Joint Gaussian Graphical Model Series – VI

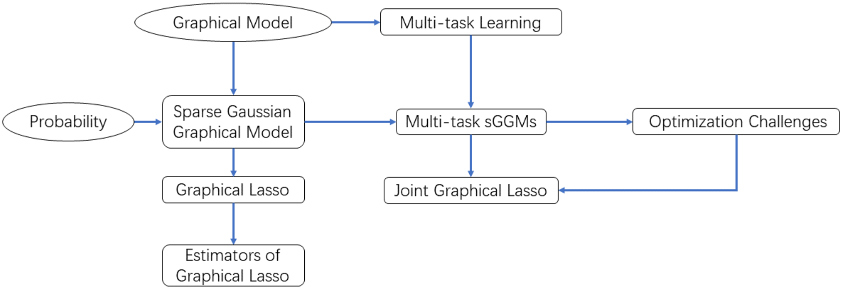
Multi-task sGGMs and its optimization challenges

Beilun Wang
Advisor: Yanjun Qi

[1]Department of Computer Science, University of Virginia
http://jointggm.org/

August 4th, 2017

Road Map


Graphical Model
Multi-task Learning
Probability
Sparse Gaussian Graphical Model
Multi-task sGGMs
Optimization Challenges
Graphical Lasso
Joint Graphical Lasso
Estimators of Graphical Lasso

Outline

1. Notation
2. Review
3. Multi-task Learning
4. Multi-task sGGMs
5. Optimization Challenge of Multi-task sGGMs
6. Joint Graphical Lasso Example

Notation

Notation

$X^{(i)}$	The i-th data matrix
$\Sigma^{(i)}$	The i-th covariance matrix.
$\Omega^{(i)}$	The i-th precision matrix.
p	The number of features.
n_i	The number of samples in the i-th data matrix.
K	The number of tasks.

Review

Review from last talk

- We introduce four estimators of sparse Gaussian Graphical Model.
- We finish most contents about sparse Gaussian Graphical Model in the last five talks.

Review of Gaussian Graphical Model

Suppose the precision matrix $\Omega = \Sigma^{-1}$.
The log-likelihood of Ω equals to $\ln \det(\Omega) - \operatorname{tr}\left(\Omega \widehat{S}\right)$

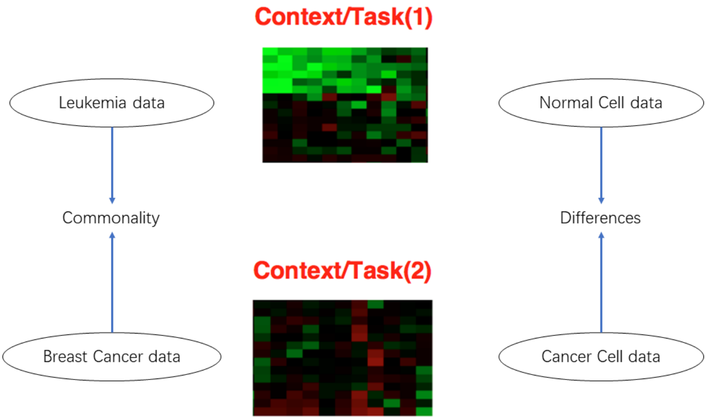
Multi-task Learning

Multi-task Learning

Multi-task Learning

Multi-task learning (MTL) is a subfield of machine learning in which multiple learning tasks are solved at the same time, while exploiting commonalities and differences across tasks.
This can result in improved learning efficiency and prediction accuracy for the task-specific models, when compared to training the models separately.

Multi-task Learning



Multi-task Learning–Linear Classifier Example

Linear Classifier

$$f(x) = \mathrm{sgn}(w^T x + b) \tag{3.1}$$

Multi-task Linear Classifiers

For the *i*-th task,

$$f_i(x) = \mathrm{sgn}((w_S^T + w_i^T)x + b) \tag{3.2}$$

Multi-task sGGMs

Multi-task sGGMs

Problem

- Input: $\{X^{(i)}\}$
- Output: $\{\Omega^{(i)}\}$
- Assumption I: Sparsity
- Assumption II: Commonalities and Differences

Multi-task sGGMs

Likelihood

$$\sum_i n_i(\ln \det(\Omega^{(i)}) - \operatorname{tr}\left(\Omega^{(i)}\widehat{S}^{(i)}\right)) \tag{4.1}$$

Likelihood with sparsity assumption

$$\sum_i n_i(\ln \det(\Omega^{(i)}) - \operatorname{tr}\left(\Omega^{(i)}\widehat{S}^{(i)}\right)) \tag{4.2}$$

$$\text{Subject to: } ||\Omega^{(i)}||_1 \leq t \tag{4.3}$$

Multi-task sGGMs

Likelihood with multi-task setting

$$\sum_i n_i(\ln\det(\Omega^{(i)}) - \operatorname{tr}\left(\Omega^{(i)}\widehat{S}^{(i)}\right)) \tag{4.4}$$

$$\text{Subject to: } ||\Omega^{(i)}||_1 \le t \tag{4.5}$$

$$P(\Omega^{(1)}, \Omega^{(2)}, \ldots, \Omega^{(K)}) \le t_2 \tag{4.6}$$

Joint Graphical Lasso [Danaher et al.(2013)Danaher, Wang, and Witten]

$$-\sum_i n_i(\ln\det(\Omega^{(i)}) + \operatorname{tr}\left(\Omega^{(i)}\widehat{S}^{(i)}\right)) + \lambda_1||\Omega^{(i)}||_1 + \lambda_2 P(\Omega^{(1)}, \Omega^{(2)}, \ldots, \Omega^{(K)}) \tag{4.7}$$

Optimization Challenge of Multi-task sGGMs

General formulation

Likelihood with multi-task setting

$$-\sum_i n_i(\ln\det(\Omega^{(i)}) + \operatorname{tr}\left(\Omega^{(i)}\widehat{S}^{(i)}\right)) \quad (5.1)$$

$$\text{Subject to: } ||\Omega^{(i)}||_1 \le t \quad (5.2)$$

$$P(\Omega^{(1)}, \Omega^{(2)}, \dots, \Omega^{(K)}) \le t_2 \quad (5.3)$$

General formulation

$$\sum_{x,z} f(x) + g(z) \quad (5.4)$$

$$\text{Subject to: } Ax + Bz = c \quad (5.5)$$

Optimization Challenge

Solution–Distributed optimization

Alternating direction method of multipliers

- ADMM problem form (with f, g convex)

$$\begin{array}{ll} \text{minimize} & f(x) + g(z) \\ \text{subject to} & Ax + Bz = c \end{array}$$

  – two sets of variables, with separable objective

- $L_\rho(x, z, y) = f(x) + g(z) + y^T(Ax + Bz - c) + (\rho/2)\|Ax + Bz - c\|_2^2$

- ADMM:

$$\begin{array}{lll} x^{k+1} & := & \text{argmin}_x\, L_\rho(x, z^k, y^k) \qquad // \textit{ x-minimization} \\ z^{k+1} & := & \text{argmin}_z\, L_\rho(x^{k+1}, z, y^k) \qquad // \textit{ z-minimization} \\ y^{k+1} & := & y^k + \rho(Ax^{k+1} + Bz^{k+1} - c) \qquad // \textit{ dual update} \end{array}$$

Optimization Challenges

- For $K > 2$ tasks, you need carefully derive the whole optimization solution.
- Each step in each iteration is still a sub-optimization problem. Sometimes, it is already difficult to solve.
- This method is at most linear Convergence.

Joint Graphical Lasso Example

JGL-group Lasso example

$$
\begin{aligned}
L_\rho(\{\boldsymbol{\Theta}\},\{\mathbf{Z}\},\{\mathbf{U}\}) = & - \sum_{k=1}^{K} n_k \left(\log\det \boldsymbol{\Theta}^{(k)} - \mathrm{trace}(\mathbf{S}^{(k)}\boldsymbol{\Theta}^{(k)}) \right) + P(\{\mathbf{Z}\}) \\
& + \frac{\rho}{2} \sum_{k=1}^{K} ||\boldsymbol{\Theta}^{(k)} - \mathbf{Z}^{(k)} + \mathbf{U}^{(k)}||_F^2,
\end{aligned}
$$

$$
P(\{\boldsymbol{\Theta}\}) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\theta_{ij}^{(k)}| + \lambda_2 \sum_{i \neq j} \sqrt{\sum_{k=1}^{K} {\theta_{ij}^{(k)}}^2}.
$$

$$
\begin{aligned}
&\text{(a)}\ \{\boldsymbol{\Theta}_{(i)}\} \leftarrow \arg\min_{\{\boldsymbol{\Theta}\}} \left\{ L_\rho\left(\{\boldsymbol{\Theta}\},\{\mathbf{Z}_{(i-1)}\},\{\mathbf{U}_{(i-1)}\}\right) \right\}. \\
&\text{(b)}\ \{\mathbf{Z}_{(i)}\} \leftarrow \arg\min_{\{\mathbf{Z}\}} \left\{ L_\rho\left(\{\boldsymbol{\Theta}_{(i)}\},\{\mathbf{Z}\},\{\mathbf{U}_{(i-1)}\}\right) \right\}. \\
&\text{(c)}\ \{\mathbf{U}_{(i)}\} \leftarrow \{\mathbf{U}_{(i-1)}\} + (\{\boldsymbol{\Theta}_{(i)}\} - \{\mathbf{Z}_{(i)}\}).
\end{aligned}
$$

JGL solution – updating $\Theta^{(i)}$

For $k = 1, \ldots, K$, update $\mathbf{\Theta}_{(i)}^{(k)}$ as the minimizer (with respect to $\mathbf{\Theta}^{(k)}$) of

$$-n_k \left(\log \det \mathbf{\Theta}^{(k)} - \operatorname{trace}(\mathbf{S}^{(k)} \mathbf{\Theta}^{(k)}) \right) + \frac{\rho}{2} || \mathbf{\Theta}^{(k)} - \mathbf{Z}_{(i-1)}^{(k)} + \mathbf{U}_{(i-1)}^{(k)} ||_F^2 .$$

Letting $\mathbf{VDV}^T$ denote the eigendecomposition of $\mathbf{S}^{(k)} - \rho \mathbf{Z}_{(i-1)}^{(k)} / n_k + \rho \mathbf{U}_{(i-1)}^{(k)} / n_k$, the solution is given (Witten & Tibshirani 2009) by $\mathbf{V}\tilde{\mathbf{D}}\mathbf{V}^T$, where $\tilde{\mathbf{D}}$ is the diagonal matrix with jth diagonal element

$$\frac{n_k}{2\rho} \left(-D_{jj} + \sqrt{D_{jj}^2 + 4\rho / n_k} \right) .$$

Set the gradient to be 0, we can get the SVD part of the solution.

JGL solution – updating $Z^{(i)}$

$$\underset{\{\mathbf{Z}\}}{\text{minimize}} \left\{ \frac{\rho}{2} \sum_{k=1}^{K} ||\mathbf{Z}^{(k)} - \mathbf{A}^{(k)}||_F^2 + P(\{\mathbf{Z}\}) \right\},$$

where

$$\mathbf{A}^{(k)} = \mathbf{\Theta}_{(i)}^{(k)} + \mathbf{U}_{(i-1)}^{(k)}.$$

$$\underset{\{\mathbf{Z}\}}{\text{minimize}} \left\{ \frac{\rho}{2} \sum_{k=1}^{K} ||\mathbf{Z}^{(k)} - \mathbf{A}^{(k)}||_F^2 + \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |Z_{ij}^{(k)}| + \lambda_2 \sum_{i \neq j} \sqrt{\sum_k {Z_{ij}^{(k)}}^2} \right\}.$$

$$\hat{Z}_{ij}^{(k)} = S(A_{ij}^{(k)}, \lambda_1/\rho) \left(1 - \frac{\lambda_2}{\rho \sqrt{\sum_{k=1}^{K} S(A_{ij}^{(k)}, \lambda_1/\rho)^2}} \right)_+ ,$$

An example for difficulty of ADMM

Algorithm 1: ADMM algorithm for the PNJGL optimization problem (6)

input: $\rho > 0, \mu > 1, t_{\max} > 0$;
Initialize: Primal variables to the identity matrix and dual variables to the zero matrix;
for $t = 1{:}t_{\max}$ **do**
  $\rho \leftarrow \mu\rho$;
  while *Not converged* **do**
    $\Theta^1 \leftarrow \text{Expand}\left(\frac{1}{2}(\Theta^2 + V + W + Z^1) - \frac{1}{2\rho}(Q^1 + n_1 S^1 + F), \rho, n_1\right)$;
    $\Theta^2 \leftarrow \text{Expand}\left(\frac{1}{2}(\Theta^1 - (V + W) + Z^2) - \frac{1}{2\rho}(Q^2 + n_2 S^2 - F), \rho, n_2\right)$;
    $Z^i \leftarrow \mathcal{T}_1\left(\Theta^i + \frac{Q^i}{\rho}, \frac{\lambda_1}{\rho}\right)$ for $i = 1, 2$;
    $V \leftarrow \mathcal{T}_q\left(\frac{1}{2}(W^T - W + (\Theta^1 - \Theta^2)) + \frac{1}{2\rho}(F - G), \frac{\lambda_2}{2\rho}\right)$;
    $W \leftarrow \frac{1}{2}(V^T - V + (\Theta^1 - \Theta^2)) + \frac{1}{2\rho}(F + G^T)$;
    $F \leftarrow F + \rho(\Theta^1 - \Theta^2 - (V + W))$;
    $G \leftarrow G + \rho(V - W^T)$;
    $Q^i \leftarrow Q^i + \rho(\Theta^i - Z^i)$ for $i = 1, 2$

Summary

- We introduce the multi-task sGGMs problem.
- We introduce the challenges of the optimization for this problem.
- We introduce the ADMM method and its drawbacks.

References I

P. Danaher, P. Wang, and D. M. Witten.
The joint graphical lasso for inverse covariance estimation across multiple classes.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2013.