# On Sparse Nonparametric Conditional Covariance Selection

**Mladen Kolar** MLADENK@CS.CMU.EDU
**Ankur P. Parikh** APPARIKH@CS.CMU.EDU
**Eric P. Xing** EPXING@CS.CMU.EDU
Machine Learning Department, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213 USA

## Abstract

We develop a penalized kernel smoothing method for the problem of selecting nonzero elements of the conditional precision matrix, known as *conditional covariance selection*. This problem has a key role in many modern applications such as finance and computational biology. However, it has not been properly addressed. Our estimator is derived under minimal assumptions on the underlying probability distribution and works well in the high-dimensional setting. The efficiency of the algorithm is demonstrated on both simulation studies and the analysis of the stock market.

## 1. Introduction

In recent years, with the advancement of large-scale data acquisition technology in various engineering, scientific, and socio-economical domains, the problem of estimating latent dependency structures underlying high-dimensional attributes has become a problem of remarkable algorithmic and theoretical interest in the machine learning and statistics community. Although a vast and rich body of work on the so-called *covariance selection* problem can be found in the recent literature (Meinshausen & Bühlmann, 2006; Friedman et al., 2008b; Ravikumar et al., 2008; Banerjee et al., 2008), current focus and progress seems to be restricted to simple scenarios where the data-likelihood function belongs to well-known parametric families, such as discrete Markov random fields or Gaussian Graphical Models, and the precision matrix is analyzed in isolation, without considering effects of other covariates that represent environmental factors. In this paper, we investigate the problem of covariance selection under situations where such assumptions are violated.

Consider the problem of gene network inference in systems biology, which is of increasing importance in drug development and disease treatment. A gene network is commonly represented as a fixed network, with edge weights denoting strength of associations between genes. Realistically, the strength of associations between genes can depend on many covariates such as blood pressure, sugar levels, and other body indicators; however, biologists have very little knowledge on how various factors affect strength of associations. Ignoring the influence of different factors leads to estimation procedures that overlook important subtleties of the regulatory networks. Consider another problem in quantitative finance, for which one wants to understand how different stocks are associated and how these associations vary with respect to external factors to help investors construct a diversified portfolio. The rule of *Diversification*, formalized by Modern Portfolio Theory (Markowitz, 1952), dictates that risk can be reduced by constructing a portfolio out of uncorrelated assets. However, it also assumes that the associations between assets are fixed (which is highly unrealistic) and a more robust approach to modeling assets would take into account how their associations change with respect to economic indicators, such as, gross domestic product (GDP), oil price or inflation rate. Unfortunately, there is very little domain knowledge on the exact relationship between economic indicators and associations between assets, which motivates the problem of *conditional covariance selection* we intend to investigate in this paper.

Let $\mathbf{X} \in \mathbb{R}^p$ denote a $p$-dimensional random vector representing genes or stock values, and $Z \in \mathbb{R}$ denote an index random variable representing some body factor or economic indicator of interest. Both of the above mentioned problems in biology and finance can be modeled as inferring non-zero partial correlations between different components of the random vector $\mathbf{X}$ conditioned on a particular value of the index variable

Appearing in *Proceedings of the 27$^{th}$ International Conference on Machine Learning*, Haifa, Israel, 2010. 

$Z = z$. We assume that the value of partial correlations change with $z$, however, the set of non-zero partial correlations is constant with respect to $z$. Let $\mathbf{\Sigma}(z) := \text{Cov}(\mathbf{X}|Z = z)$ denote the *conditional* covariance of $\mathbf{X}$ given $Z$, which we assume to be positive definite, and let $\mathbf{\Omega}(z) := \mathbf{\Sigma}(z)^{-1}$ denote the conditional precision matrix. The structure of non-zero components of the matrix $\mathbf{\Omega}(z)$ tells us a lot about associations between different components of the vector $\mathbf{X}$, since the elements of $\mathbf{\Omega}(z)$ correspond to partial correlation coefficients. One of the challenges we address in this paper is how to select non-zero components of $\mathbf{\Omega}(z)$ from noisy samples. Usually, very little is known about the relationship between the index variable $Z$ and associations between components of the random variable $\mathbf{X}$; so, in this paper, we develop a nonparametric method for estimating the non-zero elements of $\mathbf{\Omega}(z)$. Specifically, we develop a new method based on $\ell_1/\ell_2$ penalized kernel smoothing, that is able to estimate the functional relationship between the index $Z$ and components of $\mathbf{\Omega}(z)$ with minimal assumptions on the distribution $(\mathbf{X}, Z)$ and only smoothness assumption on $z \mapsto \mathbf{\Omega}(z)$. In addition to developing an estimation procedure that works with minimal assumptions, we also focus on statistical properties of the estimator in the high-dimensional setting, where the number of dimensions $p$ is comparable or even larger than the sample size. Ubiquity of high-dimensionality in many real world data forces us to carefully analyze statistical properties of the estimator, that would otherwise be apparent in a low-dimensional setting.

Our problem setting, as stated above, should be distinguished from the classical problem of covariance selection, introduced in the seminal paper by Dempster (Dempster, 1972). In the classical setting, the main goal is to select non-zero elements of the precision matrix; however, the precision matrix does not vary with respect to the index variables. As mentioned before, non-zero elements of the precision matrix correspond to partial correlation coefficients, which encode associations among sets of random variables. Due to its importance, the problem of covariance selection has drawn lots of attention from both the machine learning and statistical community, which has led to remarkable progress in both computational (Friedman et al., 2008b; Banerjee et al., 2008; Duchi et al., 2008) and statistical issues (Yuan & Lin, 2007; Ravikumar et al., 2008; Peng et al., 2009; Rothman et al., 2008; Meinshausen & Bühlmann, 2006). However, almost all of these work have been driven by the simplifying assumption of an invariant covariance structure.

Perhaps closest to the scenario we investigate in this paper, are the recent work on estimating high-dimensional time-varying graphical models (Zhou et al., 2008; Kolar et al., 2009). While their work could fit into our framework, there are a few mayor differences. In both of the papers, the distribution of $\mathbf{X}$ was explicitly given, either as a multivariate Gaussian distribution or a discrete distribution following an Ising model. Furthermore, time, which is considered to be an index variable in their work, is not random. Finally, the focus of (Zhou et al., 2008) was on point-wise estimation of covariance and precision matrices, where the loss was measured in Frobenius norm, while the correct selection of non-zero elements of the precision matrix was not investigated.

To the best of our knowledge, there are only few references for work on nonparametric models for conditional covariance and precision matrices. Yin et al., (2008) develop a kernel estimator of the conditional covariance matrix based on the local-likelihood approach. Since their approach does not perform estimation of non-zero elements in the precision matrix, it is suitable in low-dimensions. Other related work includes nonparametric estimation of the conditional variance function in longitudinal studies (see Ruppert et al., 1997; Fan & Yao, 1998, and references within). Our paper intends to fill this void in the literature.

In summary, here are the highlights of our paper. Our main contribution is a new nonparametric model for sparse conditional precision matrices, and the $\ell_1/\ell_2$ penalized kernel estimator for the proposed model. The estimation procedure was developed under minimal assumptions, with the focus on the high-dimensional setting, where the number of dimensions is potentially larger than the sample size. A modified Bayesian Information Criterion (BIC) is given that can be used to correctly identify the set of non-zero partial correlations. Finally, we demonstrate the performance of the algorithm on synthetic data and analyze the associations between the set of stocks in the S&P 500 as a function of oil price.

## 2. The Model

Let $\mathbf{X} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$ be a $p$-dimensional random vector (representing gene expressions or stock values) and let random variable $Z \in [0, 1]$ be an associated univariate index (representing a body factor or an economy index). We will estimate associations between different components of $\mathbf{X}$ conditionally on $Z$. For simplicity of presentation, we assume that the index variable can be scaled into the interval $[0, 1]$ and, furthermore, we assume that it is a scalar variable.

The kernel smoothing method, to be introduced, can be easily extended to multivariate $Z$. However, such an extension may only be practical in limited cases, due to the curse of dimensionality (Li & Liang, 2008). Throughout the paper, we assume that $\mathbb{E}[\mathbf{X}|Z=z] = 0$ for all $z \in [0,1]$. In practice, one can easily estimate the conditional mean of $\mathbf{X}$ given $Z$ using local polynomial fitting (Fan, 1993) and subtract it from $\mathbf{X}$. We denote the conditional covariance matrix of $\mathbf{X}$ given $Z$ as $\mathbf{\Sigma}(z) := \mathrm{Cov}(\mathbf{X}|Z=z) = (\sigma_{uv}(z))_{u,v\in[p]}$, where we use $[p]$ to denote the set $\{1,\ldots,p\}$. Assuming that $\mathbf{\Sigma}(z)$ is positive definite, for all $z \in [0,1]$, the conditional precision matrix is given as $\mathbf{\Omega}(z) := \mathbf{\Sigma}(z)^{-1} = (\omega_{uv}(z))_{u,v\in[p]}$. Elements $(\omega_{uv}(z))_{u,v\in[p]}$ are smooth, but unknown functions of $z$.

With the notation introduced above, the problem of conditional covariance selection, e.g., recovering the strength of association between stocks as a function of oil price, or association between gene expressions as a function of blood pressure, can be formulated as estimating the non-zero elements in the conditional precision matrix $\mathbf{\Omega}(z)$. As mentioned before, association between different components of $\mathbf{X}$ can be expressed using the partial correlation coefficients, which are directly related to the elements of precision matrix as follows; the partial correlation $\rho_{uv}(z)$ between $X_u$ and $X_v$ $(u,v \in [p])$ given $Z=z$ can be computed as

$$\rho_{uv}(z) = -\frac{\omega_{uv}(z)}{\sqrt{\omega_{uu}(z)\omega_{vv}(z)}}. \tag{1}$$

The above equation confirms that the non-zero partial correlation coefficients can be selected by estimating non-zero elements of the precision matrix. Let $S := \{(u,v) : \int_{[0,1]} \omega_{uv}^2(z)dz > 0,\ u \neq v\}$ denote the set of non-zero partial correlation coefficients, which we assume to be constant with respect to $z$, i.e., we assume that the associations are fixed, but their strength can vary with respect to the index $z$. Furthermore, we assume that the number of non-zero partial correlation coefficients, $s := |S|$, is small. This is a reasonable assumption for many problems, e.g., in biological systems a gene usually interacts with only a handful of other genes. In the following paragraphs, we relate the partial correlation coefficients to a regression problem, and present a computationally efficient method for estimating non-zero elements of the precision matrix based on this insight.

For each component $X_u$ $(u \in [p])$ we set up a regression model, where $X_u$ is the response variable, and all the other components are the covariates. Let $\mathbf{X}_{\setminus u} := \{X_v : v \neq u, v \in [p]\}$. It is a well known result (e.g., Lauritzen, 1996) that the partial correlation coefficients can be related to a regression model as follows

$$X_u = \sum_{v\neq u} X_v b_{uv}(z) + \epsilon_u(z), \quad u \in [p], \tag{2}$$

with $\epsilon_u(z)$ being uncorrelated with $\mathbf{X}_{\setminus u}$ if and only if

$$b_{uv}(z) = -\frac{\omega_{uv}(z)}{\omega_{uu}(z)} = \rho_{uv}(z)\sqrt{\frac{\omega_{vv}(z)}{\omega_{uu}(z)}}. \tag{3}$$

Observe that $\rho_{uv}(z) = \mathrm{sign}(b_{uv}(z))\sqrt{b_{uv}(z)b_{vu}(z)}$, which relates selection of the non-zero partial correlations to selection of covariates in the regression model (2). The relationship given in Eq. (3) has been used for estimation of high-dimensional Gaussian graphical models in (Meinshausen & Bühlmann, 2006), however, Eq. (3) holds for any distribution of $(\mathbf{X}, Z)$ and can be used for estimation of non-zero elements of the conditional precision matrix.

Based on the above discussion, we propose a locally weighted kernel estimator of the non-zero partial correlations. Let $\mathcal{D}^n = \{(\mathbf{x}^i, z^i)\}_{i\in[n]}$ be an independent sample of $n$ realizations of $(\mathbf{X}, Z)$. For each $u \in [p]$, we define the loss function

$$\begin{aligned}\mathcal{L}_u(\mathbf{B}_u; \mathcal{D}^n) := \\ \sum_{z\in\{z^j\}_{j\in[n]}} \sum_{i\in[n]} \big(x_u^i - \sum_{v\neq u} x_v^i b_{uv}(z)\big)^2 K_h(z-z^i) \\ + 2\lambda \sum_{v\neq u} ||b_{uv}(\cdot)||_2\end{aligned} \tag{4}$$

where $\mathbf{B}_u = (\mathbf{b}_u(z^1), \ldots, \mathbf{b}_u(z^n))$, $\mathbf{b}_u(z^j) \in \mathbb{R}^{p-1}$, $K_h(z-z^i) = K(\frac{|z-z^i|}{h})$ is a symmetric density function with bounded support that defines local weights, $h$ denotes the bandwidth, $\lambda$ is the penalty parameter and $||b_{uv}(\cdot)||_2 := \sqrt{\sum_{z\in\{z^j\}_{j\in[n]}} b_{uv}(z)^2}$. Define $\hat{\mathbf{B}}_u$ as a minimizer of the loss

$$\hat{\mathbf{B}}_u := \underset{\mathbf{B}\in\mathbb{R}^{p-1\times n}}{\mathrm{argmin}}\ \mathcal{L}_u(\mathbf{B}; \mathcal{D}^n). \tag{5}$$

Combing $\{\hat{\mathbf{B}}_u\}_{u\in[p]}$ gives an estimator

$$\hat{S} := \{(u,v) : \max\{||\hat{b}_{uv}(\cdot)||_2, ||\hat{b}_{vu}(\cdot)||_2\} > 0\} \tag{6}$$

of the non-zero elements of the precision matrix.

In Eq. (4), the $\ell_1/\ell_2$ norm is used to penalize model parameters. This norm is commonly used in the Group Lasso (Yuan & Lin, 2006). In our case, since we assume the set of non-zero elements $S$, of the precision matrix, to be fixed with respect to $z$, the $\ell_2$ norm is a natural way to shrink the whole group of coefficients $\{b_{uv}(z^i)\}_{i\in[n]}$ to zero. Note that the group consists

of the same element, say $(u, v)$, of the precision matrix for different values of $z$. Our approach should be contrasted to the approach in (Zhou et al., 2008; Kolar et al., 2009), where the set of non-zero elements of the precision matrix changes with respect to time, in which case one cannot use the $\ell_2$ norm, which would make the set of non-zero elements constant over time. Instead, the use of $\ell_1$ norm is necessary. Under our assumptions, usage of the $\ell_2$ norm results in a more efficient procedure.

The above described kernel smoothing procedure is well justified under the assumption that the elements of $\mathbf{\Omega}(z)$ are smooth but unknown functions of $z$. The loss function in Eq. (4), without the penalty term, is common in varying coefficient regression models, however, relevant coefficients are selected using a generalized likelihood ratio test, (Li & Liang, 2008). We point out that the functions $\{b_{uv}(z)\}$ are being estimated only on $\{z^i\}_{i\in[n]}$ and not on the whole support of $\text{supp}(Z) = [0, 1]$. This is justified under the assumptions that $\{z^i\}_{i\in[n]}$ are sufficiently dense on $[0, 1]$. For example, if $Z$ has the density, $f_Z(z)$, that is continuous and bounded away from zero on $[0, 1]$, then the maximal distance between any two neighboring observations is $O_p(\frac{\log n}{n})$ (Janson, 1987). Together with the assumption that $\{b_{uv}(z)\}$ are smooth functions, this implies that the approximation error, when estimating the whole curve $b_{uv}(z)$, $z \in [0, 1]$ with points $\{\hat{b}_{uv}(z),\ z \in \{z^i\}_{i\in[n]}\}$, is of smaller order than the optimal nonparametric rate of convergence, $\mathcal{O}_p(n^{-2/5})$. Finally, we note that the statistical efficiency of our procedure can be improved by exploiting the fact that the precision matrix is symmetric and jointly optimizing for $\{\mathbf{B}_u\}_{u\in[n]}$, (see Peng et al., 2009, for specific details in estimation of multivariate Gaussian graphical models).

## 3. Optimization algorithm

In this section, we detail an efficient optimization algorithm that can be used to solve the problem given in Eq. (5). Given that the optimization problem is convex, a variety of techniques can be used to solve it. A particularly efficient optimization algorithm has been devised for $\ell_1/\ell_2$ penalized problems, that is based on the group-coordinate descent and is referred to as the active-shooting algorithm (Peng et al., 2009; Friedman et al., 2010). A modification of the procedure, suitable for our objective, is outlined in Algorithm 1, which we now explain.

We point out that the group coordinate descent will converge to an optimum, since the loss function is smooth and the penalty term in Eq. (4)

**Algorithm 1** Procedure for solving Eq. (5)

**Input:** Data $\mathcal{D}^n = \{\mathbf{x}^i, z^i\}_{i\in[n]}$, initial solution $\tilde{\mathbf{B}}_u^{(0)}$
**Output:** Solution $\hat{\mathbf{B}}_u$ to Eq. (5)

1: $\mathcal{A} := \{v \in [p] \setminus u : ||\tilde{b}_{uv}^{(0)}(\cdot)||_2 > 0\}$, $t = 0$
2: **repeat**
3:  **repeat** {iterate over $v \in \mathcal{A}$}
4:  Compute $\{r_{uv}^i(z^j)\}_{i,j\in[n]}$ using Eq. (8)
5:  **if** condition (9) is satisfied **then**
6:  $\tilde{b}_{uv}(\cdot) \leftarrow 0$
7:  **else**
8:  $\tilde{b}_{uv}(\cdot) \leftarrow \text{argmin}\, \mathcal{L}_u^v\big(b_{uv}(\cdot); \mathcal{D}^n\big)$
9:  **end if**
10:  **until** convergence on $\mathcal{A}$;
11:  **forall** $v \in [p] \setminus u$ compute lines 4 through 9 once
12:  $\mathcal{A} := \{v \in [p] \setminus u : ||\tilde{b}_{uv}(\cdot)||_2 > 0\}$
13: **until** $\mathcal{A}$ did not change
14: $\hat{\mathbf{B}}_u \leftarrow \{\tilde{b}_{uv}(\cdot)\}_{v\in[p]\setminus u}$

decomposes across different rows of the matrix $\mathbf{B}_u$ (Friedman et al., 2010). Now, we derive an update for row $v$, while keeping all other rows of $\mathbf{B}_u$ fixed. Let $\{\tilde{b}_{uv}(z^j)\}_{j\in[n]}$ be a minimizer of

$$\mathcal{L}_u^v(\{b_{uv}(z^j)\}_{j\in[n]}; \mathcal{D}^n) := \sum_{z\in\{z^j\}_{j\in[n]}} \sum_{i\in[n]} \big(r_{uv}^i(z) - x_v^i b_{uv}(z)\big)^2 K_h(z - z^i) + 2\lambda\ ||b_{uv}(\cdot)||_2, \tag{7}$$

where

$$r_{uv}^i(z) = x_u^i - \sum_{v'\neq u,v} x_{v'}^i \tilde{b}_{uv'}(z) \tag{8}$$

and $\{\tilde{b}_{uv'}(z)\}$ denotes the current solution for all the other variables. Solving Eq. (7) iteratively, by cycling through rows $v \in [p] \setminus u$, will lead to an optimal solution $\hat{\mathbf{B}}_u$ of Eq. (5). By analyzing Karush-Kuhn-Tucker conditions of the optimization problem in Eq. (7), we can conclude that the necessary and sufficient condition for $\{\tilde{b}_{uv}(z^j)\}_{j\in[n]} \equiv 0$ is

$$\frac{1}{\lambda^2} \sum_{z\in\{z^j\}_{j\in[n]}} \left( \sum_{i\in[n]} x_v^i r_{uv}^i(z) K_h(z - z^i) \right)^2 \leq 1. \tag{9}$$

Eq. (9) gives a fast way to explicitly check if the row $v$ of a solution is identical to zero or not. If the condition in Eq. (9) is not satisfied, only then we need to find a minimizer of Eq. (7), which can be done by the gradient descent, since the objective is differentiable when $\{b_{uv}(z^j)\}_{j\in[n]} \not\equiv 0$.

In practice, one needs to find a solution to (5) for a large number of penalty parameters $\lambda$. Comput-

ing solutions across a large set of possible $\lambda$ values can effectively be implemented using the warm start technique (Friedman et al., 2008a). In this technique, Eq. (5) is solved for a decreasing sequence of penalty parameters $\lambda_1 > \ldots > \lambda_N$ and the initial value $\tilde{\mathbf{B}}_u^{(0)}$ provided to Algorithm 1 for $\lambda_i$ is the final solution $\hat{\mathbf{B}}_u$ for $\lambda_{i-1}$. This experimentally results in faster convergence and a more stable algorithm.

## 4. Theoretical properties

In this section, we give some theoretical properties of the estimation procedure given in Section 2. These results are given for completeness and are presented without proofs, which will be reported elsewhere. In particular, we provide conditions under which there exists a set $\hat{S} = \hat{S}(\lambda)$ of selected non-zero partial correlations, which consistently estimates $S$, the true set of non-zero partial correlations. Observe that $\hat{S}$ depends on the penalty parameter $\lambda$, so it is of practical importance to correctly select the parameter $\lambda$ for which $\hat{S}$ consistently recovers $S$. We give conditions under which the modified BIC criterion is able to identify the correct penalty parameter $\lambda$. We start by giving general regularity conditions.

The following regularity conditions are standard in the literature (Fan & Huang, 2005; Wang & Xia, 2008): **(A1)** There is an $s > 2$ such that $\mathbb{E}[||\mathbf{X}||_2^{2s}] \leq \infty$; **(A2)** The density function $f(z)$ of the random variable $Z$ is bounded away from 0 on $[0,1]$ and has bounded second order derivative; **(A3)** The matrix $\mathbf{\Omega}(z)$ is positive definite for all $z \in [0,1]$ and its elements $(\omega_{uv}(z))$ are functions that have bounded second derivatives; **(A4)** The function $\mathbb{E}[||X||_2^4 \mid Z = z]$ is bounded; **(A5)** The kernel $K(\cdot)$ is a symmetric density with compact support. In addition the standard regularity conditions, we need the following identifiability condition, which allows us to correctly identify the true model **(A6)** $\sup_{z\in[0,1]} \max_{u\neq v} |\omega_{uv}(z^i)| \leq \mathcal{O}(\frac{1}{d})$, where $d := \max_{u\in[p]} |\{v : (u,v) \in S\}|$

**Theorem 1** *Assume that the regularity conditions **(A1)-(A6)** are satisfied. Furthermore, assume that* $\mathbb{E}[\exp(tX)|Z = z] \leq \exp(\sigma^2 t^2/2)$ *for all* $z \in [0,1]$, $t \in \mathbb{R}$ *and some* $\sigma \in (0,\infty)$. *Let* $h = \mathcal{O}(n^{-1/5})$, $\lambda = \mathcal{O}(n^{7/10}\sqrt{\log p})$ *and* $n^{-9/5}\lambda \to 0$. *If* $\frac{n^{11/10}}{\sqrt{\log p}} \min_{u,v\in S} ||b_{uv}(\cdot)||_2 \to \infty$, *then* $\mathbb{P}[\hat{S} = S] \to 1$.

Assuming that $\mathbf{X}$ is a subgaussian random variable in Theorem 1 is due to technical reasons. The assumption is needed to establish exponential inequalities for the probability that each solution $\hat{\mathbf{B}}_u$ of Eq. (5) correctly identifies the set of non-zero rows of $\mathbf{B}_u$. Then consistency of $\hat{S}$ can be established by applying the union bound over the events that estimators $\{\hat{\mathbf{B}}_u\}_{u\in[p]}$ consistently estimate non-zero rows of $\{\mathbf{B}_u\}_{u\in[p]}$. For the last claim to be true when the dimension $p$ is large, e.g., $p = \mathcal{O}(\exp(n^\alpha)), \alpha > 0$, we need a good tail behavior of the distribution of $\mathbf{X}$. The statement of the theorem still holds true, even if we do not establish exponential inequalities, but only for smaller dimensions. Another commonly used regularity condition on $\mathbf{X}$ is to assume that it is bounded with probability 1, which would again allow us to establish exponential inequalities needed in the proof. Finally, we need to assume that for $(u,v) \in S$, $||b_{uv}(\cdot)||_2$ does not decay to zero too quickly. Otherwise, the element of the precision matrix would be to hard to distinguish from 0.

Next, we show that the correct penalty parameter $\lambda$ can be chosen using the modified BIC criterion of (Chen & Chen, 2008). Denote $\hat{\mathbf{B}}_{u,\lambda}$ as the solution of Eq. (5) obtained for the penalty parameter $\lambda$. We define the residual sum of squares as

$$\mathrm{RSS}_u(\lambda) := n^{-2} \sum_z \sum_{i\in[n]} \big(x_u^i - \sum_{v\neq u} x_v^i \hat{b}_{uv,\lambda}(z)\big)^2 K_h(z - z^i)$$

and the BIC-type criterion

$$\mathrm{BIC}_u(\lambda) = \log(\mathrm{RSS}_u(\lambda)) + \frac{\widehat{df}_{u,\lambda}(\log(nh) + 2\log p)}{nh},$$

where $\widehat{df}_{u,\lambda}$ denotes the number of non-zero rows of $\hat{\mathbf{B}}_{u,\lambda}$. We used the modified version of the BIC criterion, since the ordinary BIC criterion tends to include many spurious variables when the complexity of the model space is large (Chen & Chen, 2008). Now, $\lambda$ is chosen by a minimization:

$$\hat{\lambda} = \operatorname*{argmin}_{\lambda} \sum_{u\in[p]} \mathrm{BIC}_u(\lambda), \tag{10}$$

and the final estimator of the non-zero components of the precision matrix $\hat{S} = \hat{S}(\hat{\lambda})$ is obtained by combining $\{\hat{\mathbf{B}}_{u,\hat{\lambda}}\}_{u\in[p]}$. We have the following theorem.

**Theorem 2** *Assume that the conditions of Theorem 1 are satisfied. Then the tuning parameter* $\hat{\lambda}$ *obtained by minimizing criterion* (10) *asymptotically identifies the correct model, i.e.,* $\mathbb{P}[\hat{S}(\hat{\lambda}) = S] \to 1$.

## 5. Simulation results

### 5.1. Toy example

We first consider a small toy example in order to demonstrate our algorithm's performance. We draw $n$ samples, from the joint distribution of $(\mathbf{X}, Z)$ where the conditional distribution of $\mathbf{X}$ given $Z = z$ is a 5-dimensional multivariate Gaussian with mean 0 and

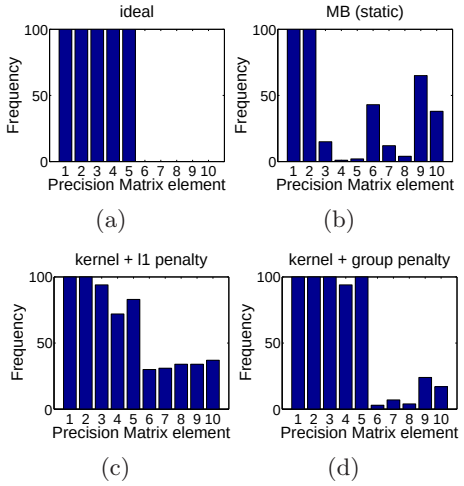



*Figure 1.* Toy example results. Each bar represents the number of times the corresponding precision matrix element was included in $\hat{S}$. Performance of the ideal algorithm is shown in Figure 1(a). Our algorithm (Figure 1(d)) gets close to this, and far outperforms both the other methods.

precision matrix $\mathbf{\Omega}(z)$, and $Z$ is uniformly distributed on $[0,1]$. The set $S = \{(1,2),(3,4),(2,4),(1,5),(3,5)\}$ denotes the non-zero elements of $\mathbf{\Omega}(z)$. We set elements $\omega_{uv}(z) = \omega_{uv}(z) = f_{uv}(z)$ for all $(u,v) \in S$, where the functions $\{f_{uv}(z)\}$ are defined as follows: **(1)** $f_{1,2} \equiv 1$ (constant), **(2)** $f_{3,4} \equiv 1$ (constant), **(3)** $f_{2,4}(z) = 1$ if $z \leq .5$ and $-1$ for $z > .5$ (piecewise constant), **(4)** $f_{1,5}(z) = 2z - 1$ (linear), **(5)** $f_{3,5}(z) = sin(2\pi z)$ (sinusoid). The diagonal elements $\omega_{uu}(z)$ $(z \in [0,1])$ are set to a constant number such that $\mathbf{\Omega}(z)$ is diagonally dominant, and hence positive definite.

We compared our method against the approach of (Meinshausen & Bühlmann, 2006) (referred to as `MB`), which assumes an invariant covariance matrix and ignores $z$, and against a simpler variant of our algorithm (called "kernel, $\ell_1$ penalty"), which replaces the group $\ell_1/\ell_2$ penalty in Eq. (4) with the $\ell_1$ penalty. Recall that the $\ell_1$ penalty does not encourage the set of non-zero elements in the precision matrix to remain fixed for all $z \in [0,1]$. Our algorithm, developed in Section 2 is referred to as "kernel, group penalty".

We average our results over 100 random trials. For each trial, $n = 300$ samples are randomly generated using the procedure described above. We counted the number of times each of the $\binom{5}{2} = 10$ possible off-diagonal elements of the precision matrix were selected as non-zeros. Figure 1 displays results as histograms. Bars 1-5 correspond to the true non-zero elements in $S$, as enumerated above, while bars 6-10 correspond to the elements that should be set to zero. Thus, in the ideal case, bars 1-5 should be estimated as non-zero for all 100 trials, while bars 6-10 should never be selected. As we can see, all algorithms select the constant elements $\omega_{12}(\cdot)$ (bar 1) and $\omega_{34}(\cdot)$ (bar 2). However, the `MB` approach fails to recover the three varying precision matrix elements and also recovers many false elements. Just using the kernel + $\ell_1$ penalty, described above, performs better, but still selects many elements not in $S$. Our algorithm, on the other hand, selects all the elements in $S$ almost all of the time, and also excludes the elements not in $S$ the vast majority of the time. This higher precision is the result of our group penalty, and gives superior performance to just using an $\ell_1$ penalty (assuming that the set of non-zero partial correlation coefficients is fixed with respect to $z$).

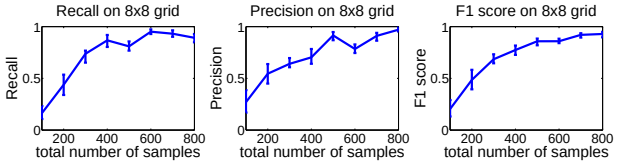



*Figure 2.* Simulation results for 8x8 grid. See section 5.2 for details.

### 5.2. Large simulations

We next tested our algorithm on a larger problem where $\mathbf{X} \in \mathbb{R}^{64}$. The components of $\mathbf{X}$ were arranged into an 8x8 grid, so that only adjacent components in the grid have has non-zero partial correlation. For all adjacent $(u,v)$, $\omega_{uv}(z) = sin(2\pi z + c_{uv})$, where $c_{uv} \sim \text{Unif}([0,1])$ is a random offset. We measure how well the algorithm recovers the true set of non-zero precision matrix elements. Both `MB` and "kernel + $\ell_1$" perform much worse than our estimator, so we do not display their performance. Performance of the "kernel + group penalty" estimator is shown in Figure 2. Even though the problem is significantly harder, after 800 samples our algorithm achieves an F1 score above 0.9.

## 6. Analyzing the stock market

We next apply our method to analyzing relationships among stocks in the S&P 500. Such an analysis would be useful to an economist studying the effect of various indicators on the market, or an investor who is seeking to minimize his risk by constructing a diverse portfolio according to Modern Portfolio Theory (Markowitz, 1952). Rather than assume static associations among stocks we believe it is more realistic to model them as a function of an economic indicator, such as oil price. We acquired closing stock prices from all stocks in the S&P 500[1] and oil prices[2] for all the days that the mar-

[1] http://www.finance.yahoo.com
[2] http://tonto.eia.doe.gov/

ket was open from Jan 1, 2003 through Dec 31, 2005. This gave us 750 samples of 469 stocks (we only considered stocks that remained in the S&P 500 during the entire time period). Instead of considering the raw prices, which often are a reflection of other factors, such as number of shares, we used the logarithm of the ratio of the price at time $t$ to the price at time $t-1$ and subtracted the mean value and divided by the standard deviation for each stock.

Our data consists of pairs $\{\mathbf{x}^i, z^i\}$, the vector of standardized stock prices and the oil price, respectively, obtained over a period of time. We analyze the data to recover the strength of associations between different stocks as a function of the oil price. Our belief is that each stock is associated with a small number of other stocks and that the set of associations is fixed over a time-period of interest, although the strengths may change. We believe this is justified since we are looking for long-term trends among stocks and want to ignore transient effects. Figure 3 illustrates the estimated network, where an edge between two nodes correspond to a non-zero element in the precision matrix. Note that the presented network is not a representation of an undirected probabilistic graphical model.

Clusters of related stocks are circled in Figure 3, and these largely confirm our intuition. Here are some of the stocks in a few of the clusters: **(1)** *Technology/semiconductors* - Hewlett Packard, Intel, Teradyne, Analog Devices etc.; **(2)** *Oil/drilling/energy* - Diamond Offshore Drilling, Baker Hughes, Halliburton, etc.; **(3)** *Manufacturing* - Alcoa, PPG Industries (coating products), International Paper Co. etc.; **(4)** *Financial* - American Express, Wells Fargo, Franklin Resources etc. It is also interesting that there exist coherent subgroups inside these clusters. For example, the "Retail stores" sector could be further divided into companies that specialize in clothes, like Gap and Limited, and those that are more general purpose department stores, like Wal-Mart and Target.

Another point of interest are two hubs (enlarged and highlighted in green in Figure 3), that connect a set of diverse stocks that do not easily categorize into an industrial sector. They correspond to JPMorgan Chase and Citigroup (two prominent financial institutions). It possible that these stocks are good indicators of the status of the market or have certain investment portfolios that contribute to their central positions in the network.

Finally, we explore the evolving nature of our edge weights as a function of oil price to demonstrate the advantages over simply assuming static partial correlations. Recall that the edge weights vary with oil price

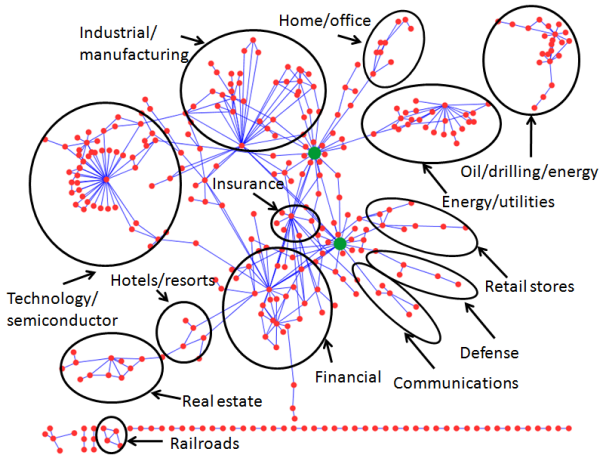
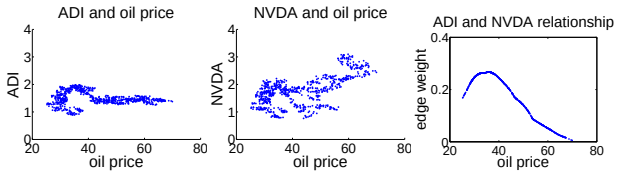



*Figure 3.* Overall stock market network that was recovered by the algorithm. Edges in the graph correspond to non-zero elements in the precision matrix. As one can see, the recovered network contains many clusters of related stocks. The green (and enlarged) hubs are described in the text.

and are proportional to the estimated partial correlation coefficients. Consider the two stocks Analog Devices (ADI), which makes signal processing solutions, and NVIDIA (NVDA), which makes graphics processing units. Ignoring the effect of the oil price, both of these companies are highly related since they belong to the semiconductor sector. However, if one analyzes the edge weights as a function of oil price, as shown in Figure 4 (a) and (b), both behave quite differently. This changing relationship is reflected by the varying strength of the edge weight between NVIDIA and Analog Devices (shown in Figure 4 (c) ). Note that when oil prices are low, the edge weight is high since Analog Devices and NVIDIA are both rising as a function of oil price. However, as oil prices increase, Analog Devices stabilizes while NVIDIA is more erratic (although it is mostly rising), so the edge weight sharply decreases. Thus, if an investor is aiming for diversification to reduce risk, he/she may be wary of investing in both of these stocks together when oil prices are low since they are highly associated, but might consider it if oil prices are high and the stocks are less associated.

## 7. Discussion

We develop a new nonparametric estimator for the problem of high-dimensional *conditional covariance selection*. Elements of the precision matrix are related to the partial correlation coefficients, whose nonzero structure tells a lot about the associations between different components of the vector $\mathbf{X}$. Our work is motivated by problems arising in biology and finance, where the associations between different variables change with respect to environmental factors.




(a) (b) (c)

*Figure 4.* This figure demonstrates how the changing edge weight between Analog Devices and NVIDIA ((c)) corroborates with the fact that Analog Devices and NVIDIA behave quite differently as a function of oil price ((a) and (b)). In (a) and (b), the y-axis is the ratio of the stock price to its price on January 1, 2003.

We believe that our method will help in extracting subtle information from noisy data, that otherwise couldn't be found using standard methods for covariance selection that analyze data in isolation, without considering various environmental factors.

## Acknowledgements

This paper is based on work supported by ONR N000140910758, NSF IIS-0713379, NSF Career DBI-0546594, NIH 1 R01 GM078622-01, and an Alfred P. Sloan Research Fellowship to EPX.

## References

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. Mach. Learn. Res.*, 9:485–516, 2008. ISSN 1533-7928.

Chen, J. and Chen, Z. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008. doi: 10.1093/biomet/asn034.

Dempster, A. P. Covariance selection. *Biometrics*, 28 (1):157–175, 1972. ISSN 0006341X.

Duchi, J., Gould, S., and Koller, D. Projected subgradient methods for learning sparse gaussians. In *Proceedings of the Twenty-fourth Conference on Uncertainty in AI (UAI)*, 2008.

Fan, J. Local Linear Regression Smoothers And Their Minimax Efficiencies. *The Annals of Statistics*, 21 (1):196–216, 1993.

Fan, J. and Huang, T. Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, 11(6):1031–1057, 2005.

Fan, J. and Yao, Q. Efficient Estimation of Conditional Variance Functions in Stochastic Regression. *Biometrika*, 85(3):645–660, 1998.

Friedman, J., Hastie, T., and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Department of Statistics, Stanford University, Tech. Rep*, 2008a.

Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostat*, 9(3):432–441, 2008b. doi: 10.1093/biostatistics/kxm045.

Friedman, J., Hastie, T., and Tibshirani, R. A note on the group lasso and a sparse group lasso. *Preprint*, 2010.

Janson, S. Maximal Spacings In Several Dimensions. *The Annals of Probability*, 15(1):274–280, 1987.

Kolar, M., Song, L., Ahmed, A., and Xing, E.P. Estimating time-varying networks. *Annals of Applied Statistics (to appear)*, 2009.

Lauritzen, S. L. *Graphical Models (Oxford Statistical Science Series)*. Oxford University Press, USA, July 1996.

Li, R. and Liang, H. Variable Selection In Semiparametric Regression Modeling. *The Annals of Statistics*, 36(1):261–286, 2008.

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

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

Peng, J., Wang, P., Zhou, N., and Zhu, J. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009. doi: 10.1198/jasa.2009.0126.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence. Nov 2008.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. Sparse permutation invariant covariance estimation. *Electronic Journal Of Statistics*, 2:494, 2008.

Ruppert, D., Wand, MP, and Holst, U. Local polynomial variance-function estimation. *Technometrics*, 39(3):262–273, 1997.

Wang, H. and Xia, Y. Shrinkage estimation of the varying coefficient model. *Manuscript*, 2008.

Yin, J., Geng, Z., Li, R., and Wang, H. Nonparametric Covariance Model. *Statistica Sinica, Forthcoming*, 2008.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

Yuan, M. and Lin, Y. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1): 19–35, March 2007. doi: 10.1093/biomet/asm018.

Zhou, S., Lafferty, J., and Wasserman, L. Time varying undirected graphs. In Servedio, Rocco A. and Zhang, Tong (eds.), *COLT*, pp. 455–466. Omnipress, 2008.