# Semi-supervised Image Classification in Likelihood Space

*Rong Duan, Wei Jiang, Hong Man*
*Stevens Institute of Technology*

# Introduction

- Semi-supervised learning
- Model Mis-specification in classification
- Log-likelihood space classification

# Terms

$\mathbf{D_k}$ Data sample $\mathbf{D_k}$$=\{X_1^{(k)}, \cdots, Xm^{(k)}\}$ ,

$\mathbf{Q}$ Training data: $\mathbf{Q} = \{Q_{label}, Q_{unlabel}\}$,

$\mathbf{Q_{label}}$ Labeled training data $\mathbf{Q_{label}}$ $=\{(D_1,1),(D_2,2)\}$,

$\mathbf{Q_{unlabel}}$ Unlabeled training data $\mathbf{Q_{unlabel}}$ $= \{(D_1,1),(D_2,2)\}$

$\mathbf{g_k(x)}$ True distributions $\mathbf{g_k(x)}$, k 2 K.

$\mathbf{f_k(x, \theta_k)}$ Assume model distribution: $\mathbf{f_k(x, \theta_k)}$

$\xi_l$ and $\varepsilon_l$ Labeled data training crosspoint and error

# Terms --- Cont’

| | |
|---|---|
| $\xi_{m_{opt}}$ and $\varepsilon_m$ | Model misspecified crosspoint and error |
| $\xi_{opt}$ and $\varepsilon_{opt}$ | Bayes optimal crosspoint and error |
| $\xi_u$ and $\varepsilon_u$ | Unlabeled data training crosspoint and error |
| $\mathbf{Z_i^{(1)}}$ and $\mathbf{Z_j^{(2)}}$ | Likelihood space : $\mathbf{Z_i^{(1)}} = [f_1(X_i^{(1)}, \theta_1), \; f_2(X_i^{(1)}, \theta_2))]$<br>$\mathbf{Z_j^{(2)}} = [f_1(X_j^{(2)}, \theta_1), \; f_2(X_j^{(2)}, \theta_2))]$ |
| $\mathbf{S_w}$ | within-class scatter matrix |
| $\mathbf{S_b}$ | between-class scatter matrix |

# Semi-supervised learning

- ***Supervised classification***: target variable is well defined and that a sufficient number of its values are labeled.
- ***Unsupervised classification***: no labeled training data are available.
- ***Semi-supervised learning*** : using large amount of ***unlabeled*** training data to help limited amount of ***labeled*** training data to improve classification performance.

# Semi-supervised learning – *Cont*'

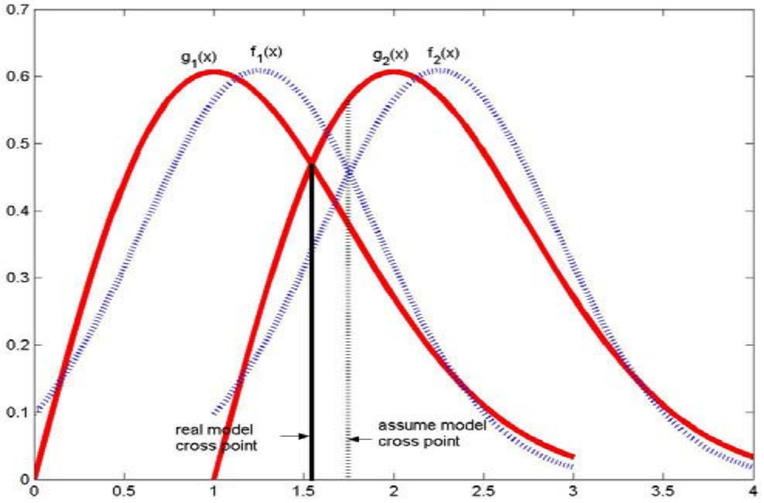
- parametric generative mixture models approach:
  - *labeled data is used initially to estimate mixture model parameters;*
  - *naive bayes classifier is used to label unlabeled data*
  - *re-estimate the mixture model parameters use The combined labeled and unlabeled data*

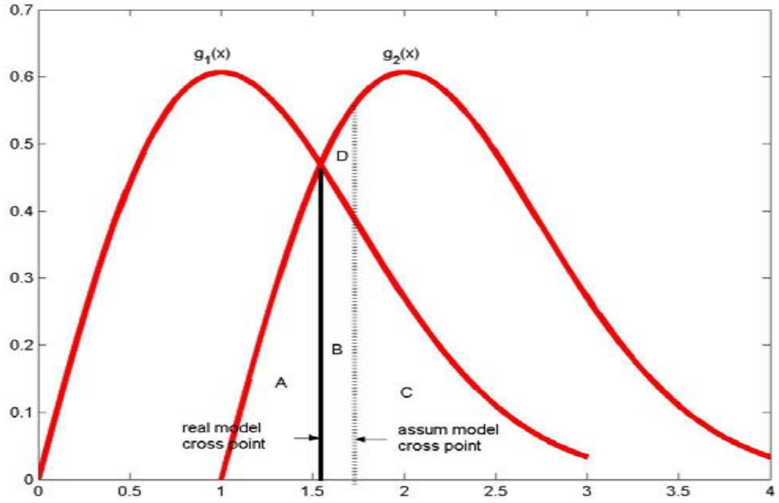
# Semi-supervised learning – *Cont'*

- The optimal probability of labeled and unlabeled data error will converge at a speed relate to the size of labeled training data, when labeled and unlabeled data are from the same structure family[5],
- Unlabeled data degrade classification performance when model misspecified

# Semi-supervised learning – *Cont’*

- Classification error: Bayes error, estimation error and Model error





$\varepsilon_{opt}$ = A + B + C

$\varepsilon_m$ = D

# Semi-supervised learning

## --- *simulation*

- Rayleigh distributed true data and mis-specify as Gaussian
- ***1st simulation:***

The labeled training data estimated cross

point $\xi_l$= $(f_1(x/(\mu_1,\sigma_1\})$ == $f_2(x/(\mu_2,\sigma_2))$ is further away from $\xi_{opt}$ than model misspecified and unlabeled data crosspoint $\xi_{(m+u)}$.

# Semi-supervised learning

--- *simulation*

- 2nd simulation:

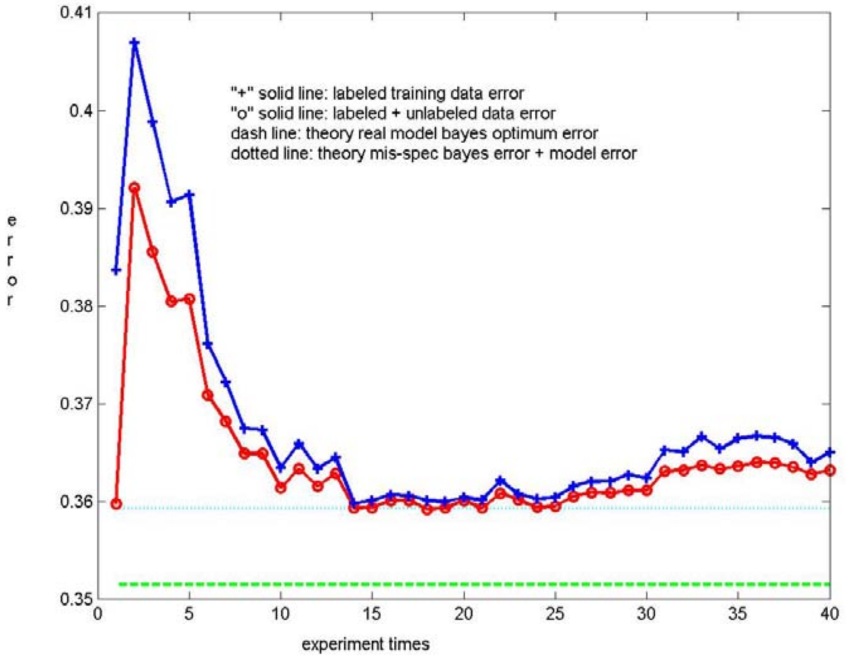
  the estimated distribution cross point is closer to $\xi_{opt}$ than $\xi_{(m+u)}$.

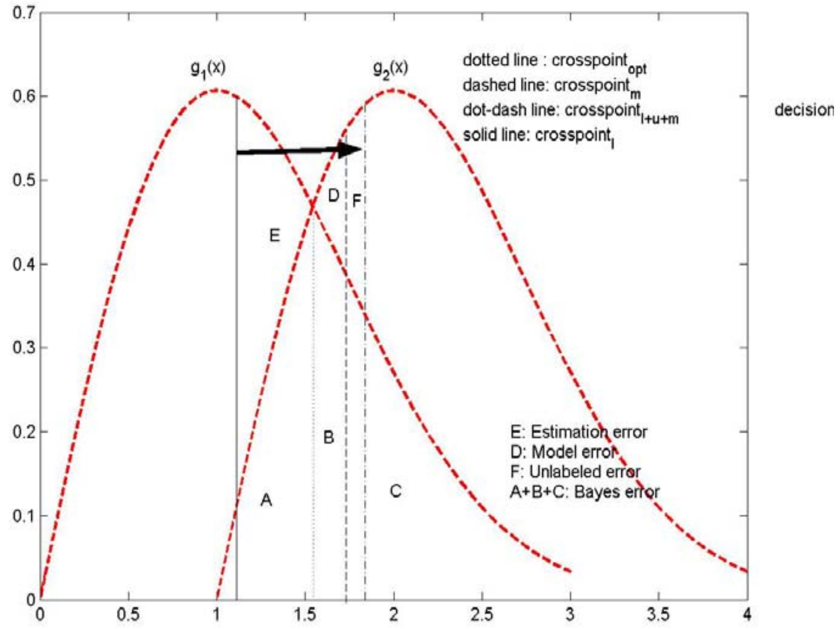
# Semi-supervised learning

## *simulation1*

Simulation 1: Dist($\xi_l$ , $\xi_{opt}$)> Dist($\xi_{(m+u)}$., $\xi_{opt}$)

$$\varepsilon_l > \varepsilon_{m_{opt}} + \varepsilon_u$$



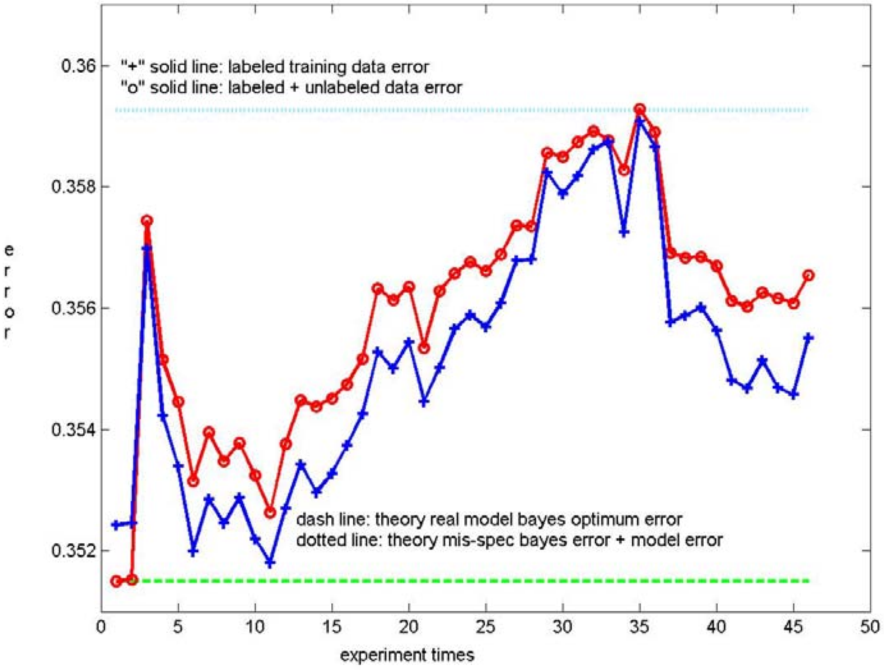


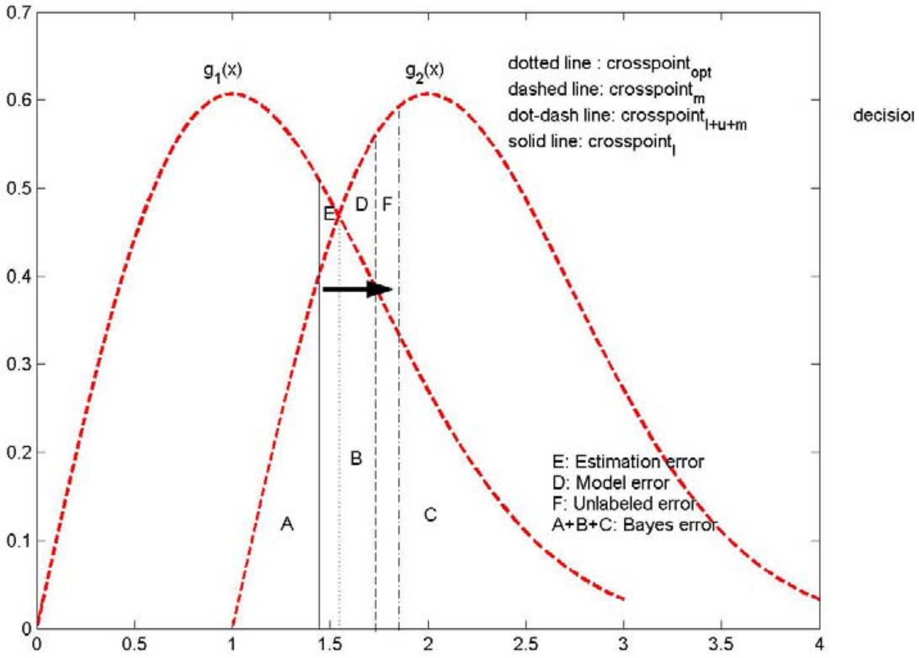
# Semi-supervised learning

## *simulation2*

Simulation 2: $\mathrm{Dist}(\xi_l\ ,\ \ \xi_{opt}) < \mathrm{Dist}(\xi_{(m+u)}.,\ \ \xi_{opt})$

$\varepsilon_l < \varepsilon_{m_{opt}} + \varepsilon_u$





# Semi-supervised learning – *simulation*

- Conclusion:

When model mis-specified , unlabeled data help to improve classification performance only when the estimation error for labeled training data is bigger than model error and unlabeled data estimation error .

$\mathrm{Dist}(\xi_l ,\} \xi_{opt}) > \mathrm{Dist}(\xi_{(m+u)},\} \xi_{opt})$

$\varepsilon_l > \varepsilon_{m_{opt}} + \varepsilon_u$

# Classification in Likelihood space

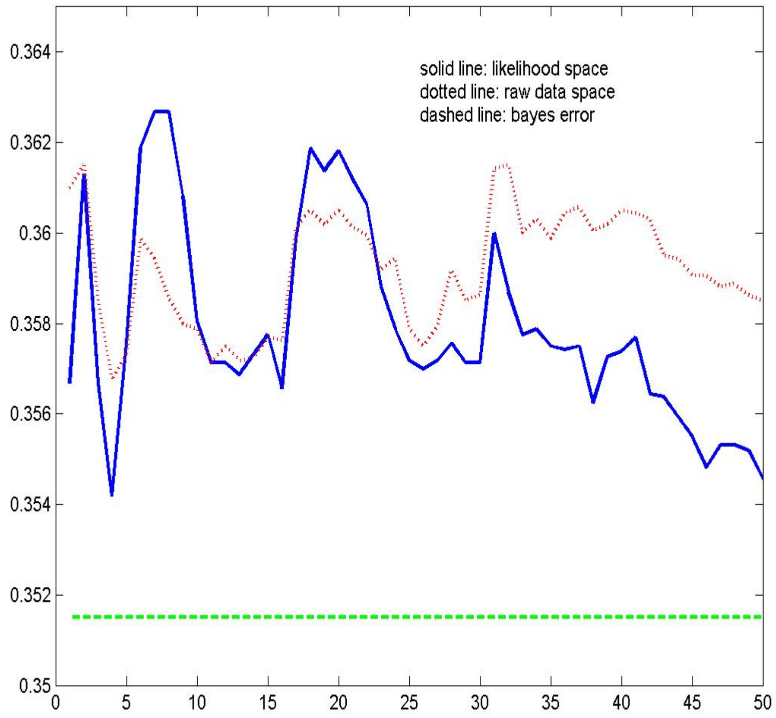
- Construct likelihood space by project the data to different classes seperatly.
- Apply Linear Discriminate Analysis to likelihood space data to classify the data.
  - $S_w = \sum(q_{\{\omega\}_i} E\{(Z-M_i)(Z-M_i)^T|i\})$
  - $S_b = \sum(q_{\{\omega\}_i}(M_i-M_0)(M_i-M_0)^{T)}$
  - The optimal LDA projection matrix:

$$\mathbf{W_{opt}=[w_1,w_2,\ldots,w_D] = arg\ max_W( tr(W^T S_b W)/tr(W^T S_w W)}$$

# Supervised Classification in likelihood space

*– simulation*

- G(x) = Rayleigh F(x) = Gaussian



## Design:

- Labeled training data size: 50:50:200
- Estimate Gaussian parameters $(\mu_1,\sigma_1)$, $(\mu_2,\sigma_2)$ from training data
- Find LDA boundary in likelihood space

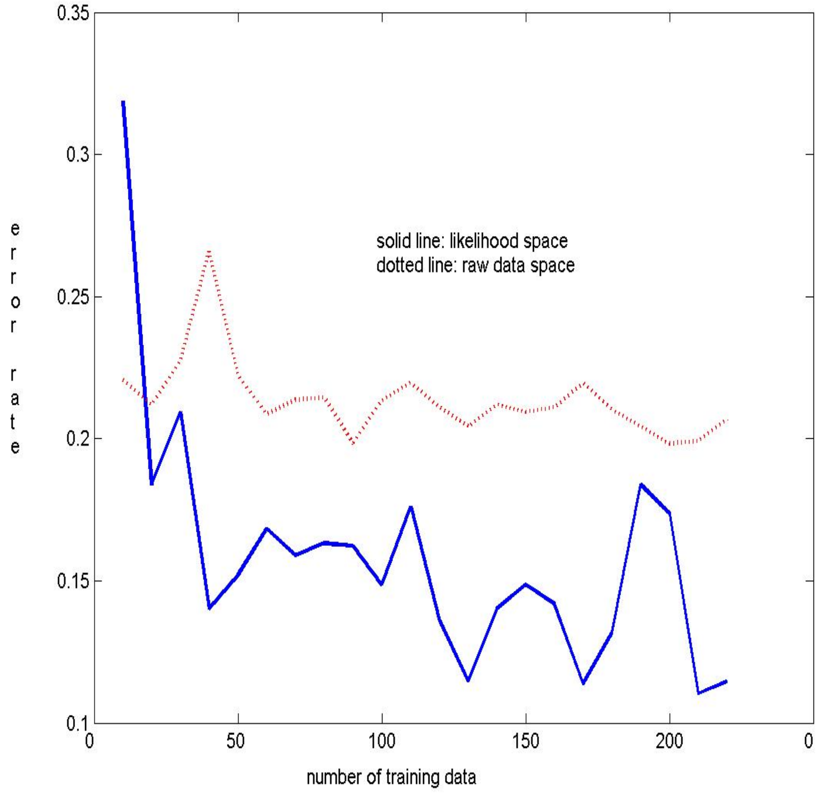
## Result:

- Green Line: Bayes Optimum error
- Blue Line: Likelihood space classification error
- Red line: raw data space classification error

## Conclusion:

- likelihood space do improve classification performance in supervised learning

# Supervised Classification in likelihood space – *SAR*



## Design:

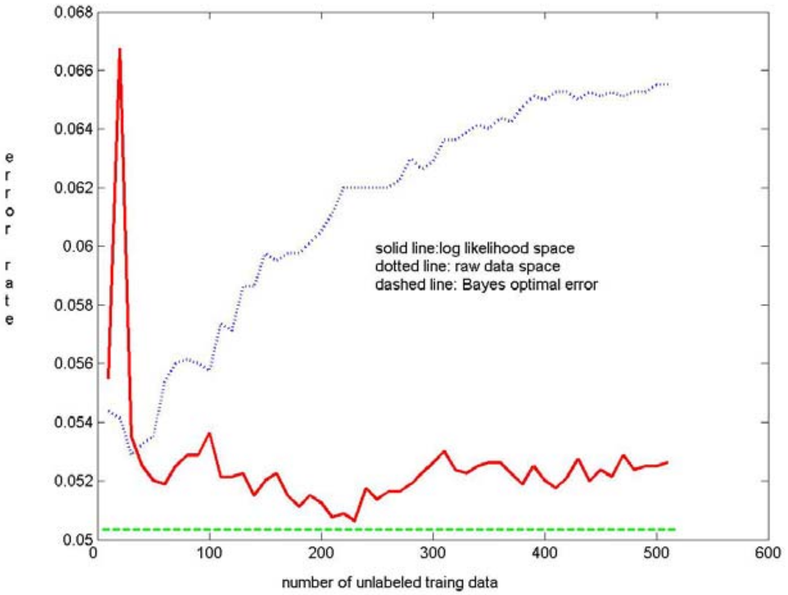
- MSTAR SAR data: T72, BMP2 2 GMMs with 5 mixtures. $q_{\omega 1} = \cdots = q_{\omega k}$
- Increase training data size by 50 each time.

## Conclusion:

- under a practical situation, accurate model assumption is difficult to obtain, and likelihood space classification has an advantage on handling model mis-specification.

# Semi-supervised Classification in likelihood space – *simulation*

- Rayleigh distributed true data and mis-specified as Gaussian



## Design:

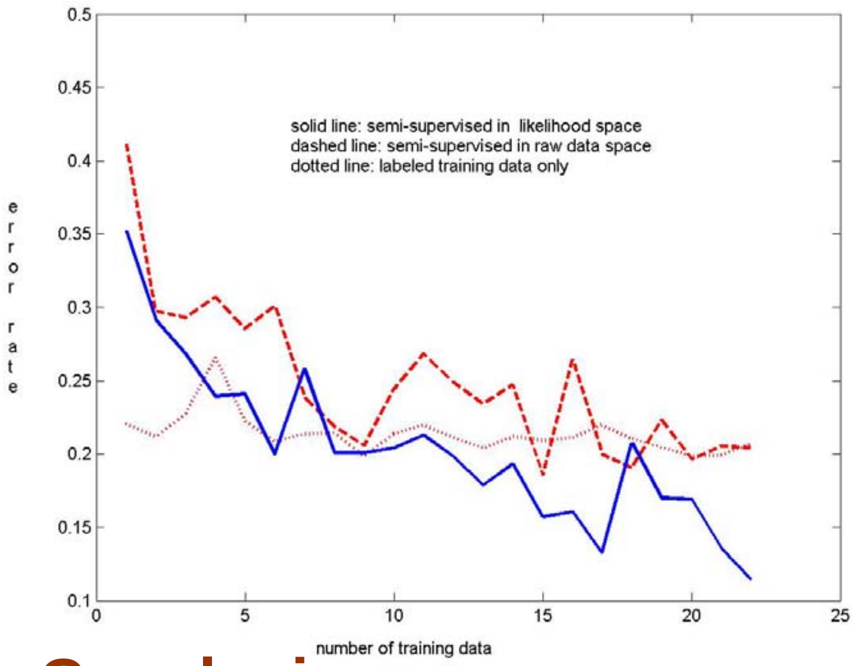
- Labeled training data size: 10:50:510, unlabeled data size 500; testing size 8000
- Estimate Gaussian parameters $(\mu_1,\sigma_1)$, $(\mu_2,\sigma_2)$ from labeled training data
- Classify unlabeled data using Bayes classifier,
- Reestimate $(\mu_1,\sigma_1),(\mu_2,\sigma_2)$ from labeled + psuedo labeled training data
- Bayes classifier in raw data space.
- LDA classifier in likelihood space

## Result:

- Green Line: Bayes Optimum error without model misspecification
- Red Line: Likelihood space classification error
- Blue line: raw data space classification error

**Conclusion:** likelihood space do improve classification performance in semi-supervised learning

# Semi-supervised Classification in likelihood space – *SAR*



## Conclusion:

likelihood space do improve classification performance in semi-supervised learning

## Design:

- Labeled training data size: 10:10:232, unlabeled data size 232-labeled training data; testing size 588
- Estimate Gaussian parameters $(\mu_1,\sigma_1)$, $(\mu_2,\sigma_2)$ from labeled training data
- Classify unlabeled data using Bayes classifier,
- Reestimate $(\mu_1,\sigma_1),(\mu_2,\sigma_2)$ from labeled + pseudo labeled training data
- Bayes classifier in raw data space.
- LDA classifier in likelihood space

## Result:

- Pink Line: raw data space classification error for labeled training data only
- Blue Line: Likelihood space classification error for label + unlabeled training data
- Red line: raw data space classification error for label + unlabeled training data

# Conclusion

- Unlabeled data may not always help to improve the semi-supervised classification performance, especially when model assumption is inaccurate.
- Projecting data samples into likelihood space and then applying LDA for classification may have better robustness with regard to model mis specification.