 \mathbf{A}) \right].$$

It is clear that for a fixed $\mathbf{C}$, $\mathcal{L}_{\mathrm{spl}}^*(\mathbf{B}, \mathbf{C})$ is a convex function in $\mathbf{B}$ (Rocha et al., 2008). Now for a fixed $\mathbf{B}$ let

$$\mathcal{L}_{\mathrm{spl},j}^*(\mathbf{B}, c_{jj}) = \frac{1}{2} \left[-2n \log c_{jj} + c_{jj}^2 a_{jj} \right] + \text{constants}$$

$$\frac{\partial Q_{\mathrm{spl}}^*(\mathbf{B}, \mathbf{C})}{\partial c_{jj}} = -\frac{n}{c_{jj}} + c_{jj} a_{jj}$$

$$\frac{\partial^2 Q_{\mathrm{spl}}^*(\mathbf{B}, \mathbf{C})}{\partial c_{jj}^2} = \frac{n}{c_{jj}^2} + a_{jj}$$

Now, note that $\partial (Q_{\mathrm{spl}}^*)^2 (\mathbf{B}, \mathbf{C}) / \partial c_{jj}^2 \geq 0$ since $a_{jj} \geq 0$.

To see that $a_{jj} \geq 0$ note that $\mathbf{A} = (\mathbf{I} - \mathbf{B}) \mathbf{Y}' \mathbf{Y} (\mathbf{I} - \mathbf{B}') = \mathbf{G}' \mathbf{G}$, where $\mathbf{G} = \mathbf{Y}(\mathbf{I} - \mathbf{B}')$ Now, $a_{jj} = \mathbf{G}_{\bullet j}' \mathbf{G}_{\bullet j} = \|\mathbf{G}_{\bullet j}\|^2 \geq 0$ $\qquad\square$

C   Proof of Lemma 4

Note that for $1 \leq i \leq p$,

$$Q_{\mathrm{con}}(\Omega) = -n \log \omega_{ii} + \frac{n}{2} \left(\omega_{ii}^2 s_{ii} + 2\omega_{ii} \sum_{j \neq i} \omega_{ij} s_{ij} \right) + \text{ terms independent of } \omega_{ii}. \qquad (26)$$

where $s_{ij} = \mathbf{Y}_i' \mathbf{Y}_j / n$. Hence,

$$\frac{\partial}{\partial \omega_{ii}} Q_{\mathrm{con}}(\Omega) = 0 \quad \Leftrightarrow \quad -\frac{1}{\omega_{ii}} + \omega_{ii} s_{ii} + \sum_{j \neq i} \omega_{ij} s_{ij} = 0$$

$$\Leftrightarrow \quad \omega_{ii} = \frac{-\sum_{j \neq i} \omega_{ij} s_{ij} + \sqrt{\left(\sum_{j \neq i} \omega_{ij} s_{ij} \right)^2 + 4 s_{ii}}}{2 s_{ii}},$$

Note that since $\omega_{ii} > 0$ the positive root has been retained as the solution.

Also, for $1 \leq i < j \leq p$,

$$Q_{\mathrm{con}}(\Omega) = n \frac{s_{ii} + s_{jj}}{2} \omega_{ij}^2 + n \left(\sum_{j' \neq j} \omega_{ij'} s_{jj'} + \sum_{i' \neq i} \omega_{i'j} s_{ii'} \right) \omega_{ij} + \lambda |\omega_{ij}| + \text{ terms independent of } \omega_{ij}.$$

$$(27)$$

It follows that

$$(T_{ij}(\Omega))_{ij} = \frac{S_{\frac{\lambda}{n}}\left(-\left(\sum_{j'\neq j}\omega_{ij'}s_{jj'} + \sum_{i'\neq i}\omega_{i'j}s_{ii'}\right)\right)}{s_{ii} + s_{jj}},$$

where S_η is the soft-thresholding operator given by $S_\eta(x) = \text{sign}(x)(|x| - \eta)_+$.

D Proof of Lemma 5

Let $\mathbf{Y}_j$ denote j^{th} column of the data matrix $\mathbf{Y}$. Then, using the identity $\sum_{k=1}^{p}\omega_{ik}s_{jk} = \omega_{ij}s_{jj} + \sum_{k\neq j}\omega_{ik}s_{jk} = \omega_{ii}s_{ij} + \sum_{k\neq i}\omega_{ik}s_{jk}$,

$$\sum_{k\neq j}\omega_{ik}s_{jk} = -\omega_{ij}s_{jj} + \omega_{ii}\left(s_{ij} + \sum_{k\neq i}\frac{\omega_{ik}}{\omega_{ii}}s_{jk}\right)$$

$$= -\omega_{ij}s_{jj} + \omega_{ii}\mathbf{Y}'_j\left(\mathbf{Y}_i + \sum_{k\neq i}\frac{\omega_{ik}}{\omega_{ii}}\mathbf{Y}_k\right)$$

$$= -\omega_{ij}s_{jj} + \omega_{ii}\mathbf{Y}'_j\mathbf{r}_i,$$

where $\mathbf{r}_i = \mathbf{Y}_i + \sum_{k\neq i}\frac{\omega_{ik}}{\omega_{ii}}\mathbf{Y}_k$ is an n-vector of residuals after regressing the ith variable on the rest. $\qquad \square$

E Proof of Lemma 6

1. Result follows easily from inspecting $\mathbf{r}_k$ and $\mathbf{r}_l$.

2. If ω_{kl} is updated to ω^*_{kl}, it follows from part 1 that among all the residual vectors, only $\mathbf{r}_k$ and $\mathbf{r}_l$ change values. The residual vector $\mathbf{r}_k$ can be updated as follows:

$$\mathbf{r}^*_k = \mathbf{r}_k + \frac{(\omega^*_{kl} - \omega_{kl})}{\omega_{kk}}\mathbf{Y}_l.$$

   Clearly, this update requires $O(n)$ operations. The vector $\mathbf{r}_l$ can be updated similarly.

3. Result follows easily from inspecting $\mathbf{r}_k$.

4. If ω_{kk} is updated to ω^*_{kk}, it follows from part 3 that among all the residual vectors, only $\mathbf{r}_k$ changes value. The residual vector $\mathbf{r}_k$ can be updated as follows:

$$\mathbf{r}^*_k = (\mathbf{r}_k - \mathbf{Y}_k)\frac{\omega_{kk}}{\omega^*_{kk}} + \mathbf{Y}_k.$$

Clearly, this update requires $O(n)$ operations. □

F Proof of Lemma 7

Proof. (**CONCORD**) Let $A = nS$ Expanding the ℓ_2-norm of the residual, we have

$$\|\omega_{ii}\mathbf{Y}_i + \sum_{j\neq i}\omega_{ij}\mathbf{Y}_j\|_2^2 = \|\sum_{j=1}^{p}\omega_{ij}\mathbf{Y}_j\|_2^2 = \|\mathbf{Y}\omega_{i\bullet}\|_2^2 = \omega_{i\bullet}'\mathbf{Y}'\mathbf{Y}\omega_{i\bullet} = \omega_{i\bullet}'\mathbf{A}\omega_{i\bullet}$$

Hence, (18) is equivalent to

$$\mathcal{L}_{\text{con}}(\Omega) = \frac{1}{2}\sum_{i=1}^{p}(-2n\log\omega_{ii} + \omega_{i\bullet}'\mathbf{A}\omega_{i\bullet}) = -n\sum_{i=1}^{p}\log\omega_{ii} + \frac{1}{2}\sum_{i=1}^{p}\omega_{i\bullet}'A\omega_{i\bullet}$$

$$= -n\log\left(\prod_{i=1}^{p}\omega_{ii}\right) + \frac{n}{2}\operatorname{tr}(\Omega\mathbf{S}\Omega)$$

$$= \frac{n}{2}\left(-\log\det\Omega_D^2 + \operatorname{tr}(\mathbf{S}\Omega^2)\right).$$

Hence, $G_{\text{con}}(\Omega) = \Omega_D$ and $H_{\text{con}}(\Omega) = \Omega^2$

(**SPACE with unit weights**) Reparameterizing (19) using the identity $-\rho^{ij}\sqrt{\omega_{jj}/\omega_{ii}} = \omega_{ij}/\omega_{ii}$, the ℓ_2-norm of the residual can be expressed as follows.

$$\|\mathbf{Y}_i + \sum_{j\neq i}\frac{\omega_{ij}}{\omega_{ii}}\mathbf{Y}_j\|_2^2 = \|\frac{1}{\omega_{ii}}(\omega_{ii}\mathbf{Y}_i + \sum_{j\neq i}\omega_{ij}\mathbf{Y}_j)\|_2^2 = \frac{1}{\omega_{ii}^2}\omega_{i\bullet}'\mathbf{A}\omega_{i\bullet}$$

Hence, (19) is equivalent to

$$\mathcal{L}_{\text{spc},1}(\Omega) = -\frac{n}{2}\log\det\Omega_D + \frac{1}{2}\sum_{i=1}^{p}\frac{1}{\omega_{ii}^2}\omega_{i\bullet}'\mathbf{A}\omega_{i\bullet}$$

$$= -\frac{n}{2}\log\det\Omega_D + \frac{n}{2}\sum_{i=1}^{p}\frac{\omega_{i\bullet}'}{\omega_{ii}}\mathbf{S}\frac{\omega_{i\bullet}}{\omega_{ii}}$$

$$= -\frac{n}{2}\log\det\Omega_D + \frac{1}{2}\operatorname{tr}(\Omega_D^{-1}\Omega\mathbf{A}\Omega\Omega_D^{-1})$$

$$= \frac{n}{2}\left(-\log\det\Omega_D + \operatorname{tr}(\mathbf{S}\Omega\Omega_D^{-2}\Omega)\right).$$

Therefore, $G_{\text{spc},1}(\Omega) = \Omega_D$ and $H_{\text{spc},1}(\Omega) = \Omega\Omega_D^{-2}\Omega$.

(SPACE with ω_{ii} weights) Similar to the analysis for SPACE1 with unit weights, the ℓ_2-norm of the residual for the SPACE2 formulation (i.e., with weights ω_{ii}) can be expressed as follows.

$$\omega_{ii}\|\mathbf{Y}_i - \sum_{j\neq i}\rho^{ij}\sqrt{\frac{\omega_{jj}}{\omega_{ii}}}\mathbf{Y}_j\|_2^2 = \omega_{ii}\left(\frac{1}{\omega_{ii}^2}\omega_{i\bullet}'\mathbf{A}\omega_{i\bullet}\right)$$

$$= \frac{1}{\omega_{ii}}\omega_{i\bullet}'\mathbf{A}\omega_{i\bullet}$$

Hence, (20) is equivalent to

$$\begin{aligned}
\mathcal{L}_{\text{spc},2}(\Omega) &= -\frac{n}{2}\log\det\Omega_D + \frac{1}{2}\sum_{i=1}^{p}\frac{1}{\omega_{ii}}\omega_{i\bullet}'\mathbf{A}\omega_{i\bullet}\\
&= -\frac{n}{2}\log\det\Omega_D + \frac{n}{2}\sum_{i=1}^{p}\frac{\omega_{i\bullet}'}{\sqrt{\omega_{ii}}}\mathbf{S}\frac{\omega_{i\bullet}}{\sqrt{\omega_{ii}}}\\
&= -\frac{n}{2}\log\det\Omega_D + \frac{n}{2}\operatorname{tr}(\Omega_D^{-1/2}\Omega\mathbf{S}\Omega\Omega_D^{-1/2})\\
&= \frac{n}{2}\left(-\log\det\Omega_D + \operatorname{tr}(\mathbf{S}\Omega\Omega_D^{-1}\Omega)\right)
\end{aligned}$$

Therefore, $G_{\text{spc},2}(\Omega) = \Omega_D$ and $H_{\text{spc},2}(\Omega) = \Omega\Omega_D^{-1}\Omega$.

(SYMLASSO) Reparameterizing (21) by $\alpha_{ii} = 1/\omega_{ii}$ and $-\rho^{ij}\sqrt{\omega_{jj}/\omega_{ii}} = \omega_{ij}/\omega_{ii}$ yields (20). It follows that $G_{\text{sym}}(\Omega) = \Omega_D$, $H_{\text{sym}}(\Omega) = \Omega\Omega_D^{-1}\Omega$.

(SPLICE) Reparameterizing (22) by $d_{ii}^2 = 1/\omega_{ii}$ and $\beta_{ij} = \rho^{ij}\sqrt{\omega_{jj}/\omega_{ii}}$ yields (20). It follows that $G_{\text{spl}}(\Omega) = \Omega_D$, $H_{\text{spl}}(\Omega) = \Omega\Omega_D^{-1}\Omega$. $\qquad\square$

G Effect of correction factor

Following steps similar to proof of Lemma 4, the update formulas for $\bar{Q}_{\text{con}}(\Omega) = \mathcal{L}_{\text{con}}(\Omega) + \lambda\sum_{i<j}|\omega_{ij}|$ of (18) can be shown to be

$$(T_{kk}(\Omega))_{kk} = \frac{-\sum_{j\neq k}\omega_{kj}s_{kj} + \sqrt{\left(\sum_{j\neq k}\omega_{kj}s_{kj}\right)^2 + 2s_{kk}}}{2s_{kk}} \tag{28}$$

$$(T_{kl}(\Omega))_{kl} = \frac{S_{\frac{\lambda}{n}}\left(-\left(\sum_{j\neq l}\omega_{kj}s_{jl} + \sum_{j\neq k}\omega_{lj}s_{jk}\right)\right)}{s_{kk} + s_{ll}} \tag{29}$$

G.1 Numerical example

Analysis on a dataset ($n = 1000$) generated from following Ω was used for this example.

$$\Omega = \begin{pmatrix} 1.0 & 0.3 & 0.0 \\ 0.3 & 1.0 & 0.3 \\ 0.0 & 0.3 & 1.0 \end{pmatrix}$$

Without penalty, i.e. $\lambda = 0$, computed solutions Ω_{con} from using CONCORD and $\Omega_{\text{uncorrected}}$ from using update formulas (28) and (29) are

$$\Omega_{\text{uncorrected}} = \begin{pmatrix} 0.675 & 0.089 & -0.015 \\ 0.089 & 0.658 & 0.117 \\ -0.015 & 0.117 & 0.668 \end{pmatrix}, \ \Omega_{\text{con}} = \begin{pmatrix} 0.974 & 0.257 & 0.007 \\ 0.257 & 0.983 & 0.344 \\ 0.007 & 0.344 & 0.978 \end{pmatrix}$$

It is clear that the estimate Ω_{con} with the correction factor performs better parameter estimation.

H Proof of Theorem 1

Khare and Rajaratnam (2014) establish convergence of the cyclic coordinatewise minimization algorithm for a general class of objective functions. The proof of convergence for CONCORD relies on showing that the corresponding objective function is a special case of the general class of objective functions considered in Khare and Rajaratnam (2014). A more detailed version of the following argument can be found in (Khare and Rajaratnam, 2014, Section 4.1). We provide the main steps here for convenience and completeness.

Let $\mathbf{y} = \mathbf{y}(\Omega) \in \mathbb{R}^{p^2}$ denote a vectorized version of Ω obtained by shifting the corresponding diagonal entry at the bottom of each column of Ω, and then stacking the columns on top of each other. Let P^i denote the $p \times p$ permutation matrix such that $P^i \mathbf{z} = (z_1, \cdots, z_{i-1}, z_{i+1}, \cdots, z_p, z_i)$ for every $\mathbf{z} \in \mathbb{R}^p$. It follows by the definition of $\mathbf{y}$ that

$$\mathbf{y} = \mathbf{y}(\Omega) = ((P^1 \Omega_{\cdot 1})^T, (P^2 \Omega_{\cdot 2})^T, \cdots, (P^p \Omega_{\cdot p})^T)^T.$$

Let $\mathbf{x} = \mathbf{x}(\Omega) \in \mathbb{R}^{\frac{p(p+1)}{2}}$ be the symmetric version of $\mathbf{y}$, obtained by removing all ω_{ij} with $i > j$ from $\mathbf{y}$. More precisely,

$$\mathbf{x} = \mathbf{x}(\Omega) = (\omega_{11}, \omega_{12}, \omega_{22}, \cdots, \omega_{1p}, \omega_{2p}, \cdots, \omega_{pp})^T.$$

Let $\tilde{P}$ be the $p^2 \times \frac{p(p+1)}{2}$ matrix such that every entry of $\tilde{P}$ is either 0 or 1, exactly one entry in each row of $\tilde{P}$ is equal to 1, and $\mathbf{y} = \tilde{P}\mathbf{x}$. Let $\tilde{S}$ be a $p^2 \times p^2$ block diagonal matrix with p diagonal blocks, and the ith diagonal block is equal to $\tilde{S}^i := \frac{1}{2}P^i S(P^i)^T$, where $S = \frac{1}{n}\mathbf{Y}^T\mathbf{Y}$. It follows that

$$\frac{1}{2}\sum_{i=1}^{p}\Omega_{\cdot i}^T S\Omega_{\cdot i} = \frac{1}{2}\sum_{i=1}^{p}\Omega_{\cdot i}^T(P^i)^T P^i S(P^i)^T P^i\Omega_{\cdot i} \;=\; \frac{1}{2}\sum_{i=1}^{p}(P^i\Omega_{\cdot i})^T(P^i S(P^i)^T)(P^i\Omega_{\cdot i})$$
$$= \; \mathbf{y}^T\tilde{S}\mathbf{y}$$
$$= \; \mathbf{x}^T\tilde{P}^T\tilde{S}\tilde{P}\mathbf{x}. \tag{30}$$

Note that for every $1 \leq i \leq p$, the matrix $\tilde{S}^i = \frac{1}{2}P^i S(P^i)^T$ is positive semi-definite. Let $\tilde{S}^{1/2}$ denote the $p^2 \times p^2$ block diagonal matrix with p diagonal blocks, such that the ith diagonal block is given by $(\tilde{S}^i)^{1/2}$. Let $E = \tilde{S}^{1/2}\tilde{P}$. It follows by (30) that

$$\frac{1}{2}\sum_{i=1}^{p}\Omega_{\cdot i}^T S\Omega_{\cdot i} = (E\mathbf{x})^T(E\mathbf{x}). \tag{31}$$

By the definition of $\mathbf{x}(\Omega)$, we obtain

$$\omega_{ii} = x_{\frac{i(i+1)}{2}} \tag{32}$$

for every $1 \leq i \leq p$. Let

$$S_0 = \left\{ j : 1 \leq j \leq \frac{p(p+1)}{2}, \; j \neq \frac{i(i+1)}{2} \text{ for any } 1 \leq i \leq p \right\},$$

and

$$\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^{\frac{p(p+1)}{2}} : x_j \geq 0 \text{ for every } j \in S_0^c\}.$$

It follows by (11), (31) and (32) that the CONCORD algorithm can be viewed as a cyclic coordinatewise minimization algorithm for minimizing the function

$$Q_{con}(\mathbf{x}) = n\left\{\mathbf{x}^T E^T E\mathbf{x} - \sum_{i \in S_0^c}\log x_i + \frac{\lambda}{n}\sum_{j \in S_0}|x_j|\right\}, \tag{33}$$

subject to $\mathbf{x} \in \mathcal{X}$. For every $1 \leq i \leq p(p+1)/2$, there exist $1 \leq k, l \leq p$ such that $x_i = \omega_{kl}$. Note that $\|E_{\cdot i}\|^2 = \frac{S_{kk}+S_{ll}}{2} > 0$. It also follows from (Khare and Rajaratnam, 2014, Lemma 4.1) that for every $\xi \in \mathbb{R}$, the set $R_\xi := \{\mathbf{x} \in \mathcal{X} : Q_{con}(\mathbf{x}) \leq \xi\}$ is bounded in the sense that for every $i \in S_0$, x_i is uniformly bounded above and below, and for every $i \in S_0^c$, x_i

is uniformly bounded above and below (from zero). It follows by (Khare and Rajaratnam, 2014, Theorem 3.1) that the sequence of iterates produced by the CONCORD algorithm converges.

I Application to breast cancer data

Gene Symbol	CONCORD	SYMLASSO	SPACE1	SPACE2	Reference
HNF3A (FOXA1)	+	+	+	+	Koboldt and Others (2012), Albergaria et al. (2009), Davidson et al. (2011), Lacroix and Leclercq (2004), Robinson et al. (2011)
TONDU	+	+	+	+	
FZD9	+	+	+	+	Katoh (2008), Rø nneberg et al. (2011)
KIAA0481	+	+	+	+	[Gene record discontinued]
KRT16	+	+	+		Glinsky et al. (2005), Joosse et al. (2012), Pellegrino et al. (1988)
KNSL6 (KIF2C)	+			+	Eschenbrenner et al. (2011), Shimo et al. (2007, 2008)
FOXC1	+	+	+	+	Du et al. (2012), Sizemore and Keri (2012), Wang et al. (2012), Ray et al. (2011), Tkocz et al. (2012)
PSA	+	+		+	Kraus et al. (2010), Mohajeri et al. (2011), Sauter et al. (2004), Yang et al. (2002)
GATA3	+	+	+	+	Koboldt and Others (2012), Davidson et al. (2011), Albergaria et al. (2009), Eeckhoute et al. (2007), Jiang et al. (2010), Licata et al. (2010), Yan et al. (2010)
C20ORF1 (TPX2)	+				Maxwell et al. (2011), Bibby et al. (2009)
E48		+	+	+	
ESR1				+	Zheng et al. (2012)

Table 6: Summary of the top hub genes identified by each of the four methods, CONCORD, SYMLASSO, SPACE1 & SPACE2: Genes indicated by '+' denote the 10 most highly connected genes for each of the methods. References are provided at the end of this supplemental section.

J Application to portfolio optimization

J.1 Constituents of Dow Jones Industrial Average

Symbol	Description	Return (%)	Risk (%)	SR
AA	Alcoa Inc.	9.593	41.970	0.109
AXP	American Express Company	18.706	38.913	0.352
BA	The Boeing Company	13.417	32.685	0.258
BAC	Bank of America Corporation	13.182	48.588	0.168
CAT	Caterpillar Inc.	19.042	35.050	0.401
CSCO	Cisco Systems, Inc.	22.650	44.565	0.396
CVX	Chevron Corporation	15.486	26.716	0.392
DD	E. I. du Pont de Nemours and Company	10.591	30.537	0.183
DIS	The Walt Disney Company	12.312	32.800	0.223
GE	General Electric Company	12.449	31.667	0.235
HD	The Home Depot, Inc.	17.266	34.422	0.356
HPQ	Hewlett-Packard Company	10.769	40.727	0.142
IBM	International Business Machines Corporation	18.715	29.944	0.458
INTC	Intel Corporation	18.325	41.543	0.321
JNJ	Johnson & Johnson	13.664	22.087	0.392
JPM	JPMorgan Chase & Co.	18.292	42.729	0.311
KO	The Coca-Cola Company	10.617	24.092	0.233
MCD	McDonald's Corp.	14.457	26.114	0.362
MMM	3M Company	12.596	25.353	0.300
MRK	Merck & Co. Inc.	12.385	29.616	0.249
MSFT	Microsoft Corporation	18.612	33.904	0.401
PFE	Pfizer Inc.	14.376	29.060	0.323
PG	Procter & Gamble Co.	13.262	24.241	0.341
T	AT&T, Inc.	11.231	28.781	0.217
TRV	The Travelers Companies, Inc.	14.726	31.706	0.307
UTX	United Technologies Corp.	18.618	28.760	0.474
VZ	Verizon Communications Inc.	11.403	27.728	0.231
WMT	Wal-Mart Stores Inc.	15.495	27.955	0.375
XOM	Exxon Mobil Corporation	15.466	25.764	0.406

Table 7: Dow Jones Industrial Average component stocks and their respective realized returns, realized risk and Sharpe ratios. The risk-free rate is set at 5%.

J.2 Investment periods

k	Date Range	k	Date Range	k	Date Range	k	Date Range
1	95/02/18-95/03/17	59	99/07/31-99/08/27	117	04/01/10-04/02/06	175	08/06/21-08/07/18
2	95/03/18-95/04/14	60	99/08/28-99/09/24	118	04/02/07-04/03/05	176	08/07/19-08/08/15
3	95/04/15-95/05/12	61	99/09/25-99/10/22	119	04/03/06-04/04/02	177	08/08/16-08/09/12
4	95/05/13-95/06/09	62	99/10/23-99/11/19	120	04/04/03-04/04/30	178	08/09/13-08/10/10
5	95/06/10-95/07/07	63	99/11/20-99/12/17	121	04/05/01-04/05/28	179	08/10/11-08/11/07
6	95/07/08-95/08/04	64	99/12/18-00/01/14	122	04/05/29-04/06/25	180	08/11/08-08/12/05
7	95/08/05-95/09/01	65	00/01/15-00/02/11	123	04/06/26-04/07/23	181	08/12/06-09/01/02
8	95/09/02-95/09/29	66	00/02/12-00/03/10	124	04/07/24-04/08/20	182	09/01/03-09/01/30
9	95/09/30-95/10/27	67	00/03/11-00/04/07	125	04/08/21-04/09/17	183	09/01/31-09/02/27
10	95/10/28-95/11/24	68	00/04/08-00/05/05	126	04/09/18-04/10/15	184	09/02/28-09/03/27
11	95/11/25-95/12/22	69	00/05/06-00/06/02	127	04/10/16-04/11/12	185	09/03/28-09/04/24
12	95/12/23-96/01/19	70	00/06/03-00/06/30	128	04/11/13-04/12/10	186	09/04/25-09/05/22
13	96/01/20-96/02/16	71	00/07/01-00/07/28	129	04/12/11-05/01/07	187	09/05/23-09/06/19
14	96/02/17-96/03/15	72	00/07/29-00/08/25	130	05/01/08-05/02/04	188	09/06/20-09/07/17
15	96/03/16-96/04/12	73	00/08/26-00/09/22	131	05/02/05-05/03/04	189	09/07/18-09/08/14
16	96/04/13-96/05/10	74	00/09/23-00/10/20	132	05/03/05-05/04/01	190	09/08/15-09/09/11
17	96/05/11-96/06/07	75	00/10/21-00/11/17	133	05/04/02-05/04/29	191	09/09/12-09/10/09
18	96/06/08-96/07/05	76	00/11/18-00/12/15	134	05/04/30-05/05/27	192	09/10/10-09/11/06
19	96/07/06-96/08/02	77	00/12/16-01/01/12	135	05/05/28-05/06/24	193	09/11/07-09/12/04
20	96/08/03-96/08/30	78	01/01/13-01/02/09	136	05/06/25-05/07/22	194	09/12/05-10/01/01
21	96/08/31-96/09/27	79	01/02/10-01/03/09	137	05/07/23-05/08/19	195	10/01/02-10/01/29
22	96/09/28-96/10/25	80	01/03/10-01/04/06	138	05/08/20-05/09/16	196	10/01/30-10/02/26
23	96/10/26-96/11/22	81	01/04/07-01/05/04	139	05/09/17-05/10/14	197	10/02/27-10/03/26
24	96/11/23-96/12/20	82	01/05/05-01/06/01	140	05/10/15-05/11/11	198	10/03/27-10/04/23
25	96/12/21-97/01/17	83	01/06/02-01/06/29	141	05/11/12-05/12/09	199	10/04/24-10/05/21
26	97/01/18-97/02/14	84	01/06/30-01/07/27	142	05/12/10-06/01/06	200	10/05/22-10/06/18
27	97/02/15-97/03/14	85	01/07/28-01/08/24	143	06/01/07-06/02/03	201	10/06/19-10/07/16
28	97/03/15-97/04/11	86	01/08/25-01/09/21	144	06/02/04-06/03/03	202	10/07/17-10/08/13
29	97/04/12-97/05/09	87	01/09/22-01/10/19	145	06/03/04-06/03/31	203	10/08/14-10/09/10
30	97/05/10-97/06/06	88	01/10/20-01/11/16	146	06/04/01-06/04/28	204	10/09/11-10/10/08
31	97/06/07-97/07/04	89	01/11/17-01/12/14	147	06/04/29-06/05/26	205	10/10/09-10/11/05
32	97/07/05-97/08/01	90	01/12/15-02/01/11	148	06/05/27-06/06/23	206	10/11/06-10/12/03
33	97/08/02-97/08/29	91	02/01/12-02/02/08	149	06/06/24-06/07/21	207	10/12/04-10/12/31
34	97/08/30-97/09/26	92	02/02/09-02/03/08	150	06/07/22-06/08/18	208	11/01/01-11/01/28
35	97/09/27-97/10/24	93	02/03/09-02/04/05	151	06/08/19-06/09/15	209	11/01/29-11/02/25
36	97/10/25-97/11/21	94	02/04/06-02/05/03	152	06/09/16-06/10/13	210	11/02/26-11/03/25
37	97/11/22-97/12/19	95	02/05/04-02/05/31	153	06/10/14-06/11/10	211	11/03/26-11/04/22
38	97/12/20-98/01/16	96	02/06/01-02/06/28	154	06/11/11-06/12/08	212	11/04/23-11/05/20
39	98/01/17-98/02/13	97	02/06/29-02/07/26	155	06/12/09-07/01/05	213	11/05/21-11/06/17
40	98/02/14-98/03/13	98	02/07/27-02/08/23	156	07/01/06-07/02/02	214	11/06/18-11/07/15
41	98/03/14-98/04/10	99	02/08/24-02/09/20	157	07/02/03-07/03/02	215	11/07/16-11/08/12
42	98/04/11-98/05/08	100	02/09/21-02/10/18	158	07/03/03-07/03/30	216	11/08/13-11/09/09
43	98/05/09-98/06/05	101	02/10/19-02/11/15	159	07/03/31-07/04/27	217	11/09/10-11/10/07
44	98/06/06-98/07/03	102	02/11/16-02/12/13	160	07/04/28-07/05/25	218	11/10/08-11/11/04
45	98/07/04-98/07/31	103	02/12/14-03/01/10	161	07/05/26-07/06/22	219	11/11/05-11/12/02
46	98/08/01-98/08/28	104	03/01/11-03/02/07	162	07/06/23-07/07/20	220	11/12/03-11/12/30
47	98/08/29-98/09/25	105	03/02/08-03/03/07	163	07/07/21-07/08/17	221	11/12/31-12/01/27
48	98/09/26-98/10/23	106	03/03/08-03/04/04	164	07/08/18-07/09/14	222	12/01/28-12/02/24
49	98/10/24-98/11/20	107	03/04/05-03/05/02	165	07/09/15-07/10/12	223	12/02/25-12/03/23
50	98/11/21-98/12/18	108	03/05/03-03/05/30	166	07/10/13-07/11/09	224	12/03/24-12/04/20
51	98/12/19-99/01/15	109	03/05/31-03/06/27	167	07/11/10-07/12/07	225	12/04/21-12/05/18
52	99/01/16-99/02/12	110	03/06/28-03/07/25	168	07/12/08-08/01/04	226	12/05/19-12/06/15
53	99/02/13-99/03/12	111	03/07/26-03/08/22	169	08/01/05-08/02/01	227	12/06/16-12/07/13
54	99/03/13-99/04/09	112	03/08/23-03/09/19	170	08/02/02-08/02/29	228	12/07/14-12/08/10
55	99/04/10-99/05/07	113	03/09/20-03/10/17	171	08/03/01-08/03/28	229	12/08/11-12/09/07
56	99/05/08-99/06/04	114	03/10/18-03/11/14	172	08/03/29-08/04/25	230	12/09/08-12/10/05
57	99/06/05-99/07/02	115	03/11/15-03/12/12	173	08/04/26-08/05/23	231	12/10/06-12/10/26
58	99/07/03-99/07/30	116	03/12/13-04/01/09	174	08/05/24-08/06/20		

Table 8: Investment periods in YY/MM/DD format

J.3 Details of minimum variance portfolio rebalancing

The investment period during which a set of portfolio weights are held constant is also referred to as the "holding period". The number of trading days in the k-th investment period, L_k, may vary if rebalancing time points are chosen to coincide with either calendar months, weeks or fiscal quarters. Let t index the number of an arbitrary day during the entire investment horizon. The number of trading days T_j in the first j investment periods is given by

$$T_j = \sum_{k=1}^{j} L_k, \tag{34}$$

where $j = 1, 2, \ldots, K$ with $T_0 = 0$. We consider holding N_{est} constant for all investment periods, $k = 1, 2, \ldots$. For convenience, denote by k_t the investment period that trading day t belongs to: i.e., $k_t = k(t) := \{k : t \in [T_{k-1}, T_k]\}$.

The algorithm for the minimum variance portfolio rebalancing strategy (MVR) can now be described as follows: At the beginning of time period k, that is after T_{k-1} days, compute an estimate of the covariance matrix $\widehat{\Sigma}_k$ for period k from N_{est} past returns: i.e., $\{r_t : t \in [T_{k-1} - N_{\text{est}} + 1, T_{k-1}]\}$. Then, compute a new set of portfolio weights $w_k = (\mathbf{1}^T \widehat{\Sigma}_k^{-1} \mathbf{1})^{-1} \widehat{\Sigma}_k^{-1} \mathbf{1}$, and hold this portfolio constant until the T_k-th trading day. The process is then repeated for the next holding period.

J.4 Details of cross-validation

Consider the matrix of returns $\mathbf{R}$ for all the stocks in the portfolio in the estimation horizon preceding the start of the investment period $(k-1)$.

$$\mathbf{R} = ((r_{ti})), \text{ where } i \in \{1, \ldots, p\}, \ t \in \{T_{k-1} - N_{\text{est}} + 1, \ldots, T_{k-1}\}.$$

Hence, $\mathbf{R}$ is an N_{est}-by-p matrix, and the column vector $\mathbf{R}_j$ is an N_{est}-vector of returns for the j-th stock.

Now denote by $\Omega(\lambda) = ((\omega_{ij}(\lambda)))_{1 \leq i,j \leq p}$ an estimate of Ω obtained by ℓ_1-regularization methods such as Glasso or CONCORD. The use of λ makes explicit the dependence of these estimation methods on the penalty parameter λ. The data are the over the estimation horizon is divided into m-folds. The penalty parameter is chosen so as to minimize the out

of sample predictive risk (PR) given by

$$PR(\lambda) = \sum_{m=1}^{M} \left\{ \frac{1}{N_m} \sum_{i=1}^{p} \|\mathbf{R}_i^{(m)} - \sum_{j \neq i} \beta_{ij}^{(\backslash m)}(\lambda) \mathbf{R}_j^{(m)}\|_2^2 \right\},$$

where $\mathbf{R}_i^{(m)}$ is the vector of returns for stock i in fold m, and where N_m is the number of observations in the m-th fold. The regression coefficient $\beta_{ij}^{(\backslash m)}(\lambda)$ is determined as follows: $\beta_{ij}^{(\backslash m)}(\lambda) = -\frac{\omega_{ij}^{(\backslash m)}(\lambda)}{\omega_{ii}^{(\backslash m)}(\lambda)}$, with $\Omega^{(\backslash m)}(\lambda)$ based on using all the available data within a given estimation horizon except for fold m. The optimal choice of penalty parameter λ^* is then determined as follows:

$$\lambda^* = \arg \inf_{\lambda \geq 0} PR(\lambda).$$

J.5 Performance metrics

For comparison purposes with (Won et al., 2012), we use the following quantities to assess the performance of the five MVR strategies. The formulas for these metrics are given below.

- *Realized return*: The average return of the portfolio over the entire investment horizon.

$$r_p = \frac{1}{T} \sum_{t=1}^{T} r_t' w_{k_t}$$

- *Realized risk*: The risk (standard error) of the portfolio over the entire investment horizon.

$$\sigma_p = \left[\frac{1}{T} \sum_{t=1}^{T} (r_t' w_{k_t} - r_p)^2 \right]^{1/2}$$

- *Realized Sharpe ratio (SR)*: The realized excess return of the portfolio over the risk-free rate per unit realized risk for the entire investment horizon.

$$SR = \frac{r_p - r_f}{\sigma_p} \tag{35}$$

- *Turnover*: The amount of new portfolio assets purchased or sold over each trading period. The turnover for the k-th investment period when the portfolio weights w_k are

12

held constant is given by

$$TO(k) = \sum_{i=1}^{p} \left| w_{ik} - \left(\prod_{t=T_{k-1}+1}^{T_{k-1}+L_k} (1 + r_{it}) \right) w_{i(k-1)} \right| \tag{36}$$

with $w_{i0} = 0$ for all $i = 1, \ldots, p$.

- *Size of the short side* The proportion of the negative weights to the sum of the absolute weights of each portfolio. The short side for the k-th investment period is given by

$$SS(k) = \frac{\sum_{i=1}^{p} |\min(w_{ik}, 0)|}{\sum_{i=1}^{p} |w_{ik}|}$$

The average and standard error of the short sides over the all investment periods is

$$\overline{SS} = \frac{1}{K} \sum_{k=1}^{K} SS(k), \quad \hat{\sigma}_{SS} = \left[\frac{1}{K} \sum_{k=1}^{K} (SS(k) - \overline{SS})^2 \right]^{1/2}$$

- *Normalized wealth growth*: Accumulated wealth derived from the portfolio over the trading period when the initial budget is normalized to one. Note that both transaction costs and borrowing costs are taken into account. Let $W(t-1)$ denote the wealth of the portfolio after the $(t-1)$-th trading day. Then, the wealth of the portfolio after the t-th trading day is given by

$$W(t) = \begin{cases} W(t-1)\left(1 + r'_t w_{k_t} - TC(k_t) - BC(k_t)\right), & t = T_{k_t-1} + 1 \\ W(t-1)\left(1 + r'_t w_{k_t}\right), & t \neq T_{k_t-1} + 1 \end{cases},$$

where $TC(k)$ and $BC(k)$ are transaction costs (of trading stocks) and borrowing costs (of capital for taking short positions on stocks), respectively. On the first day of each trading period, we adjust the return for these trading costs. Denote the transaction cost rate by r_c, then the transaction cost incurred at the beginning of period k is given by

$$TC(k) = r_c \cdot TO(k). \tag{37}$$

The borrowing cost rate, $BC(k)$, depends on the short side of the portfolio weights during the $(k-1)$-th period. Denote the borrowing daily percentage by r_b, then the

13

N_{est}	Sample	Glasso	CONCORD	CondReg	LedoitWolf	DJIA
35	**17.08** (33.86)	13.10 (16.57)	13.29 (17.04)	13.62 (17.74)	12.33 (**15.58**)	8.51 (18.96)
40	**16.66** (26.52)	13.13 (16.57)	13.34 (17.02)	13.39 (17.74)	11.78 (**15.46**)	8.51 (18.96)
45	11.13 (23.19)	12.74 (16.52)	**13.05** (17.04)	**13.05** (17.77)	10.99 (**15.43**)	8.51 (18.96)
50	9.90 (20.95)	12.89 (16.39)	**13.21** (17.04)	13.08 (17.65)	11.25 (**15.36**)	8.51 (18.96)
75	11.61 (17.45)	11.28 (15.57)	**13.10** (17.04)	12.77 (17.15)	10.56 (**15.10**)	8.51 (18.96)
150	9.40 (15.41)	10.28 (14.97)	**13.20** (17.08)	12.76 (16.30)	10.63 (**14.66**)	8.51 (18.96)
225	10.49 (14.98)	10.38 (14.89)	**13.58** (17.10)	12.92 (16.04)	11.04 (**14.52**)	8.51 (18.96)
300	10.41 (14.95)	10.37 (14.95)	**13.66** (17.16)	12.85 (16.07)	10.94 (**14.52**)	8.51 (18.96)

Table 9: Realized returns of different investment strategies corresponding to different estimators with various N_{est} (realized risks are given in parentheses). The maximum annualized returns and risks are highlighted in bold.

N_{est}	Sample	Glasso	CONCORD	CondReg	LedoitWolf
35	8.42 (3.19)	0.45 (0.12)	**0.38** (**0.10**)	0.39 (0.27)	1.40 (0.38)
40	5.81 (2.28)	0.41 (0.12)	**0.34** (**0.10**)	0.37 (0.26)	1.29 (0.36)
45	4.58 (1.65)	0.39 (0.12)	**0.31** (**0.10**)	0.36 (0.23)	1.20 (0.35)
50	3.74 (1.19)	0.39 (0.13)	**0.28** (**0.09**)	0.36 (0.25)	1.11 (0.33)
75	2.03 (0.67)	0.50 (0.19)	**0.21** (**0.08**)	0.43 (0.29)	0.86 (0.29)
150	0.87 (0.32)	0.73 (0.27)	**0.14** (**0.07**)	0.40 (0.22)	0.54 (0.23)
225	0.57 (0.24)	0.56 (0.22)	**0.11** (**0.07**)	0.31 (0.13)	0.41 (0.18)
300	0.44 (0.21)	0.44 (0.23)	**0.09** (**0.07**)	0.24 (0.11)	0.33 (0.17)

Table 10: Average turnovers for various estimation horizons, N_{est} (standard errors are given in parentheses). The minimum average and standard error values for each row are highlighted in bold.

borrowing cost rate is given by

$$BC(k) = ((1 + r_b)^{L_{k-1}} - 1) \sum_{i=1}^{p} |\min(w_{i(k-1)}, 0)|. \tag{38}$$

K Proof of Theorem 2

The result follows by noting the following straightforward facts

1. The existence of a minimizer follows by the convexity of Q_{con}.

2. By assumptions (A0) and (A1), for any $\eta > 0$, $\{\widehat{\alpha}_{n,ii}\}_{1 \leq i \leq p_n}$ are uniformly bounded away from zero and infinity with probability larger than $1 - O(n^{-\eta})$.

3. When the diagonal entries are fixed at $\{\widehat{\alpha}_{n,ii}\}_{1 \leq i \leq p_n}$, then the objective function Q_{con} (reparameterized from ω^o to θ) is same as the objective function of SPACE with

N_{est}	Sample	Glasso	CONCORD	CondReg	LedoitWolf
35	41.13 (3.18)	0.66 (0.84)	**0.05** (**0.14**)	1.75 (5.00)	20.50 (6.64)
40	38.64 (3.47)	0.64 (0.75)	**0.05** (**0.14**)	1.78 (5.04)	20.45 (6.63)
45	36.89 (4.26)	0.90 (0.85)	**0.05** (**0.14**)	1.84 (4.95)	20.31 (6.61)
50	35.46 (4.38)	1.35 (1.19)	**0.04** (**0.11**)	2.17 (5.44)	20.33 (6.66)
75	30.89 (5.37)	8.67 (3.76)	**0.04** (**0.11**)	4.91 (7.38)	20.13 (6.83)
150	25.65 (6.25)	23.48 (4.68)	**0.02** (**0.07**)	9.07 (6.31)	19.60 (6.82)
225	23.68 (6.69)	23.36 (6.27)	**0.01** (**0.05**)	10.71 (3.22)	19.26 (6.91)
300	22.45 (6.90)	22.42 (6.87)	**0.00** (**0.02**)	9.95 (2.93)	18.85 (7.10)

Table 11: Average short sides for various estimation horizons, N_{est} (standard errors are given in parentheses). The minimum average and standard error values for each row are highlighted in bold.

N_{est}	Sample	Glasso	CONCORD	CondReg	LedoitWolf
35	567.958 (214.05)	22.635 (5.62)	**18.642** (**4.53**)	20.757 (17.46)	91.316 (25.19)
40	394.508 (149.90)	20.660 (5.70)	**16.858** (**4.40**)	20.013 (16.78)	85.661 (24.16)
45	315.340 (108.87)	19.899 (5.80)	**15.470** (**4.22**)	19.419 (15.27)	80.524 (23.39)
50	260.887 (81.13)	20.146 (6.39)	**14.081** (**4.06**)	19.695 (16.04)	76.154 (22.43)
75	150.242 (45.87)	30.942 (10.92)	**10.516** (**3.17**)	25.191 (19.19)	63.481 (20.94)
150	75.700 (27.88)	61.495 (18.40)	**6.596** (**2.24**)	26.788 (12.83)	46.680 (17.78)
225	56.242 (22.09)	54.117 (18.82)	**5.155** (**1.80**)	22.973 (6.08)	39.441 (15.72)
300	46.904 (20.09)	47.118 (20.72)	**4.404** (**1.67**)	18.823 (5.16)	35.065 (14.89)

Table 12: Average trading costs in basis points for various estimation horizons, N_{est} (standard errors are given in parentheses). Borrowing rate is taken to be 7% APR and transaction cost rate is taken to be 0.5% of principal for each transaction. The minimum transaction cost for each row is highlighted in bold.
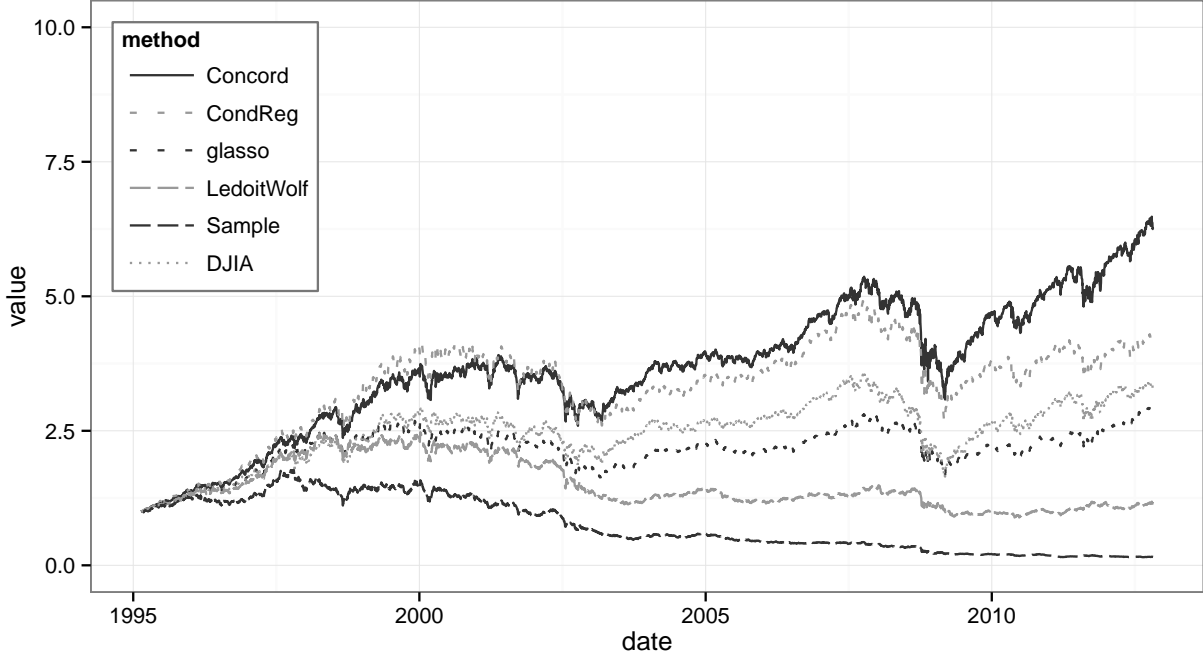
Figure 3: Normalized wealth growth after adjusting for transaction costs (0.5% of principal) and borrowing costs (interest rate of 7% APR) with $N_{\text{est}} = 75$.

weights $w_i = \widehat{\alpha}^2_{n,ii}$ (which are uniformly bounded), except that the penalty term is now $\sum_{1 \leq i < j \leq p_n} \lambda_n \sqrt{\widehat{\alpha}_{n,ii}\widehat{\alpha}_{n,jj}}\theta_{ij}$, instead of $\sum_{1 \leq i < j \leq p_n} \lambda_n \theta_{ij}$ as in Q_{spc}.

4. Since $\bar{\theta}_{n,ij} = \frac{\bar{\omega}_{n,ij}}{\sqrt{\widehat{\alpha}_{n,ii}\widehat{\alpha}_{n,jj}}}$, using the uniform boundedness of $\{\widehat{\alpha}_{n,ii}\}_{1 \leq i \leq p_n}$, there exists a constant C_1 such that for any $\eta > 0$,

$$\|\widehat{\omega}^o_n - \bar{\omega}^o_n\|_2 \leq C_1 \|\widehat{\theta}^o_n - \bar{\theta}^o_n\|_2$$

holds with probability larger than $1 - O(n^{-\eta})$.

5. For $1 \leq i < j \leq p_n$, $\text{sign}(\hat{\omega}_{n,ij}) = \text{sign}(\hat{\theta}_{n,ij})$, since they differ by a positive multiplicative constant.

6. When the penalty term in SPACE is replaced by $\sum_{1 \leq i < j \leq p_n} \lambda_n \sqrt{\widehat{\alpha}_{n,ii}\widehat{\alpha}_{n,jj}}\theta_{ij}$, the uniform boundedness of $\{\widehat{\alpha}_{n,ii}\}_{1 \leq i \leq p_n}$ implies that Theorems 1, 2 and 3 of Peng et al. (2009) hold with trivial modifications at appropriate places. The result now follows immediately using these theorems along with the above assertions. $\square$

*R*emark: Note that Theorem 2 on the consistency of CONCORD has been formulated as to exactly parallel the result given for SPACE by Peng et al. (2009). An accurate estimator
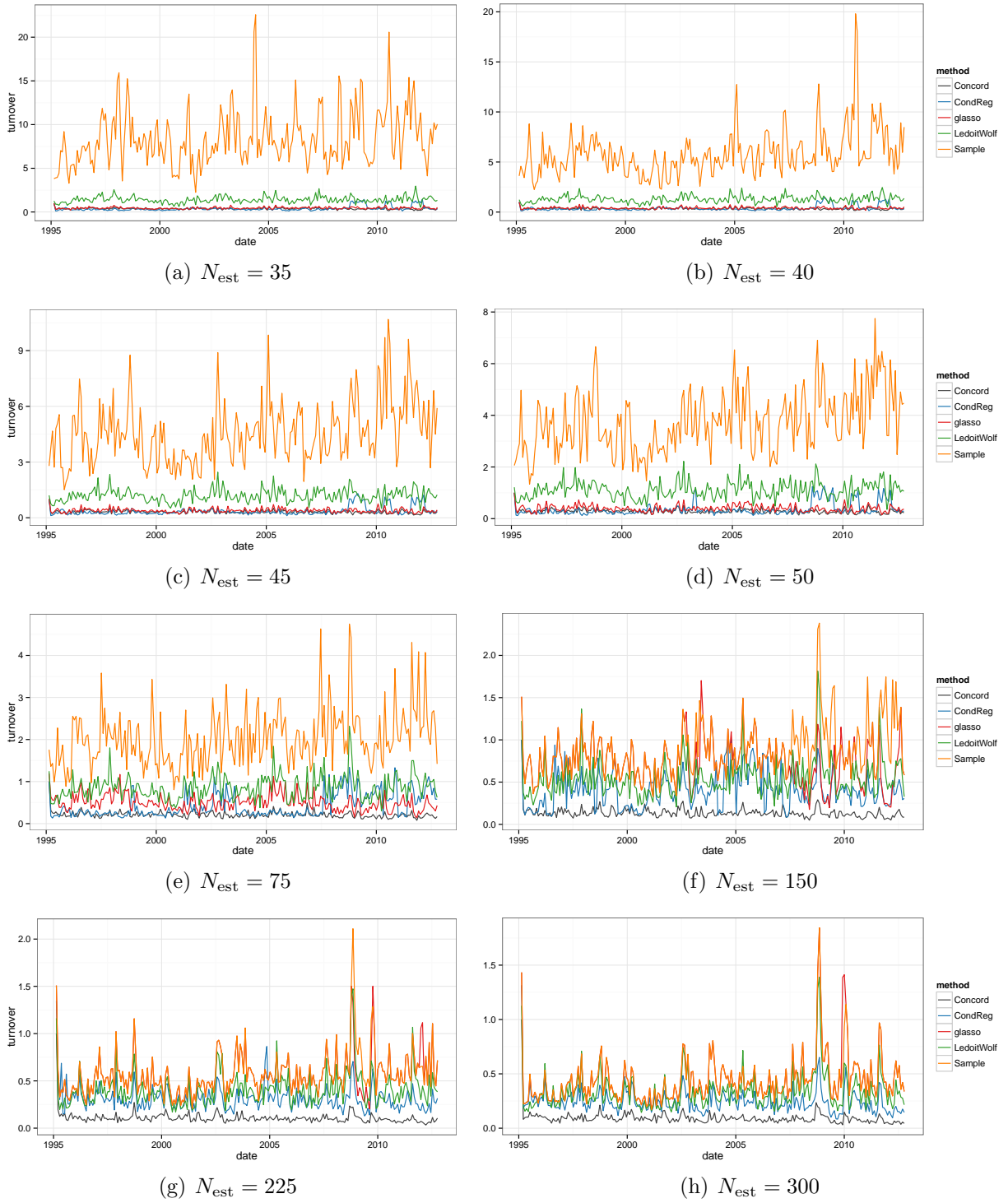
16

(a) $N_{\mathrm{est}} = 35$

(b) $N_{\mathrm{est}} = 40$

(c) $N_{\mathrm{est}} = 45$

(d) $N_{\mathrm{est}} = 50$

(e) $N_{\mathrm{est}} = 75$

(f) $N_{\mathrm{est}} = 150$

(g) $N_{\mathrm{est}} = 225$

(h) $N_{\mathrm{est}} = 300$

Figure 4: Turnover in percentage points.

(a) $N_{\text{est}} = 35$

(b) $N_{\text{est}} = 40$

(c) $N_{\text{est}} = 45$

(d) $N_{\text{est}} = 50$

(e) $N_{\text{est}} = 75$

(f) $N_{\text{est}} = 150$

(g) $N_{\text{est}} = 225$

(h) $N_{\text{est}} = 300$
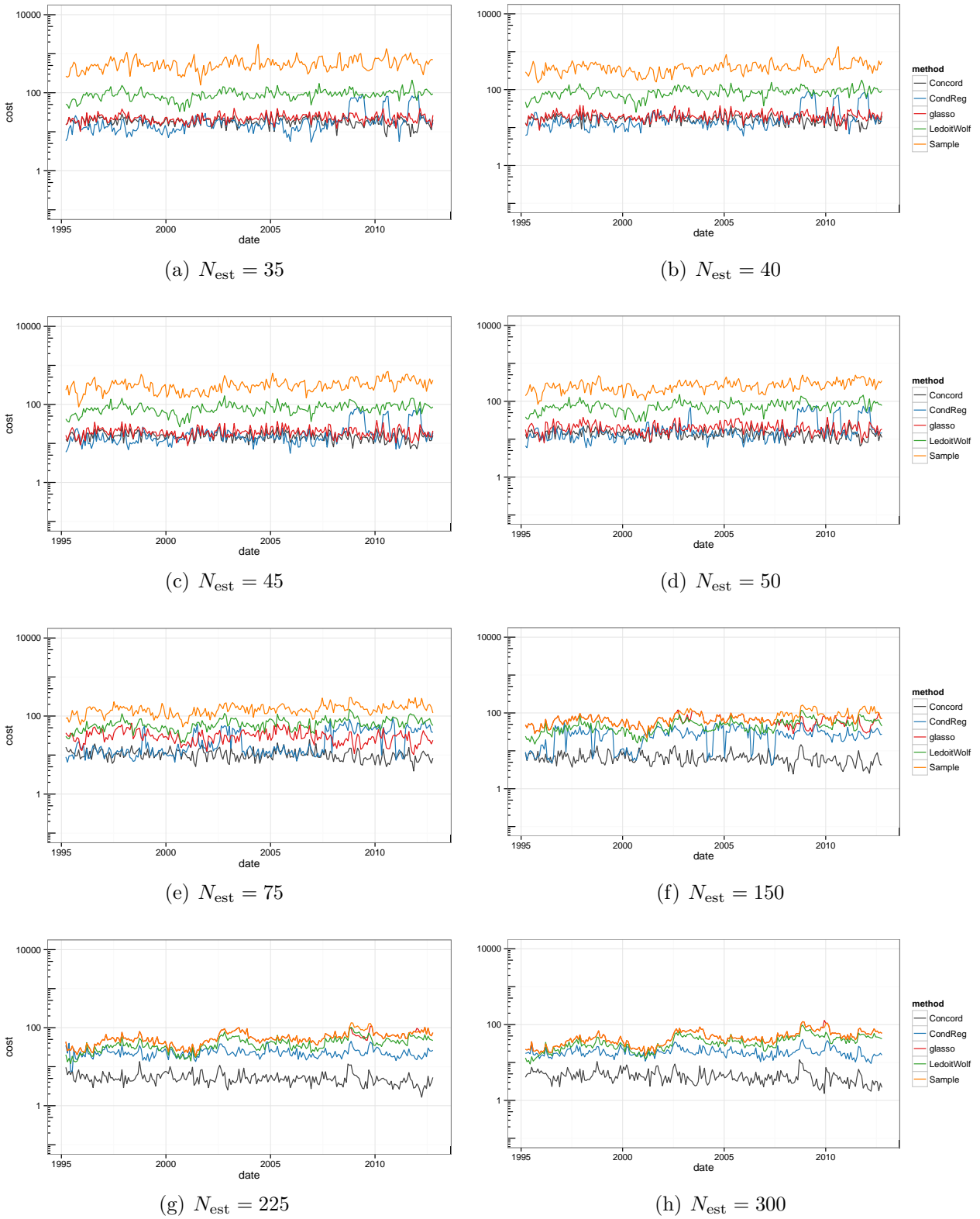
Figure 5: Trading costs in basis points for each trading period. Borrowing rate is taken to be 7% APR and transaction cost rate is taken to be 0.5% APR. The y-axes are log-scaled.

of $\bar{\omega}_{ii}$ when $p_n > n$ can be obtained by using the inverse of the sample conditional variance of each variable. In practice, however, once can simply use the diagonal estimates given by CONCORD, and there is no need for recourse to external estimates. Note also that CONCORD estimates themselves always exist, regardless of the sample size, and with certainty will lead to estimates, even when $p_n > n$. This property follows directly from the convergence of the CONCORD algorithm.

L   Joint convexity of the SYMLASSO in the Ω parameterization

We will show that the SYMLASSO objective function in (7) is jointly convex if we reparameterize in terms of Ω (see also Lee and Hastie (2014)). However, the SYMLASSO objective function is not in general strictly convex if $n < p$, and hence the convergence of the coordinatewise descent algorithm is not guaranteed. It follows from the proof of Lemma 7 that the SYMLASSO objective function (in terms of Ω) is given by

$$
\begin{aligned}
Q_{\mathrm{sym}}(\Omega) &= \frac{n}{2}\left[-\log|\Omega_D| + tr(S\Omega\Omega_D^{-1}\Omega)\right] + \lambda \sum_{1 \le i < j \le p} |\omega_{ij}| \\
&= \frac{n}{2}\left[-\sum_{i=1}^{p} \log \omega_{ii} + \frac{1}{\omega_{ii}}\omega_{i\bullet}^T S \omega_{i\bullet}\right] + \lambda \sum_{1 \le i < j \le p} |\omega_{ij}|.
\end{aligned}
$$

To prove the convexity of $Q_{\mathrm{sym}}(\Omega)$, we first prove the following lemma.

Lemma 8. *Consider the function f on $\mathbb{R}_+ \times \mathbb{R}^k$ defined by $f(\mathbf{a}) = \frac{\mathbf{a}^T A \mathbf{a}}{a_1}$. If A is positive semi-definite, then f is a convex function.*

Proof It follows by straightforward manipulations that

$$
f(\mathbf{a}) = A_{11}a_1 + 2\sum_{j=2}^{k+1} A_{1j}a_j + \frac{\mathbf{a}_{-1}^T A_{-1}\mathbf{a}_{-1}}{a_1}, \tag{39}
$$

where $\mathbf{a}_{-1} := (a_j)_{j=2}^{k+1}$ and A_{-1} is the principle submatrix of A obtained by excluding the first row and the first column. Since the first two terms above are clearly convex functions of $\mathbf{a}$, it suffices to prove that the third term $\frac{\mathbf{a}_{-1}^T A_{-1}\mathbf{a}_{-1}}{a_1}$ is a convex function of $\mathbf{a}$. Again, by straightforward manipulations, it follows that the Hessian matrix of this term is given by

$$
H = \frac{2}{a_1^3}\begin{pmatrix} \mathbf{a}_{-1}^T A_{-1}\mathbf{a}_{-1} & -(a_1 A_{-1}\mathbf{a}_{-1})^T \\ -a_1 A_{-1}\mathbf{a}_{-1} & a_1^2 A_{-1} \end{pmatrix}.
$$

Hence, for any $\mathbf{b} \in \mathbb{R}^{k+1}$ (with $\mathbf{b}_{-1} := (b_j)_{j=2}^{k+1}$), it follows that

$$\mathbf{b}^T H \mathbf{b}$$
$$= \frac{2}{a_1^3} \left(b_1^2 \mathbf{a}_{-1}^T A_{-1} \mathbf{a}_{-1} - 2 b_1 a_1 \mathbf{b}_{-1}^T A_{-1} \mathbf{a}_{-1} + a_1^2 \mathbf{b}_{-1}^T A_{-1} \mathbf{b}_{-1} \right). \qquad (40)$$

Since A_{-1} is positive semi-definite, it follows that if $\mathbf{b}_{-1}^T A_{-1} \mathbf{b}_{-1} = 0$, then $A_{-1} \mathbf{b}_{-1} = 0$. In this case

$$\mathbf{b}^T H \mathbf{b} = \frac{2}{a_1^3} \left(b_1^2 \mathbf{a}_{-1}^T A_{-1} \mathbf{a}_{-1} \right) \geq 0.$$

If $\mathbf{b}_{-1}^T A_{-1} \mathbf{b}_{-1} > 0$, then it follows by (40) that

$$\mathbf{b}^T H \mathbf{b}$$
$$= \frac{2 b_1^2}{a_1^3} \left(\mathbf{a}_{-1}^T A_{-1} \mathbf{a}_{-1} - \frac{(\mathbf{b}_{-1}^T A_{-1} \mathbf{a}_{-1})^2}{\mathbf{b}_{-1}^T A_{-1} \mathbf{b}_{-1}} \right) + \frac{2}{a_1^3} \left(a_1 \sqrt{\mathbf{b}_{-1}^T A_{-1} \mathbf{b}_{-1}} - b_1 \frac{\mathbf{b}_{-1}^T A_{-1} \mathbf{a}_{-1}}{\sqrt{\mathbf{b}_{-1}^T A_{-1} \mathbf{b}_{-1}}} \right)^2$$
$$\geq 0.$$

The last statement follows by noting that $\left(\mathbf{a}_{-1}^T A_{-1} \mathbf{a}_{-1} \right) \left(\mathbf{b}_{-1}^T A_{-1} \mathbf{b}_{-1} \right) \geq (\mathbf{b}_{-1}^T A_{-1} \mathbf{a}_{-1})^2$ (using the positive semi-definiteness of A_{-1} and the Cauchy-Schwarz inequality). Hence H is a positive semi-definite matrix, which combined with (39) implies that f is a convex function. □

It follows by the above lemma that $\frac{1}{\omega_{ii}} \omega_{i\bullet}^T S \omega_{i\bullet}$ is a convex function in $\omega_{i\bullet}$ (and hence Ω) for every $1 \leq i \leq p$. Since $-\log x$ and $|x|$ are convex functions, it follows that $Q_{\text{sym}}(\Omega)$ is a convex function.

M Examples where the Incoherence condition (A3) is satisfied

We now present two lemmas which outline settings where the Incoherence condition (A3) is satisfied. The first lemma shows that (A3) is satisfied if the true correlations are sufficiently small. This lemma can be regarded as a parallel result to (Zhao and Yu, 2006, Corollary 2), which shows that the irrepresentable condition for lasso regression is satisfied if the entries of $\frac{1}{n} X_n^T X_n$ (X_n being the regression design matrix) are bounded by $\frac{c}{2q_n-1}$ for some $0 \leq c < 1$.

Lemma 9. *Let*

$$d_n := \max_{1 \leq i \leq p_n} |\{j : \bar{\omega}_{n,ij} \neq 0\}|.$$

The incoherence condition (A3) is satisfied if

$$\frac{|\bar{\Sigma}_{n,ij}|}{\sqrt{\bar{\Sigma}_{n,ii}\bar{\Sigma}_{n,jj}}} \leq \frac{\sqrt{2}\delta\lambda_{min}}{\sqrt{q_n d_n}\lambda_{max}},$$

for every $n \geq 1$ and $1 \leq i \neq j \leq p_n$.

Proof: It can be shown by straightforward algebraic manipulations that

$$\bar{\mathcal{L}}''_{\mathcal{A}_n,\mathcal{A}_n}(\bar{\Omega}_n) = U_n^T V_n U_n,$$

where V_n is a p_n-block diagonal matrix with the ith diagonal block given by $\bar{\Sigma}_n$ without the ith row and column, and U_n is an appropriate $p_n(p_n - 1) \times q_n$ orthogonal matrix with 0 and 1 elements. Each column of U_n has exactly two 1's. Hence for any $\mathbf{x} \in \mathbb{R}^{q_n}$, it follows that $\mathbf{x}^T U_n^T U_n \mathbf{x} = 2\mathbf{x}^T \mathbf{x}$. It follows that the smallest eigenvalue of $U_n^T V_n U_n$ is bounded below by $\frac{2}{\lambda_{max}}$. Consequently, the largest eigenvalue of $(U_n^T V_n U_n)^{-1}$ is bounded above by $\frac{\lambda_{max}}{2}$.

Since the diagonal entries of $\bar{\Sigma}_n$ are uniformly bounded above by $\frac{1}{\lambda_{min}}$, it follows that

$$|\bar{\Sigma}_{n,kl}| \leq \frac{\sqrt{2}\delta}{\sqrt{q_n d_n}\lambda_{max}},$$

for every $n \geq 1$ and $1 \leq k \neq l \leq p_n$. Note that for every $(i,j) \notin \mathcal{A}_n$, $\bar{\mathcal{L}}''_{ij,\mathcal{A}_n}(\bar{\Omega}_n)$ has at most $2d_n$ non-zero entries. Hence, we get that

$$\left\|\bar{\mathcal{L}}''_{ij,\mathcal{A}_n}(\bar{\Omega}_n)\right\| \leq \sqrt{2d_n} \times \frac{\sqrt{2}\delta}{\sqrt{q_n d_n}\lambda_{max}} = \frac{2\delta}{\sqrt{q_n}\lambda_{max}}.$$

Finally, we note from the discussion above that

$$\left|\bar{\mathcal{L}}''_{ij,\mathcal{A}_n}(\bar{\Omega}_n)\left[\bar{\mathcal{L}}''_{\mathcal{A}_n,\mathcal{A}_n}(\bar{\Omega}_n)\right]^{-1}\text{sign}(\bar{\omega}^o_{\mathcal{A}_n})\right|$$

$$\leq \left\|\bar{\mathcal{L}}''_{ij,\mathcal{A}_n}(\bar{\Omega}_n)\right\| \left\|\left[\bar{\mathcal{L}}''_{\mathcal{A}_n,\mathcal{A}_n}(\bar{\Omega}_n)\right]^{-1}\right\| \left\|\text{sign}(\bar{\omega}^o_{\mathcal{A}_n})\right\|$$

$$\leq \frac{2\delta}{\sqrt{q_n}\lambda_{max}} \times \frac{\lambda_{max}}{2} \times \sqrt{q_n}$$

$$= \delta.$$

Hence (A3) is satisfied. $\qquad\square$

The next lemma shows that the Incoherence condition (A3) holds if the true $\bar{\Omega}_n$'s are tridiagonal matrices satisfying some mild conditions. This lemma can be regarded as a parallel result to (Zhao and Yu, 2006, Corollary 3).

Lemma 10. *Suppose that $\bar{\Omega}_n$ is a tridiagonal matrix with all diagonal entries equal to 1 and the non-zero off-diagonal entries equal to ρ_n, for every $n \geq 1$. If $\rho := \sup_n |\rho_n|$ satisfies*

$$\frac{8\rho}{(1-\rho^2)(2-\rho^4/2)} \leq \delta,$$

then (A3) is satisfied.

Proof: Using standard results for inverse of tridiagonal matrices, it follows that

$$\bar{\Sigma}_{n,ij} = \frac{\rho_n^{|i-j|}}{1-\rho_n^2},$$

for every $1 \leq i, j \leq p_n$. Note that $\mathcal{A}_n = \{(i-1, i) : 2 \leq i \leq p_n\}$, and $|\mathcal{A}_n| = p_n - 1$. Hence, $\bar{\mathcal{L}}''_{\mathcal{A}_n, \mathcal{A}_n}(\bar{\Omega}_n)$ is a tridiagonal matrix (with the ith row corresponding to the edge $(i, i+1)$), with

$$\bar{\mathcal{L}}''_{i(i+1),i(i+1)}(\bar{\Omega}_n) = \bar{\Sigma}_{n,ii} + \bar{\Sigma}_{n,(i+1)(i+1)} = \frac{2}{1-\rho_n^2},$$

for every $1 \leq i \leq p_n - 1$, and

$$\bar{\mathcal{L}}''_{i(i+1),(i+1)(i+2)}(\bar{\Omega}_n) = \bar{\Sigma}_{n,i(i+2)} = \frac{\rho_n^2}{1-\rho_n^2},$$

for every $1 \leq i \leq p_n - 2$. Again, using standard results for inverse of tridiagonal matrices, it follows that

$$\left(\bar{\mathcal{L}}''_{\mathcal{A}_n, \mathcal{A}_n}(\bar{\Omega}_n)\right)^{-1}_{i(i+1),j(j+1)} = \frac{(1-\rho_n^2)(\rho_n^2/2)^{|i-j|}}{2-\rho_n^4/2},$$

for every $1 \leq i, j \leq p_n - 1$. Using the fact that $\sum_{i=0}^{\infty} a^i = \frac{1}{1-a}$ for $|a| < 1$, we conclude that each entry in $\left(\bar{\mathcal{L}}''_{\mathcal{A}_n, \mathcal{A}_n}\right)^{-1}(\bar{\Omega}_n) \operatorname{sign}(\bar{\omega}^o_{\mathcal{A}_n})$ is bounded above in absolute value by $\frac{2}{2-\rho_n^4/2}$. Moreover, if $i < j$ and $(i, j) \notin \mathcal{A}_n$, then $\bar{\mathcal{L}}''_{ij,\mathcal{A}_n}(\bar{\Omega}_n)$ has at most four non-zero entries (entries corresponding to the edges $(i-1, i), (i, i+1), (j-1, j)$ and $(j, j+1)$, if applicable). All of these non-zero entries are bounded above in absolute value by $\frac{|\rho_n|}{1-\rho_n^2}$. It follows that for every $(i, j) \notin \mathcal{A}_n$,

$$\left| \bar{\mathcal{L}}''_{ij,\mathcal{A}_n}(\bar{\Omega}_n) \left[\bar{\mathcal{L}}''_{\mathcal{A}_n, \mathcal{A}_n}(\bar{\Omega}_n) \right]^{-1} \operatorname{sign}(\bar{\omega}^o_{\mathcal{A}_n}) \right|$$

$$\leq \frac{4|\rho_n|}{1-\rho_n^2} \times \frac{2}{2-\rho_n^4/2}$$

$$= \frac{8|\rho_n|}{(1-\rho_n^2)(2-\rho_n^4/2)}$$

$$\leq \frac{8|\rho|}{(1-\rho^2)(2-\rho^4/2)}$$

22

$$\leq \quad \delta.$$

Hence (A3) is satisfied. □

N   Non-convergence of SPACE

We provide a simple example where the SPACE algorithm (with uniform weights) does not converge, and the iterates alternate between two matrices. A sample of $n = 4$ *i.i.d.* vectors was generated from the $\mathcal{N}(\mathbf{0}, \Sigma)$ distribution with Σ as in (5). The standardized data is as follows:

$$\begin{pmatrix} 0.659253 & -0.635923 & 0.492419 \\ 0.994414 & -1.015863 & 1.115863 \\ -1.150266 & 1.141668 & -1.135115 \\ -0.503401 & 0.510117 & -0.473166 \end{pmatrix}. \tag{41}$$

The SPACE algorithm was implemented with choice of weights $w_i = 1$ and $\lambda = 0.2$. Again, after the first few iterations, it turns out that successive SPACE iterates alternate between

$$\begin{pmatrix} 1.432570 & 1.416740 & -2.132500 \\ 1.416740 & 3552.598070 & 0.000000 \\ -2.132500 & 0.000000 & 89.163310 \end{pmatrix} \text{ and } \begin{pmatrix} 3552.565950 & 1.416720 & 0.000000 \\ 1.416720 & 1.404240 & 2.100770 \\ 0.000000 & 2.100770 & 123.137260 \end{pmatrix},$$

thereby also establishing non-convergence of the SPACE algorithm in the case when the weights $w_i = 1$. Note that some of the elements in the two matrices above are vastly different. The sparsity pattern is also different, thereby yielding two different partial correlation graphs.

References

Albergaria, A., Paredes, J., Sousa, B., Milanezi, F., Carneiro, V., Bastos, J., Costa, S., Vieira, D., Lopes, N., Lam, E. W., Lunet, N., and Schmitt, F. (2009). Expression of FOXA1 and GATA-3 in breast cancer: the prognostic significance in hormone receptor-negative tumours. *Breast Cancer Research*, 11(3):R40.

Bibby, R. A., Tang, C., Faisal, A., Drosopoulos, K., Lubbe, S., Houlston, R., Bayliss, R., and Linardopoulos, S. (2009). A cancer-associated aurora A mutant is mislocalized and misregulated due to loss of interaction with TPX2. *The Journal of Biological Chemistry*, 284(48):33177–84.

Davidson, B., Stavnes, H. T., Holth, A., Chen, X., Yang, Y., Shih, I.-M., and Wang, T.-L. (2011). Gene expression signatures differentiate ovarian/peritoneal serous carcinoma from breast carcinoma in effusions. *Journal of Cellular and Molecular Medicine*, 15(3):535–44.

Du, J., Li, L., Ou, Z., Kong, C., Zhang, Y., Dong, Z., Zhu, S., Jiang, H., Shao, Z., Huang, B., and Lu, J. (2012). FOXC1, a target of polycomb, inhibits metastasis of breast cancer cells. *Breast Cancer Research and Treatment*, 131(1):65–73.

Eeckhoute, J., Keeton, E. K., Lupien, M., Krum, S. A., Carroll, J. S., and Brown, M. (2007). Positive cross-regulatory loop ties GATA-3 to estrogen receptor alpha expression in breast cancer. *Cancer Research*, 67(13):6477–83.

Eschenbrenner, J., Winsel, S., Hammer, S., Sommer, A., Mittelstaedt, K., Drosch, M., Klar, U., Sachse, C., Hannus, M., Seidel, M., Weiss, B., Merz, C., Siemeister, G., and Hoffmann, J. (2011). Evaluation of activity and combination strategies with the microtubule-targeting drug sagopilone in breast cancer cell lines. *Frontiers in Oncology*, 1:44.

Glinsky, G. V., Berezovska, O., and Glinskii, A. B. (2005). Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. *The Journal of clinical investigation*, 115(6):1503–21.

Jiang, S., Katayama, H., Wang, J., Li, S. A., Hong, Y., Radvanyi, L., Li, J. J., and Sen, S. (2010). Estrogen-induced aurora kinase-A (AURKA) gene expression is activated by GATA-3 in estrogen receptor-positive breast cancer cells. *Hormones & Cancer*, 1(1):11–20.

Joosse, S. A., Hannemann, J., Spötter, J., Bauche, A., Andreas, A., Müller, V., and Pantel, K. (2012). Changes in Keratin Expression during Metastatic Progression of Breast Cancer: Impact on the Detection of Circulating Tumor Cells. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 18(4):993–1003.

Katoh, M. (2008). WNT signaling in stem cell biology and regenerative medicine. *Current Drug Targets*, 9(7):565–70.

Khare, K. and Rajaratnam, B. (2014). Convergence of cyclic coordinate l1 minimization. *Preprint, Department of Statistics, Stanford University (soon to be available on arxiv).*

Koboldt, D. C. and Others (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70.

Kraus, T. S., Cohen, C., and Siddiqui, M. T. (2010). Prostate-specific antigen and hormone receptor expression in male and female breast carcinoma. *Diagnostic Pathology*, 5:63.

Lacroix, M. and Leclercq, G. (2004). About GATA3, HNF3A, and XBP1, three genes co-expressed with the oestrogen receptor-alpha gene (ESR1) in breast cancer. *Molecular and Cellular Endocrinology*, 219(1-2):1–7.

Lee, J. D. and Hastie, T. J. (2014). Learning the structure of mixed graphical models. *to appear in Journal of Computational and Graphical Statistics.*

Licata, L. A., Hostetter, C. L., Crismale, J., Sheth, A., and Keen, J. C. (2010). The RNA-binding protein HuR regulates GATA3 mRNA stability in human breast cancer cell lines. *Breast Cancer Research and Treatment*, 122(1):55–63.

Maxwell, C. A., Bentez, J., Gmez-Bald, L., Osorio, A., Bonifaci, N., Fernndez-Ramires, R., Costes, S. V., Guin, E., Chen, H., Evans, G. J. R., Mohan, P., Catal, I., Petit, A., Aguilar, H., Villanueva, A., Aytes, A., Serra-Musach, J., Rennert, G., Lejbkowicz, F., Peterlongo, P., Manoukian, S., Peissel, B., Ripamonti, C. B., Bonanni, B., Viel, A., Allavena, A., Bernard, L., Radice, P., Friedman, E., Kaufman, B., Laitman, Y., Dubrovsky, M., Milgrom, R., Jakubowska, A., Cybulski, C., Gorski, B., Jaworska, K., Durda, K., Sukiennicki, G., Lubiski, J., Shugart, Y. Y., Domchek, S. M., Letrero, R., Weber, B. L., Hogervorst, F. B. L., Rookus, M. A., Collee, J. M., Devilee, P., Ligtenberg, M. J., van der Luijt, R. B., Aalfs, C. M., Waisfisz, Q., Wijnen, J., van Roozendaal, C. E. P., Easton, D. F., Peock, S., Cook, M., Oliver, C., Frost, D., Harrington, P., Evans, D. G., Lalloo, F., Eeles, R., Izatt, L., Chu, C., Eccles, D., Douglas, F., Brewer, C., Nevanlinna, H., Heikkinen, T., Couch, F. J., Lindor, N. M., Wang, X., Godwin, A. K., Caligo, M. A., Lombardi, G., Loman, N., Karlsson, P., Ehrencrona, H., von Wachenfeldt, A., Bjork Barkardottir, R., Hamann, U., Rashid, M. U., Lasa, A., Calds, T., Andrs, R., Schmitt, M., Assmann, V., Stevens, K., Offit, K., Curado, J., Tilgner, H., Guig, R., Aiza, G., Brunet, J., Castellsagu, J., Martrat, G., Urruticoechea, A., Blanco, I., Tihomirova, L., Goldgar, D. E., Buys, S., John, E. M.,

Miron, A., Southey, M., Daly, M. B., Schmutzler, R. K., Wappenschmidt, B., Meindl, A., Arnold, N., Deissler, H., Varon-Mateeva, R., Sutter, C., Niederacher, D., Imyamitov, E., Sinilnikova, O. M., Stoppa-Lyonne, D., Mazoyer, S., Verny-Pierre, C., Castera, L., de Pauw, A., Bignon, Y.-J., Uhrhammer, N., Peyrat, J.-P., Vennin, P., Fert Ferrer, S., Collonge-Rame, M.-A., Mortemousque, I., Spurdle, A. B., Beesley, J., Chen, X., Healey, S., Barcellos-Hoff, M. H., Vidal, M., Gruber, S. B., Lzaro, C., Capell, G., McGuffog, L., Nathanson, K. L., Antoniou, A. C., Chenevix-Trench, G., Fleisch, M. C., Moreno, V., Pujana, M. A., HEBON, EMBRACE, SWE-BRCA, BCFR, GEMO Study Collaborators, and kConFab (2011). Interplay between brca1 and rhamm regulates epithelial apicobasal polarization and may influence risk of breast cancer. *PLoS Biol*, 9(11):e1001199.

Mohajeri, A., Zarghami, N., Pourhasan Moghadam, M., Alani, B., Montazeri, V., Baiat, A., and Fekhrjou, A. (2011). Prostate-specific antigen gene expression and telomerase activity in breast cancer patients: possible relationship to steroid hormone receptors. *Oncology Research*, 19(8-9):375–80.

Pellegrino, M. B., Asch, B. B., Connolly, J. L., and Asch, H. L. (1988). Differential expression of keratins 13 and 16 in normal epithelium, benign lesions, and ductal carcinomas of the human breast determined by the monoclonal antibody Ks8.12. *Cancer Research*, 48(20):5831–6.

Ray, P. S., Bagaria, S. P., Wang, J., Shamonki, J. M., Ye, X., Sim, M.-S., Steen, S., Qu, Y., Cui, X., and Giuliano, A. E. (2011). Basal-like breast cancer defined by FOXC1 expression offers superior prognostic value: a retrospective immunohistochemical study. *Annals of Surgical Oncology*, 18(13):3839–47.

Rø nneberg, J. A., Fleischer, T., Solvang, H. K., Nordgard, S. H., Edvardsen, H., Potapenko, I., Nebdal, D., Daviaud, C., Gut, I., Bukholm, I., Naume, B. r., Bø rresen Dale, A.-L., Tost, J., and Kristensen, V. (2011). Methylation profiling with a panel of cancer related genes: association with estrogen receptor, TP53 mutation status and expression subtypes in sporadic breast cancer. *Molecular Oncology*, 5(1):61–76.

Robinson, J. L. L., Macarthur, S., Ross-Innes, C. S., Tilley, W. D., Neal, D. E., Mills, I. G., and Carroll, J. S. (2011). Androgen receptor driven transcription in molecular apocrine breast cancer is mediated by FoxA1. *The EMBO Journal*, 30(15):3019–27.

Sauter, E. R., Lininger, J., Magklara, A., Hewett, J. E., and Diamandis, E. P. (2004). Association of kallikrein expression in nipple aspirate fluid with breast cancer risk. *International Journal of Cancer*, 108(4):588–91.

Shimo, A., Nishidate, T., Ohta, T., Fukuda, M., Nakamura, Y., and Katagiri, T. (2007). Elevated expression of protein regulator of cytokinesis 1, involved in the growth of breast cancer cells. *Cancer Science*, 98(2):174–81.

Shimo, A., Tanikawa, C., Nishidate, T., Lin, M.-L., Matsuda, K., Park, J.-H., Ueki, T., Ohta, T., Hirata, K., Fukuda, M., Nakamura, Y., and Katagiri, T. (2008). Involvement of kinesin family member 2C/mitotic centromere-associated kinesin overexpression in mammary carcinogenesis. *Cancer Science*, 99(1):62–70.

Sizemore, S. T. and Keri, R. A. (2012). The Forkhead Box Transcription Factor FOXC1 Promotes Breast Cancer Invasion by Inducing Matrix Metalloprotease 7 (MMP7) Expression. *The Journal of Biological Chemistry*, 287(29):24631–40.

Tkocz, D., Crawford, N. T., Buckley, N. E., Berry, F. B., Kennedy, R. D., Gorski, J. J., Harkin, D. P., and Mullan, P. B. (2012). BRCA1 and GATA3 corepress FOXC1 to inhibit the pathogenesis of basal-like breast cancers. *Oncogene*, 31(32):3667–3678.

Wang, J., Ray, P. S., Sim, M.-S., Zhou, X. Z., Lu, K. P., Lee, A. V., Lin, X., Bagaria, S. P., Giuliano, A. E., and Cui, X. (2012). FOXC1 regulates the functions of human basal-like breast cancer cells by activating NF-κB signaling. *Oncogene*.

Yan, W., Cao, Q. J., Arenas, R. B., Bentley, B., and Shao, R. (2010). GATA3 inhibits breast cancer metastasis through the reversal of epithelial-mesenchymal transition. *The Journal of Biological Chemistry*, 285(18):14042–14051.

Yang, Q., Nakamura, M., Nakamura, Y., Yoshimura, G., Suzuma, T., Umemura, T., Tamaki, T., Mori, I., Sakurai, T., and Kakudo, K. (2002). Correlation of prostate-specific antigen promoter polymorphisms with clinicopathological characteristics in breast cancer. *Anticancer Research*, 22(3):1825–8.

Zheng, Y., Huo, D., Zhang, J., Yoshimatsu, T. F., Niu, Q., and Olopade, O. I. (2012). Microsatellites in the Estrogen Receptor (ESR1, ESR2) and Androgen Receptor (AR) Genes and Breast Cancer Risk in African American and Nigerian Women. *PLoS ONE*, 7(7):e40494.