Robust Inverse Covariance Estimation under Noisy Measurements

Jun-Kun Wang WANGJIM123@GMAIL.COM
Intel-NTU, National Taiwan University, Taiwan

Shou-de Lin SDLIN@CSIE.NTU.EDU.TW
Intel-NTU, National Taiwan University, Taiwan

Abstract

This paper proposes a robust method to estimate the inverse covariance under noisy measurements. The method is based on the estimation of each column in the inverse covariance matrix independently via robust regression, which enables parallelization. Different from previous linear programming based methods that cannot guarantee a positive semi-definite covariance matrix, our method adjusts the learned matrix to satisfy this condition, which further facilitates the tasks of forecasting future values. Experiments on time series prediction and classification under noisy condition demonstrate the effectiveness of the approach.

1. Introduction

Inverse covariance estimation of Gaussian variables in high dimensional setting has shown its importance in classification, structure learning, and time series prediction. A class of typical solutions is based on l_1 penalized log-likelihood, which encourages sparsity on its entries. Many efficient optimization methods and theories have been developed along this line (Yuan & Lin, 2007; Friedman et al., 2008; Banerjee et al., 2008; d'Aspremont et al., 2008; Rothman et al., 2008; Duchi et al., 2008; Ravikumar et al., 2011; Hsieh et al., 2011; 2013). Another class of solutions is based on estimating each column of inverse covariance via linear programming (Meinshausen & Buhlmann, 2006; Friedman et al., 2008; Yuan, 2010; Cai et al., 2011). An advantage of the linear-programing based methods is that the algorithms can be executed in parallel easily as the estimation of each column in the covariance matrix can be done independently. Our work is thus aligned to estimation via linear programming.

To our knowledge, people have not yet considered estimation of inverse covariance under noisy measurements. Noisy measurements in the features are common in many real-world applications utilizing sensor and networking data. Forecasting or classification may not achieve a satisfactory result without considering the existence of noise. Therefore, this paper focuses on designing a robust method in estimating the inverse covariance under noisy measurements.

The second contribution of our work is to provide a generative counterpart of Gaussian conditional random field (GCRF). GCRF models relations between two sets of random variables (Sohn & Kim, 2012; Yuan & Zhang, 2012; Wytock & Kolter, 2013). A promising empirical time series forecasting by GCRF has been studied (Wytock & Kolter, 2013). The objective of these works are based on minimizing penalized l_1 negative log-likelihood. Yet, such MLE based methods need to consider all the variables together to maximize the distribution during estimation, preventing the usage of a divide-and-conquer strategy. This may cause serious performance concern since, for instance, modeling thousands of variables requires the estimation of millions of entries in an inverse covariance matrix. Therefore, in this paper we choose to focus on another direction that facilitates the usage of divide-and-conquer strategy for better efficiency. Our generative counterpart jointly models the covariance of the input and output variables by a series of regression, and uses conditional distribution of Gaussian variables to obtain the model.

More importantly, the proposed strategy allows us to seamlessly leverage techniques in robust optimization (Ben-Tal et al., 2009) for estimation under noisy measurements. As can be seen through our experiments, with robust optimization techniques it is possible to significantly boost the forecasting and classification performance under noisy environment. For the l_1 penalized likelihood approach, since it is already a semi-definite programming problem. After robust optimization integrated into the approach, the resulting algorithm may become intractable. Though Banerjee et al.

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

(2008) and d'Aspremont et al. (2008) point out that the dual of l_1 MLE objective can be interpreted as a worst-case maximum likelihood with additive noise in the the sample covariance, the approach is not equivalent to model the additive noise in the features. Assuming the additive noise in the features is more general than assuming it in the sample covariance, since the former contains additional multiplicative component of noise in the sample covariance.

Yet, previous approaches of estimation by regression (Meinshausen & Buhlmann, 2006; Friedman et al., 2008; Yuan, 2010; Cai et al., 2011) have concerns that they do not guarantee the estimated matrix to be positive semi-definite. Previous works (Meinshausen & Buhlmann, 2006; Friedman et al., 2008; Yuan, 2010; Cai et al., 2011) focus on recovering the graph or the entries of a matrix, so they do not constrain the solutions to be positive semi-definite. Being positive semidefinite is a necessary condition for the learned matrix to be a valid (inverse) covariance (Yates & Goodman, 2004). Most of the sampling methods for multivariate Gaussian distribution require performing the Cholesky factorization of the given covariance (Barr & Slezak, 1972; Law & Kelton, 1991). If a matrix is not positive semi-definite, the factorization cannot be conducted, which prohibits the corresponding sampling that is required to perform a prediction task. Here we provide a method to guarantee the positive semi-definite property.

To summarize, our contributions, 1) to our knowledge, we are the first to study estimating inverse covariance under the existence of noise in the features. 2) We are the first to deal with a major concern in the previous works of using linear programming methods such that their solutions do not guarantee positive semi-definiteness, and thus prevent further sampling from the distribution. 3) We show that in Gaussian regression (i.e. GCRF), our generative approach performs better than discriminative approach under noisy condition.

2. Preliminaries

We now provide the background of our method. In Subsection 2.1, we give the notation used in the paper. Since our method is based on linear regression to estimate the inverse covariance, we describe a related work (Yuan, 2010) along this line in Subsection 2.2. Then we exploit the estimated covariance to perform classification and time series prediction. The corresponding models we use are linear discriminant analysis (LDA) for classification and GCRF for prediction, which will be introduced respectively in Subsection 2.3 and 2.4. Both models involve estimating inverse covariance.

2.1. Notation

In the following, we denote x_{-i} as the vector after removing the ith entry of x. Similarly, $\Sigma_{-i,-j}$ is the submatrix after removing the ith row and jth column of Σ. $\Sigma_{i,-j}$ is a vector that represents the ith row of Σ with its jth entry removed. $\Sigma_{-i,j}$ is a vector that represents the jth column of Σ with its ith entry removed.

2.2. Estimating inverse covariance via linear programming

We begin with stating two basic formula (Yates & Goodman, 2004). The first is about conditional distribution: if a random vector $x \in \mathbb{R}^d$ follows a multivariate Gaussian distribution $N(\mu, \Sigma)$, then the conditional distribution of x_i given the remaining variables x_{-i} still follows multivariate Gaussian distribution:

$$\begin{aligned} x_i | x_{-i} \sim N(\mu_i + &\Sigma_{i,-i}\Sigma^{-1}_{-i,-i}(x_{-i} - \mu_{-i}), \\ &\Sigma_{i,i} - \Sigma_{i,-i}\Sigma^{-1}_{-i,-i}\Sigma_{-i,i}). \end{aligned} \tag{1}$$

The second is inverse block formula of symmetric matrix which connects the covariance with its inverse counterpart. Denote $\Phi = \Sigma^{-1}$. By the inverse block formula of symmetric matrices for the first variable x_1, we have:

$$\begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,-1} \\ \Sigma_{-1,1} & \Sigma_{-1,-1} \end{pmatrix}^{-1} = \\ \begin{pmatrix} (\Sigma_{1,1} - \Sigma_{1,-1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,1})^{-1} & -\Sigma^{-1}_{-1,-1}\Sigma_{-1,1}\Phi_{1,1} \\ -\Sigma^{-1}_{-1,-1}\Sigma_{-1,1}\Phi_{1,1} & - \end{pmatrix}, \tag{2}$$

where $\Phi_{1,1} = (\Sigma_{1,1} - \Sigma_{1,-1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,1})^{-1}$. We can generalize the case for each variable: the ith column of Φ can be written as

$$\begin{aligned} \Phi_{i,i} &= (\Sigma_{i,i} - \Sigma_{i,-i}\Sigma^{-1}_{-i,-i}\Sigma_{-i,i})^{-1} \\ \Phi_{-i,i} &= -\Sigma^{-1}_{-i,-i}\Sigma_{-i,i}\Phi_{i,i}, \end{aligned} \tag{3}$$

Yuan (2010) points out that (1) can be viewed as an equivalent regression problem:

$$x_i = c_i + w^T_{(i)} x_{-i} + \epsilon_i, \tag{4}$$

where scalar $c_i = \mu_i - \Sigma_{i,-i}\Sigma^{-1}_{-i,-i}\mu_{-i}$, vector $w_i = \Sigma_{i,-i}$ $\Sigma^{-1}_{-i,-i}$ in $\mathbb{R}^{d-1}$, and $\epsilon_i \sim N(0, \Sigma_{i,i} - \Sigma_{i,-i}\ \Sigma^{-1}_{-i,-i}\ \Sigma_{-i,i})$. Combine (3) with (4), we get

$$\begin{aligned} \Phi_{i,i} &= (Var(\epsilon_i))^{-1} \\ \Phi_{-i,i} &= -w_{(i)}(Var(\epsilon_i))^{-1}, \end{aligned} \tag{5}$$

Thus, the inverse covariance of Φ can be estimated by regressing x_i over x_{-i}. Yuan (2010) suggests using Dantzig selector (Cands & Tao, 2005) to estimate the regression coefficients. Once the estimation of $w_{(i)}$ is obtained, $Var(\epsilon_i)$

can be estimated using the average squared error of residuals over samples:

$$\begin{aligned}\widehat{Var(\epsilon_i)} =& \frac{1}{m}\|x_i - x_{-i}^T w_{(i)}\|_2^2 = \\ & S_{i,i} - 2w_{(i)}^T S_{-i,i} + w_{(i)}^T S_{-i,-i} w_{(i)},\end{aligned} \tag{6}$$

where S represents the sample covariance.

The estimator of Φ obtained above may not be symmetric. Denote the obtained estimator as $\widetilde{\Phi}$. Yuan (2010) adjusts $\widetilde{\Phi}$ by seeking a symmetric matrix as an approximation.

$$\underset{\Phi \text{ is symmetric}}{\text{minimize}} \ \|\Phi - \widetilde{\Phi}\|_{1,\infty}. \tag{7}$$

Thus, estimating inverse covariance via linear programming provides a counterpart to l_1 MLE based approach:

$$\underset{\Phi}{\text{minimize}} = -log|\Phi| + tr(S\Phi) + \lambda\|\Phi\|_1 \tag{8}$$

where $\|.\|_1$ is the element-wise l_1 norm, and λ is the regularization parameter that controls the sparsity.

2.3. Gaussian conditional random field

GCRF models the conditional probability of a set of random variables $y \in \mathbb{R}^p$ given another set of random variables $x \in \mathbb{R}^d$, and assumes the variables follow multivariate Gaussian distribution,

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0_d} \\ \mathbf{0_p} \end{pmatrix}, \Sigma\right), \tag{9}$$

where $\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_{yy} \end{pmatrix} = \begin{pmatrix} \Phi_{xx} & \Phi_{xy} \\ \Phi_{xy}^T & \Phi_{yy} \end{pmatrix}^{-1}$.

By the conditional distribution of multivariate Gaussian variables, we have

$$y|x \sim N(-\Phi_{yy}^{-1}\Phi_{xy}^T x, \Phi_{yy}^{-1}). \tag{10}$$

Note that we have assumed variables y and x are zero mean and can be normalized in practice. We can rewrite (10) further to its exponential form,

$$P(y|x) = \frac{1}{z(x)} exp(-\frac{1}{2}y^T\Phi_{yy}y - x^T\Phi_{xy}y), \tag{11}$$

where

$$z(x) = c|\Phi_{yy}|^{-1/2} exp(\frac{1}{2}x^T\Phi_{xy}\Phi_{yy}^{-1}\Phi_{xy}^T x) \tag{12}$$

is the partition function. Let m be the number of samples, $X \in \mathbb{R}^{m\times d}$ be the input data matrix, and $Y \in \mathbb{R}^{m\times p}$ be the output data matrix, the negative log-likelihood is

$$\begin{aligned} f(\Phi_{yy}, \Phi_{xy}) = -log|\Phi_{xy}| + & tr(S_{yy}\Phi_{yy} + 2S_{yx}\Phi_{xy} \\ & + \Phi_{yy}^{-1}\Phi_{xy}^T S_{xx}\Phi_{xy}), \end{aligned} \tag{13}$$

where $S_{yy} \in \mathbb{R}^{p\times p}$ is the sample covariance of output variables, $S_{xx} \in \mathbb{R}^{d\times d}$ is the sample covariance of input variables, and $S_{yx} \in \mathbb{R}^{p\times d}$ is sample covariance between output and input variables. Typically, l_1 regularization is used to enforce the sparsity of the model parameters, Φ_{yy} and Φ_{xy}. Thus, the objective of GCRF becomes

$$\underset{\Phi_{yy},\Phi_{xy}}{\text{minimize}} f(\Phi_{yy}, \Phi_{xy}) + \lambda(\|\Phi_{yy}\|_1 + \|\Phi_{xy}\|_1). \tag{14}$$

2.4. Linear discriminant analysis

Consider classifying the binary class $k = \{1, -1\}$. Denote $x \in \mathbb{R}^d$ as the feature of a sample. Let π_k be a prior probability of class k. LDA assumes the features conditioned on the class follow multivariate Gaussian distribution (Hastie et al., 2009; Murphy, 2012),

$$f_k(x) = \frac{1}{(2\pi)^{d/2}|\Phi_k|^{-1/2}} exp(-\frac{1}{2}(x-\mu_k)^T\Phi_k(x-\mu_k)). \tag{15}$$

Assume both class have equal inverse covariance, $\Phi_k = \Phi$. By maximizing a posterior, the label is assigned by the class that has the maximum linear discriminant score:

$$\delta(k) = x^T\widehat{\Phi}\widehat{\mu_k} - \frac{1}{2}\widehat{\mu}_k^T\widehat{\Phi}\widehat{\mu}_k + log\widehat{\pi_k}, \tag{16}$$

where $\widehat{\pi_k}$ is the fraction of class k in the training set, $\widehat{\mu}_k$ is the mean of features in class k.

3. Our method

3.1. Inverse covariance estimation by robust regression

Following the analysis in Subsection 2.2, we can now consider estimating the inverse covariance under noisy measurements. To estimate a column of the matrix, our model solves the following objective:

$$\underset{w_{(i)}\in\mathbb{R}^{d-1}}{\text{minimize}}\{\underset{\Delta\in\mathbb{U}}{\text{maximize}}\|X_i - (X_{-i}+\Delta)w_{(i)}\|_2\}, \tag{17}$$

where $\Delta \in \mathbb{R}^{m\times(d-1)}$ is the measurement errors, and $\mathbb{U}$ is the uncertainty set, or the set of admissible disturbances of the data matrix $X_{-i} \in \mathbb{R}^{m\times(d-1)}$. For example, if the first column of inverse covariance is estimated, which corresponds to the covariance of the first variable with the others, the observed vector of the first variable, say $X_1 \in \mathbb{R}^m$, is regressed over $X_{-i} \in \mathbb{R}^{m\times(d-1)}$. Once $w_{(i)}$ is obtained, the column is formed according to (5). Such min-max objective is common in robust optimization literature (Ben-Tal et al., 2009; Xu et al., 2008; 2009) , which minimize the worst case losses in perturbation set.

We now explain the perturbation set. We assume the random variables can be divided into groups, and within each group the error pattern is similar. Across groups the error bound can be significantly different. Note that each group of variables maintains a matrix to represent its error pattern. To model the measurement error, we propose to optimize subject to the following perturbation set:

$$\|\Delta_g\|_2 \leq c_g \tag{18}$$

where g is the group index and Δ_g of which the i_{th} column is Δ_i (which represents the measurement errors for the i_{th} variable over samples) if the i_{th} variable belongs to group g, or 0 otherwise. For example, suppose group $g = 1$, consists of variables $R1$ and $R2$, then the first and the second column of Δ_g is Δ_1 and Δ_2, while the other columns are all zeros. The bound c_g in (18) has to be specified first. In contrast to assuming the distribution of the noise, specifying the perturbation bound is a more natural way. Since the range of measurement errors of an instrument is usually reported in the manual. c_g can be computed easily based on the information. We can place the variables readings from the sensors with similar measurement errors in a group, and here we make a reasonable assumption that a variable belongs only to one group.

The estimation is obtained by solving (17) subjected to the uncertainty set (18). To solve this, we leverage the work of Yang & Xu (2013). They provide a unified method that solves

$$\underset{w}{\text{minimize}}\{\underset{\Delta\in\mathbb{U}}{\text{maximize}}\|y-(X+\Delta)w\|_2\}, \tag{19}$$

under a general uncertainty set:

$$\begin{aligned}\mathbb{U} =&\{W_1\Delta^{(1)} + \cdots + W_t\Delta^{(t)}|\\ &\forall g \in G_i, \|\Delta_g^{(i)}\| \leq c_g\}.\end{aligned} \tag{20}$$

where $\Delta^{(t)}$ is the t_{th} type of perturbation set, matrix $W_t \in \mathbb{R}^{m\times m}$ is fixed or specified according to the relation over samples, G_i is the set of groups under perturbation set i, and c_g provides the bound of the disturbances for g_{th} group of G_i.

Proposition 1:(Yang & Xu, 2013) *Suppose that $t = 1$, $W_1 = I$, $G_1 = \{g_1, \ldots, g_k\}$ and $g_i \cap g_j = \emptyset$ for any $i \neq j$, then the robust regression problem (19-20) is equivalent to:*

$$\underset{w}{\text{minimize}}\|y - Xw\|_2 + \sum_i^k c_{g_i}\|w_{g_i}\|_2, \tag{21}$$

which is the non-overlapped group lasso (Yuan & Lin, 2006) and there exists efficient method to solve it (Roth & Fischer, 2008).

In our case, the perturbation (18) satisfies the conditions in Proposition 1. Thus, the robust optimization (17-18) is equivalent to

$$\underset{w_{(i)}}{\text{minimize}}\|X_i - X_{-i}w_{(i)}\|_2 + \sum_i^k c_{g_i}\|w_{g_i}\|_2. \tag{22}$$

After obtaining $w_{(i)}$ and using (5) to form the estimation of the column, (7) or (25) is used to symmetrize the matrix. The above analysis suggests that estimating the inverse covariance under noisy condition can be obtained by conducting group lasso for each column. One major advantage is that the estimation of each column is independent of other columns, which enables further parallelization. Though we consider non-overlapped structure here, the method can be generalized to overlapped structure, that is, the equivalent regression becomes overlapped group lasso (Jacob et al., 2009).

3.2. A generative counterpart of Gaussian conditional random field

Recall that the model of GCRF consists of two inverse covariance Φ_{yy} and Φ_{yx}; the former represents inverse covariance within output variables $y \in \mathbb{R}^p$ and the latter represents inverse covariance between output and input variables $x \in \mathbb{R}^d$. We can view training as the process of estimating the inverse covariance matrices that consists of input and output variables, $\Phi = \begin{pmatrix}\Phi_{xx} & \Phi_{xy}\\ \Phi_{xy}^T & \Phi_{yy}\end{pmatrix}$. Thus, the method proposed in the previous subsection can be exploited.

Objective (23) is solved when estimating the inverse covariance of an output variable y_i with the other variables, including the other output variables and all the input variables. Denote the data matrix $Z \in \mathbb{R}^{m\times(d+p-1)}$ with each row representing a sample (or instance) of the random variables $[x_1, \ldots, x_d, y_1, \ldots, y_{i-1}, y_{i+1}, y_p]$.

$$\underset{w\in\mathbb{R}^{d+p-1}}{\text{minimize}}\|y_i - Zw\|_2 + \sum_i^k c_{g_i}\|w_{g_i}\|_2, \tag{23}$$

where $y_i \in \mathbb{R}^m$ represents the samples of ith output variable. Similarly, when estimating the covariance of an input variable x_i with other variables, the following objective is solved:

$$\underset{w\in\mathbb{R}^{d+p-1}}{\text{minimize}}\|x_i - Zw\|_2 + \sum_i^k c_{g_i}\|w_{g_i}\|_2, \tag{24}$$

where matrix $Z \in \mathbb{R}^{m\times(d+p-1)}$ becomes the instances of $[x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d, y_1, \ldots, y_p]$. Then, to make the estimated inverse covariance symmetric, $\widehat{\Phi}$ can be set by

$$\begin{aligned}&\widehat{\Phi}_{i,j} = \widehat{\Phi}_{j,i} =\\ &\quad\widehat{\Phi}_{i,j}I(|\widehat{\Phi}_{i,j}| < |\widehat{\Phi}_{j,i}|) + \widehat{\Phi}_{j,i}I(|\widehat{\Phi}_{j,i}| < |\widehat{\Phi}_{i,j}|).\end{aligned} \tag{25}$$

However, the procedure does not guarantee that the inverse covariance is positive semi-definite. Note that all the

Algorithm 1 Adjusting the inverse covariance that guarantees positive semi-definiteness

```
Input: $\Phi_{\text{tmp}}$ an estimated symmetric inverse covariance by regression (may not be positive semi-definite), $\alpha$ is initialized $\in (0, 1]$, and $\beta \in (0, 1)$.
$D = \Phi_{\text{tmp}} - I$
repeat
    Compute the Cholesky factorization of $I + \alpha D$.
    if $I + \alpha D$ is not positive definite then
        $\alpha = \beta\alpha$
    end if
until $I + \alpha D$ is positive definite
```

previous approaches of estimation by regression share the same concern (Meinshausen & Buhlmann, 2006; Friedman et al., 2008; Yuan, 2010; Cai et al., 2011). Being positive semidefinite is a condition that the learned matrix is a valid (inverse) covariance (Yates & Goodman, 2004). The concern is critical, since when performing forecast, the predicted value is sampled according to (10). Here, we provide a way to adjust the learned covariance to become positive semi-definite, so that sampling from the learned distribution is admissible.

Our method is described in Algorithm 1, which is inspired by the work of (Hsieh et al., 2011). Hsieh et al. propose an efficient optimization for l_1 maximum likelihood estimation of inverse covariance. The optimization alternatively finds the descent direction and chooses a step size. The initial estimated matrix is set to a positive definite matrix (for example, an identity matrix). At every step, step size is calculated such that the estimated matrix is positive definite. Here, denote Φ_{tmp} as the estimated symmetric matrix after conducting series of robust regression. Let D be the matrix of an identity matrix subtracted from Φ_{tmp}. Algorithm 1 finds an α such that the final estimated matrix $I + \alpha D$ is positive (semi-)definite. Experiments show that choosing the identity matrix as the pivot is sufficient to achieve good performance in prediction and classification.

Proposition 2: *Algorithm 1 guarantees the final estimate to be positive definite.*
Proof: Since $I \succ 0$ and $\Phi_{\text{tmp}} - I$ is symmetric, the proposition is an immediate consequence of the following Lemma.
Lemma (Hsieh et al., 2011)**:** *For any $X \succ 0$ and a symmetric D, there exists an $\alpha' > 0$ such that for all $\alpha \leq \alpha'$: $X + \alpha D \succ 0$*

3.3. Comparison to related works

Loh and Wainwright (2012) consider high dimensional regression under noisy measurements, then Chen and Caramanis (2013) further extend their work. Our work differs from theirs as their works require the assumption of the noise (e.g. sub-Gaussian distribution) and the knowledge of its sufficient statistics, while ours does not. In our work, only the perturbation bound is required which is a less strict constraint and conceivably more general. Another difference is that they focus on successful recovery of true parameter, whereas ours concerns the practical application in prediction and classification.

Mazumder and Hastie (2012) and Hsieh (2012) propose to threshold the sample covariance before running the l_1 penalized likelihood approach. They show that the connected components of the threshold sample covariance by a regularization parameter is equal to the connected components of the learned estimator after solving the optimization problem with that regularization parameter. The result is that the original problem can be decomposed into several smaller problems, reducing the computational cost. Yet, when the number of variables in the original problem gets large, scalability issue arises since the number of entries in the subproblem still scales quadratically with the number of variables. Moreover, thresholding may result in unequally sized connected components, while the computation bottleneck would still be on the largest components. In contrast, the size of each regression in our model grows linearly with the number of variables.

4. Experiment

To show the merits of our work, the robust inverse covariance estimation is compared to several baselines in time series forecast and data classification. We use grid search to tune the regularization parameters. For our method, denote c as a vector whose entries are perturbation bound c_g, the regularization vector is searched by c times $[10^{-8}, 10^{-7}, \ldots, 10^{2}]$ over the grid.

4.1. Prediction

To predict time series, the input features contain previous three historic values, so the size of input variables of GCRF is three times the number of time series being modeled. After estimating the inverse covariance, predicting the time series is performed by sampling according to (10).

We conduct the experiments on three datasets.
1) Stock: The data are downloaded from Yahoo Finance, which contains daily stock prices. Companies among $S\&P$ 100 that have moderate variance ($\sigma \leq 15$) in year 2012 over the time are selected, which results in a set of 60 companies. The last 30 trading days are reserved for testing; the second to last 30 days are for validation; and the remaining data are for training.
2) Temperature (medium variable size): The data

are downloaded from National Oceanic and Atmospheric Administration (NOAA) [1], which contains daily air temperature. We choose the grid $(20°N, 50°N)$ and $(230°E, 290°E)$ with moderate variance ($\sigma \leq 15$), which results in 73 time series tracked. As the result, the size of inverse covariance in our generative model is $(73 \times 4)^2 \approx 85,000$ variables.

3) Temperature (large variable size): The temperature data in between $(30°N, 70°N)$ with moderate variance ($\sigma \leq 15$) are chosen, which results in 401 time series and about 2.5 millions entries in the estimated matrix. The test and validation sets are the last 30 days and the second to last 30 days of the year respectively; and the remaining days are for training.

To simulate noisy measurements, we consider two types of perturbation in the time series: uniform and Gaussian perturbation. To specify the noise level, the variance of each time series over time is first calculated. Furthermore, for the two smaller datasets, every random 10 variables are grouped together, assuming they have the same noise level. For the large dataset, every random 20 variables are grouped together. For uniform perturbation, the range of uniform distribution is randomly chosen between $\pm(0.1, 1)$ times the average variance in each group. For Gaussian perturbation, the standard deviation of the noise is set randomly to k times the average variance in each group, where k is a random value between (0.1, 1.1). After specifying the noise level and distribution, noise is sampled from the distribution and added to the time series in each group.

We compare our method with l_1 maximizing condition likelihood approach (Wytock & Kolter, 2013), denoted WK13. Since, to our knowledge, we are the first to consider estimating inverse covariance under perturbation, there is no prior work yet, we use the work of (Wytock & Kolter, 2013) as baseline. For each type of noise, we replicate each dataset 5 times, and the number inside the parenthesis in the tables is standard deviation. For clean data, we use lasso for estimating each column of inverse covariance and adjust the estimation with Algorithm 1 to obtain the final positive semi-definite matrix. We choose root mean square error (RMSE) as the measure for prediction.

Table 1. Forecasting results (RMSE) on clean data (no noise is added).

Data	WK13	Ours
stock	**2.6314**	2.6828
temp. (medium variable size)	**2.3917**	2.7966
temp. (large variable size)	2.4119	**1.9832**

Table 2. Forecasting results (RMSE) under uniform perturbation.

Data	WK13	Ours
stock	3.6296 (0.0876)	**3.1300** (0.0613)
temp. (medium variable size)	3.6697 (0.3561)	**3.0286** (0.0447)
temp. (large variable size)	4.7019 (0.6669)	**2.1671** (0.0220)

Table 3. Forecasting results (RMSE) under Gaussian perturbation.

Data	WK13	Ours
stock	5.7316 (0.1611)	**3.1751** (0.0888)
temp. (medium variable size)	6.2661 (0.4859)	**3.2863** (0.2468)
temp. (large variable size)	8.0439 (0.8086)	**2.2704** (0.0472)

Tables 1 to 3 show the final results. Without noisy measurements, l_1 conditional likelihood model performs better than our generative counterpart in **stock** and **temperature (medium variable size)** datasets. This is reasonable, since a discriminative approach usually outperforms the generative counterpart in prediction given limited amount of data. Under noise, our generative approach through robust regression performs significantly better than its discriminative counterpart. It may be due to that current l_1 conditional likelihood approach (Sohn & Kim, 2012; Yuan & Zhang, 2012; Wytock & Kolter, 2013) does not model the noise in the features. Yet, we see that the conditional likelihood approach degrades much more in Gaussian perturbation than in uniform perturbation. It seems that the conditional likelihood approach mistakenly models the noise as variables of interest through the input sample covariance. A challenge for maximizing conditional likelihood approach under noisy environment is to explicitly model the components of the signal and noise in the sample covariance, which inevitably leads to the assumption of certain distribution of noise and consequently lead to a less-general model. It is not a concern in our approach, because robust optimization does not assume the distribution of measurement error. For example, we only assume the perturbation bounds, and therefore suggest a more natural way to estimate the covariance matrix under general noisy condition.

4.2. Training time comparison

We also compare the training time of our approach with the l_1 discriminative likelihood approach of Wytock and Kolter (2013). Our experiment is run on a machine with dual core 2.66 GHz (INTEL E5500) and 32GB memory. We assume that there could exist at least k numbers of cores, where k is larger than the number of variables to be learned, thus we report the average time required to estimate one column of

[1] File name: air.sig995.2012.nc

Table 4. Training time in seconds.

Data	WK13	Our regress.	Our Algor.1
stock (clean)	34.99	0.30	2.44
stock (uniform)	5.26	5.66	2.27
stock (gaussian)	7.67	5.93	2.43
medium size			
temp. (clean)	308.57	0.47	3.72
temp. (uniform)	19.30	6.03	3.91
temp. (gaussian)	17.12	6.20	3.92
large size			
temp. (clean)	$\approx 8.2\times 10^3$	8.94	446.83
temp. (uniform)	$\approx 7.5\times 10^4$	13.23	293.88
temp. (gaussian)	$\approx 1.5\times 10^5$	13.35	275.58

the matrix and the time to adjust the matrix to become positive semi-definite. The running time in seconds is reported in Table 4. The advantage of estimation by regression is shown with a large number of variables are modeled, as the size of each regression grows linearly with number of variables; while the entries in the MLE based approach grows quadratically.

4.3. Classification

We conduct the experiments using four datasets, all are available on http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/.
1) Heart: For predicting the presence of the heart disease according to 13 attributes of potential patients. There are 270 samples in the dataset.
The following three datasets are about predicting whether a person suffers cancer:
2) Breast-cancer: There are 10 attributes and 683 samples.
3) Duke breast-cancer: There are 7129 attributes and 44 samples.
4) Colon-cancer: There are 2000 attributes and 62 samples.
For the last two datasets, we perform two sample t-test to select the first 100 significant attributes for feature reduction.

Same as forecasting, we consider uniform and Gaussian perturbation. For **heart** and **breast-cancer** datasets, the features are roughly divided into two equally sized groups. For **duke breast-cancer** and **colon-cancer** datasets, ten groups are formed with each consisting of ten features. For uniform perturbation, the range of uniform distribution is randomly chosen between $\pm(0.1, 2)$ times the average variance in each group. For Gaussian perturbation, the standard deviation of the noise is set randomly to k times the average variance in each group, where k is a random value between $(0.1, 1)$. After specifying the noise level and distribution, noise is sampled from the distribution and added to the time

Table 5. Classification results (MCC, F−measure, ACC) on clean data (no noise is added) over 5 random split.

Data	QUIC	Yuan
heart (MCC)	0.7070	0.7297
heart (F-mea)	0.8203	0.8409
heart (ACC)	0.8519	0.8630
breast. (MCC)	0.8944	0.9197
breast. (F-mea)	0.9327	0.9474
breast. (ACC)	0.9547	0.9635
duke. (MCC)	0.9633	0.9266
duke. (F-mea)	0.9788	0.9556
duke. (ACC)	0.9800	0.9600
colon. (MCC)	0.7767	0.7986
colon. (F-mea)	0.8570	0.8752
colon. (ACC)	0.8923	0.8923

Table 6. Classification results (MCC, F−measure, ACC) under Uniform perturbation.

Data	QUIC	Yuan	Ours
heart (MCC)	0.6618	0.6640	**0.6964**
heart (F-mea)	0.8043	0.8110	**0.8315**
heart (ACC)	0.8407	0.8333	**0.8481**
breast. (MCC)	0.8364	0.8023	**0.8651**
breast. (F-mea)	0.8854	0.8763	**0.9172**
breast. (ACC)	0.9255	0.9197	**0.9445**
duke. (MCC)	0.7841	0.6932	**0.8532**
duke. (F-mea)	0.8652	0.8392	**0.9111**
duke. (ACC)	0.8800	0.8400	**0.9200**
colon. (MCC)	**0.7760**	0.6626	**0.7772**
colon. (F-mea)	0.8267	0.7974	**0.8518**
colon. (ACC)	0.8462	0.8308	**0.8923**

Table 7. Classification results (MCC, F−measure, ACC) under Gaussian perturbation.

Data	QUIC	Yuan	Ours
heart (MCC)	0.6936	0.6641	**0.7074**
heart (F-mea)	0.8223	0.8070	**0.8333**
heart (ACC)	0.8481	0.8407	**0.8556**
breast. (MCC)	0.8488	0.8548	**0.8743**
breast. (F-mea)	0.8987	0.9018	**0.9170**
breast. (ACC)	0.9314	0.9343	**0.9431**
duke. (MCC)	**0.8165**	0.6608	0.6975
duke. (F-mea)	**0.8929**	0.8114	0.8256
duke. (ACC)	**0.9000**	0.8200	0.8400
colon. (MCC)	0.5904	0.6389	**0.7444**
colon. (F-mea)	0.7212	0.7644	**0.8396**
colon. (ACC)	0.8000	0.8308	**0.8769**

series in each group.

Once the inverse covariance is estimated, the testing data are classified based on the corresponding linear discriminant score. We compare our method with 1) the work of Hsieh (2011), which is a l_1 MLE based approach, denoted QUIC, and 2) the work of Yuan (2010), which is introduced in Subsection 2.2. For each dataset, we random split data 5 times that 80 percent of data are for cross-validation and the remaining for testing. The noise is added on the 5 replicated data for each dataset. We report the average results over the 5 test sets. Because there is imbalance between the positive and negative instances, choosing accuracy as the measure may not be adequate. We use classification accuracy (ACC) as well as F-measure and Mathews correlation coefficient (MCC) as the measures: MCC $= \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$. The experiment results are shown on Tables 5 to 7.

Table 5 shows the classification results before noise is added. The effectiveness of our robust estimation can be seen when noise is added (Table 6 and 7). Our method is comparable to or outperforms QUIC (Hsieh et al., 2011) with only one exception in the duke cancer data. Since both ours and the work of Yuan (2010) estimate inverse covariance by regression, better performance demonstrates the effectiveness of performing robust regression.

5. Conclusion

This work introduces a new research direction to estimate the inverse covariance under perturbation using regression-based techniques, which is very different from most of the existing works of efficiently solving the common l_1 penalized likelihood objective. By converting the problem into linear-programming based problem, we are able to exploit the robust optimization techniques to handle this problem; each column of the matrix is estimated by the equivalent group lasso. To guarantee the positive semi-definite property of the learned matrix, we further provided an algorithm to adjust the matrix through choosing the step size as in a coordinate descent manner. We showed the promising empirical results of our approach in forecasting and classification. An immediate future work will be to deal with missing values. Since the estimation has been converted to several regression problems, the proposed framework allows us to leverage existing works of regression that concern missing data. Furthermore, a unified solution that simultaneously deals with noisy and missing data is possible along this direction. We also want to study the statistical properties such as recovering the graph or the true inverse covariance given the existence of perturbation. The codes to reproduce the experiments are available on the first author's page https://sites.google.com/site/wangjim123.

Acknowledgements

We thank Ting-Wei Lin for providing technical support and the anonymous reviewers for comments. This work was sponsored by AOARD grant number No. FA2386-13-1-4045 and National Science Council, National Taiwan University and Intel Corporation under Grants NSC102-2911-I-002-001 and NTU103R7501 and grant 102-2923-E-002 -007 -MY2, 102-2221-E-002 -170, 101-2628-E-002 -028 -MY2.

References

Banerjee, O., Ghaoui, L. El, and d'Aspremont, A. Model selection through sparse maximun likelihood. *Journal of Machine Learning Research*, 9:485–516, 2008.

Barr, D.R. and Slezak, N.L. A comparison of multivariate normal generators. *Communications of the ACM*, 15(12): 1048–1049, 1972.

Ben-Tal, A., Ghaoui, L. El, and Nemirovski, A.S. *Robust Optimization*. Princeton University Press, 2009.

Cai, T., Liu, W., and Lou, X. A constrained l_1 minimization approach to sparse precision matrix estimation. *Journal of American Statistical Association*, 106:594–607, 2011.

Cands, E. J. and Tao, T. The dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2313–2351, 2005.

Chen, Y. and Caramanis, Constantine. Noisy and missing data regression: Distribution-oblivious support recovery. In *The Proceedings of the International Conference on Machine Learning (ICML) 30*, 2013.

d'Aspremont, A., Banerjee, O., and Ghaoui, L. El. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and its Applications*, 30(1): 56–66, 2008.

Duchi, J., Gould, S., and Koller, D. Projected subgradient methods for learning sparse gaussians. In *Conference on Uncertainty in Artificial Intelligence (UAI) 24*, 2008.

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

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition*. Springer, 2009.

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

Hsieh, C.-J., Dhillon, I., Ravikumar, P., and Banerjee, A. A divide-and-conquer method for sparse inverse covariance estimation. In *Advances in Neural Information Processing Systems (NIPS) 25*, 2012.

Hsieh, C.-J., Sustik, M., Dhillon, I., Ravikumar, P., and Poldrack, R. Big & quic: Sparse inverse covariance estimation for a million variables. In *Advances in Neural Information Processing Systems (NIPS) 26*, 2013.

Jacob, L., Obozinski, G., and Vert, J.-P. Group lasso with overlap and graph lasso. In *The Proceedings of the International Conference on Machine Learning (ICML) 26*, 2009.

Law, A.M. and Kelton, W.D. *Simulation Modeling and Analysis*. McGraw-Hill College; 2 edition, 1991.

Loh, P. and Wainwright, M.J. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Annals of Statistics*, 40(3):1637–1664, 2012.

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

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

Murphy, K.P. *Machine Learning: a Probabilistic Perspective*. The MIT Press, 2012.

Ravikumar, P., Wainwright, M., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing l_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Roth, Volker and Fischer, Bernd. The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *The Proceedings of the International Conference on Machine Learning (ICML) 25*, 2008.

Rothman, A., Bickel, P., Levina, E., and zhou, J. Sparse permutation invariant covariance matrices. *Electronic Journal of Statistics*, 2:494–515, 2008.

Sohn, K. and Kim, S. Joint estimation of structured sparsity and and output structure in multiple-output regression via inverse-covariance regularization. In *International Conference on Artificial Intelligence and Statistics (AISTATS) 15*, 2012.

Wytock, M. and Kolter, Z. Sparse gaussian conditional random fields: algorithms, theory, and application to energy forecasting. In *International Conference on Machine Learning (ICML) 30*, 2013.

Xu, H., Caramanis, C., and Mannor, S. Robust regression and lasso. In *Advances in Neural Information Processing Systems (NIPS) 21*, 2008.

Xu, H., Caramanis, C., and Mannor, S. Robustness and regularization of support vector machines. In *Journal of Machine Learning Research, 10:1485-1510*, 2009.

Yang, W. and Xu, H. A unified robust regression model for lasso-like algorithms. In *International Conference on Machine Learning (ICML) 30*, 2013.

Yates, R.D. and Goodman, D. *Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers*. Wiley, 2004.

Yuan, M. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11:2261–2286, 2010.

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*, 94:19–35, 2007.

Yuan, X.-T. and Zhang, T. Partial gaussian graphical model estimation. In *CoRR, abs/1209.6419*, 2012.