Generalized Pseudolikelihood Methods for Inverse Covariance Estimation

Alnur Ali
Machine Learning Dept.
Carnegie Mellon University
alnurali@cmu.edu

Kshitij Khare
Dept. of Statistics
University of Florida
kdkhare@stat.ufl.edu

Sang-Yun Oh
Dept. of Stats. and Applied Prob.
UC Santa Barbara
syoh@pstat.ucsb.edu

Bala Rajaratnam
Dept. of Statistics
UC Davis
brajaratnam01@gmail.com

Abstract

We introduce PseudoNet, a new *pseudolikelihood*-based estimator of the inverse covariance matrix, that has a number of useful statistical and computational properties. We show, through detailed experiments with synthetic as well as real-world finance and wind power data, that PseudoNet outperforms related methods in terms of estimation error and support recovery, making it well-suited for use in a downstream application, where obtaining low estimation error can be important. We also show, under regularity conditions, that PseudoNet is consistent. Our proof assumes the existence of accurate estimates of the diagonal entries of the underlying inverse covariance matrix; we additionally provide a two-step method to obtain these estimates, even in a high-dimensional setting, going beyond the proofs for related methods. Unlike other pseudolikelihood-based methods, we also show that PseudoNet does not *saturate*, *i.e.*, in high dimensions, there is *no* hard limit on the number of nonzero entries in the PseudoNet estimate. We present a fast algorithm as well as *screening rules* that make computing the PseudoNet estimate over a range of tuning parameters tractable.

1 INTRODUCTION

We consider the problem of obtaining a sparse estimate of the inverse covariance matrix in a high-dimensional setup, where the number of variables (*i.e.*, features) p is possibly much larger than the number of data samples n. This is an important problem in modern statistics as well as across a variety of applications.

In high dimensions (*i.e.*, when $p \gg n$), it makes sense to obtain an estimate by maximizing an ℓ_1-penalized Gaussian likelihood (see, *e.g.*, Yuan and Lin (2007); Banerjee et al. (2008); Friedman et al. (2008); Rothman et al. (2008)) — although other penalities are certainly possible. This is, of course, a massive area of research, and a number of estimators for, as well as extensions to, this basic Gaussian setup have been proposed over the years, including the seminal graphical lasso algorithm (GLasso) of Friedman et al. (2008). *Pseudolikelihood*-based estimators (Besag, 1974) take a somewhat different approach, in that they can be seen as minimizing the sum of ℓ_1-penalized regression (*i.e.*, lasso) problems, which more directly exploits the connection between the inverse covariance matrix and partial correlations; see, *e.g.*, Meinshausen and Bühlmann (2006); Rocha et al. (2008); Peng et al. (2009); Friedman et al. (2010); Khare et al. (2015); Ali et al. (2016). Pseudolikelihood-based estimators are thus, in a sense, more flexible in moving beyond the usual Gaussian setup.

Under the assumption that the data-generating process is multivariate normal, it is a well-known fact that the random variables i and j are conditionally independent given the remaining variables if and only if the (i, j) entry in the underlying inverse covariance matrix is zero (see, *e.g.*, Lauritzen (1996)); as a result, much work has looked at producing estimates that accurately recover the underlying support (*i.e.*, the set of nonzero entries), which makes these estimates more interpretable. On the other hand, we often want to use an estimate later in our workflow, in which case low estimation error (as measured by a suitable matrix norm) is perhaps a more useful criterion for evaluating an estimate. Asymptotically, the SPACE and CONCORD pseudolikelihood-based esti-

Appearing in Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Flordia, USA. JMLR: W&CP volume 54.

mators of Peng et al. (2009) and Khare et al. (2015), respectively, have been shown to be consistent (in a Frobenius norm sense) under certain conditions; however, carefully checking the conditions required by the consistency proofs in these papers reveals that they presume the existence of accurate estimates of the diagonal entries of the underlying inverse covariance matrix. A natural choice here is to simply use the diagonal entries of the sample inverse covariance matrix, but such estimates unfortunately do not exist when $p > n$, and alternatives are not immediately apparent.

Returning to the issue of interpretability of pseudolikelihood-based estimates, we raise a basic question: are the estimates given by pseudolikelihood-based methods well-defined (*i.e.*, unique)? We elaborate below (see Section 1.3), but the short answer to this question for now is that the estimates given by many pseudolikelihood-based methods, including SPACE, CONCORD, the SPLICE estimator of Rocha et al. (2008), as well as the Symmetric Lasso estimator of Friedman et al. (2010), may not be unique, and in fact many of these methods may not even converge to a particular estimate — which can be problematic from an interpretability point of view. For example, in a finance application, we may wish to understand which assets are correlated, in order to assemble a diversified portfolio (Markowitz, 1952); if the outcome of an estimation procedure is not unique, then which assets should we use?

Furthermore, given the connection between pseudolikelihood-based methods and the lasso, we recall a basic result from lasso theory, which states that the lasso can *saturate*, meaning that when $p > n$, there exists a lasso estimate with at most n nonzero entries (equivalently, selected variables) (Rosset et al., 2004; Zou and Hastie, 2005; Tibshirani, 2013); this behavior can be quite limiting from the points of view of interpretability as well as estimation error. It is therefore natural to ask: do estimates given by existing pseudolikelihood-based methods also saturate? We show that several estimators, including SPACE, CONCORD, and SPLICE, unfortunately can saturate, which establishes an analogous result for undirected graphical models (see Section 4.3).

1.1 Overview of contributions

In this paper, we introduce a new, more flexible pseudolikelihood-based estimator of the inverse covariance matrix, which we call PseudoNet, that addresses all the aforementioned issues with existing pseudolikelihood-based methods, and additionally possesses a number of other useful statistical and computational properties. We give a brief summary below.

Computational aspects and uniqueness. We present a fast algorithm for computing the PseudoNet estimate by leveraging recent advances in convex optimization, and show that our algorithm converges at a geometric ("linear") rate to the (global) solution of a convex optimization problem that defines the PseudoNet estimate. Furthermore, this solution is unique, as the objective in the optimization problem is strictly convex. This contrasts with a number of other pseudolikelihood-based methods (Rocha et al., 2008; Peng et al., 2009; Friedman et al., 2010; Khare et al., 2015; Oh et al., 2014), which do not provide unique estimates, making interpretation difficult, and are either not guaranteed to converge or converge at a slower rate (*e.g.*, the CONCORD estimator).

We also derive *screening rules* for PseudoNet, which make the optimization problem much faster to solve by omitting some of the variables (Banerjee et al., 2008; Tibshirani et al., 2012; Mazumder and Hastie, 2012). These rules can be implemented as simple checks based on the optimality conditions of the PseudoNet optimization problem; in some cases, we are able to reduce the size of the optimization problem by 90%.

Estimation error. We show that PseudoNet significantly outperforms the closely related CONCORD estimator, which we build upon, in terms of estimation error (as measured by several matrix norms), while also outperforming CONCORD in terms of support recovery. As mentioned above, although the literature often emphasizes support recovery, obtaining an estimate with low estimation error is perhaps more useful in situations where our estimate will be used by a downstream application.

Consistency. We also show, under standard regularity conditions, that PseudoNet is consistent at a rate of $\sqrt{(\log p)/n}$. The consistency proofs for the related pseudolikelihood-based estimators SPACE and CONCORD assume the existence of accurate estimates of the diagonal entries of the underlying inverse covariance matrix, but do not provide a method for obtaining these estimates when $p > n$. In this paper, we go further and give a two-step method that obtains accurate diagonal estimates, even when $p > n$; this result is therefore also useful in the consistency proofs for SPACE (Peng et al., 2009, Theorem 3) and CONCORD (Khare et al., 2015, Theorem 2).

Saturation. We show that the PseudoNet estimate does *not* saturate, meaning that when $p \gg n$, the number of variables selected by PseudoNet can be greater than np (out of $p(p-1)/2$ total variables), which is not true for several other pseudolikelihood-based estimators (Rocha et al., 2008; Peng et al., 2009; Khare et al.,

2015); establishing this result involves generalizing an analogous claim for the (standard) lasso as in, *e.g.*, Rosset et al. (2004); Tibshirani (2013). This result is useful from the points of view of the estimation error as well as the interpretability of the PseudoNet estimate.

Non-Gaussian data. Lastly, we illustrate, through numerical examples with real finance and wind power data, that PseudoNet deals effectively with non-Gaussian data, outperforming several strong baselines. This is due, in part, to the precise form of the objective in the PseudoNet optimization problem, which dispenses with the assumption that the true distribution is normal, and is helpful in moving beyond the usual Gaussian setup.

1.2 Outline

An outline for the rest of this paper is as follows. In the next subsection, we survey related work. In Section 2, we describe the PseudoNet estimator and its screening rules. In Section 3, we present an empirical evaluation of PseudoNet, as well as several baselines, on synthetic and real-world data. We present all of our theoretical results on PseudoNet's statistical and computational properties in Section 4; all of our proofs are given in the supplement. We conclude in Section 5.

1.3 Related work

The literature on high-dimensional sparse inverse covariance estimation is quite vast; we do not claim to give a complete treatment of it here, and instead highlight work most related to our own. Yuan and Lin (2007); Banerjee et al. (2008); Friedman et al. (2008); Rothman et al. (2008) first proposed estimating the inverse covariance matrix by maximizing an ℓ_1-penalized Gaussian likelihood; Friedman et al. (2008) proposed the GLasso, a fast algorithm for computing an estimate in this framework. In a related but distinct line of work, a number of pseudolikelihood-based estimators have been proposed; pseudolikelihood-based methods take a somewhat different perspective, in that they can be seen as roughly minimizing a series of ℓ_1-penalized regression problems, making them arguably simpler to analyze and extend than other approaches. The seminal *neighborhood selection* method of Meinshausen and Bühlmann (2006), which fits a lasso regression of each variable on the rest, is an example; a drawback of neighborhood selection is that the estimate may not be symmetric, so a post-processing step is required.

In a nice step forward, Peng et al. (2009) introduced the SPACE estimator, and showed that it is symmetric and also consistent, under suitable regularity conditions. Unfortunately, SPACE is not guaranteed to converge (it is easy to find examples where the iterates produced by SPACE alternate between two values), and furthermore the SPACE estimate may not be unique (Khare et al., 2015); additionally, the consistency proof for SPACE assumes that accurate estimates for the diagonal entries of the underlying inverse covariance matrix are available, even when $p > n$, without giving a method to obtain them. Inspired by SPACE, Friedman et al. (2010) introduced the Symmetric Lasso estimator, which is also symmetric, but is not guaranteed to converge, be unique, or be consistent (Khare et al., 2015, Lemma 2). The SPLICE estimator of Rocha et al. (2008) has some useful computational properties, but does not have any of these guarantees either (Khare et al., 2015, Lemma 3).

Building on SPACE, the CONCORD estimator (Khare et al., 2015; Oh et al., 2014) recently made useful progress: CONCORD is symmetric, like SPACE, but is additionally guaranteed to converge at a rate of $O(1/k^2)$, where k here is the number of iterations, and is also consistent. On the downside, as we show later in this paper, CONCORD's consistency proof assumes accurate diagonal estimates even when $p > n$, its estimate may not be unique when $p > n$, and it can saturate (*i.e.*, when $p \gg n$, the CONCORD estimate can select at most np out of $p(p-1)/2$ total variables).

2 THE PseudoNet ESTIMATOR

Assume that we are given n samples $X_{1\cdot}, \ldots, X_{n\cdot} \in \mathbf{R}^p$, drawn i.i.d. from some unknown distribution that, without a loss of generality, we take to have mean zero and covariance matrix $\Sigma^0 \in \mathbf{S}^p_{++}$ (the space of $p \times p$ positive definite matrices). We want to estimate the underlying inverse covariance matrix $\Omega^0 = (\Sigma^0)^{-1}$ with a small number of nonzero entries.

We define the PseudoNet estimate, which gives a sparse estimate of the underlying inverse covariance matrix, as the solution of the following convex optimization problem:

$$\begin{aligned}\underset{\Omega \in \mathbf{R}^{p\times p}}{\text{minimize}} \quad & -(1/2)\textstyle\sum_{i=1}^p \log(\Omega_{ii}^2) \\ & +(1/2)\textstyle\sum_{i=1}^p \left\| \Omega_{ii}X_i + \sum_{j\neq i}\Omega_{ij}X_j \right\|_2^2 \\ & +\lambda_1 \textstyle\sum_{i\neq j}|\Omega_{ij}| + (\lambda_2/2)\|\Omega\|_F^2,\end{aligned}$$

where $\lambda_1, \lambda_2 > 0$ are tuning parameters, and $\|\cdot\|_F$ is the Frobenius norm. After some manipulations, we can put the above optimization problem into the following matrix form, which is useful for much of the rest of the paper:

$$\begin{aligned}\underset{\Omega \in \mathbf{R}^{p\times p}}{\text{minimize}} \quad & -(1/2)\log\det(\Omega^2_{\text{diag}}) + (n/2)\,\mathbf{Tr}\,S\Omega^2 \\ & +\lambda_1\|\Omega_{\text{off}}\|_1 + (\lambda_2/2)\|\Omega\|_F^2. \end{aligned} \tag{1}$$

Here, $\Omega_{\text{diag}} \in \mathbf{R}^{p\times p}$ is a matrix of the diagonal entries of Ω, with its off-diagonal entries set to zero; $S \in \mathbf{R}^{p\times p}$ is the sample covariance matrix, *i.e.*, $S = \frac{1}{n}X^TX$, and $X \in \mathbf{R}^{n\times p}$ is a data matrix; $\Omega_{\text{off}} \in \mathbf{R}^{p\times p}$ is a matrix of the off-diagonal entries of Ω, with its diagonal entries set to zero; and $\|\cdot\|_1$ is the elementwise ℓ_1 norm.

Note that we do not make the assumption here that the underlying data-generating process is, *e.g.*, multivariate normal, which is helpful in moving beyond the usual Gaussian setup; nonetheless, the objective of the PseudoNet optimization problem in matrix form (1) does bear some resemblance to an ℓ_1-penalized Gaussian likelihood. In fact, the PseudoNet optimization problem (1) generalizes the (standard) ℓ_1-penalized Gaussian maximum likelihood problem (by design), when (1) is written as

$$\begin{aligned}\underset{\Omega\in\mathbf{R}^{p\times p}}{\text{minimize}} \quad & -(1/2)\log\det F(\Omega) + (n/2)\,\mathbf{Tr}\,SG(\Omega) \\ & +\lambda_1\|H(\Omega)\|_1 + (\lambda_2/2)\|\Omega\|_F^2,\end{aligned}$$

for some operators $F, G, H : \mathbf{R}^{p\times p} \to \mathbf{R}^{p\times p}$. (Taking F as $\Omega \mapsto \Omega_{\text{diag}}^2$, G as $\Omega \mapsto \Omega^2$, and H as $\Omega \mapsto \Omega_{\text{off}}$ recovers the PseudoNet optimization problem (1).) Now taking F, G, H all as $\Omega \mapsto \Omega$, with $\lambda_2 = 0$, recovers the GLasso optimization problem (Friedman et al., 2008, Equation 1). Furthermore, the framework above also generalizes several pseudolikelihood-based approaches; *e.g.*, taking F as $\Omega \mapsto \Omega_{\text{diag}}$, G as $\Omega \mapsto \Omega\Omega_{\text{diag}}^{-1}\Omega$, H as $\Omega \mapsto \Omega_{\text{off}}$, and $\lambda_2 = 0$ recovers the (nonconvex) SPACE optimization problem (Peng et al., 2009, Equation 2), and taking F as $\Omega \mapsto \Omega_{\text{diag}}^2$, G as $\Omega \mapsto \Omega^2$, H as $\Omega \mapsto \Omega_{\text{off}}$, and $\lambda_2 = 0$ recovers the CONCORD optimization problem (Khare et al., 2015, Equation 8), revealing a close connection between the PseudoNet and CONCORD optimization problems.

Although simple in appearance, the squared Frobenius norm penalty in the PseudoNet optimization problem (1) gives PseudoNet a number of statistical and computational advantages (that are not always simple to show) over many other pseudolikelihood-based approaches, including the ones just mentioned. Statistically, owing to this penalty, PseudoNet is able to obtain much better estimation error than CONCORD (see Sections 3, 4.2, and 4.2.1), which is again useful when our estimate will be used by a downstream application; PseudoNet's estimates also tend to be more stable than CONCORD's. We can understand this intuitively, by considering the relationship between the *elastic net* (Zou and Hastie, 2005) and the (standard) lasso optimization problems: the elastic net augments the objective in the lasso optimization problem with a ridge penalty, which is seen as giving a sparse estimate with better prediction error than the associated lasso estimate — taking a pseudolikelihood-based approach makes it natural to incorporate these ridge penalties into each regression (sub)problem in order to obtain a sparse estimate of the inverse covariance matrix with low estimation error.

The elastic net is also an elegant solution to the issue of saturation in the lasso (*i.e.*, when $p > n$, the number of variables selected by the lasso can be at most n). Even though pseudolikelihood-based estimators and the lasso are connected in many ways, it is still natural to wonder if pseudolikelihood-based estimators can also saturate, since the objectives in the defining optimization problems for many pseudolikelihood-based estimators include terms that go beyond pure lasso regressions? We show later (see Section 4.3) that several pseudolikelihood-based estimators (specifically, SPLICE, SPACE, and CONCORD) indeed can saturate — and that the squared Frobenius norm penalty in the PseudoNet optimization problem (1) is what prevents it from saturating. This is a useful result for PseudoNet, from the points of view of estimation error as well as interpretability.

Finally, the choices of F, G, H that we make in the framework above in order to arrive at the PseudoNet optimization problem (1) ensure that (1) is convex; further imposing the squared Frobenius norm penalty guarantees that the objective in (1) is strictly convex, and hence the PseudoNet estimate is always unique (as mentioned above, convexity as well as uniqueness are not guaranteed for many pseudolikelihood-based estimators). Computationally, the squared Frobenius norm penalty allows us to derive a fast algorithm for computing the PseudoNet estimate (which we do next) that converges to the unique, global solution of the PseudoNet optimization problem (1) at a geometric rate (see Section 4.1), and is much faster than CONCORD (see Section 3).

Next, we turn to deriving a fast algorithm for computing the PseudoNet estimate. Rewriting (1) as the sum of a smooth function g and a nonsmooth function h, *i.e.*, letting $f(\Omega)$ be the objective in (1), we have that $f(\Omega) = g(\Omega) + h(\Omega)$, with $h(\Omega) = \lambda_1\|\Omega_{\text{off}}\|_1$ and

$$g(\Omega) = -\frac{1}{2}\log\det(\Omega_{\text{diag}}^2) + \frac{n}{2}\,\mathbf{Tr}\,S\Omega^2 + \frac{\lambda_2}{2}\|\Omega\|_F^2. \tag{2}$$

The presence of the nonsmooth term h here makes the PseudoNet optimization problem (1) difficult to solve using, say, an interior point method. On the other hand, h does admit a computationally efficient *proximal operator* (Parikh and Boyd, 2013), *i.e.*,

$$\begin{aligned}\mathbf{prox}_{th}(V) &= \underset{Z\in\mathbf{R}^{p\times p}}{\text{argmin}}\left(h(Z) + \frac{1}{2t}\|Z - V\|_F^2\right) \\ \implies [\mathbf{prox}_{th}(V)]_{ij} &= \begin{cases} V_{ij} - t & V_{ij} > t \\ 0 & |V_{ij}| \le t \\ V_{ij} + t & V_{ij} < -t, \end{cases}\end{aligned} \tag{3}$$

for $i, j = 1, \ldots, p$, some $V \in \mathbf{R}^{p \times p}$, and a constant $t > 0$; (3) is known as the *soft-thresholding* operator. A proximal gradient method is thus a natural choice here; *i.e.*, on each iteration of the algorithm, we take a step in the direction of the negative gradient of g, and then apply (3). Provided that the gradient of g is Lipschitz and the step sizes are chosen appropriately, proximal gradient methods in general obtain a convergence rate of $O(1/k)$, where k here is the number of iterations. However, we are able to obtain a much better (*i.e.*, geometric) rate of convergence, owing to the strong convexity of (1), as we show later in Section 4.1.

To complete the specification of the proximal gradient method, we give the gradient and Hessian of the smooth term g in (2):

$$\nabla g(\Omega) = -\Omega_{\text{diag}}^{-1} + (n/2)(S\Omega + \Omega S) + \lambda_2 \Omega \quad (4)$$

$$\nabla^2 g(\Omega) = \sum_{i=1}^{p} (1/\Omega_{ii}^2)(e_i e_i^T \otimes e_i e_i^T) + (n/2)(S \otimes I + I \otimes S) + \lambda_2 I_{p^2}, \quad (5)$$

where $\otimes$ is the Kronecker product, and e_i is the ith standard basis vector in $\mathbf{R}^p$. A complete specification of our proximal gradient method, along with a way to choose the tuning parameters, is in the supplement.

2.1 Omitting predictors via screening rules

We often want to solve the PseudoNet optimization problem (1) over a grid of (λ_1, λ_2) values, and then choose a suitable estimate. By leveraging the nature of the PseudoNet optimization problem, we are able to derive *sequential strong* screening rules here (Tibshirani et al., 2012), which are well-suited for this as they omit variables from the PseudoNet optimization problem as we solve it over many tuning parameters.

Tibshirani et al. (2012) introduced sequential strong screening rules as a framework for deriving screening rules that drop variables as we solve a sequence of convex optimization problems; these optimization problems are required to have an objective that can be expressed as the sum of a smooth loss and a potentially nonsmooth penalty. Sequential strong rules are based on the optimality conditions for the optimization problem in question, as well as the assumption that the gradient of the smooth loss is *nonexpansive*, *i.e.*, that it has a Lipschitz constant equal to one; thus, strong rules might commit *violations*, *i.e.*, they might suggest that a variable could be dropped when it is actually nonzero at the solution. Consequently, we check the optimality conditions after applying sequential strong rules; we do so in our numerical experiments, and never observe a violation (see Sections 3.1 and the supplement). We state our rules in Lemma 2.1; an algorithmic specification is in the supplement.

Lemma 2.1 (Screening rules)**.** *Let $\lambda_1^{(1)} \geq \cdots \geq \lambda_1^{(r)}$ and $\lambda_2^{(1)} \geq \cdots \geq \lambda_2^{(s)}$ be nonincreasing sequences of tuning parameters. Also let $\hat{\Omega}^{net}(\lambda_1^{(k)}, \lambda_2^{(\ell)})$ be the solution of the PseudoNet optimization problem* (1)*, for a particular $\lambda_1^{(k)}$, $\lambda_2^{(\ell)}$, $k = 1, \ldots, r$, $\ell = 1, \ldots, s$. Write the components of the gradient of the smooth parts of the objective in* (1) *evaluated at $\hat{\Omega}^{net}(\lambda_1^{(k-1)}, \lambda_2^{(\ell)})$ as*

$$C_{ij}(\lambda_1^{(k-1)}, \lambda_2^{(\ell)}) = (S_{ii} + S_{jj} + \lambda_2)\hat{\Omega}_{ij}^{net}(\lambda_1^{(k-1)}, \lambda_2^{(\ell)}) + \sum_{j' \neq j}^{p} \hat{\Omega}_{ij'}^{net}(\lambda_1^{(k-1)}, \lambda_2^{(\ell)}) S_{jj'} + \sum_{i' \neq i}^{p} \hat{\Omega}_{i'j}^{net}(\lambda_1^{(k-1)}, \lambda_2^{(\ell)}) S_{ii'},$$

for $i, j = 1, \ldots, p$, $i \neq j$. Now assume the C_{ij} here are nonexpansive, i.e., $\left| C_{ij}(\lambda_1^{(k)}, \lambda_2^{(\ell)}) - C_{ij}(\lambda_1^{(k-1)}, \lambda_2^{(\ell)}) \right| \leq \left| \lambda_1^{(k)} - \lambda_1^{(k-1)} \right|$. Then we have that

$$\left| C_{ij}(\lambda_1^{(k-1)}, \lambda_2^{(\ell)}) \right| < 2\lambda_1^{(k)} - \lambda_1^{(k-1)} \quad (6)$$

implies that $\hat{\Omega}_{ij}^{net}(\lambda_1^{(k)}, \lambda_2^{(\ell)}) = 0$; i.e., the entries satisfying this condition can be omitted from the PseudoNet optimization problem (1) *for $\lambda_1^{(k)}, \lambda_2^{(\ell)}$.*

3 NUMERICAL EXAMPLES

3.1 Synthetic data

We begin by discussing our synthetic examples; in these, we directly compare to CONCORD, which is the method most related to ours. We generated synthetic data as follows. First, we generated a random, sparse, diagonally dominant $p \times p$ (ground truth) matrix Ω^0, by following the procedure in Oh et al. (2014); Khare et al. (2015); Peng et al. (2009); Ali et al. (2016); we investigated $p \in \{1000, 3000\}$. Then, we drew n samples from a multivariate normal distribution with mean zero and covariance matrix $(\Omega^0)^{-1}$, which were input into PseudoNet and CONCORD; we investigated $n \in \{0.2p, 0.4p, 0.8p\}$ and $\lambda_1, \lambda_2 \in \{2^{-10}, 2^{-9.5}, \ldots, 1, 2^{0.5}\}$, *i.e.*, a 22×22 grid. Finally, we computed a method's false and true positive rates, by counting the number of nonzero entries in the method's estimate $\hat{\Omega}$ that were zero and nonzero, respectively, in Ω^0; we also computed the estimation error, *i.e.*, $\|\Omega^0 - \hat{\Omega}\|$, in several matrix norms. To summarize the variable selection accuracy and estimation errors across λ_1, λ_2, we computed the area under the curve (AUC) (Oh et al., 2014; Khare et al., 2015; Ali et al., 2016); to summarize the estimation errors, we computed the median across λ_1, λ_2. We repeated this entire process 50 times; thus, Table 1 reports the medians and interquartile ranges (IQRs) across these 50 trials, for $p = 3000$ ($p = 1000$ is in the supplement).

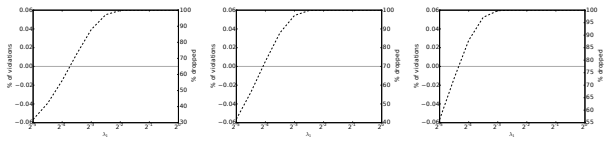

Figure 1: Percentages of dropped variables excluding diagonal entries (dashed line, right vertical axes) and violations (solid line, left vertical axes) for PseudoNet's screening rules ($\lambda_2 = 1$, $p = 3000$); first column is $n = 0.2p$, second is $n = 0.4p$, third is $n = 0.8p$. The rules never commit a violation.

PseudoNet outperforms CONCORD in AUC and estimation error across all sample sizes and norms (as well as on each trial individually). PseudoNet's estimation error, in particular, is significantly lower than CONCORD's. We also see that PseudoNet's wallclock times as well as most of its interquartile ranges (IQRs) are generally lower than CONCORD's, and that the estimates produced by PseudoNet are quite stable.

We also investigate the efficacy of PseudoNet's screening rules; using the same synthetic data, we measure the (median across 50 trials) percentages of variables the rules suggest dropping (excluding diagonal entries), as well as the percentages of violations. Figure 1 presents the results: the rules drop more variables as λ_1 increases (expected), but never commit violations.

3.2 Minimum variance portfolio optimization

Next, we evaluate PseudoNet, as well as several other methods, in the context of a finance application. We consider the problem of *minimum variance portfolio optimization*, *i.e.*, we must allocate our wealth across p assets so that our overall risk is minimized; we model risk here as $x^T \hat{\Sigma} x$, where $x \in \mathbf{R}^p$ is an allocation vector ($x_i > 0$ corresponds to a long position, while $x_i < 0$ corresponds to a short position), and $\hat{\Sigma}$ is an estimate of the underlying covariance matrix. This leads to the following (convex) optimization problem: $\underset{x \in \mathbf{R}^p}{\text{minimize}}\ x^T \hat{\Sigma} x$ subject to $\mathbf{1}^T x = 1$, which admits the analytical solution $x = (\mathbf{1}^T \hat{\Sigma}^{-1} \mathbf{1})^{-1} \hat{\Sigma}^{-1} \mathbf{1}$. We solve a minimum variance portfolio optimization problem (instead of, say, a *mean/variance problem* (Markowitz, 1952)) to isolate the impact of the estimate $\hat{\Omega} = \hat{\Sigma}^{-1}$.

We obtained the closing prices of the 30 constituent stocks of the Dow Jones Industrial Average (DJIA) from February 18, 1995 through October 26, 2012 (roughly 17 years) from `http://finance.yahoo.com`. We divided the data into $T = 261$ consecutive time periods (of roughly 20 days each). The H days preceding each trading period, commonly referred to as the *estimation horizon*, were used to compute the estimate $\hat{\Omega}$; 10-fold cross-validation was used to choose λ_1, λ_2. The trading period was then used to evaluate the methods.

We investigated $H \in \{35, 40, 45, 50, 75, 150, 225, 300\}$.

We primarily evaluated each method using *realized risk*, *i.e.*, $r = \left((1/T)\sum_{t=1}^{T}\left(x_t^T p_t - \bar{p}\right)^2\right)^{1/2}$, where $x_t, p_t \in \mathbf{R}^p$ are the portfolio allocation and price change vectors for period t, respectively, and $\bar{p}$ is the *realized return*, *i.e.*, $\bar{p} = (1/T)\sum_{t=1}^{T} x_t^T p_t$, as well as the (commonly used) *Sharpe ratio*, *i.e.*, $(\bar{p} - p_{\text{free}})/r$, where p_{free} is the risk-free rate (we set $p_{\text{free}} = 5\%$); intuitively, realized risk measures the instability (*i.e.*, riskiness) of a trading strategy, and the Sharpe ratio trades off the (risk-free rate adjusted) returns and risk.

We compared PseudoNet with CONCORD, the sample covariance matrix (denoted Sample), the GLasso, the condition number-regularized inverse covariance matrix estimator of Won et al. (2013) (CondReg), the Ledoit-Wolf estimator (Ledoit and Wolf, 2003) (Ledoit), as well as the DJIA itself (*i.e.*, an index fund). Table 2 presents the results. When the estimation horizon is small, *i.e.*, when $H \in \{35, 40, 45, 50, 75\}$, PseudoNet achieves the lowest risk, which is a useful feature when markets fluctuate; PseudoNet is always within 4% of the lowest risk when the estimation horizon is larger. Additionally, PseudoNet achieves significantly lower risk than CONCORD across all estimation horizons. These reductions in risk also translate into better Sharpe ratios for PseudoNet: PseudoNet achieves the highest Sharpe ratio four (out of eight) times, which is more than any other method. When PseudoNet does not achieve the highest Sharpe ratio, it is usually within 5% of the best Sharpe ratio. We also plot the cumulative wealth (in \$) achieved by an estimator (for $H = 300$) in Figure 2. PseudoNet achieves the highest cumulative wealth despite not (directly) optimizing for returns (\$8.75 for PseudoNet versus \$8.72 for CONCORD) while incurring less risk: PseudoNet also preserves the most wealth during the 2008–2009 financial crisis (\$4.64 for PseudoNet versus \$4.43 for CONCORD and \$4.23 for CondReg). Further details are in the supplement.

Due to space constraints, we present our application of PseudoNet to wind power data in the supplement.

4 THEORY

4.1 Linear convergence

We begin by showing that the proximal gradient method used to compute the PseudoNet estimate, converges to the unique, global solution of the PseudoNet optimization problem (1) at a geometric ("linear") rate; this constrasts with a number of other pseudolikelihood-based methods, which do not provide unique estimates (Rocha et al., 2008; Peng et al., 2009;

		$n = 600$		$n = 1200$		$n = 2400$	
		PseudoNet	CONCORD	PseudoNet	CONCORD	PseudoNet	CONCORD
AUC	Median	**0.64**	0.63	**0.75**	0.71	**0.86**	0.84
	IQR	0.01	0.01	0.00	0.01	0.01	0.01
Squared Frobenius norm	Median	**15495.27**	49063.26	**12913.39**	42021.80	**8639.99**	30054.52
	IQR	83.60	75.39	4.46	78.99	21.98	34.91
ℓ_2 operator norm	Median	**2.17**	4.48	**2.01**	4.19	**1.99**	4.43
	IQR	0.00	0.00	0.00	0.01	0.00	0.00
Elementwise ℓ_1 norm	Median	**72178.79**	148152.12	**87484.12**	187895.91	**84109.25**	195442.01
	IQR	114.88	89.36	28.19	150.31	66.62	112.74
Elementwise ℓ_∞ norm	Median	**1.10**	2.38	**0.83**	1.77	**0.49**	0.95
	IQR	0.00	0.01	0.00	0.01	0.00	0.01
Wallclock time (secs.)	Median	**1861.35**	3657.65	**580.11**	1208.06	**124.72**	236.40
	IQR	7.86	36.14	1.48	7.43	0.06	2.14

Table 1: Median and interquartile range for PseudoNet and CONCORD's areas under the curves (AUCs), estimation errors in several matrix norms, and wallclock times ($p = 3000$). Higher median AUC is better, lower median estimation error and wallclock time is better; best in **bold**. PseudoNet outperforms CONCORD across all sample sizes and metrics.

		$H = 35$	40	45	50	75	150	225	300
PseudoNet	Risk	**15.23**	**15.04**	**15.21**	**15.01**	**15.06**	15.07	15.12	15.25
	Sharpe	**0.52**	**0.50**	0.43	0.47	**0.48**	0.47	0.50	**0.55**
CONCORD	Risk	17.03	17.02	17.04	17.02	17.04	17.09	17.10	17.16
	Sharpe	0.48	0.48	**0.47**	**0.49**	0.47	**0.48**	0.50	0.50
Sample	Risk	33.86	26.52	23.19	20.95	17.45	15.41	14.98	14.95
	Sharpe	0.36	0.44	0.26	0.23	0.38	0.29	0.37	0.36
GLasso	Risk	16.55	16.54	16.56	16.36	15.61	14.99	14.87	14.95
	Sharpe	0.49	0.49	0.47	0.47	0.42	0.36	0.36	0.36
CondReg	Risk	17.83	17.76	17.64	17.61	17.20	16.37	16.07	16.10
	Sharpe	0.48	0.48	0.45	0.46	0.46	0.47	**0.52**	0.49
Ledoit	Risk	15.58	15.46	15.43	15.36	15.10	**14.66**	**14.52**	**14.52**
	Sharpe	0.47	0.44	0.39	0.41	0.37	0.38	0.42	0.41
DJIA	Risk	18.96	18.96	18.96	18.96	18.96	18.96	18.96	18.96
	Sharpe	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19

Table 2: Realized risks and Sharpe ratios for various estimators and estimation horizons H in the portfolio optimization example. Lower realized risk is better (PseudoNet is best 5/8 times), and higher Sharpe ratio is better (PseudoNet is best 4/8 times); best in **bold**.

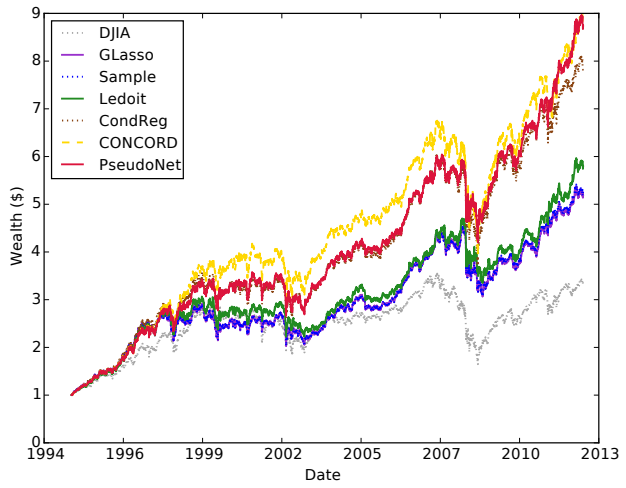

Figure 2: Cumulative wealth for various estimators in the portfolio optimization example ($H = 300$); higher is better. PseudoNet achieves the highest cumulative wealth.

Friedman et al., 2010; Khare et al., 2015; Oh et al., 2014), making interpretation difficult, are not guaranteed to converge (Rocha et al., 2008; Peng et al., 2009; Friedman et al., 2010), or converge at a slower rate (Khare et al., 2015; Oh et al., 2014).

Lemma 4.1 (Linear convergence)**.** *Suppose* $(\Omega^{(i)})_{i=0}^k$ *is a sequence of PseudoNet iterates with nonincreasing objective value. Let* $\hat{\Omega}^{net}$ *be the solution of the PseudoNet optimization problem* (1)*. Then* $\|\Omega^{(i)} - \hat{\Omega}^{net}\|_F \leq (1-c)^i \|\Omega^{(0)} - \hat{\Omega}^{net}\|_F$*,* $i = 1, \ldots, k$*, where* $c = \lambda_2/L$*, and* L *is the Lipschitz constant for the gradient of the smooth term* ∇g *in* (2)*.*

4.2 Consistency

Next, we show, under suitable regularity conditions, that PseudoNet is consistent at a rate of $\sqrt{(\log p)/n}$. Previous consistency results on pseudolikelihood-based estimators assume the existence of accurate estimates of the diagonal entries of the underlying inverse covariance matrix Ω^0; however, no method for obtaining such estimates is provided in these papers when $p > n$ (Khare et al., 2015; Peng et al., 2009). Below, we provide a two-step method that obtains accurate diagonal estimates, which are required for PseudoNet's (as well as CONCORD's and SPACE's) consistency proofs; this is done in Theorem 4.3. We give the regularity conditions required to establish the consistency of PseudoNet in the supplement; the assumptions are similar to those required in Khare et al. (2015), which are in turn similar to those in Peng et al. (2009), except that here we must additionally control how the new tuning parameter λ_2 grows with n. The following theorem presents our consistency result for PseudoNet.

Theorem 4.2 (Consistency)**.** *Assume the conditions stated in the supplement. Let* $p = O(n^\kappa)$ *for a constant* $\kappa > 0$*, and let* $\hat{\Omega}^{net}$ *be the PseudoNet estimate given by the solution of the PseudoNet optimization problem* (1)*. Then, we have, with probability at least* $1 - O(n^{-\beta})$ *for a constant* $\beta > 0$*: (a) signed support recovery:* $\mathbf{sign}\,\hat{\omega}_{ij}^{net} = \mathbf{sign}\,\omega_{ij}^0$*,* $i, j = 1, \ldots, p$*, where* $\hat{\omega}^{net}$ *and and* ω^0 *are the vectorizations of* $\hat{\Omega}^{net}$ *and* Ω^0*, respectively (i.e., the concatenations of the columns of these matrices), and we take* $\mathbf{sign}\,0 = 0$*; (b) estimation error:* $\|\hat{\omega}^{net} - \omega^0\|_2 \leq c_1 \lambda_1 \sqrt{q_n}$*, for a constant* $c_1 > 0$*, where* q_n *is a quantity such that* $\lambda_1\sqrt{q_n} \to 0$ *as* $n \to \infty$ *(see the supplement for details).*

4.2.1 Accurate diagonal estimates

The following theorem gives a way to obtain consistent estimates of the diagonal entries of the underlying inverse covariance matrix; the result is also useful in the consistency proofs for CONCORD (Khare et al., 2015, Theorem 2) and SPACE (Peng et al., 2009, Theorem 3), where consistent estimates are assumed, but a method to obtain them is not given, resolving an important gap in the literature. Our two-step method first performs a lasso regression (with tuning parameter λ_1) of each diagonal element on the remaining variables to identify subsets of relevant variables, and second estimates each diagonal element with the variance of the residuals given by the linear regression of each diagonal element on its subset of relevant variables (see the supplement for a discussion).

Theorem 4.3 (Accurate diagonal estimates via two-step method)**.** *Assume the conditions stated in the supplement. Now, for $j = 1, \ldots, p$, let $\hat{\mathcal{A}}_n^j$ be the set of indices corresponding to the nonzero coefficients obtained by fitting a lasso regression of the jth diagonal element on the remaining variables (with tuning parameter λ_1). Also, let $\hat{\omega}_{diag,j}$ be the sample variance of the jth diagonal element conditioned on the variables in $\hat{\mathcal{A}}_n^j$. Then, for every $\beta > 0$, there exists a constant $c_2 > 0$ such that $\|\hat{\omega}_{diag} - \omega^0_{diag}\|_\infty \leq c_2 d_n \sqrt{(\log n)/n}$, with probability at least $1 - O(n^{-\beta})$, where ω^0_{diag} means the diagonal entries of Ω^0, and d_n denotes the maximum number of nonzero entries in any row of Ω^0.*

4.3 Saturation

Lastly, we show that the PseudoNet estimate does not saturate (*i.e.*, when $p \gg n$, the number of variables selected by PseudoNet can be greater than np out of $p(p-1)/2$ total variables), while the SPLICE, SPACE, and CONCORD estimates can saturate; this is rather limiting for these latter estimators from the points of view of estimation error as well as interpretability.

To do this, we first introduce some notation that makes the statements of these results, as well as their proofs, more concise. We use **vech** to mean the *half-vectorization* operator, *i.e.*, the concatenation of the lower triangle of its (matrix) argument, excluding diagonal entries. We use **card** to count the number of nonzero entries in its argument. Also, we say that the columns of a wide matrix $A \in \mathbf{R}^{k \times \ell}$ (*i.e.*, $\ell > k$) are in *general position* if the affine span of any $m \leq k$ signed columns of A, *i.e.*, $s_{i_1} A_{i_1}, \ldots, s_{i_m} A_{i_m}$, where each s_j, $j = i_1, \ldots, i_m$ is fixed to either $+1$ or -1, does not contain any of the points $\pm A_j$, $j \neq i_1, \ldots i_m$.

Theorem 4.4 (Saturation results for PseudoNet and CONCORD)**.** *Let $A \in \mathbf{R}^{np \times p(p-1)/2}$ be a matrix containing the columns of the data matrix X arranged in a particular fashion (details in the supplement). Also, let $\hat{\Omega}^{net}$ be the PseudoNet estimate, i.e., the solution of the PseudoNet optimization problem (1), and let $\hat{\Omega}^{con}$ be a CONCORD estimate; so, we have $\mathbf{vech}\, \hat{\Omega}^{net}, \mathbf{vech}\, \hat{\Omega}^{con} \in \mathbf{R}^{p(p-1)/2}$. Assume that $p \gg n$. Then, the PseudoNet estimate does not saturate, i.e., $\mathbf{card}\, \mathbf{vech}\, \hat{\Omega}^{net} \leq p(p-1)/2$, and there exists a CONCORD estimate that saturates, i.e., $\mathbf{card}\, \mathbf{vech}\, \hat{\Omega}^{con} \leq np$. Furthermore, if the columns of the matrix A are in general position, then all CONCORD estimates saturate.*

The analogous results for SPLICE and SPACE, computed using iterative algorithms (see the supplement), follow by using arguments similar to those given in the proof of Theorem 4.4.

Corollary 4.5 (Saturation results for SPLICE and SPACE)**.** *Let $\hat{\Omega}^{spl,(i)}$ and $\hat{\Omega}^{spc,(i)}$ denote SPLICE and SPACE estimates at the end of iteration i, respectively; so, we have $\mathbf{vech}\, \hat{\Omega}^{spl,(i)}, \mathbf{vech}\, \hat{\Omega}^{spc,(i)} \in \mathbf{R}^{p(p-1)/2}$. Assume that $p \gg n$. Then, there exist SPLICE and SPACE estimates at the end of iteration i that saturate, i.e., $\mathbf{card}\, \mathbf{vech}\, \hat{\Omega}^{spl,(i)} \leq np$ and $\mathbf{card}\, \mathbf{vech}\, \hat{\Omega}^{spc,(i)} \leq np$.*

5 DISCUSSION

We introduced PseudoNet, a new, more flexible pseudolikelihood-based estimator of the inverse covariance matrix; PseudoNet can be viewed as generalizing several Gaussian likelihood-based, as well as pseudolikelihood-based, estimators in ways that give PseudoNet a number of statistical and computational advantages. As a whole, we believe these statistical and computational properties represent a useful step forward in the design of pseudolikelihood-based estimators of the inverse covariance matrix.

Acknowledgements. AA was supported by the DoE Computational Science Graduate Fellowship DE-FG02-97ER25308. SYO was supported in part by Laboratory Directed Research and Development (LDRD) funding from Berkeley Lab, provided by the Director, Office of Science of the DoE, under Contract No. DE-AC02-05CH11231, and is also affiliated with the Computational Research Division at Lawrence Berkeley National Laboratory. The work of BR was partially supported by US Air Force Office of Scientific Research grant award number FA9550-13-1-0043, US National Science Foundation under Grant Nos. DMS-CMG 1025465, AGS-1003823, DMS-1106642, DMS-CAREER-1352656, Defense Advanced Research Projects Agency DARPA-YFAN66001-111-4131, and SMC-DBNKY.

References

Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. The multiple quantile graphical model. In *Advances in Neural Information Processing Systems*, 2016. Available at `http://arxiv.org/pdf/1607.00515.pdf`.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B*, 36(2):192–236, 1974.

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

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Applications of the lasso and grouped lasso to the estimation of sparse graphical models. Available at `http://statweb.stanford.edu/~tibs/ftp/ggraph.pdf`, 2010.

Kshitij Khare, Sang-Yun Oh, and Bala Rajaratnam. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B*, 77(4):803–825, 2015.

Steffen Lauritzen. *Graphical Models*. Oxford University Press, 1996.

Olivier Ledoit and Michael Wolf. Honey, I shrunk the sample covariance matrix. *UPF Economics and Business Working Paper*, (691), 2003.

Harry Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.

Rahul Mazumder and Trevor Hastie. Exact covariance thresholding into connected components for large-scale graphical lasso. *Journal of Machine Learning Research*, 13:781–794, 2012.

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

Sang-Yun Oh, Onkar Dalal, Kshitij Khare, and Bala Rajaratnam. Optimization methods for sparse pseudolikelihood graphical model selection. In *Advances in Neural Information Processing Systems*, pages 667–675. 2014.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013.

Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

Guilherme Rocha, Peng Zhao, and Bin Yu. A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (SPLICE). Available at `https://www.stat.berkeley.edu/~binyu/ps/rocha.pseudo.pdf`, 2008.

Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5:941–973, 2004.

Adam Rothman, Peter Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B*, 74(2):245–266, 2012.

Ryan J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

Joong Won, Johan Lim, Seung Kim, and Bala Rajaratnam. Condition number-regularized covariance estimation. *Journal of the Royal Statistical Society: Series B*, 75(3):427–450, 2013.

Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1): 19–35, 2007.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67:301–320, 2005.