Joint Statistical Modeling of Multiple High Dimensional Data

University of North Carolina at Chapel Hill

Wonyul Lee and Yufeng Liu

Current Challenges in Statistical Learning

December 12, 2011

Outline

1. Motivation and Problem
2. Multivariate response regression with inverse covariance
3. Asymptotic properties
4. Numerical examples

Glioblastoma multiforme (GBM) Cancer Data

- The primary form of brain tumor
- 305 samples, 21694 gene expressions, 535 micro-RNAs, CN, SNP,...

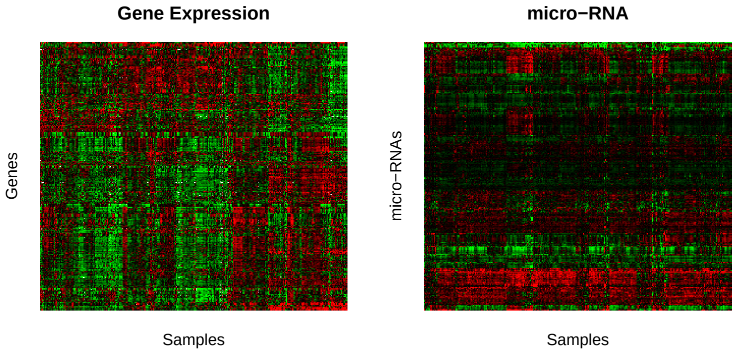


Figure: Heatmaps of gene expression data and micro-RNA data

Glioblastoma multiforme (GBM) Cancer Data

- Goal
  - Regression models
    - micro-RNAs($\mathbf{X}$) $\rightarrow$ Gene expressions($\mathbf{Y}$)
    - micro-RNAs($\mathbf{Y}$) $\leftarrow$ Gene expressions($\mathbf{X}$)
  - Dependence structure in one data set given the other
- Challenge
  - A large number of responses and covariates

Multivariate Response Regression

- Training sample: $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1,...,n}$
- $\mathbf{x}_i \in \mathbf{R}^p$, $\mathbf{y}_i = (y_{i1}, ..., y_{im}) \in \mathbf{R}^m$
- $\mathbf{y}_i | \mathbf{x}_i$ follows a multivariate Gaussian distribution

$$\mathbf{y}_i = \mathbf{B}^T \mathbf{x}_i + \epsilon_i \quad \text{for } i = 1, ..., n,$$

  where $\epsilon \sim \mathbf{N}(\mathbf{0}, \mathbf{\Sigma})$.
- Our goal is to estimate $\mathbf{B}$ and $\mathbf{C} = \mathbf{\Sigma}^{-1}$

Multivariate Response Regression

- $\mathbf{Y}$: $n \times m$ response matrix, $\mathbf{X}$: $n \times p$ predictor matrix.

$$n \log \det(\mathbf{C}) - \operatorname{tr}\left\{(\mathbf{Y} - \mathbf{XB})\mathbf{C}(\mathbf{Y} - \mathbf{XB})^T\right\}$$

- Log-likelihood of $(\mathbf{B}, \mathbf{C})$ given $\mathbf{X}$
- $n > p, m$: maximum likelihood estimator
- $n < p, m$: Penalized approach

Penalized Maximum Likelihood Estimator

- Lee and Liu (2010)

$$\underset{\mathbf{B},\mathbf{C}}{\operatorname{argmin}}[-n \log \det(\mathbf{C}) + \operatorname{tr}\left\{(\mathbf{Y}-\mathbf{XB})\mathbf{C}(\mathbf{Y}-\mathbf{XB})^T\right\} \\ + \lambda_1 \sum_{j,k} w_{jk} \left|\beta_{jk}\right| + \lambda_2 \sum_{s \neq t} v_{st} \left|c_{st}\right|]$$

- Rothman, Levina and Zhu (2010): L_1 penalties, focus on **B**

Is a single Gaussian model reasonable?

- Apply the method to our real data
- Verhaak et al. (2010)

  - Four subtypes of GBM patients

  - based on gene expressions
- A mixture of several Gaussian distributions

Glioblastoma multiforme (GBM) Cancer Data

- Four subtypes: Classical, Mesenchymal, Neural, and Proneural

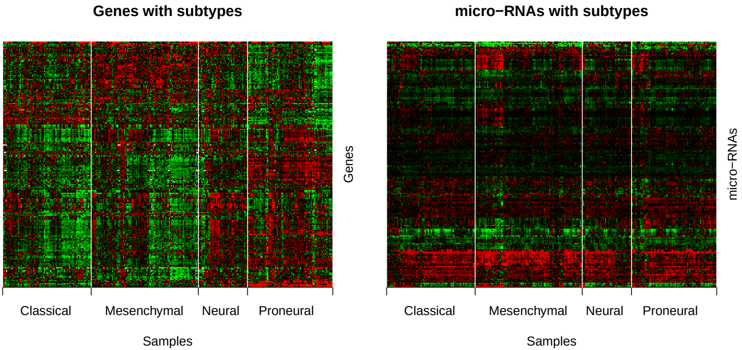


Figure: Heatmaps of gene expression data and micro-RNA data

Mixture of Gaussian Models

- G different groups.

$$\mathbf{y}_i^{(g)} = \mathbf{B}^{(g)\,T}\mathbf{x}_i^{(g)} + \epsilon_i^{(g)} \quad \text{for } i = 1, ..., n_g; g = 1, ..., G.$$

- $\mathbf{B}^{(g)}$ is an unknown $p \times m$ parameter matrix.
- $\epsilon_i^{(g)} \sim \mathbf{N}(0, \mathbf{\Sigma}^{(g)})$
- $\mathbf{C}^{(g)} = (\mathbf{\Sigma}^{(g)})^{-1}$
- Group label g is given.

Mixture of Gaussian Models

- Can model each group separately.
- The groups have shared information with similar structure.
- Model G groups jointly.
- Identify the common and unique structure on $\{\mathbf{B}^{(g)}, \mathbf{C}^{(g)}\}$

Group penalty

- $\boldsymbol{\beta}_{jk} = (\beta_{jk}^{(1)}, ..., \beta_{jk}^{(G)})^T$
- $p(\boldsymbol{\beta}_{jk}) = p(\beta_{jk}^{(1)}, ..., \beta_{jk}^{(G)})$
- Yuan and Lin (2006): Group Lasso

$$p(\boldsymbol{\beta}_{jk}) = ||\boldsymbol{\beta}_{jk}||_2 = \sqrt{{\beta_{jk}^{(1)}}^2 + ... + {\beta_{jk}^{(G)}}^2}$$

- Turlach et al.(2005), Zhang et al.(2008), Zhao et al.(2009)

$$p(\boldsymbol{\beta}_{jk}) = ||\boldsymbol{\beta}_{jk}||_\infty = \max(|\beta_{jk}^{(1)}|, ..., |\beta_{jk}^{(G)}|)$$

- Select variables in an "all-in-all-out" fashion

Group penalty

- No flexibility of selecting variables within a group.
- For a gene expression ($\mathbf{y}$) and a micro-RNA ($\mathbf{x}$),
  - Classical, Mesenchymal: $\mathbf{x} \Rightarrow \mathbf{y}$
  - Neural, Proneural: $\mathbf{x} \nRightarrow \mathbf{y}$
- Need flexibility

Hierarchical Group Penalty

- $\boldsymbol{\beta}_{jk} = (\beta_{jk}^{(1)}, \ldots, \beta_{jk}^{(G)})^T$
- Zhou and Zhu (2010)

$$p(\boldsymbol{\beta}_{jk}) = \sqrt{|\beta_{jk}^{(1)}| + \ldots + |\beta_{jk}^{(G)}|} \approx \sum_{g=1}^{G} \frac{1}{(\sum_{g=1}^{G} |\beta_{jk}^{(g),0}|)^{1/2}} \left|\beta_{jk}^{(g)}\right|,$$

where $\beta_{jk}^{(g),0}$ is close to the solution.

- All coefficients in $\boldsymbol{\beta}_{jk}$ have the same weight.
- Allow sparsity within group.

Penalized MLE with Hierarchical Group Penalty

$$\sum_{g=1}^{G} \left[-n_g \log \det(\mathbf{C}^{(g)}) + \text{tr} \left\{ (\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})\mathbf{C}^{(g)}(\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})^T \right\} \right]$$

- Two groups of matrices to be estimated: $\{\mathbf{B}^{(g)}\}$ and $\{\mathbf{C}^{(g)}\}$
- Two plug-in methods
  1. $\{\hat{\mathbf{B}}^{(g)}\} \to \{\mathbf{C}^{(g)}\}$
  2. $\{\hat{\mathbf{C}}^{(g)}\} \to \{\mathbf{B}^{(g)}\}$
- One joint method : $\{\mathbf{B}^{(g)}, \mathbf{C}^{(g)}\}$ together

Two Plug-in Methods

1 Plug-in Hierarchical LASSO (PHL) estimator

$$\underset{(\mathbf{B}^{(g)})_{g=1}^{G}}{\operatorname{argmin}} \sum_{g=1}^{G} \operatorname{tr}\left\{(\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})\hat{\mathbf{C}}^{(g)}(\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})^T\right\}$$
$$+\lambda_1 \sum_{j,k}\left(\sum_{g=1}^{G}\left|\beta_{jk}^{(g)}\right|\right)^{1/2}.$$

- Need $\hat{\mathbf{C}}^{(g)}$ to plug in.
- $\{\hat{\mathbf{B}}^{(g),0}\}$: initial estimates of $\{\mathbf{B}^{(g)}\}$ (LASSO).

$$\mathbf{S}^{(g)} = \frac{1}{n_g}(\mathbf{Y}^{(g)} - \mathbf{X}\hat{\mathbf{B}}^{(g),0})(\mathbf{Y}^{(g)} - \mathbf{X}\hat{\mathbf{B}}^{(g),0})^T$$

- Estimate $\{\mathbf{C}^{(g)}\}$ using GLASSO with $\{\mathbf{S}^{(g)}\}$

Two Plug-in Methods

2 Plug-in Hierarchical Graphical LASSO (PHGL) estimator

$$\underset{(\mathbf{C}^{(g)})_{g=1}^{G}}{\operatorname{argmin}} \sum_{g=1}^{G} \left\{ -n_g \log\det(\mathbf{C}^{(g)}) + n_g \operatorname{tr}(\mathbf{S}^{(g)}\mathbf{C}^{(g)}) \right\} + \lambda_2 \sum_{s \neq t} \left(\sum_{g=1}^{G} \left| c_{st}^{(g)} \right| \right)^{1/2}.$$

- Need $\hat{\mathbf{B}}^{(g)}$ to plug in.
- $\{\hat{\mathbf{B}}^{(g)}\}$: LASSO.

Doubly Penalized Sparse (DPS) Estimator

$$\underset{(\mathbf{B}^{(g)},\mathbf{C}^{(g)})_{g=1}^{G}}{\operatorname{argmin}} \sum_{g=1}^{G} \left\{ -l_g(\mathbf{B}^{(g)}, \mathbf{C}^{(g)}) + \lambda_1 \sum_{jk} \left(\sum_{g=1}^{G} \left| \beta_{jk}^{(g)} \right| \right)^{1/2} + \lambda_2 \sum_{s \neq t} \left(\sum_{g=1}^{G} \left| c_{st}^{(g)} \right| \right)^{1/2} \right\},$$

where

$$l_g(\mathbf{B}^{(g)}, \mathbf{C}^{(g)}) = n_g \log\det(\mathbf{C}^{(g)}) - \operatorname{tr}\left\{ (\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})\mathbf{C}^{(g)}(\mathbf{Y}^{(g)} - \mathbf{X}^{(g)}\mathbf{B}^{(g)})^T \right\}$$

- The first penalty term : hierarchical sparsity among $\{\mathbf{B}^{(g)}\}$
- The second penalty term : hierarchical sparsity among $\{\mathbf{C}^{(g)}\}$

Asymptotic Properties

- $n \to \infty$
- $\{\mathbf{B}^{*,(g)}\}$ and $\{\mathbf{C}^{*,(g)}\}$: true parameter matrices
- $\boldsymbol{\beta}^* = (\text{Vec}(\mathbf{B}^{*,(1)})^T, ..., \text{Vec}(\mathbf{B}^{*,(G)})^T)^T$
- $\mathbf{c}^* = (\text{Vec}(\mathbf{C}^{*,(1)})^T, ..., \text{Vec}(\mathbf{C}^{*,(G)})^T)^T$
- Assumption

$$\frac{1}{n}\mathbf{X}^{(g)\,T}\mathbf{X}^{(g)} \to A^{(g)} \text{ as } n \to \infty,$$

where $A^{(g)}$ is a positive definite matrix; $g = 1, ..., G$.

Asymptotic Properties of the PHL solution

Theorem

If $\lambda_1 n^{-\frac{1}{2}} \to 0$ *and* $\hat{\mathbf{C}}^{(g)}$ *is a consistent estimator of* $\mathbf{C}^{*,(g)}$,

1. *(Consistency)* $\parallel \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \parallel = O_p(\frac{1}{\sqrt{n}})$
2. *(Sparsity) If* $\lambda_1 n^{-\frac{1}{4}} \to \infty$, $\lim_n P(\hat{\beta}_{jk}^{(g)} = 0) = 1$ *if* $\beta_{jk}^{*,(g)} = 0$.

Asymptotic Properties of the PHGL solution

Theorem

If $\lambda_2 n^{-\frac{1}{2}} \to 0$ *and* $\hat{\mathbf{B}}^{(g)}$ *is a consistent estimator of* $\mathbf{B}^{*,(g)}$,

1. *(Consistency)* $\| \hat{\mathbf{c}} - \mathbf{c}^* \| = O_p(\frac{1}{\sqrt{n}})$
2. *(Sparsity) If* $\lambda_2 n^{-\frac{1}{4}} \to \infty$, $\lim_n P(\hat{c}_{jk}^{(g)} = 0) = 1$ *if* $c_{jk}^{*,(g)} = 0$.

Asymptotic Properties of the DPS solution

Theorem

If $\lambda_1 n^{-\frac{1}{2}} \to 0$ *and* $\lambda_2 n^{-\frac{1}{2}} \to 0$,

1. *(Consistency)*

$$\| (\hat{\boldsymbol{\beta}}^T, \hat{\mathbf{c}}^T)^T - (\boldsymbol{\beta}^{*T}, \mathbf{c}^{*T})^T \| = O_p(\frac{1}{\sqrt{n}}),$$

2. *(Sparsity) If* $\lambda_1 n^{-\frac{1}{4}} \to \infty$, $\lim_n P(\hat{\beta}_{jk}^{(g)} = 0) = 1$ *if* $\beta_{jk}^{*,(g)} = 0$;
3. *(Sparsity) If* $\lambda_2 n^{-\frac{1}{4}} \to \infty$, $\lim_n P(\hat{c}_{jk}^{(g)} = 0) = 1$ *if* $c_{jk}^{*,(g)} = 0$.

Simulated Example

- $G = 3, m = 20, p = 20, n = 40$
- Common structure across groups

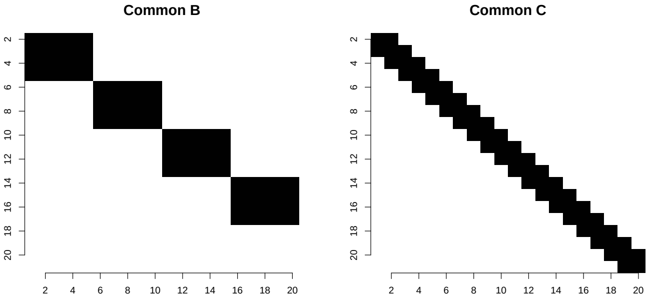


Figure: Black: nonzero parameters, White: zero parameters

- Add unique nonzero parameters to each group.

$$\rho = \frac{\text{number of unique nonzero parameters}}{\text{number of common nonzero parameters}}$$

Simulated Example

- Prediction Error

$$PE = \frac{1}{nmG} \sum_{g=1}^{G} \| \mathbf{Y}^{(g)} - \hat{\mathbf{Y}}^{(g)} \|_F^2$$

- Entropy Loss

$$EL = \frac{1}{G} \sum_{g=1}^{G} \left[\mathrm{tr}((\mathbf{C}^{(g)})^{-1} \hat{\mathbf{C}}^{(g)}) - \log(|(\mathbf{C}^{(g)})^{-1} \hat{\mathbf{C}}^{(g)}|) - m \right]$$

- Frobenius Loss

$$FL = \frac{1}{G} \sum_{g=1}^{G} \| \mathbf{C}^{(g)} - \hat{\mathbf{C}}^{(g)} \|_F^2 \; / \; \| \mathbf{C}^{(g)} \|_F^2$$

Simulated Example

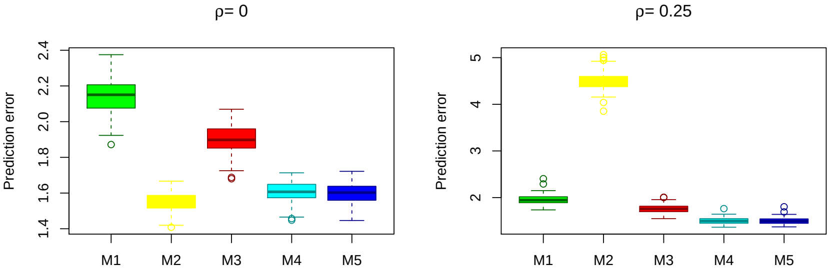


Figure: **M4 and M5 are ours**

- M1: Model each group separately. (penalized MLE with L_1 penalties)
- M2: Combine all groups. (penalized MLE with L_1 penalies)
- M3: Applying LASSO separately to each response in each group
- **M4: Plug-in method with hierarchical penalty for $\{\mathbf{B}^{(g)}\}$**
- **M5: Joint method with two hierarchical penalties.**

Simulated Example

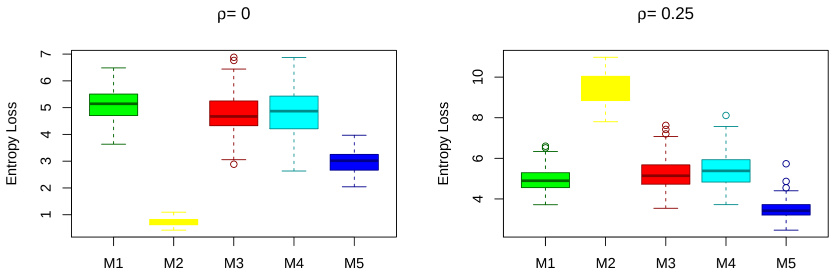


Figure: **M4 and M5 are ours**

- M1: Model each group separately. (penalized MLE with L_1 penalties)
- M2: Combine all groups. (penalized MLE with L_1 penalties)
- M3: Applying GLASSO separately to each group
- **M4: Plug-in method with hierarchical penalty for $\{\mathbf{C}^{(g)}\}$**
- **M5: Joint method with two hierarchical penalties.**

Simulated Example

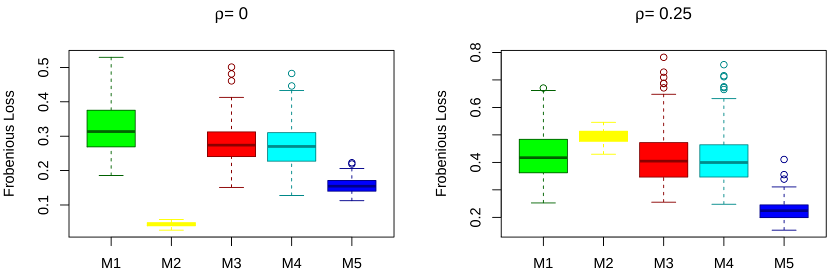
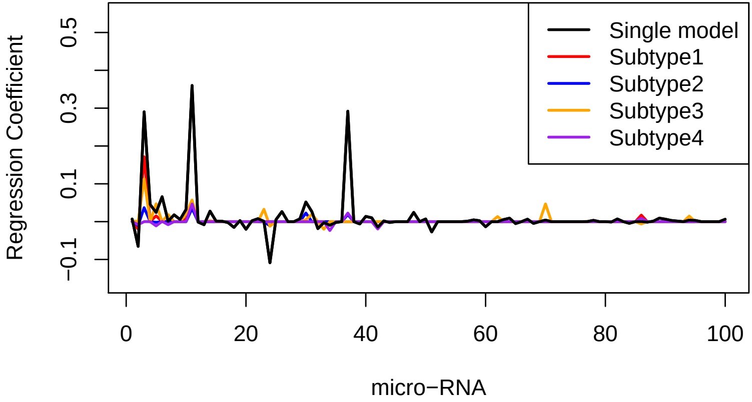


Figure: **M4 and M5 are ours**

- M1: Model each group separately. (penalized MLE with L_1 penalties)
- M2: Combine all groups. (penalized MLE with L_1 penalties)
- M3: Applying GLASSO separately to each group
- **M4: Plug-in method with hierarchical penalty for $\{\mathbf{C}^{(g)}\}$**
- **M5: Joint method with two hierarchical penalties.**

GBM example

100 microRNAs ($\mathbf{X}$) and 20 gene expressions ($\mathbf{Y}$)



GBM example

100 microRNAs (**X**) and 20 gene expressions (**Y**)

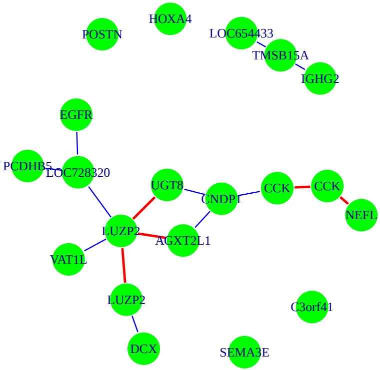
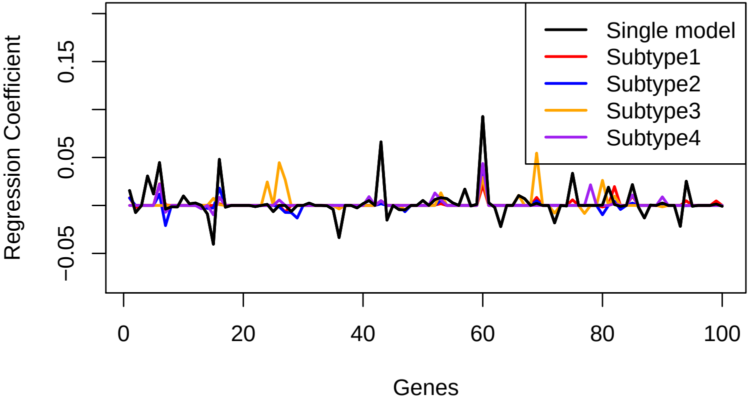


Figure: A graphical model of gene expressions based on $\{\hat{C}^{(g)}\}$

GBM example

20 microRNAs (**Y**) and 100 gene expressions (**X**)



GBM example

20 microRNAs (**Y**) and 100 gene expressions (**X**)

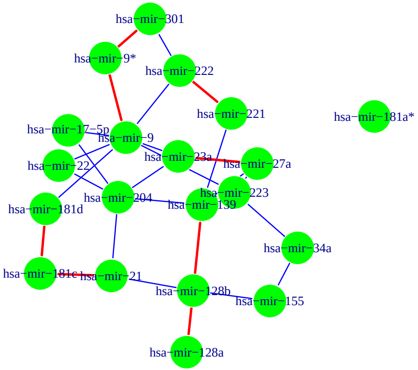


Figure: A graphical model of micro-RNAs based on $\{\hat{C}^{(g)}\}$

Future Work

- Asymptotic properties when $p, m \to \infty$
- Improve computational efficiency
- More comprehensive study on real data (GBM data)

Thank you very much !!