Robust Inverse Covariance Estimation under Noisy Measurements

Jun-Kun Wang, Shou-De Lin

Intel-NTU, National Taiwan University

ICML 2014

Table of contents

1. Introduction
2. Background
   - Related works: neighborhood estimation
3. This work
   - Robust inverse covariance estimation
   - Generative counterpart of GCRF
   - Positive semi-definiteness Guarantee
4. Experiment
   - Time series prediction
   - Classification
5. Conclusion

What inverse covariance estimation can do?

1) **graphical model/ structure learning**. The non-zeros pattern in a matrix has a one-to-one correspondence of edges in a Markov network.

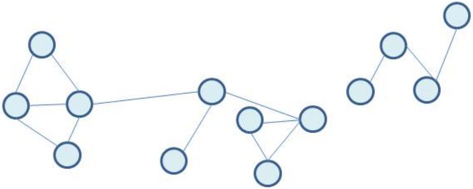


Figure: structure learning of a Markov network

What inverse covariance estimation can do?

2) **time series prediction**: Gaussian conditional random field (GCRF)
The l_1 penalized likelihood, Wytock and Kolter (2013).

$$\underset{\Phi_{yy},\Phi_{xy}}{\text{minimize}} \; -log|\Phi_{xy}| + tr(S_{yy}\Phi_{yy} + 2S_{yx}\Phi_{xy} + \Phi_{yy}^{-1}\Phi_{xy}^{T}S_{xx}\Phi_{xy}) + \lambda(\|\Phi_{yy}\|_1 + \|\Phi_{xy}\|_1). \quad (1)$$

where Φ_{yy} reveals the relations within output variables y, and Φ_{xy} reveals the relations between input and output variables.

After learning the inverse covariance, prediction is made by sampling

$$y|x \sim N(-\Phi_{yy}^{-1}\Phi_{xy}^{T}x, \Phi_{yy}^{-1}). \quad (2)$$

What inverse covariance estimation can do?

3) **classification**: Linear discriminant analysis (LDA) (Hastie et al. (2009) and Murphy (2013))
LDA assumes the features conditioned on the class follow multivariate Gaussian distribution.

By maximizing a posterior, the label is assigned by the class that has the maximum linear discriminant score:

$$\delta(k) = x^T \Phi \widehat{\mu_k} - \frac{1}{2} \widehat{\mu}_k^T \Phi \widehat{\mu}_k + log \widehat{\pi_k}, \tag{3}$$

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.

Related works for estimating inverse covariance

1) l_1 **penalized negative log-likelihood**:

$$\underset{\Phi}{\text{minimize}} \;\; = -log|\Phi| + tr(S\Phi) + \lambda|\Phi| \qquad (4)$$

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).

2) **neighborhood estimation: (The paper belongs to this category.)**

Estimate each column by linear regression/programing.

Meinshausen and Buhlmann (2006); Friedman et al. (2008); Yuan (2010); Cai et al. (2011).

Neighborhood estimation

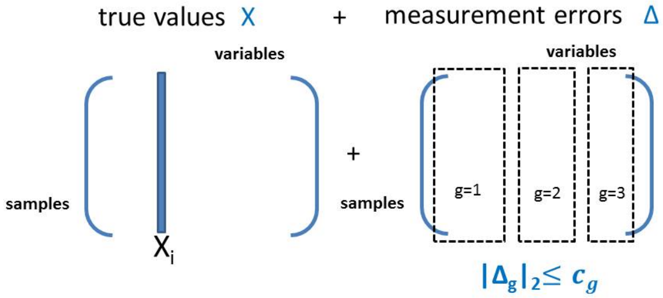
To estimate a column $i \in \{1, \ldots, d\}$ of the inverse covariance.
step 1: running a regression.

$$x_i = c_i + w_{(i)}^T x_{-i} + \epsilon_i. \tag{5}$$

step 2: forming a column.

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

step 3: adjusting the matrix to be symmetric.



Robust optimization:

$$\underset{w_{(i)}\in\mathbb{R}^{d-1}}{\text{minimize}}\{\underset{\Delta\in\mathbb{U}}{\text{maximize}}\|X_i-(X_{-i}+\Delta)w_{(i)}\|_2\}, \tag{7}$$

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 $X_{-i} \in \mathbb{R}^{m\times(d-1)}$.

Subproblem about estimating the covariance of each input variable with the others

$$\underset{w_{(i)}\in\mathbb{R}^{d-1}}{\text{minimize}}\{\underset{\Delta\in\mathbb{U}}{\text{maximize}}\|X_i-(X_{-i}+\Delta)w_{(i)}\|_2\}, \tag{8}$$

To model the measurement error, we propose to optimize under the following uncertainty set:

$$\|\Delta_g\|_2 \leq c_g \tag{9}$$

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.

The robust optimization is equivalent to the following objective, by a proposition in Yang and Xu (2013).

Equivalent to group lasso

$$\underset{w\in\mathbb{R}^d}{\text{minimize}}\|X_{-i}-X_iw\|_2+\sum_{i}^{k}c_{g_i}\|w_{g_i}\|_2, \tag{10}$$

which is the non-overlapped group lasso (Yuan and Lin, 2006) and exists efficient methods to solve it (Meier et al., 2008 and Roth and Fischer, 2008).

Generative counterpart of GCRF

Recall the GCRF. The l_1 penalized likelihood, Wytock and Kolter (2013).

$$\underset{\Phi_{yy},\Phi_{xy}}{\text{minimize}} \; -log|\Phi_{xy}| + tr(S_{yy}\Phi_{yy} + 2S_{yx}\Phi_{xy} + \Phi_{yy}^{-1}\Phi_{xy}^{T}S_{xx}\Phi_{xy}) + \lambda(\|\Phi_{yy}\|_1 + \|\Phi_{xy}\|_1). \tag{11}$$

We propose a generative counterpart that enables parallelization.

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.

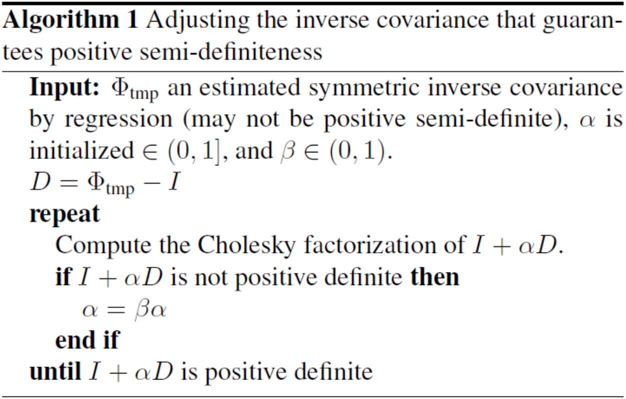
To guarantee positive semi-definiteness

Concern

The estimators are not guarantee to be positive semi-definite. (Meinshausen and Buhlmann (2006); Friedman et al. (2008); Yuan (2010); Cai et al. (2011) also share the same concern).

Most of the sampling methods for multivariate Gaussian distribution require performing the Cholesky factorization of the given estimated covariance.
When prediction, predicted values are sampled from (Barr and Slezak (1972); Law and Kelton (1991)).

$$y|x \sim N(-\Phi_{yy}^{-1}\Phi_{xy}^{T}x, \Phi_{yy}^{-1}). \tag{12}$$



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
```

Lemma Hsieh et al. (2011)**:** *For any $X \succ 0$ and a symmetric D, there exists an $\alpha' > 0$ such that for all $\alpha \leq \alpha'$: $X + \alpha D \succ 0$.*

To estimate a column $i \in \{1, \dots, d\}$ of the inverse covariance.

step 1: running a group lasso.

$$\underset{w_{(i)}}{\text{minimize}} \|X_i - X_{-i} w_{(i)}\|_2 + \sum_{i}^{k} c_{g_i} \|w_{g_i}\|_2. \tag{13}$$

step 2: forming a column.

$$\begin{aligned} \Phi_{i,i} &= (Var(\epsilon_i))^{-1} \\ \Phi_{-i,i} &= -w_{(i)}(Var(\epsilon_i))^{-1} \\ \text{where} \widehat{Var(\epsilon_i)} &= \frac{1}{m}\|x_i - w_{(i)}^T x_{-i}\|_2^2 = \\ & S_{i,i} - 2w_{(i)}^T S_{-i,i} + w_{(i)}^T S_{-i,-i} w_{(i)}. \end{aligned} \tag{14}$$

step 3: adjusting the matrix to be symmetric.

step 4: adjusting the matrix to be positive semi-definite by the proposed algorithm .

Experiment sketch

Robust method of inverse convariance esimation under noisy measurements in sensor data

- Time series predicton: GCRF.
- Classification: LDA.

GCRF Time series prediction: experiment setup

- Preprocessing: We choose time series that have moderate variance ($\sigma \leq 15$). The input features contain the values of time series previous three days (AR 3).
- Datasets:
  1) Stock: *S&P* 100 in year 2012. 60 times series.
  2) Temperature (medium variable size): NOAA 73 time series.
  3) Temperature (large variable size): NOAA 401 time series.
- Baseline: WK13 (ICML 13), an l_1 penalized approach for GCRF .

GCRF Time series prediction: experiment setup

To simulate the noisy measurements, we add some artificial noise in the data.

1. Every 10 time series are randomly grouped and the noise in the series of a group will be given the same perturbation bound.

2. Specifying the noise level and noise distribution.

The average variance of time series over time in a group is first calculated, and for

Uniform noise:

The range of uniform noise is randomly chosen between $\pm(0.1, 1)$ times the average variance in each group.

Gaussian noise:

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).

GCRF Time series prediction: experiment setup

Our method only require the bound, not the actual value. We provide the information to our method based on the sufficient statistics of the distribution of the noise we add.

1. **Uniform noise**: range of the distribution.
2. **Gaussian noise**: standard deviation of the distribution.

$$\|\Delta_g\|_2 \leq c_g \tag{15}$$

$$\underset{w_{(i)}}{\text{minimize}} \|X_i - X_{-i} w_{(i)}\|_2 + \sum_i^k c_{g_i} \|w_{g_i}\|_2. \tag{16}$$

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.

GCRF Time series prediction: experiment result

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.

Note: For clean data, our method use lasso to estimate the inverse convariance.

Table: 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**

GCRF Time series prediction: experiment result

Table: 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)

GCRF Time series prediction: experiment result

Table: 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)

LDA classification: experiment setup

- Datasets: heart: instances:270 feature dimension:13
  colon_cancer: instances:62 feature dimension:2000
  breast_cancer: instances:683 feature dimension:10
  duke_breast_cancer: instances:44 feature dimension:7129
- Settings: For each dataset, we randomly 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.
- Baselines:
  1) QUIC: A l_1 penalized likelihood approach (An MRF, Hsieh et al. (2011))
  2) Yuan: A regression based approach (Yuan (2010))

LDA classification: experiment results

Table: Classification results (Accuracy) on clean data (no noise is added).

Data	QUIC	Yuan
heart (ACC)	0.8519	0.8630
breast. (ACC)	0.9547	0.9635
duke. (ACC)	0.9800	0.9600
colon. (ACC)	0.8923	0.8923

LDA classification: experiment results

Table: Classification results (Accuracy) under uniform and Gaussian perturbation.

Data	QUIC	Yuan	Ours
heart (uniform noise)	0.8407	0.8333	**0.8481**
breast. (uniform noise)	0.9255	0.9197	**0.9445**
duke. (uniform noise)	0.8800	0.8400	**0.9200**
colon. (uniform noise)	0.8462	0.8308	**0.8923**
heart (Gaussian noise)	0.8481	0.8407	**0.8556**
breast. (Gaussian noise)	0.9314	0.9343	**0.9431**
duke. (Gaussian noise)	**0.9000**	0.8200	0.8400
colon. (Gaussian noise)	0.8000	0.8308	**0.8769**

Contributions

1) Study inverse covariance estimation under the existence of additive noise in the features.

2) Guarantee the estimator to be positive semi-definite.

3) Show the effectiveness of our method in classification and time series prediction.

Future works

1. Dealing with noisy measurements and missing value simultaneouely.
2. Recovering the Markov network.

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

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

Ben-Tal, A., Ghaoui, L. E., and Nemirovski, A. (2009). *Robust Optimization*. Princeton University Press.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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