ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2015)
77, *Part* 4, *pp.* 803–825

A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees

Kshitij Khare,

University of Florida, Gainesville, USA

and Sang-Yun Oh and Bala Rajaratnam

Stanford University, USA

[Received June 2013. Revised July 2014]

Summary. 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 parametric likelihoods, or regularized regression/pseudolikelihoods, with the latter having the distinct advantage that they do not explicitly assume Gaussianity. As none of the popular methods proposed for solving pseudolikelihood-based objective functions have provable convergence guarantees, it is not clear whether corresponding estimators exist or are even computable, or if they actually yield correct partial correlation graphs. We propose a new pseudolikelihood-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 by 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 and real data, timing comparisons and numerical convergence is demonstrated. We also present a novel unifying framework that places all graphical pseudolikelihood methods as special cases of a more general formulation, leading to important insights.

Keywords: Convergence guarantee; Generalized pseudolikelihood; Gene regulatory network; Graphical model selection; Partial correlation graph; Soft thresholding; Sparse inverse covariance estimation

1. Introduction

One of the hallmarks of modern day statistics is the advent of high dimensional data sets 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

Address for correspondence: Bala Rajaratnam, Department of Statistics, Stanford University, Sequoia Hall, 390 Serra Mall, Stanford, CA 94305-4065, USA.
E-mail: brajarat@stanford.edu

1369–7412/15/77803

in the data set. Hence $\mathbf{S}$ is not a useful estimator for Σ for high dimensional data sets, where often either $p \gg n$ ('large p, small n') or when p is comparable with 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 just mentioned, the sample size is often not sufficiently large to obtain a good estimator.

For many real life applications, the quantity of interest is the inverse covariance or 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 $\Omega = \Sigma^{-1} = ((\omega_{ij}))_{1 \leqslant i,j \leqslant p}$ 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 jth variables is zero (under Gaussianity, this reduces to the statement that the ith and jth 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) pointed out, among 26 examples of published networks compiled by Newman (2003), 24 networks had edge density less than 4%.

Various 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, 2010; Peng *et al.*, 2009). 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 has been established, a positive definite estimate can be easily obtained by using efficient methods (see Hastie *et al.* (2009) and 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 set-up, this approach was pursued by Banerjee *et al.* (2008) and others. Friedman *et al.* (2008) proposed the *graphical lasso* algorithm 'Glasso' for the above minimization problem, which 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 that was 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 neighbourhoods that link each variable to others. Peng *et al.* (2009) improved this neighbourhood selection method (the NS algorithm) 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 neighbourhoods.

In particular, the sparse partial correlation estimation (SPACE) 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

(a) updating partial correlations by a joint lasso regression and
(b) separately updating the partial variances.

As indicated above, it also accounts for the symmetry in Ω and is computationally efficient. Peng *et al.* (2009) showed 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 that were presented in Peng *et al.* (2009), the authors found that empirically the SPACE algorithm

seems to converge very fast. It is, however, not clear whether 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 that were suggested in Peng *et al*. (2009). Motivated by SPACE, Friedman *et al*. (2010) presented a co-ordinatewise 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 whether the symmetric lasso will converge.

In this paper, we present a new method called the convex correlation selection method and algorithm (CONCORD) 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 co-ordinatewise descent. The convexity is strict if $n \geqslant p$, in which case standard results guarantee the convergence of the co-ordinatewise 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 sequence of iterates of the co-ordinatewise descent algorithm to a global minimum. In Section 4, by exploiting the quadratic forms that are 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 algorithm estimators are asymptotically consistent in high dimensional settings under regularity assumptions that are identical to those of 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 the CONCORD algorithm yields an estimator $\hat{\Omega}$ that is well defined and is always computable. The strengths of the method are further illustrated in the simulations and real data analysis that are presented in Section 5. A comparison of the relevant properties of various algorithms proposed in the literature is provided in Table 1 (NS by Meinshausen and Bühlmann (2006), SPACE by Peng *et al*. (2009), the symmetric lasso algorithm SYMLASSO by Friedman *et al*. (2010), the pseudolikelihood inverse covariance estimation algorithm SPLICE by Rocha *et al*. (2008) and CONCORD). Table 1 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 various pseudolikelihood-based graphical model selection procedures as special cases. This general formulation facilitates a direct comparison between the above pseudolikelihood-based methods and gives deep insights into their respective strengths and weaknesses.

Table 1. Comparison of regression-based graphical model selection methods†

Property	*Method*				
	NS	*SPACE*	*SYMLASSO*	*SPLICE*	*CONCORD*
Symmetry		+	+	+	+
Convergence guarantee (fixed n)	NA				+
Asymptotic consistency ($n, p \to \infty$)	+	+			+

†A '+' sign indicates that a specified method has the given property. A blank space indicates the absence of a property. 'NA' stands for 'not applicable'.

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 analyses other regression-based or pseudolikelihood methods that have been proposed. Section 3 introduces the convex correlation selection method and presents a general framework that unifies recently proposed pseudolikelihood methods. Section 4 establishes convergence of CONCORD to a global minimum, even if $n < p$. Section 5 illustrates the performance of the CONCORD algorithm on simulated and real data. Comparisons with 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 convex correlation selection approach. Concluding remarks are given in Section 7. The on-line supplemental document contains proofs of some of the results in the paper.

2. The SPACE algorithm and convergence properties

Let the random vector $\mathbf{Y}^k = (y_1^k, y_2^k, \ldots, y_p^k)'$, $k = 1, 2, \ldots, n$, denote independent and identically distributed (IID) observations from a multivariate distribution with mean vector $\mathbf{0}$ and covariance matrix Σ. Let $\Omega = \Sigma^{-1} = ((\omega_{ij}))_{1 \leqslant i,\, j \leqslant p}$ denote the inverse covariance matrix, and let $\boldsymbol{\rho} = (\rho^{ij})_{1 \leqslant i < j \leqslant p}$ where $\rho^{ij} = -\omega_{ij}/\sqrt{(\omega_{ii}\omega_{jj})}$ denotes the partial correlation between the ith and jth variable for $1 \leqslant i \neq j \leqslant p$. 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 $\mathbf{Y}_i = (y_i^1, y_i^2, \ldots, y_i^n)'$.

2.1. The SPACE algorithm

Peng *et al.* (2009) proposed a novel iterative algorithm called SPACE to estimate the partial correlations $\{\rho^{ij}\}_{1 \leqslant i < j \leqslant p}$ and the partial covariances $\{\omega_{ii}\}_{1 \leqslant i \leqslant p}$ corresponding to Ω. This algorithm is summarized in the on-line supplemental section A.

2.2. Convergence properties of SPACE

From empirical studies, Peng *et al.* (2009) found that the SPACE algorithm converges quickly. As mentioned in Section 1, it is not immediately clear whether convergence can be established theoretically. In an effort to understand such properties, we now place the SPACE algorithm in a useful optimization framework. (See the on-line supplemental section A for a proof.)

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

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

Although lemma 1 identifies SPACE as an iterative partial minimization algorithm, the existing theory for iterative partial minimization (see for example Zangwill (1969), Jensen *et al.* (1991) and Lauritzen (1996)) 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 whether 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.

2.1.1. Example 1

Consider the following population covariance and inverse covariance matrices:

$$\Omega = \begin{pmatrix} 3.0 & 2.1 & 0.0 \\ 2.1 & 3.0 & 2.1 \\ 0.0 & 2.1 & 3.0 \end{pmatrix},$$
$$\Sigma = \Omega^{-1} = \begin{pmatrix} 8.500 & -11.667 & 8.167 \\ -11.667 & 16.667 & -11.667 \\ 8.167 & -11.667 & 8.500 \end{pmatrix}. \tag{2}$$

A sample of $n = 100$ IID vectors was generated from the corresponding $\mathcal{N}(\mathbf{0}, \Sigma)$ distribution. The data were 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:

$$\begin{pmatrix} 29.009570 & 27.266460 & 0.000000 \\ 27.266460 & 51.863320 & 24.680140 \\ 0.000000 & 24.680140 & 26.359350 \end{pmatrix},$$
$$\begin{pmatrix} 28.340040 & 27.221520 & -0.705390 \\ 27.221520 & 54.255190 & 24.569900 \\ -0.705390 & 24.569900 & 25.753040 \end{pmatrix}, \tag{3}$$

thereby establishing non-convergence of the SPACE algorithm in this example (see also Fig. 1(a)). Note that the two matrices in expression (3) have different sparsity patterns. A similar example of non-convergence of SPACE with uniform weights is provided in the on-line supplemental section Q.

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

2.2.2. 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 data sets (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) that algorithms SPACE1 (SPACE with uniform weights) and SPACE2 (SPACE with partial variance weights) do not converge within 1500 iterations. When they do converge, the mean numbers of iterations are 22.3 for SPACE1 and 14.1 for SPACE2 (since the original implementation of SPACE by Peng *et al.* (2009) was programmed to stop after three iterations, we modified the implementation to allow for more iterations 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 behaviour. Our simulations suggest that the convergence problem is exacerbated as the condition number of Ω increases.

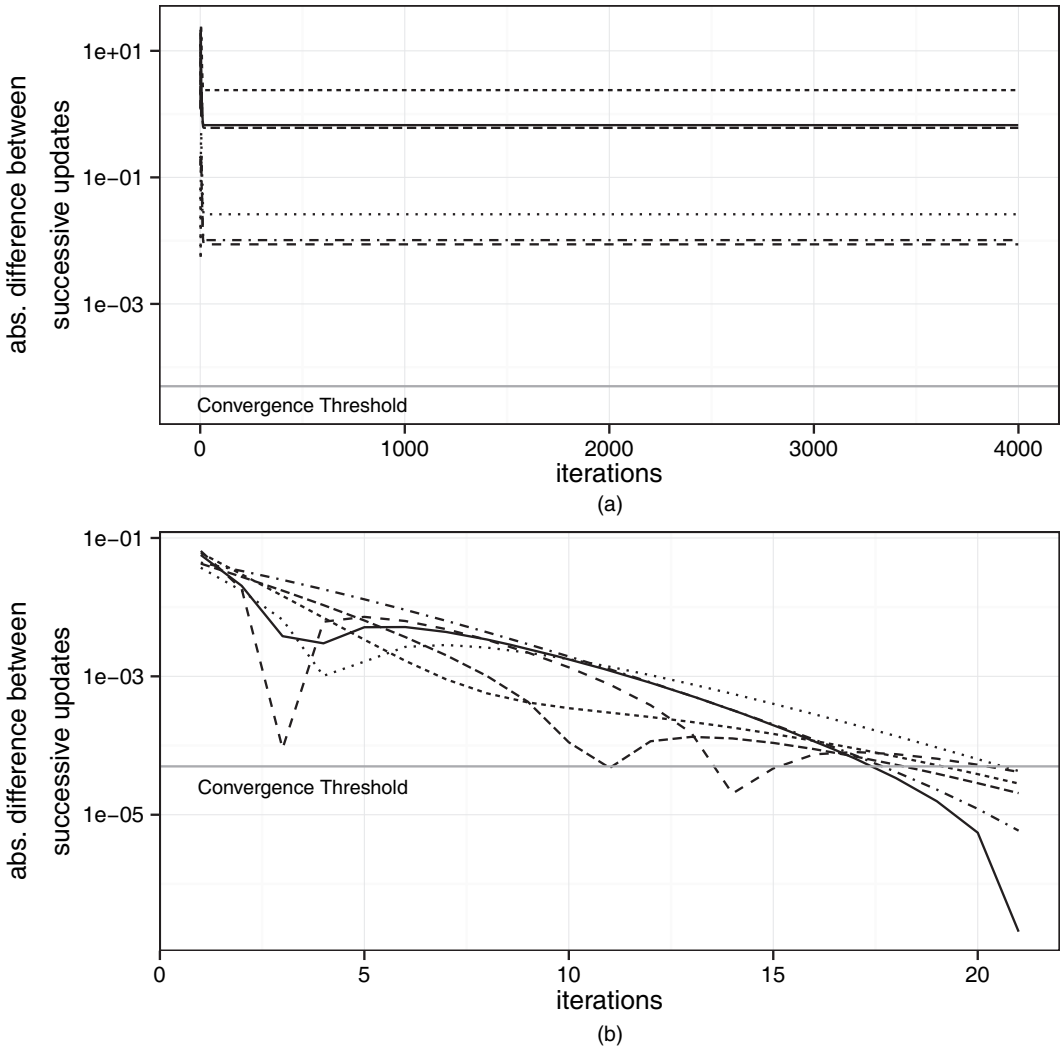

Fig. 1. Illustrations of the non-convergence of SPACE and the convergence of CONCORD (the y-axes are on the log-scale; for SPACE, the log-absolute difference between entries of successive estimates becomes constant (thus indicating non-convergence) (———, pvar.1; - - - - - -, pvar.2; — — —, pvar.3, — — —, pcor.1; · · · · · ·, pcor.2; -·-·-·-, pcor.3): (a) SPACE algorithm (partial variance weights) applied to the data set in example 1; (b) CONCORD algorithm applied to the data set in example 1

2.3. Symmetric lasso

The symmetric lasso algorithm was proposed as a useful alternative to SPACE in recent work by Friedman *et al.* (2010). The symmetric lasso minimizes the (negative) pseudolikelihood

$$Q_{\mathrm{sym}}(\boldsymbol{\alpha}, \breve{\Omega}) = \frac{1}{2} \sum_{i=1}^{p} \left\{ n \log(\alpha_{ii}) + \frac{1}{\alpha_{ii}} \|\mathbf{Y}_i + \sum_{j \neq i} \omega_{ij} \alpha_{ii} \mathbf{Y}_j\|^2 \right\} + \lambda \sum_{1 \leqslant i < j \leqslant p} |\omega_{ij}|. \qquad (4)$$

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 0. A comparison of equations (1) and (4) shows a deep connection between SPACE (with $w_i = \omega_{ii}$) and symmetric lasso objective functions.

Table 2. Number of simulations (out of 100) that do not converge within 1500 iterations, NC, for select values of penalty parameter ($\lambda^* = \lambda/n$)†

Results for SPACE1 ($w_i = 1$)			*Results for 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

†The average percentages of non-zeros, NZ, in Ω are also shown.

In particular, the $Q_{\mathrm{sym}}(\boldsymbol{\alpha}, \breve{\Omega})$ objective function in equation (4) is a reparameterization of equation (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 equation (4). The minimization of the objective function in equation (4) is performed by co-ordinatewise descent on $(\boldsymbol{\alpha}, \breve{\Omega})$. The 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 that are used in the symmetric lasso.

Lemma 2. The symmetric lasso objective in equation (4) is a non-convex function of $(\boldsymbol{\alpha}, \breve{\Omega})$.

The proof of lemma 2 is given in the on-line supplemental section B. The arguments in the proof of lemma 2 demonstrate that the objective function that is used in the symmetric lasso is not convex, or even biconvex in the parameterization that is used above. However, it can be shown that the SYMLASSO algorithm objective function is jointly convex in the elements of Ω (see Lee and Hastie (2014) and the on-line supplemental section O). It is straightforward to check that the co-ordinatewise descent algorithms for both parameterizations are exactly the same. However, unless a function is strictly convex, there are no general theoretical guarantees of convergence for the corresponding co-ordinatewise descent algorithm. Indeed, when $n < p$, the SYMLASSO objective function is not strictly convex. Therefore, it is not clear whether the co-ordinate descent algorithm converges in general. We conclude this section by remarking that both SPACE and the 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 sparse pseudolikelihood inverse covariance estimates algorithm SPLICE 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 pseudolikelihood 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. $\beta_{jj} = 0$. On the basis of the above properties, the l_1-penalized pseudolikelihood objective function of the SPLICE algorithm (without the constant term) is given by

$$Q_{\mathrm{spl}}(\mathbf{B},\mathbf{D}) = \frac{n}{2}\sum_{i=1}^{p}\log(d_{ii}^2) + \frac{1}{2}\sum_{i=1}^{p}\frac{1}{d_{ii}^2}\left\|\mathbf{Y}_i - \sum_{j\neq i}\beta_{ij}\mathbf{Y}_j\right\|^2 + \lambda\sum_{i<j}|\beta_{ij}|. \quad (5)$$

To optimize equation (5) with respect to **B** and **D**, Rocha *et al.* (2008) also proposed an iterative algorithm that alternates between maximizing **B** fixing **D**, followed by maximizing **D** fixing **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.

(a) The SPLICE objective function $Q_{\mathrm{spl}}(\mathbf{B},\mathbf{D})$ is not jointly convex in $(\mathbf{B},\mathbf{D})$.
(b) Under the transformation $\mathbf{C}=\mathbf{D}^{-1}$, $Q_{\mathrm{spl}}(\mathbf{B},\mathbf{C})$ is biconvex.

The proof of lemma 3 is given in on-line supplemental section C. The convergence properties of the SPLICE algorithm are 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 pseudolikelihood framework for sparse partial covariance estimation

The two pseudolikelihood-based approaches, SPACE and the 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 established (under suitable regularity assumptions) consistency of SPACE estimators for distributions with sub-Gaussian tails. However, none of the existing pseudolikelihood-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 pseudolikelihood framework which preserves all the attractive properties of the SPACE and SYMLASSO algorithms and, at the same time, leads to theoretical guarantees of convergence. It is not clear immediately, however, 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

(a) a guarantee of existence of a global minimum,
(b) a better chance of convergence by using convex optimization algorithms and
(c) a 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 behaviour. 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 co-ordinatewise descent algorithm. Indeed, when $n < p$, the SYMLASSO objective function is not strictly convex, and it is not clear whether the corresponding co-ordinatewise descent algorithm converges.

In this section, we introduce a new approach for estimating Ω, called the convex correlation selection method and algorithm CONCORD, that aim 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 convex correlation selection method and then proceed to derive the details of the corresponding co-ordinatewise 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 co-ordinatewise descent algorithm does indeed converge to a global minimum. Computational complexity and running time comparisons for CONCORD are given in the on-line supplemental section E and Section 5.1 respectively. Subsequently, large sample properties of the resulting estimator are established in Section 6 to provide asymptotic guarantees in the regime when both the dimension p and the sample size n tend to ∞. Thereafter, the performance of CONCORD on simulated data and on real data from biomedical and financial applications is demonstrated. Such analysis serves to establish that CONCORD preserves all the attractive properties of existing pseudolikelihood methods and additionally provides the crucial theoretical guarantee of convergence and existence of a well-defined solution.

3.1. The CONCORD objective function

To develop a convex formulation of the pseudolikelihood graphical model selection problem let us first revisit the formulation of the SPACE objective function (1) with arbitrary weights w_i instead of ω_{ii}:

$$Q_{\mathrm{spc}}(\Omega) = \frac{1}{2}\sum_{i=1}^{p}\left\{-n\log(\omega_{ii}) + w_i\left\|\mathbf{Y}_i - \sum_{j\neq i}\rho^{ij}\sqrt{\left(\frac{\omega_{jj}}{\omega_{ii}}\right)}\mathbf{Y}_j\right\|_2^2\right\} + \lambda\sum_{1\leqslant i<j\leqslant p}|\omega^{ij}|. \tag{6}$$

Now note that this objective function is not jointly convex in the elements of Ω, since

(a) 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 and
(b) the penalty term is on the partial correlations $\rho^{ij} = -\omega_{ij}/\sqrt{(\omega_{ii}\omega_{jj})}$ and is hence not a jointly convex function of the elements of Ω.

Now note the following relationship for the regression term:

$$\begin{aligned} w_i\left\|\mathbf{Y}_i - \sum_{j\neq i}\rho^{ij}\sqrt{\left(\frac{\omega_{jj}}{\omega_{ii}}\right)}\mathbf{Y}_j\right\|_2^2 &= w_i\left\|\mathbf{Y}_i + \sum_{j\neq i}\frac{\omega_{ij}}{\omega_{ii}}\mathbf{Y}_j\right\|_2^2 \qquad \left(\text{therefore } \rho^{ij} = \frac{-\omega_{ij}}{\sqrt{(\omega_{ii}\omega_{jj})}}\right) \\ &= w_i\left\|\frac{1}{\omega_{ii}}(\omega_{ii}\mathbf{Y}_i + \sum_{j\neq i}\omega_{ij}\mathbf{Y}_j)\right\|_2^2 \\ &= \frac{w_i}{\omega_{ii}^2}\left\|\sum_{j=1}^{p}\omega_{ij}\mathbf{Y}_j\right\|_2^2 \\ &= \frac{w_i}{\omega_{ii}^2}(\omega_{\cdot i}'\mathbf{Y}'\mathbf{Y}\omega_{\cdot i}). \end{aligned}$$

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

$$w_i\left\|\mathbf{Y}_i - \sum_{j\neq i}\rho^{ij}\sqrt{\left(\frac{\omega_{jj}}{\omega_{ii}}\right)}\mathbf{Y}_j\right\|_2^2 = \omega_{\cdot i}'\mathbf{Y}'\mathbf{Y}\omega_{\cdot i} \geqslant 0. \tag{7}$$

Expression (7) 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:

$$Q_{\mathrm{con}}(\Omega) =: \mathcal{L}_{\mathrm{con}}(\Omega) + \lambda \sum_{1 \leqslant i < j \leqslant p} |\omega_{ij}|$$
$$=: -\sum_{i=1}^{p} n \log(\omega_{ii}) + \frac{1}{2} \sum_{i=1}^{p} \left\| \omega_{ii} \mathbf{Y}_i + \sum_{j \neq i} \omega_{ij} \mathbf{Y}_j \right\|_2^2 + \lambda \sum_{1 \leqslant i < j \leqslant p} |\omega_{ij}|. \quad (8)$$

The function $\mathcal{L}_{\mathrm{con}}(\Omega)$ can be regarded as a pseudolikelihood function in the spirit of Besag (1975). Since $-\log(x)$ and $|x|$ are convex functions, and $\Sigma_{i=1}^{p} \|\omega_{ii}\mathbf{Y}_i + \Sigma_{j \neq i} \omega_{ij}\mathbf{Y}_j\|^2$ is a positive semidefinite quadratic form in Ω, it follows that $Q_{\mathrm{con}}(\Omega)$ is a jointly convex function of Ω (but not necessarily strictly convex). As we shall see later, this particular formulation helps us to 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 equation (6) has been replaced by n in expression (8). The point is elaborated further in remark 4. We now proceed to derive the details of the co-ordinatewise descent algorithm for minimizing $Q_{\mathrm{con}}(\Omega)$.

3.2. A co-ordinatewise minimization algorithm for minimizing $Q_{\mathrm{con}}(\Omega)$

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

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

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 Section 1, 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 optimize $Q_{\mathrm{con}}(\Omega)$. For $1 \leqslant i \leqslant j \leqslant p$, define the function $T_{ij} : \mathcal{M} \to \mathcal{M}$ by

$$T_{ij}(\Omega) = \underset{\{\tilde{\Omega} : (\tilde{\Omega})_{kl} = \omega_{kl} \, \forall (k,l) \neq (i,j)\}}{\arg\min} Q_{\mathrm{con}}(\tilde{\Omega}). \quad (9)$$

For each (i, j), $T_{ij}(\Omega)$ gives the matrix where all the elements of Ω are left as they are except the (i, j)th element. The (i, j)th element is replaced by the value that minimizes $Q_{\mathrm{con}}(\Omega)$ 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 equation (9) can be computed in closed form. In particular,

$$(T_{ii}(\Omega))_{ii} = \frac{-\sum_{j \neq i} \omega_{ij} s_{ij} + \sqrt{\left\{ \left(\sum_{j \neq i} \omega_{ij} s_{ij} \right)^2 + 4 s_{ii} \right\}}}{2 s_{ii}}, \qquad \text{for } 1 \leqslant i \leqslant p, \quad (10)$$

and

$$(T_{ij}(\Omega))_{ij} = \frac{S_{\lambda/n}\left\{ -\left(\sum_{j' \neq j} \omega_{ij'} s_{jj'} + \sum_{i' \neq i} \omega_{i'j} s_{ii'} \right) \right\}}{s_{ii} + s_{jj}}, \qquad \text{for } 1 \leqslant i < j \leqslant p, \quad (11)$$

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

The proof is given in the on-line supplemental section F. An important contribution of lemma 4 is that it gives the necessary ingredients for designing a co-ordinate descent approach to min-

imizing the CONCORD objective function. More specifically, equation (10) can be used to update the partial variance terms, and equation (11) can be used to update the partial covariance terms. The co-ordinatewise descent algorithm for CONCORD is summarized in algorithm 2 in the on-line supplemental section D. The computational complexity of the CONCORD algorithm is $\min\{O(np^2), O(p^3)\}$. Hence CONCORD is competitive with the SPACE and the symmetric lasso algorithms (see the on-line supplemental section E for details). 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 for $i = 1, \ldots, p$ as

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

Further, the ith component of a Bayes information criterion type of score can be defined as

$$\mathrm{BIC}_i(\lambda) = n \log\{\mathrm{RSS}_i(\lambda)\} + \log(n) |\{j : j \neq i, \omega_{ij,\lambda} \neq 0\}|.$$

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

3.3. A unifying framework for pseudolikelihood-based graphical model selection

In this section, we provide a unifying framework which formally connects the five pseudolikelihood formulations that are considered in this paper, namely algorithms SPACE1, SPACE2, SYMLASSO, SPLICE and CONCORD (counting two choices for weights in the SPACE algorithm as two different formulations). Recall that the random vectors $\mathbf{Y}^k = (y_1^k, y_2^k, \ldots, y_p^k)'$, $k = 1, 2, \ldots, n$, denote IID observations from a multivariate distribution with mean vector $\mathbf{0}$ and covariance matrix Σ, the precision matrix is given by $\Omega = \Sigma^{-1} = ((\omega_{ij}))_{1 \leqslant i,j \leqslant p}$, and $\mathbf{S}$ denotes the sample covariance matrix. Let Ω_D denote the diagonal matrix with ith diagonal entry given by ω_{ii}. Lemma 5 below formally identifies the relationship between all five of the regression-based pseudolikelihood methods.

Lemma 5.

(a) The (negative) pseudolikelihood functions of CONCORD, SPACE1, SPACE2, SYMLASSO and SPLICE formulations can be expressed in matrix form as shown in Table 3 (up to reparameterization).

(b) All five pseudolikelihoods above correspond to a unified or generalized form of the Gaussian log-likelihood function

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

where $G(\Omega)$ and $H(\Omega)$ are functions of Ω. The functions G and H which characterize the pseudolikelihood formulations corresponding to CONCORD, SPACE1, SPACE2, SYMLASSO and SPLICE are as follows:

$$\begin{aligned} G_{\mathrm{con}}(\Omega) &= \Omega_D^2, & H_{\mathrm{con}}(\Omega) &= \Omega^2; \\ G_{\mathrm{spc},1}(\Omega) &= \Omega_D, & H_{\mathrm{spc},1}(\Omega) &= \Omega \Omega_D^{-2} \Omega; \\ G_{\mathrm{spc},2}(\Omega) = G_{\mathrm{sym}}(\Omega) &= G_{\mathrm{spl}}(\Omega) = \Omega_D, & H_{\mathrm{spc},2}(\Omega) &= H_{\mathrm{sym}}(\Omega) = H_{\mathrm{spl}}(\Omega) = \Omega \Omega_D^{-1} \Omega. \end{aligned}$$

The proof of lemma 5 is given in the on-line supplemental section I. Lemma 5 gives various useful insights into the different pseudolikelihoods that have been proposed for the inverse covariance estimation problem. The following remarks discuss these insights.

Table 3. Pseudo-log-likelihood for graphical models in both regression and matrix forms

Function	*Regression form*	*Matrix form*	*Expression*
$\mathcal{L}_{\mathrm{con}}(\Omega)$	$\frac{1}{2}\sum_{i=1}^{p}\left\{-n\log(\omega_{ii}^2)+\left\|\omega_{ii}\mathbf{Y}_i+\sum_{j\neq i}\omega_{ij}\mathbf{Y}_j\right\|_2^2\right\}$	$\frac{n}{2}\{-\log\|\Omega_D^2\|+\mathrm{tr}(\mathbf{S}\Omega^2)\}$	(12)
$\mathcal{L}_{\mathrm{spc},1}(\Omega_D,\boldsymbol{\rho})$	$\frac{1}{2}\sum_{i=1}^{p}\left\{-n\log(\omega_{ii})+\left\|\mathbf{Y}_i-\sum_{j\neq i}\rho^{ij}\sqrt{\left(\frac{\omega_{jj}}{\omega_{ii}}\right)}\mathbf{Y}_j\right\|_2^2\right\}$	$\frac{n}{2}\{-\log\|\Omega_D\|+\mathrm{tr}(\mathbf{S}\Omega\Omega_D^{-2}\Omega)\}$	(13)
$\mathcal{L}_{\mathrm{spc},2}(\Omega_D,\boldsymbol{\rho})$	$\frac{1}{2}\sum_{i=1}^{p}\left\{-n\log(\omega_{ii})+\omega_{ii}\left\|\mathbf{Y}_i-\sum_{j\neq i}\rho^{ij}\sqrt{\left(\frac{\omega_{jj}}{\omega_{ii}}\right)}\mathbf{Y}_j\right\|_2^2\right\}$	$\frac{n}{2}\{-\log\|\Omega_D\|+\mathrm{tr}(\mathbf{S}\Omega\Omega_D^{-1}\Omega)\}$	(14)
$\mathcal{L}_{\mathrm{sym}}(\boldsymbol{\alpha},\Omega_F)$	$\frac{1}{2}\sum_{i=1}^{p}\left\{n\log(\alpha_{ii})+(1/\alpha_{ii})\left\|\mathbf{Y}_i+\sum_{j\neq i}\omega_{ij}\alpha_{ii}\mathbf{Y}_j\right\|^2\right\}$	$\frac{n}{2}\{-\log\|\Omega_D\|+\mathrm{tr}(\mathbf{S}\Omega\Omega_D^{-1}\Omega)\}$	(15)
$\mathcal{L}_{\mathrm{spl}}(\mathbf{B},\mathbf{D})$	$\frac{1}{2}\sum_{i=1}^{p}\left\{n\log(d_{ii}^2)+(1/d_{ii}^2)\left\|\mathbf{Y}_i-\sum_{j\neq i}\beta_{ij}\mathbf{Y}_j\right\|_2^2\right\}$	$\frac{n}{2}\{-\log\|\Omega_D\|+\mathrm{tr}(\mathbf{S}\Omega\Omega_D^{-1}\Omega)\}$	(16)

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

Remark 2. $\Omega_D^{-1}\Omega$ is a rescaling of Ω to make all the diagonal elements 1 (hence the sparsities 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 $\mathrm{tr}(\mathbf{S}\Omega)$ is approximated by $\mathrm{tr}(\mathbf{S}\Omega\Omega_D^{-1}\Omega)$. 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 $\mathrm{tr}(\mathbf{S}\Omega)$ is perturbed by an off-diagonal matrix $\mathbf{C}$, resulting in an expression of the form $\mathrm{tr}\{\mathbf{S}\Omega(\mathbf{I}+\mathbf{C})\}$.

Remark 3. Conceptually, the sole source of difference between the three regularized versions of the objective functions of the 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. The convex correlation selection method approximates the normal likelihood by approximating the $\log|\Omega|$ term by $\log|\Omega_D^2|$, and $\mathrm{tr}(\mathbf{S}\Omega)$ by $\mathrm{tr}(\mathbf{S}\Omega^2)$. 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,

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

and justifies the appearance of 'n' as compared with '$n/2$' in the CONCORD objective in expression (8). In the on-line supplemental section J, we illustrate the usefulness of this correction based on the insight from our unification framework, and we 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 $Q_{\mathrm{con}}(\Omega)$ is not differentiable. Also, if $n<p$, then $Q_{\mathrm{con}}(\Omega)$ is not necessarily strictly convex.

Hence, the global minimum may not be unique and, as discussed below, the convergence of the co-ordinatewise minimization algorithm to a global minimum does not follow from existing theory. Although $Q_{\mathrm{con}}(\Omega)$ is not differentiable, it can be expressed as a sum of a smooth function of Ω and a separable function of Ω (namely $\lambda \Sigma_{1 \leqslant i < j \leqslant p} |\omega_{ij}|$). Tseng (1988, 2001) proved that, under certain conditions, every cluster point of the sequence of iterates of the co-ordinatewise 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 co-ordinatewise minimization algorithm applied to the CONCORD objective function converges to a global minimum. A proof of this result can be found in the on-line supplemental section K.

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

Remark 5. If $n \geqslant 2$, 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.

With theory in hand, we now proceed to illustrate numerically the convergence properties that have been established above. When CONCORD is applied to the data set in example 1, convergence is achieved (see Fig. 1(b) in Section 2.2), whereas SPACE does not converge (see Fig. 1(a)).

5. Applications

5.1. Simulated data

5.1.1. Timing comparison

We now proceed to compare the timing performance of CONCORD with the graphical lasso algorithm Glasso and the two versions of SPACE. The algorithm names SPACE1 and SPACE2 denote SPACE estimates using uniform weights and partial variance weights respectively. We first consider the setting $p = 1000$ and $n = 200$. For this simulation study, a $p \times p$ positive definite matrix Ω (with $p = 1000$) with condition number 10 was used. Thereafter, 50 independent data sets were generated, each consisting of $n = 200$ IID samples from an $\mathcal{N}_p(0, \Sigma = \Omega^{-1})$ distribution. For each data set, 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 number of iterations was smaller). The timing results (averaged over the 100 data sets) in the top part of Table 4 show wall clock times until convergence (in seconds) for Glasso, CONCORD, SPACE1 and SPACE2.

We can see that, in the $p = 1000$, $n = 200$, setting, CONCORD is uniformly faster than its competitors. Note that 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 4 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

Table 4. Timing comparison for $p = 1000, 3000$ and varying n†

Results for Glasso			*Results for CONCORD*			*Results for SPACE1* ($w_i = 1$)			*Results for SPACE2* ($w_i = \omega_{ii}$)		
λ	*NZ (%)*	*Time (s)*	λ^*	*NZ (%)*	*Time (s)*	λ^*	*NZ (%)*	*Time (s)*	λ^*	*NZ (%)*	*Time (s)*
$p = 1000, n = 200$											
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

Results for Glasso			*Results for CONCORD*			*Results for Glasso*			*Results for CONCORD*		
λ	*NZ (%)*	*Time (s)*	λ^*	*NZ (%)*	*Time (s)*	λ	*NZ (%)*	*Time (s)*	λ^*	*NZ (%)*	*Time (s)*
$p = 3000, n = 600$						$p = 3000, n = 900$					
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

†SPACE is run until convergence or 50 iterations (whichever number of iterations is smaller). Note that SPACE1 and SPACE2 are much slower than CONCORD and Glasso in wall clock time, for the $p = 1000$ simulations. 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 is the percentage of non-zero entries in the corresponding estimator.

of iterations until convergence, the order of computations for a single iteration and also the implementational details (such as the choice of software and efficiency of the code). Note that the order of computations for a single iteration is the 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 implementational details and the larger number of iterations that are 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 that were 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 data sets were generated, each consisting of $n = 600$ IID samples from an $\mathcal{N}_p(0, \Sigma = \Omega^{-1})$ distribution. The same exercise was repeated with $n = 900$. The timing results (averaged over the 100 data sets) in the bottom part of Table 4 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 the $n = 900$ cases, CONCORD was around 10 times faster than Glasso.

In conclusion, these simulation results in this subsection illustrate that CONCORD is much faster compared with 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 from `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 are 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 are drawn from a multivariate t-distribution to illustrate the potential benefit of using penalized regression methods (convex correlation selection) outside the Gaussian setting.

For this study, using a similar approach to that 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 this Ω for each sample size $n = 200$, $n = 400$ and $n = 800$, 50 data sets, each having an IID multivariate t-distribution with mean 0 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 = 0.15. Table 5 provides the median of the AUCs (divided by 0.15 to normalize to 1), as well as the interquartile ranges IQR over the 50 data sets for $n = 200$, $n = 400$ and $n = 800$.

Table 5 shows that CONCORD has a much better model selection performance compared with Glasso. Moreover, it turns out that CONCORD has a higher AUC than Glasso for every single one of the 150 data sets (50 each for $n = 200, 400, 800$). We note that CONCORD not only recovers the sparsity structure more accurately in general but also has much less variation.

Remark 5. We need to simulate 50 data sets for each of the three sample sizes $n = 200, 400, 800$. For each of these data sets, an algorithm must 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 three iterations. However, given the possible non-convergence issues that are associated with SPACE, it is not clear whether 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

Table 5. Median and IQR of area under the curve, AUC, for 50 simulations†

Solver	*Results for n = 200*		*Results for n = 400*		*Results for n = 800*	
	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

†Each simulation yields an ROC curve from which AUC is computed for FPR in the interval [0, 0.15] and normalized to 1.

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 convex correlation selection method on a real data set. To facilitate comparison, we consider data from a breast cancer study (Chang *et al*., 2005) on which SPACE was illustrated. This data set contains expression levels of 24481 genes on 248 patients with breast cancer. The data set 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 data set. In particular, survival analysis via univariate Cox regression with patient survival times is used to select a subset of genes that are closely associated with breast cancer. A choice of p-value less than 0.0003 yields a reduced data set with 1107 genes. This subset of the data is then mean centred 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). For this, we focus on identifying the 10 most highly connected genes ('hub' genes) identified by each partial correlation graph estimation method. Table 6 in the on-line supplemental section L summarizes the top 10 hub genes obtained by CONCORD, SYMLASSO, SPACE1 and SPACE2. That table also gives references from the biomedical literature that place these genes in the context of breast cancer. These references illustrate that most of the genes identified 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) suggested 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) (`http://www.ncbi.nlm.nih.gov/pubmed/22110403`) 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 *versus* the top 10), CONCORD identifies additional genes that have not been 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 convex correlation selection is a useful addition to the graphical models literature as it is competitive with other methods in terms of model selection accuracy, timing and 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 set-

ting where a stable estimate of the covariance matrix is often required. We follow closely the exposition of the problem that was given in Won *et al*. (2013). A portfolio of financial instruments constitutes a collection of both risky and risk-free assets that are 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 correspond 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 theory, this risk is taken to be the standard deviation of the portfolio (Markowitz, 1952). As noted in Luenberger (1997) and 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 mean variance portfolio theory. 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, compared with 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 constructing a minimum variance portfolio. The performance of each of the methods and the associated strategies will be compared over a sustained period of time 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, $r_t^{\mathrm{T}} = (r_{1t}, r_{2t}, \ldots, r_{pt})$. The portfolio weights $w_k^{\mathrm{T}} = (w_{1k}, w_{2k}, \ldots, w_{pk})$ denote the weight of asset $i = 1, \ldots, p$ in the portfolio for the kth 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 $\mathbf{1}^{\mathrm{T}} w_k = 1$, where $\mathbf{1}$ denotes the vector of all 1s. Note that the risk of a given portfolio as measured by the standard deviation of its return is simply $(w_k^{\mathrm{T}} \Sigma w_k)^{1/2}$.

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

$$\text{minimize } w_k^{\mathrm{T}} \Sigma w_k \qquad \text{subject to } \mathbf{1}^{\mathrm{T}} w_k = 1. \tag{17}$$

As definition (17) is a simple quadratic programme, it has an analytic solution given by $w_k^* = (\mathbf{1}^{\mathrm{T}} \Sigma^{-1} \mathbf{1})^{-1} \Sigma^{-1} \mathbf{1}$. The solution depends on the theoretical covariance matrix Σ. In practice, the parameter Σ must 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 on-line supplemental section M.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 index. The Dow Jones industrial average is a composite blue chip index consisting of 30 stocks (note that the Kraft Foods data were removed in our analysis because of their limited data span). (Kraft Foods were a component stock of the Dow Jones industrial average from September 22nd, 2008, to September 13th, 2012. From September 14th, 2012, Kraft Foods were replaced with the United Health Group.) Table 7 in the on-line supplemental section M.1 lists the 29 component stocks that were used in our analysis.

Rebalancing time points were chosen to be every 4 weeks starting from February 18th, 1995, to October 26th, 2012 (approximately 17 years), and are shown in Table 8 in the on-line supplemental section M.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, the graphical lasso of Friedman *et al.* (2008), CONCORD, the condition-number-regularized estimator CondReg of Won *et al.* (2013) and the Ledoit–Wolf estimator of Ledoit and Wolf (2004). We consider various choices of N_{est}, in particular $N_{\mathrm{est}} \in \{35, 40, 45, 50, 75, 150, 225, 300\}$, in our analysis. Once a choice for N_{est} has been made, it is kept constant for all the 231 investment periods.

For l_1-penalized regression methods, such as Glasso and CONCORD, a value for the penalty parameter must be chosen. For the purposes of this study, cross-validation was performed within each estimation horizon to minimize the residual sum of squares from out-of-sample prediction averaged over all stocks. Further details are given in the on-line supplemental section M.4. The condition-number-regularized and Ledoit–Wolf estimators each use different criteria to perform cross-validation. The readers is referred to Won *et al.* (2013) and Ledoit and Wolf (2004) for details on the cross-validation procedure for these methods. For comparison with Won *et al.* (2013), we use the following quantities to assess the performance of the five minimum variance rebalancing strategies: the 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 the on-line supplemental section M.5.

Table 6 gives the realized Sharpe ratio of all MVR strategies for the various choices of estimation horizon N_{est}. The column DJIA stands for the passive index tracking strategy that tracks

Table 6. Realized Sharpe ratio of various investment strategies corresponding to different estimators with various N_{est}†

N_{est}	*Ratios for the following methods:*					
	Sample	*Graphical 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

†The maximum annualized Sharpe ratios for each row, and others within 1% of this maximum, are highlighted in italics.

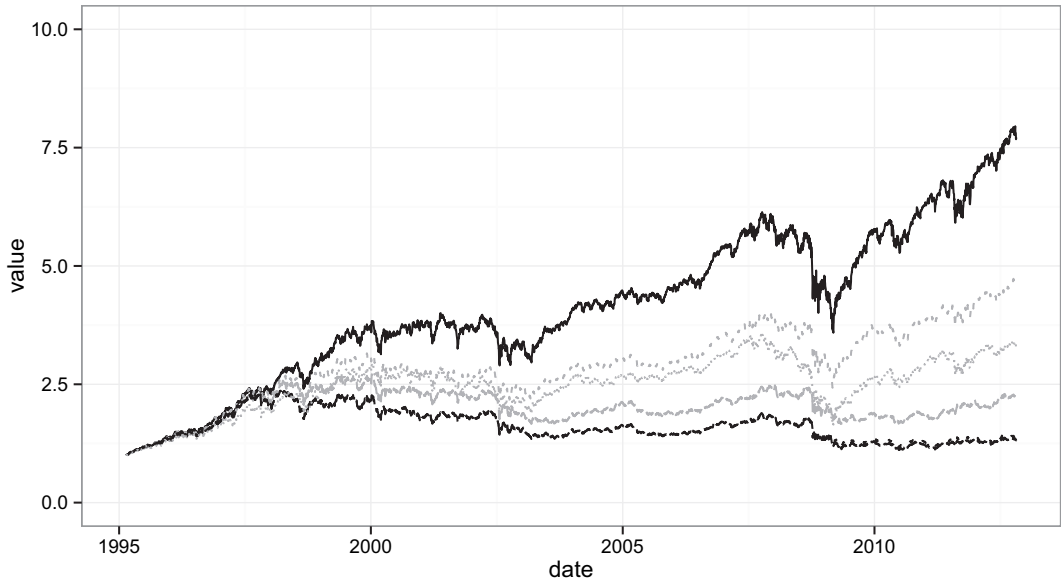

Fig. 2. Normalized wealth growth after adjusting for transaction costs (0.5% of the principal) and borrowing costs (interest rate of 7% annualized percentage rate) with $N_{\text{est}} = 225$: ———, CONCORD; - - - - -, CondReg; - - - - -, graphical lasso; — —, Ledoit–Wolf; — —, Sample; · · · · ·, Dow Jones industrial average

the Dow Jones industrial average index. It is clear from Table 6 that CONCORD performs uniformly well across different choices of estimation horizons.

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

We also note that trading costs that are 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 the on-line supplemental section M.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.

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 $\{\bar{\Omega}_n\}_{n \geqslant 1}$ denote the sequence of true inverse covariance matrices. As in Peng *et al.* (2009), for consistency, 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.

Assumption 1 (accurate diagonal estimates). There are estimates $\{\hat{\alpha}_{n,ii}\}_{1 \leqslant i \leqslant p_n}$ such that, for any $\eta > 0$, there is a constant $C > 0$ such that

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

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

The theory that follows is valid when the estimates $\{\hat{\alpha}_{n,ii}\}_{1\leqslant i\leqslant p_n}$ and the estimates of the off-diagonal entries are obtained from the same data set. When $\limsup_{n\to\infty} p_n/n < 1$, Peng *et al.* (2009) showed that the diagonal entries of S^{-1} can be used as estimates of the diagonal entries of Ω. However, no such general recipe was 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 ∞.

For vectors $\omega^{\mathrm{o}} \in \mathbb{R}^{p_n(p_n-1)/2}$ and $\omega^{\mathrm{d}} \in \mathbb{R}_+^{p_n}$, the notation $\mathcal{L}_n(\omega^{\mathrm{o}}, \omega^{\mathrm{d}})$ stands for $\mathcal{L}_{\mathrm{con}}/n$ ($\mathcal{L}_{\mathrm{con}}$ is defined in expression (8)) evaluated at a matrix with off-diagonal entries ω^{o} and diagonal entries ω^{d}. Let $\bar{\omega}_n^{\mathrm{o}} = ((\bar{\omega}_{n,ij}))_{1\leqslant i<j\leqslant p_n}$ denote the vector of off-diagonal entries of $\bar{\Omega}_n$, and $\hat{\boldsymbol{\alpha}}_{p_n} \in \mathbb{R}_+^{p_n}$ denotes the vector with entries $\{\hat{\alpha}_{n,ii}\}_{1\leqslant i\leqslant p_n}$. Let $\mathcal{A}_n$ denote the set of non-zero entries in the vector $\bar{\omega}_n^{\mathrm{o}}$, and let $q_n = |\mathcal{A}_n|$. Let $\bar{\Sigma}_n = \bar{\Omega}_n^{-1}$ denote the true covariance matrix for every $n \geqslant 1$. The following standard assumptions are required.

Assumption 2 (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.

Assumption 3 (sub-Gaussianity). The random vectors $\mathbf{Y}^1, \ldots, \mathbf{Y}^n$ are IID sub-Gaussian for every $n \geqslant 1$, i.e. there is a constant $c > 0$ such that, for every $\mathbf{x} \in \mathbb{R}^{p_n}$, $E[\exp(\mathbf{x}'\mathbf{Y}^i)] \leqslant \exp(c\mathbf{x}'\bar{\Sigma}_n\mathbf{x})$ and, for every $i, j > 0$, there exists $\eta_j > 0$ such that $E[\exp\{t(Y_j^i)^2\}] < K$ whenever $|t| < \eta_j$. Here K is independent of i and j.

Assumption 4 (incoherence condition). There exists $\delta < 1$ such that, for all $(i, j) \notin \mathcal{A}_n$,

$$|\bar{\mathcal{L}}''_{ij,\mathcal{A}_n}(\bar{\Omega}_n)\, \bar{\mathcal{L}}''_{\mathcal{A}_n,\mathcal{A}_n}(\bar{\Omega}_n)^{-1} \mathrm{sgn}(\bar{\omega}^{\mathrm{o}}_{\mathcal{A}_n})| \leqslant \delta,$$

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

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

Conditions analogous to assumption 4 have been used in Zhao and Yu (2006), Peng *et al.* (2009) and Meinshausen and Bühlmann (2006) to establish high dimensional model selection consistency. In the context of lasso regression, Zhao and Yu (2006) showed that such a condition (which they referred to as an irrepresentable condition) is almost necessary and sufficient for model selection consistency and they provided some examples when this condition is satisfied. We provide some examples of situations where condition 4 is satisfied, along the lines of Zhao and Yu (2006), in the on-line supplemental section P.

Define $\bar{\theta}_n^{\mathrm{o}} = ((\bar{\theta}_{n,ij}))_{1\leqslant i<j\leqslant p_n} \in \mathbb{R}^{p_n(p_n-1)/2}$ by $\bar{\theta}_{n,ij} = \bar{\omega}_{n,ij}/\sqrt{(\hat{\alpha}_{n,ii}\hat{\alpha}_{n,jj})}$ for $1 \leqslant i < j \leqslant p_n$. 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 1–4 are satisfied. Suppose that $p_n = O(n^{\kappa})$ for some $\kappa > 0$, $q_n = o[\sqrt{\{n/\log(n)\}}]$, $\sqrt{\{q_n \log(n)/n\}} = o(\lambda_n)$, $\lambda_n\sqrt{\{n/\log(n)\}} \to \infty$, $s_n/\sqrt{q_n}\lambda_n \to \infty$ and $\sqrt{q_n}\lambda_n \to 0$, as $n \to \infty$. Then there is a constant C such that, for any $\eta > 0$, the following events hold with probability at least $1 - O(n^{-\eta})$.

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

The proof of theorem 2 is provided in the on-line supplemental section N.

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 to retain their strengths. We first placed the highly useful SPACE method in an optimization framework, which in turn allowed us to identify SPACE with a specific objective function. These and other insights led to the formulation of the CONCORD objective function. It was then shown that the CONCORD objective function is comprised of quadratic forms, is convex and can be regarded as a penalized pseudolikelihood. A co-ordinatewise descent algorithm that minimizes this objective, via closed form iterates, was proposed and subsequently analysed. The convergence of this co-ordinatewise descent algorithm was 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. The large sample properties of CONCORD establish consistency of the method as both the sample size and dimension tend to ∞. The performance of CONCORD was also illustrated via simulations and was shown to be competitive in terms of graphical model selection accuracy and timing. CONCORD was then applied to a biomedical data set and to a finance data set, leading to novel findings. Last, but not least, a framework that unifies all pseudolikelihood methods was 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 the graphical lasso) and rather use only pseudolikelihood methods. We note that CONCORD is attractive over the Glasso algorithm 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 method could be complementary in practice.

Acknowledgements

KK was supported in part by National Science Foundation grant DMS-1106084. SO and BR were supported in part by the National Science Foundation under grants DMS-0906392, DMS-CG 1025465, AGS-1003823, DMS-1106642, DMS CAREER-1352656 and grants NSA H98230-11-1-0194, DARPA-YFA N66001-111-4131 and SMC-DBNKY. Mark Hayes and SMC are gratefully acknowledged for useful discussions.

References

Banerjee, O., El Ghaoui, L. and D'Aspremont, A. (2008) Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, **9**, 485–516.

Besag, J. (1975) Statistical analysis of non-lattice data. *Statistician*, **24**, 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. *J. Biol. Chem.*, **284**, 33177–33184.

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*, **20**, 2242–2250.

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, Cambridge.
Chang, H. Y., Nuyten, D. S. A., Sneddon, J. B., Hastie, T., Tibshirani, R., Sørlie, T., Dai, H., He, Y. D., van't Veer, L. J., Bartelink, H., van de Rijn, M., Brown, P. O. and van de Vijver, M. J. (2005) Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc. Natn. Acad. Sci. USA*, **102**, 3738–3743.
Fawcett, T. (2006) An introduction to ROC analysis. *Pattn Recogn Lett.*, **27**, 861–874.
Friedman, J., Hastie, T. and Tibshirani, R. (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 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, Stanford.
Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J. M., Cusick, M. E., Roth, F. P. and Vidal, M. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, 88–93.
Hastie, T., Tibshirani, R. and Friedman, J. H. (2009) *The Elements of Statistical Learning*. New York: Springer.
Jensen, S. R. T., Johansen, S. R. and Lauritzen, S. L. (1991) Globally convergent algorithms for maximizing likelihood function. *Biometrika*, **78**, 867–877.
Jeong, H., Mason, S. P., Barabasi, A.-L. and Oltvai, Z. N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.
Lauritzen, S. L. (1996) *Graphical Models*. New York: Oxford University Press.
Ledoit, O. and Wolf, M. (2004) A well-conditioned estimator for large-dimensional covariance matrices. *J. Multiv. Anal.*, **88**, 365–411.
Lee, J. D. and Hastie, T. J. (2014) Learning the stucture of mixed graphical models. *J. Computnl Graph. Statist.*, to be published.
Luenberger, D. G. (1997) *Investment Science*. New York: Oxford University Press.
Markowitz, H. (1952) Portfolio selection. *J. Finan.*, **7**, 77–91.
Maxwell, C. A., Benítez, J., Gómez-Bald, L, Osorio, A., Bonifaci, N., Fernández-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., Caldés, T., Andrés, 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., Sinilnikoya, 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**, no. 11, article e1001199.
Mazumder, R. and Hastie, T. (2012) Exact covariance thresholding into connected components for large-scale graphical lasso. *J. Mach. Learn. Res.*, **13**, 781–794.
Meinshausen, N. and Bühlmann, P. (2006) High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, **34**, 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, Cambridge.
Newman, M. E. J. (2003) The structure and function of complex networks. *SIAM Rev.*, **45**, 167–256.
Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009) Partial correlation estimation by joint sparse regression models. *J. Am. Statist. Ass.*, **104**, 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, University of California at Berkeley, Berkeley.
Speed, T. P. and Kiiveri, H. T. (1986) Gaussian Markov distributions over finite graphs. *Ann. Statist.*, **14**, 138–150.
Tseng, P. (1988) Coordinate ascent for maximizing nondifferentiable concave functions. *Technical Report*. Massachusetts Institute of Technology, Cambridge.
Tseng, P. (2001) Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optimizn Theor. Appl.*, **109**, 475–494.

Won, J.-H., Lim, J., Kim, S.-J. and Rajaratnam, B. (2013) Condition-number-regularized covariance estimation. *J. R. Statist. Soc.* B, **75**, 427–450.
Xu, P.-F., Guo, J. and He, X. (2011) An improved iterative proportional scaling procedure for Gaussian graphical models. *J. Computnl Graph. Statist.*, **20**, 417–431.
Zangwill, W. I. (1969) *Nonlinear Programming: a Unified Approach*. Englewood Cliffs: Prentice-Hall.
Zhao, P. and Yu, B. (2006) On model selection consistency of lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplemental section'.