Machine Learning Applications in Genetics and Genomics

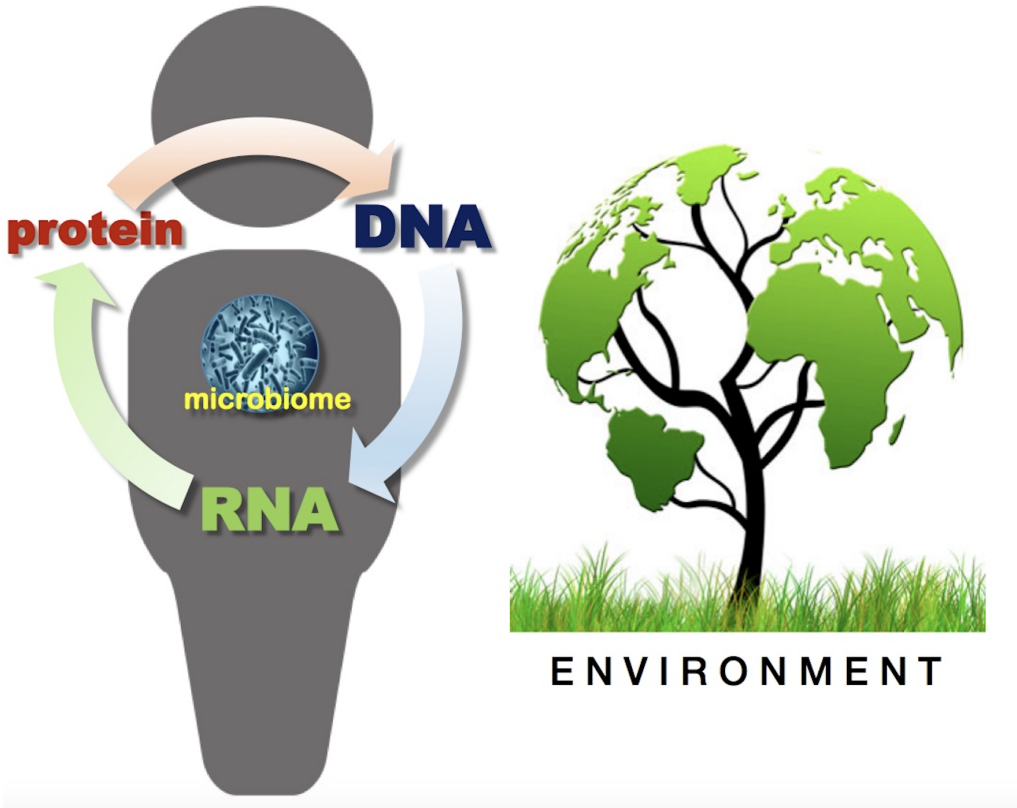
Jing Ma

Public Health Sciences Division
Fred Hutch Cancer Research Center
jingma@fredhutch.org

SLAB LAB
February 5, 2018

Part I:
Improved Variance Component Score Tests of Gene-Environment Interaction

Gene-environment interaction



Gene-environment interaction

$$y_i = \alpha + \beta_1 \underbrace{E_i}_{\text{environment}} + \beta_2' \overbrace{G_i}^{\text{genetic markers}} + \gamma' \underbrace{G_i E_i}_{\text{interaction}} + \varepsilon_i,$$

Gene-environment interaction

$$y_i = \alpha + \beta_1 \underbrace{E_i}_{\text{environment}} + \beta_2' \overbrace{G_i}^{\text{genetic markers}} + \gamma' \underbrace{G_i E_i}_{\text{interaction}} + \varepsilon_i,$$

- y: health outcome (continuous)

Gene-environment interaction

$$y_i = \alpha + \beta_1 \underbrace{E_i}_{\text{environment}} + \boldsymbol{\beta}_2' \overbrace{G_i}^{\text{genetic markers}} + \boldsymbol{\gamma}' \underbrace{G_i E_i}_{\text{interaction}} + \varepsilon_i,$$

- y: health outcome (continuous)
- $\varepsilon_i \sim N(0, \sigma^2)$

Gene-environment interaction

$$y_i = \alpha + \beta_1 \underbrace{E_i}_{\text{environment}} + \beta_2' \overbrace{G_i}^{\text{genetic markers}} + \gamma' \underbrace{G_i E_i}_{\text{interaction}} + \varepsilon_i,$$

- y: health outcome (continuous)
- $\varepsilon_i \sim N(0, \sigma^2)$
- Want to test $H_0 : \gamma = 0$

Connection with linear mixed models

$$y_i = \alpha + \beta_1 E_i + \boldsymbol{\beta}_2' G_i + \underbrace{\boldsymbol{\gamma}'}_{\text{random effects}} G_i E_i + \varepsilon_i,$$

[1] Liu et al. Biometrics. 2007

Connection with linear mixed models

$$y_i = \alpha + \beta_1 E_i + \boldsymbol{\beta}_2' G_i + \underbrace{\boldsymbol{\gamma}'}_{\text{random effects}} G_i E_i + \varepsilon_i,$$

- Exercise: find the connection between lease squares and linear mixed models[1].

[1] Liu et al. Biometrics. 2007

Connection with linear mixed models

$$y_i = \alpha + \beta_1 E_i + \boldsymbol{\beta}_2' G_i + \underbrace{\boldsymbol{\gamma}'}_{\text{random effects}} G_i E_i + \varepsilon_i,$$

- Exercise: find the connection between lease squares and linear mixed models[1].
- Random effects $\gamma_j \sim N(0, \tau)$.

[1] Liu et al. Biometrics. 2007

Connection with linear mixed models

$$y_i = \alpha + \beta_1 E_i + \boldsymbol{\beta}_2' G_i + \underbrace{\boldsymbol{\gamma}'}_{\text{random effects}} G_i E_i + \varepsilon_i,$$

- Exercise: find the connection between lease squares and linear mixed models[1].
- Random effects $\gamma_j \sim N(0, \tau)$.
- No interaction $H_0 : \boldsymbol{\gamma} = 0 \Leftrightarrow H_0 : \tau = 0$.

[1] Liu et al. Biometrics. 2007

Variance component test

The VC score statistic is

$$Q = (y - \hat{y}_0)' K (y - \hat{y}_0) \sim \sum_{i=1}^{n} \lambda_i \chi^2_{1,i}$$

Variance component test

The VC score statistic is

$$Q = (y - \hat{y}_0)' K (y - \hat{y}_0) \sim \sum_{i=1}^{n} \lambda_i \chi^2_{1,i}$$

- $\hat{y}_0$ is the predicted mean of y.

Variance component test

The VC score statistic is

$$Q = (y - \hat{y}_0)' K (y - \hat{y}_0) \sim \sum_{i=1}^{n} \lambda_i \chi^2_{1,i}$$

- $\hat{y}_0$ is the predicted mean of y.
- K depends on G and E.

Variance component test

The VC score statistic is

$$Q = (y - \hat{y}_0)' K (y - \hat{y}_0) \sim \sum_{i=1}^{n} \lambda_i \chi^2_{1,i}$$

- $\hat{y}_0$ is the predicted mean of y.
- K depends on G and E.
- $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\sigma^2 P_0 K P_0$.

Variance component test

The VC score statistic is

$$Q = (y - \hat{y}_0)' K (y - \hat{y}_0) \sim \sum_{i=1}^{n} \lambda_i \chi^2_{1,i}$$

- $\hat{y}_0$ is the predicted mean of y.
- K depends on G and E.
- $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\sigma^2 P_0 K P_0$.
- $P_0 = \mathbf{I}_n - H$ where H is the hat (projection) matrix.

Estimation

Null model:

$$\begin{aligned} y &= \alpha + \beta_1 E + G\boldsymbol{\beta}_2 + \varepsilon \\ &= X\boldsymbol{\beta} + \varepsilon \end{aligned}$$

- $\boldsymbol{\beta}$ can be estimated via OLS or ridge regression.

Estimation

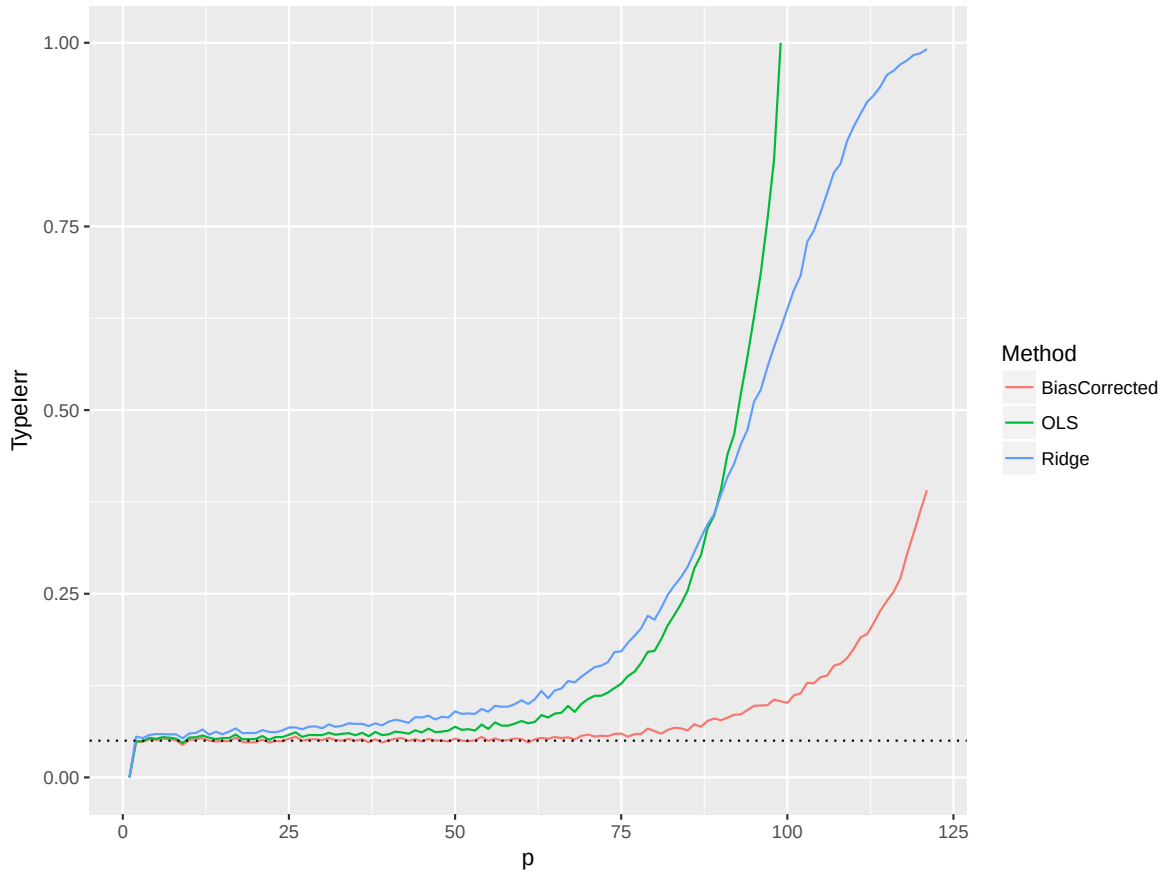
Null model:

$$\begin{aligned} y &= \alpha + \beta_1 E + G\boldsymbol{\beta}_2 + \varepsilon \\ &= X\boldsymbol{\beta} + \varepsilon \end{aligned}$$

- $\boldsymbol{\beta}$ can be estimated via OLS or ridge regression.
- σ^2 can be estimated from RSS.

Type I error - simulation

- $n = 100, \beta_1 = 1, \beta_{2,1} = 1, \sigma^2 = 1.$



Why OLS and Ridge fail

- OLS does not do well when p is large.

[2] Lin et al. Biometrics. 2016

Why OLS and Ridge fail

- OLS does not do well when p is large.
- The ridge estimator[2] of β is biased!

$$\hat{\beta} = \arg\min \left\{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\}$$

[2] Lin et al. Biometrics. 2016

Why OLS and Ridge fail

- OLS does not do well when p is large.
- The ridge estimator[2] of β is biased!

$$\hat{\beta} = \arg\min \left\{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\}$$

- The bias is not critical when all signals are weak, but causes inflated type I errors when some signals are strong.

[2] Lin et al. Biometrics. 2016

Ridge estimator is biased

The ridge regression

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \arg\min \left\{ \|y - X\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\} \\ &= \left(X'X + \lambda \mathbf{I}_p\right)^{-1} X'y\end{aligned}$$

Ridge estimator is biased

The ridge regression

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \arg\min \left\{ \|y - X\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\} \\ &= \left(X'X + \lambda \mathbf{I}_p\right)^{-1} X'y\end{aligned}$$

Let $P_X = X'(XX')^{-}X$ and $\boldsymbol{\theta}^0 = P_X \boldsymbol{\beta}^0$. The bias in $\hat{\boldsymbol{\beta}}_j$ is

$$\underbrace{\mathbb{E}[\hat{\boldsymbol{\beta}}_j] - \boldsymbol{\theta}_j^0}_{\text{estimation bias}} + \underbrace{\boldsymbol{\theta}_j^0 - \boldsymbol{\beta}_j^0}_{\text{projection bias}}$$

Bias corrected ridge estimator

- The estimation bias is dominated by the choice of λ.

[3] Buhlmann. Bernoulli. 2013

Bias corrected ridge estimator

- The estimation bias is dominated by the choice of λ.
- The projection bias

$$B_j = (P_X \beta^0)_j - \beta_j^0 = (P_X)_{jj}\beta_j^0 - \beta_j^0 + \sum_{k \neq j} (P_X)_{jk}\beta_k^0.$$

[3] Buhlmann. Bernoulli. 2013

Bias corrected ridge estimator

- The estimation bias is dominated by the choice of λ.
- The projection bias

$$B_j = (P_X \beta^0)_j - \beta_j^0 = (P_X)_{jj} \beta_j^0 - \beta_j^0 + \sum_{k \neq j} (P_X)_{jk} \beta_k^0.$$

- The bias corrected ridge estimator[3] is

$$\hat{\beta}_j^{corr} = \hat{\beta}_j - \sum_{k \neq j} (P_X)_{jk} \hat{\beta}_k^{init}.$$

[3] Buhlmann. Bernoulli. 2013

Bias corrected ridge estimator

- The estimation bias is dominated by the choice of λ.
- The projection bias

$$B_j = (P_X \beta^0)_j - \beta_j^0 = (P_X)_{jj} \beta_j^0 - \beta_j^0 + \sum_{k \neq j} (P_X)_{jk} \beta_k^0.$$

- The bias corrected ridge estimator[3] is

$$\hat{\beta}_j^{corr} = \hat{\beta}_j - \sum_{k \neq j} (P_X)_{jk} \hat{\beta}_k^{init}.$$

- $\hat{\beta}^{init}$ can be from (scaled) lasso.

[3] Buhlmann. Bernoulli. 2013

Property of the bias corrected estimator

$$\hat{\beta}_j^{corr} = Z_j + \delta_j, \quad j = 1, \ldots, p$$

- Z_j's are normal.

Property of the bias corrected estimator

$$\hat{\beta}_j^{corr} = Z_j + \delta_j, \quad j = 1, \ldots, p$$

- Z_j's are normal.
- $\delta_j = (P_X)_{jj}\beta_j^0 - \sum_{k \neq j}(P_X)_{jk}(\hat{\beta}_k^{init} - \beta_k^0) + \mathbb{E}[\hat{\beta}_j] - \theta_j^0$.

Property of the bias corrected estimator

$$\hat{\beta}_j^{corr} = Z_j + \delta_j, \quad j = 1, \ldots, p$$

- Z_j's are normal.
- $\delta_j = (P_X)_{jj}\beta_j^0 - \sum_{k \neq j}(P_X)_{jk}(\hat{\beta}_k^{init} - \beta_k^0) + \mathbb{E}[\hat{\beta}_j] - \theta_j^0$.
- $\hat{\beta}^{corr}$ is easy to compute.

The matrix P_0

- $P_0 = \mathbf{I}_n - X(X'X + \lambda \mathbf{I}_p)^{-1}X'$ when the estimator is $\hat{\boldsymbol{\beta}}^R$.

The matrix P_0

- $P_0 = \mathbf{I}_n - X(X'X + \lambda \mathbf{I}_p)^{-1}X'$ when the estimator is $\hat{\beta}^R$.
- What is P_0 if using $\hat{\beta}^{corr}$?

$$\begin{aligned}
\hat{\beta}^{corr} =& \hat{\beta}^R - \{P_X - \text{diag}(P_X)\}\hat{\beta}^{init} \\
=& \hat{\beta}^R - \{P_X - \text{diag}(P_X)\}\text{diag}(W)\hat{\beta}^R \\
=& \left[\mathbf{I}_p - \{P_X - \text{diag}(P_X)\}\text{diag}(W)\right]\hat{\beta}^R
\end{aligned}$$

The matrix P_0

- $P_0 = \mathbf{I}_n - X(X'X + \lambda \mathbf{I}_p)^{-1}X'$ when the estimator is $\hat{\boldsymbol{\beta}}^R$.
- What is P_0 if using $\hat{\boldsymbol{\beta}}^{corr}$?

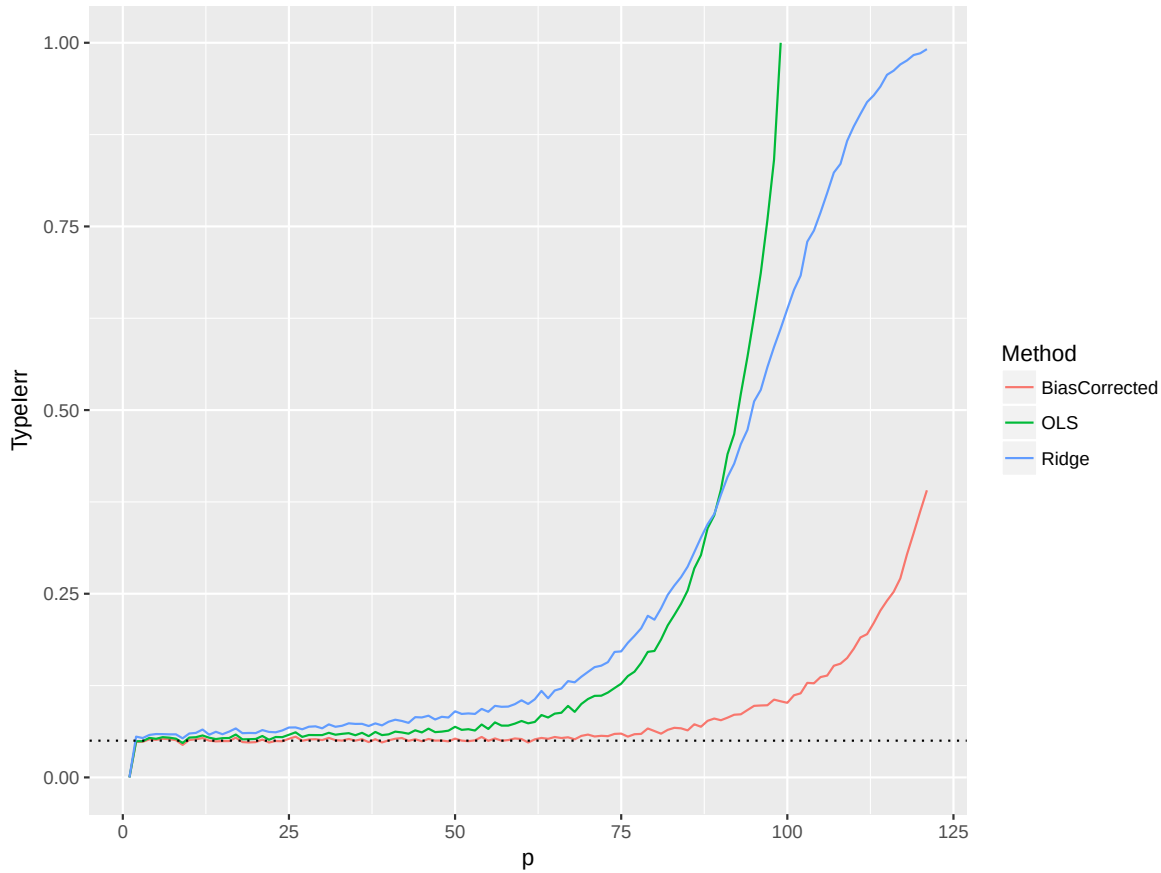
$$\begin{aligned}\hat{\boldsymbol{\beta}}^{corr} =& \hat{\boldsymbol{\beta}}^R - \{P_X - \text{diag}(P_X)\}\hat{\boldsymbol{\beta}}^{init} \\ =& \hat{\boldsymbol{\beta}}^R - \{P_X - \text{diag}(P_X)\}\text{diag}(W)\hat{\boldsymbol{\beta}}^R \\ =& \left[\mathbf{I}_p - \{P_X - \text{diag}(P_X)\}\text{diag}(W)\right]\hat{\boldsymbol{\beta}}^R\end{aligned}$$

- New P_0 with $\hat{\boldsymbol{\beta}}^{corr}$ is

$$\mathbf{I}_n - X\left[\mathbf{I}_p - \{P_X - \text{diag}(P_X)\}\text{diag}(W)\right](X'X + \lambda \mathbf{I}_p)^{-1}X'.$$

Type I error - simulation

- $n = 100, \beta_1 = 1, \beta_{2,1} = 1, \sigma^2 = 1.$



Open questions

$$Q = (y - \hat{y}_0)' K (y - \hat{y}_0) \sim \sum_{i=1}^{n} \lambda_i \chi^2_{1,i}$$

- Other bias correction procedures that allow valid type I error control for $p \sim n$, and possibly $p > n$.

Open questions

$$Q = (y - \hat{y}_0)' K (y - \hat{y}_0) \sim \sum_{i=1}^{n} \lambda_i \chi^2_{1,i}$$

- Other bias correction procedures that allow valid type I error control for $p \sim n$, and possibly $p > n$.
  - Improved fit of the residual $y - \hat{y}_0$.

Open questions

$$Q = (y - \hat{y}_0)' K (y - \hat{y}_0) \sim \sum_{i=1}^{n} \lambda_i \chi^2_{1,i}$$

- Other bias correction procedures that allow valid type I error control for $p \sim n$, and possibly $p > n$.
  - Improved fit of the residual $y - \hat{y}_0$.
  - Or even better, estimate directly $X'(y - \hat{y}_0)$.

Open questions

$$Q = (y - \hat{y}_0)' K (y - \hat{y}_0) \sim \sum_{i=1}^{n} \lambda_i \chi^2_{1,i}$$

- Other bias correction procedures that allow valid type I error control for $p \sim n$, and possibly $p > n$.
  - Improved fit of the residual $y - \hat{y}_0$.
  - Or even better, estimate directly $X'(y - \hat{y}_0)$.
- What if y is binary?

Open questions

$$Q = (y - \hat{y}_0)' K (y - \hat{y}_0) \sim \sum_{i=1}^{n} \lambda_i \chi^2_{1,i}$$

- Other bias correction procedures that allow valid type I error control for $p \sim n$, and possibly $p > n$.
  - Improved fit of the residual $y - \hat{y}_0$.
  - Or even better, estimate directly $X'(y - \hat{y}_0)$.
- What if y is binary?
- Kernel machine based interaction testing.

Open questions

$$Q = (y - \hat{y}_0)' K (y - \hat{y}_0) \sim \sum_{i=1}^{n} \lambda_i \chi^2_{1,i}$$

- Other bias correction procedures that allow valid type I error control for $p \sim n$, and possibly $p > n$.
  - Improved fit of the residual $y - \hat{y}_0$.
  - Or even better, estimate directly $X'(y - \hat{y}_0)$.
- What if y is binary?
- Kernel machine based interaction testing.
- Beyond VC score test.

Another look at the model

Suppose E_i is binary (e.g. smoker / non-smoker).

$$
\begin{aligned}
y_i &= \alpha + \beta_1 E_i + \boldsymbol{\beta}_2' G_i + \boldsymbol{\gamma}' G_i E_i + \varepsilon_i, \\
&= \alpha + \beta_1 E_i + (\boldsymbol{\beta}_2 + \boldsymbol{\gamma} E_i)' G_i \\
&= \begin{cases} \alpha + \boldsymbol{\beta}_2' G_i, & \text{if } E_i = 0 \\ \alpha + \beta_1 + (\boldsymbol{\beta}_2 + \boldsymbol{\gamma})' G_i, & \text{if } E_i = 1. \end{cases}
\end{aligned}
$$

Two sample inference of HD linear regressions

$$y = \alpha_d + X\beta^{(d)} + \varepsilon_d, \quad d = 1, 2$$

Two sample inference of HD linear regressions

$$y = \alpha_d + X\beta^{(d)} + \varepsilon_d, \quad d = 1, 2$$

- $H_{0,j} : \beta_j^{(1)} = \beta_j^{(2)}, j = 1, \ldots, p.$

Two sample inference of HD linear regressions

$$y = \alpha_d + X\beta^{(d)} + \varepsilon_d, \quad d = 1, 2$$

- $H_{0,j} : \beta_j^{(1)} = \beta_j^{(2)}, j = 1, \ldots, p.$
- Inverse regression X_j on (y, X_{-j}) (Xia et al. Stat. Sin. 2017)

Two sample inference of HD linear regressions

$$y = \alpha_d + X\beta^{(d)} + \varepsilon_d, \quad d = 1, 2$$

- $H_{0,j} : \beta_j^{(1)} = \beta_j^{(2)}, j = 1, \ldots, p.$
- Inverse regression X_j on (y, X_{-j}) (Xia et al. Stat. Sin. 2017)
- Alternative: direct comparison of $\beta_j^{(1)}$ and $\beta_j^{(2)}$?

From linear to logistic regressions

- First approach:

$$\text{logit}P(y_i = 1) = \alpha_d + X_i' \beta^{(d)}, \quad d = 1, 2$$

From linear to logistic regressions

- First approach:

$$\text{logit}P(y_i = 1) = \alpha_d + X_i' \beta^{(d)}, \quad d = 1, 2$$

- Second approach:

$$y = \dot{g}(\alpha_d + X\beta^{(d)}) + \varepsilon_d, \quad d = 1, 2$$

where $g(u) = \log(e^u + e^{-u})$ and ε_d is sub-Gaussian.

Part II:

CHIME: Clustering of HIgh-dimensional Gaussian Mixtures with EM

Clustering

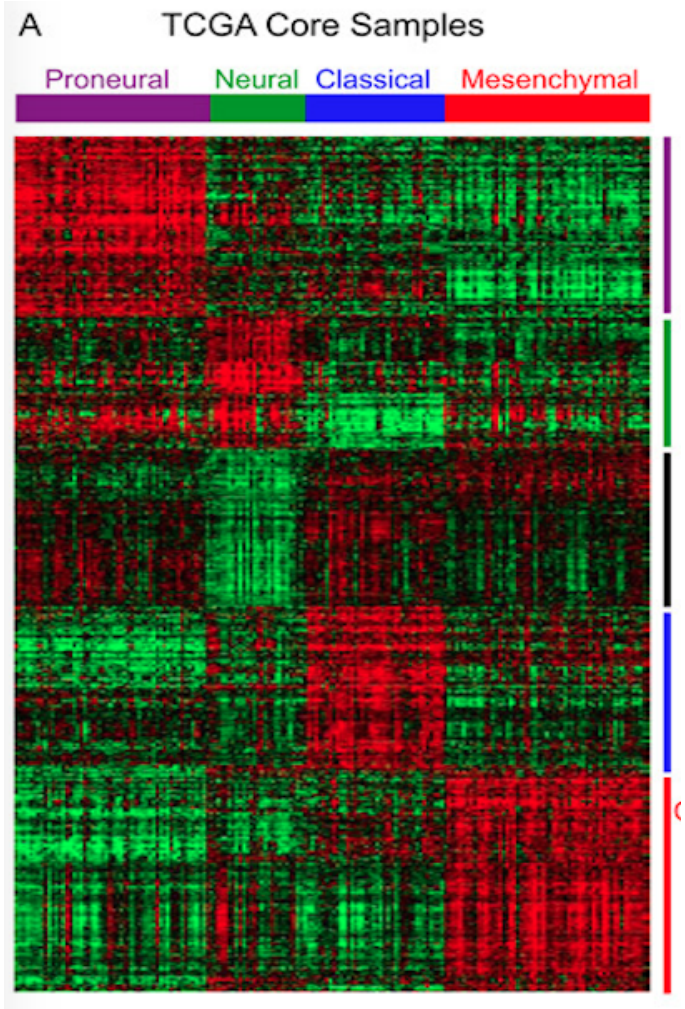


Fig: Verhaak et al. Cancer Cell, 2010

Clustering - existing algorithms

- K-means / K-median

Clustering - existing algorithms

- K-means / K-median
- Hierarchical clustering

Clustering - existing algorithms

- K-means / K-median
- Hierarchical clustering
- Expectation-Maximization (EM) algorithm

Clustering - existing algorithms

- K-means / K-median
- Hierarchical clustering
- Expectation-Maximization (EM) algorithm
- ...

Clustering - existing algorithms

- K-means / K-median
- Hierarchical clustering
- Expectation-Maximization (EM) algorithm
- ...

However, theoretical performance of the clustering algorithm is not fully understood.

Gaussian mixture model

- General form (2-class):

$$y^{(1)}, \ldots y^{(n)} \text{ i.i.d } \sim \begin{cases} 1, & \text{with probability } 1-\omega; \\ 2, & \text{with probability } \omega. \end{cases}$$

$$\boldsymbol{z}^{(i)} \mid y^{(i)} = d \; i.i.d. \sim N_p(\boldsymbol{\mu}_d, \Sigma); \quad d = 1, 2.$$

Gaussian mixture model

- General form (2-class):

$$y^{(1)}, \ldots y^{(n)} \text{ i.i.d } \sim \begin{cases} 1, & \text{with probability } 1-\omega; \\ 2, & \text{with probability } \omega. \end{cases}$$

$$\boldsymbol{z}^{(i)} \mid y^{(i)} = d \; i.i.d. \sim N_p(\boldsymbol{\mu}_d, \Sigma); \quad d = 1, 2.$$

- Observations: $\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, \ldots, \boldsymbol{z}^{(n)}$.

Gaussian mixture model

- General form (2-class):

$$y^{(1)}, \ldots y^{(n)} \text{ i.i.d } \sim \begin{cases} 1, & \text{with probability } 1-\omega; \\ 2, & \text{with probability } \omega. \end{cases}$$

$$\boldsymbol{z}^{(i)} \mid y^{(i)} = d \; i.i.d. \sim N_p(\boldsymbol{\mu}_d, \Sigma); \quad d = 1, 2.$$

- Observations: $\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, \ldots, \boldsymbol{z}^{(n)}$.
- Goal: Cluster $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(n)}$ into two groups with statistical guarantees.

Gaussian mixture model

- When p is small, we solve for MLE to maximize

$$L(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log\left\{ f(\boldsymbol{z}^{(i)}|\boldsymbol{\mu}_1, \Sigma) P(y^{(i)} = 1) + f(\boldsymbol{z}^{(i)}|\boldsymbol{\mu}_2, \Sigma) P(y^{(i)} = 2) \right\}.$$

[4] Dempster et al. JRSSB. 1977

Gaussian mixture model

- When p is small, we solve for MLE to maximize

$$L(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log\left\{ f(\boldsymbol{z}^{(i)}|\boldsymbol{\mu}_1, \Sigma) P(y^{(i)} = 1) + f(\boldsymbol{z}^{(i)}|\boldsymbol{\mu}_2, \Sigma) P(y^{(i)} = 2) \right\}.$$

- Drawbacks: $L(\boldsymbol{\theta})$ is not convex; MLE is challenging for large p.

[4] Dempster et al. JRSSB. 1977

Gaussian mixture model

- When p is small, we solve for MLE to maximize

$$L(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log\left\{ f(\boldsymbol{z}^{(i)}|\boldsymbol{\mu}_1, \Sigma) P(y^{(i)} = 1) + f(\boldsymbol{z}^{(i)}|\boldsymbol{\mu}_2, \Sigma) P(y^{(i)} = 2) \right\}.$$

- Drawbacks: $L(\boldsymbol{\theta})$ is not convex; MLE is challenging for large p.
- Solution: Expectation-Maximization (EM) algorithm[4].

[4] Dempster et al. JRSSB. 1977

Linear discriminant analysis

- Suppose we know the true parameters ω, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and Σ.

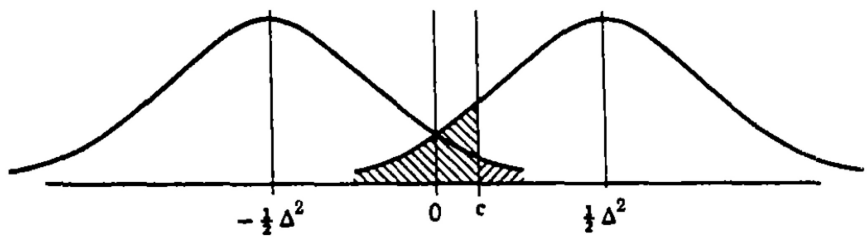


Fig: Mis-classification error of LDA[5]

[5] $\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is the SNR.

Linear discriminant analysis

- Suppose we know the true parameters ω, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and Σ.
- The discriminating direction $\boldsymbol{\beta} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

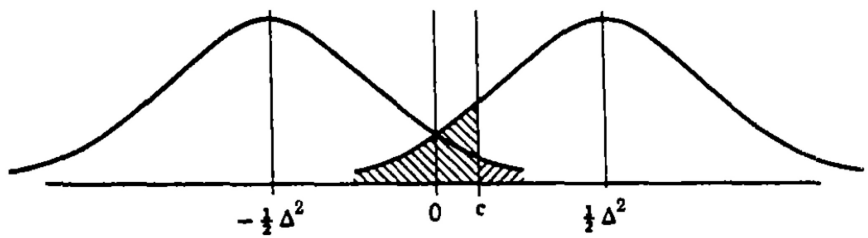


Fig: Mis-classification error of LDA[5]

[5] $\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is the SNR.

Linear discriminant analysis

- Suppose we know the true parameters $\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and Σ.
- The discriminating direction $\boldsymbol{\beta} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.
- Then the optimal classification rule

$$C_{opt}(\boldsymbol{z}) = \begin{cases} 1, & \{\boldsymbol{z} - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2\}'\boldsymbol{\beta} \geq \log(\frac{\omega}{1-\omega}) \\ 2, & \{\boldsymbol{z} - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2\}'\boldsymbol{\beta} < \log(\frac{\omega}{1-\omega}). \end{cases}$$

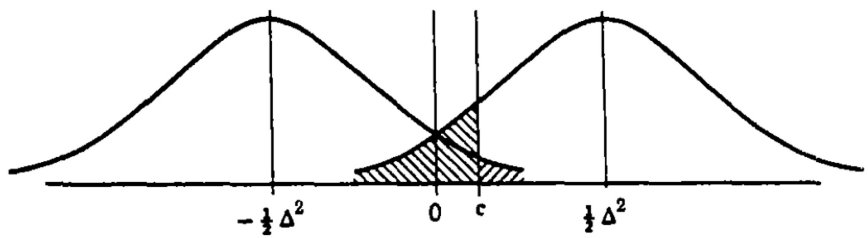


Fig: Mis-classification error of LDA[5]

[5] $\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is the SNR.

Linear programming discriminant

- If we know the sample labels $y^{(1)}, \ldots y^{(n)}$, estimate $\boldsymbol{\mu}_d$ by

$$\hat{\boldsymbol{\mu}}_d = \frac{1}{n_d} \sum_{i=1}^{n} \boldsymbol{z}^{(i)} I(y^{(i)} = d), \quad d = 1, 2,$$

and

$$\hat{\Sigma} = \frac{1}{n}(n_1 \hat{\Sigma}_1 + n_2 \hat{\Sigma}_2).$$

[6] Cai and Liu. JASA. 2011

Linear programming discriminant

- If we know the sample labels $y^{(1)}, \ldots y^{(n)}$, estimate $\boldsymbol{\mu}_d$ by

$$\hat{\boldsymbol{\mu}}_d = \frac{1}{n_d} \sum_{i=1}^{n} \boldsymbol{z}^{(i)} I(y^{(i)} = d), \quad d = 1, 2,$$

  and

$$\hat{\Sigma} = \frac{1}{n}(n_1 \hat{\Sigma}_1 + n_2 \hat{\Sigma}_2).$$

- Assuming sparse $\boldsymbol{\beta}$, apply the LPD[6] to get

$$\hat{\boldsymbol{\beta}} = \arg\min\{\|\boldsymbol{\beta}\|_1 : \|\hat{\Sigma}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)\|_\infty \leq \lambda_n\}.$$

[6] Cai and Liu. JASA. 2011

The EM algorithm

- We combine the above ideas to iteratively estimate $\boldsymbol{\theta} = (\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\beta})$.

The EM algorithm

- We combine the above ideas to iteratively estimate $\boldsymbol{\theta} = (\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\beta})$.
- The conditional log-likelihood

$$
\begin{aligned}
Q_n(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}) &= \mathbb{E}_n[\log L(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}, \boldsymbol{z})] \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{d=1}^{2} P(y^{(i)} = d \mid \tilde{\boldsymbol{\theta}}) \log f(\boldsymbol{z}^{(i)} | \boldsymbol{\mu}_d, \Sigma)
\end{aligned}
$$

The EM algorithm

- Initialization $\boldsymbol{\theta}^{(0)} = \{\omega^{(0)}, \boldsymbol{\mu}_d^{(0)}, \boldsymbol{\beta}^{(0)}\}$; $\kappa \in [1/2, 3/4]$; $\lambda_n^{(0)}$; T_{stop}.

The EM algorithm

- Initialization $\boldsymbol{\theta}^{(0)} = \{\omega^{(0)}, \boldsymbol{\mu}_d^{(0)}, \boldsymbol{\beta}^{(0)}\}$; $\kappa \in [1/2, 3/4]$; $\lambda_n^{(0)}$; T_{stop}.
- E-step: Evaluate $Q_n(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$.

The EM algorithm

- Initialization $\boldsymbol{\theta}^{(0)} = \{\omega^{(0)}, \boldsymbol{\mu}_d^{(0)}, \boldsymbol{\beta}^{(0)}\}$; $\kappa \in [1/2, 3/4]$; $\lambda_n^{(0)}$; T_{stop}.
- E-step: Evaluate $Q_n(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$.
- M-step:

$$(\omega^{(t+1)}, \boldsymbol{\mu}_d^{(t+1)}, \Sigma^{(t+1)}) = \arg\max Q_n(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}),$$

$$\lambda_n^{(t+1)} = \kappa\lambda_n^{(t)} + C\sqrt{\log p/n},$$

$$\boldsymbol{\beta}^{(t+1)} = \arg\min\{\|\boldsymbol{\beta}\|_1 : \|\Sigma^{(t+1)}\boldsymbol{\beta} - (\boldsymbol{\mu}_1^{(t+1)} - \boldsymbol{\mu}_2^{(t+1)})\|_\infty \leq \lambda_n^{(t+1)}\}.$$

The EM algorithm

- Initialization $\boldsymbol{\theta}^{(0)} = \{\omega^{(0)}, \boldsymbol{\mu}_d^{(0)}, \boldsymbol{\beta}^{(0)}\}$; $\kappa \in [1/2, 3/4]$; $\lambda_n^{(0)}$; T_{stop}.
- E-step: Evaluate $Q_n(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$.
- M-step:

$$(\omega^{(t+1)}, \boldsymbol{\mu}_d^{(t+1)}, \Sigma^{(t+1)}) = \arg\max Q_n(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}),$$
$$\lambda_n^{(t+1)} = \kappa\lambda_n^{(t)} + C\sqrt{\log p/n},$$
$$\boldsymbol{\beta}^{(t+1)} = \arg\min\{\|\boldsymbol{\beta}\|_1 : \|\Sigma^{(t+1)}\boldsymbol{\beta} - (\boldsymbol{\mu}_1^{(t+1)} - \boldsymbol{\mu}_2^{(t+1)})\|_\infty \leq \lambda_n^{(t+1)}\}.$$

- Upon convergence, output $\hat{\omega}, \hat{\boldsymbol{\mu}}_d, \hat{\boldsymbol{\beta}} \leftarrow \omega^{(T_{stop})}, \boldsymbol{\mu}_d^{(T_{stop})}, \boldsymbol{\beta}^{(T_{stop})}$.

Upper bound

Theorem (Cai, M, Zhang. 2018)

Assume $\|\boldsymbol{\beta}\|_0 \leq s$. Under certain technical conditions, the output $\boldsymbol{\beta}^{(T_{stop})}$ satisfies with high probability

$$\|\boldsymbol{\beta}^{(T_{stop})} - \boldsymbol{\beta}\|_2 \lesssim \kappa^{T_{stop}} \|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}\|_2 + \sqrt{\frac{s \log p}{n}}.$$

Consequently, if $T_{stop} \gtrsim \log n$, then

$$\|\boldsymbol{\beta}^{(T_{stop})} - \boldsymbol{\beta}\|_2 \lesssim \sqrt{\frac{s \log p}{n}}.$$

Remarks

The results in Wang et al. ('15) ($\Sigma = \sigma^2 \mathbf{I}_p$) show

$$\|\boldsymbol{\beta}^{(T_{stop})} - \boldsymbol{\beta}\|_2 \lesssim \sqrt{\frac{s \log p \cdot \log n}{n}}.$$

The proposed classifier

- Given the estimated $\hat{\omega}, \hat{\boldsymbol{\mu}}_d, \hat{\boldsymbol{\beta}}$, the sample $\boldsymbol{z}$ is classified as

$$\hat{C}(\boldsymbol{z}) = \begin{cases} 1, & \{\boldsymbol{z} - (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2\}'\hat{\boldsymbol{\beta}} \geq \log(\frac{\hat{\omega}}{1-\hat{\omega}}) \\ 2, & \{\boldsymbol{z} - (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2\}'\hat{\boldsymbol{\beta}} < \log(\frac{\hat{\omega}}{1-\hat{\omega}}). \end{cases}$$

The proposed classifier

- Given the estimated $\hat{\omega}, \hat{\boldsymbol{\mu}}_d, \hat{\boldsymbol{\beta}}$, the sample $\boldsymbol{z}$ is classified as

$$\hat{C}(\boldsymbol{z}) = \begin{cases} 1, & \{\boldsymbol{z} - (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2\}'\hat{\boldsymbol{\beta}} \geq \log(\frac{\hat{\omega}}{1-\hat{\omega}}) \\ 2, & \{\boldsymbol{z} - (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2\}'\hat{\boldsymbol{\beta}} < \log(\frac{\hat{\omega}}{1-\hat{\omega}}). \end{cases}$$

- The mis-clustering error is defined as

$$R(\hat{C}) = \min_{\pi \in \mathbb{P}_2} \mathbb{E}[I(\hat{C}(\boldsymbol{z}) \neq \pi(y))],$$

where $\mathbb{P}_2 = \{\pi : [1, 2] \to [1, 2]\}$ is a set of permutation function.

Mis-clustering error

Theorem (Cai, M, Zhang. 2018)

Under the same conditions of Theorem 1 and with $T_{stop} \gtrsim \log n$*, the classifier* $\hat{C}$ *with mis-clustering error* $R(\hat{C})$*, satisfies w.h.p.*

$$R(\hat{C}) - R_{opt} \lesssim \frac{s \log p}{n}.$$

Simulation

Competitors

- KM: k-means

Simulation

Competitors

- KM: k-means
- SKM: sparse k-means (Witten and Tibshirani '12)

Simulation

Competitors

- KM: k-means
- SKM: sparse k-means (Witten and Tibshirani '12)
- SHP: sparse clustering with HARDT-PRICE (Azizyan et al. '14)

Simulation

Competitors

- KM: k-means
- SKM: sparse k-means (Witten and Tibshirani '12)
- SHP: sparse clustering with HARDT-PRICE (Azizyan et al. '14)
- PCCM: penalized clustering with common covariances (Zhou et al. '09)

Simulation

Competitors

- KM: k-means
- SKM: sparse k-means (Witten and Tibshirani '12)
- SHP: sparse clustering with HARDT-PRICE (Azizyan et al. '14)
- PCCM: penalized clustering with common covariances (Zhou et al. '09)

Benchmark

- LPD: supervised linear program discriminant rule (Cai and Liu '11)

Simulation

Competitors

- KM: k-means
- SKM: sparse k-means (Witten and Tibshirani '12)
- SHP: sparse clustering with HARDT-PRICE (Azizyan et al. '14)
- PCCM: penalized clustering with common covariances (Zhou et al. '09)

Benchmark

- LPD: supervised linear program discriminant rule (Cai and Liu '11)
- Oracle: Fisher's LDA with true parameters

Simulation

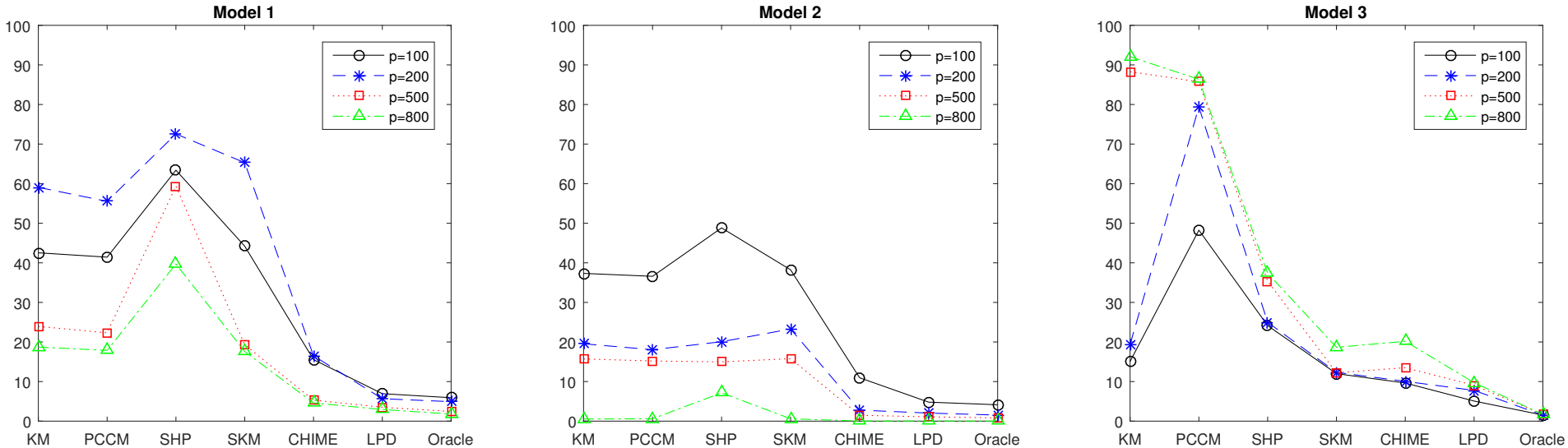


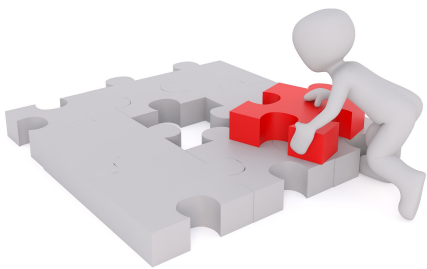
Fig: Clustering errors based on $n = 200$ test samples and 100 replications.

Application to clustering GBM data

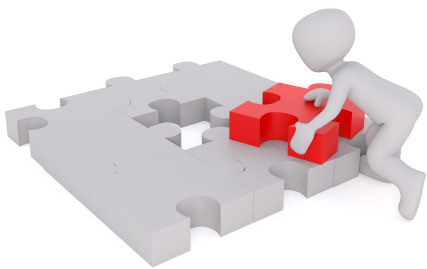
	CHIME		KM		PCCM		SHP		SKM	
Class	1	2	1	2	1	2	1	2	1	2
Neural	26	0	26	0	26	0	12	14	25	1
Mesenchymal	2	54	7	49	5	51	10	46	6	50

Table: Clustering results for the GBM gene expression data with $p = 200$ genes and 82 samples

Summary

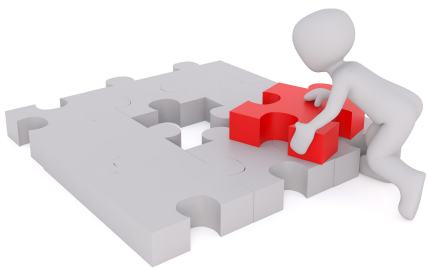


Summary



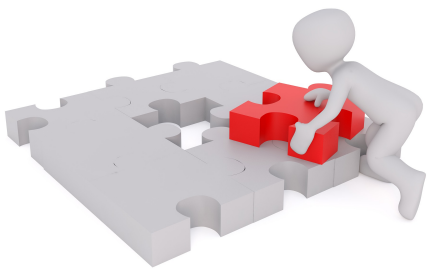
- Knowing labels doesn't improve the convergence rate of estimation and classification.

Summary



- Knowing labels doesn't improve the convergence rate of estimation and classification.
- Not covered in this talk
  - Lower bound of estimation and clustering error is in the same order of the respective upper bound.

Summary



- Knowing labels doesn't improve the convergence rate of estimation and classification.
- Not covered in this talk
  - Lower bound of estimation and clustering error is in the same order of the respective upper bound.
  - Extensions to multi-class GMM and/or unequal covariance matrices are available.

Collaborators

- Nanxun Ma (UW)
- Michael Wu (Hutch)
- Linjun Zhang (U Penn)
- Tony Cai (U Penn)