

# Classtering: Joint Classification and Clustering with Mixture of Factor Analysers

Emanuele Sansone, Andrea Passerini, Francesco G. B. De Natale



UNIVERSITY OF TRENTO - Italy

# Motivation

## Semi-supervised learning (SSL)

Classification

Clustering

- Problem of label propagation
- Cluster assumption

Discriminative

Generative

$$f : y \rightarrow c$$

$$p(c|y) = \frac{p(y|c)p(c)}{Z}$$

# Motivation

## Semi-supervised learning (SSL)

Classification

Clustering

Discriminative

Generative

$$f : y \rightarrow c$$

$$p(c|y) = \frac{p(y|c)p(c)}{Z}$$

- Problem of label propagation
- Cluster assumption

Desired:

- No wrong label propagation
- Relaxing the cluster assumption

Desired:

- Model inter- and intra-class variabilities
- Achieve possibly “good performance”

# Motivation

## Semi-supervised learning (SSL)

Classification

Clustering

Discriminative

Generative

$$f : y \rightarrow c$$

$$p(c|y) = \frac{p(y|c)p(c)}{Z}$$

- Problem of label propagation
- Cluster assumption

Desired:

- No wrong label propagation
- Relaxing the cluster assumption

Desired:

- Model inter- and intra-class variabilities
- Achieve possibly “good performance”

Discover the structure of data while preserving the discrimination among classes

# Motivation

## Why jointly addressing classification and clustering?

**Medicine:** discrimination between healthy and pathological cases is often hard (lack of complete understanding of the pathology, data collection)



Healthy vs. pathological case + Different forms of disease

# Motivation

## Why jointly addressing classification and clustering?

**Medicine:** discrimination between healthy and pathological cases is often hard (lack of complete understanding of the pathology, data collection)



Healthy vs. pathological case + Different forms of disease

## Why not using two-stage approaches?

1. Clustering - Classification
2. Classification - Clustering

# Motivation

## Why jointly addressing classification and clustering?

**Medicine:** discrimination between healthy and pathological cases is often hard (lack of complete understanding of the pathology, data collection)

Healthy vs. pathological case + Different forms of disease

## Why not using two-stage approaches?

1. Clustering - Classification
2. Classification - Clustering

Clustering and Classification with limited amount of supervised information

# Model

## **Assumptions:**

1. Class-conditional densities are well approximated by a Gaussian mixture
2. i.i.d. samples
3. Data lie on a manifold



# Model

## **Assumptions:**

1. Class-conditional densities are well approximated by a Gaussian mixture
2. i.i.d. samples
3. Data lie on a manifold

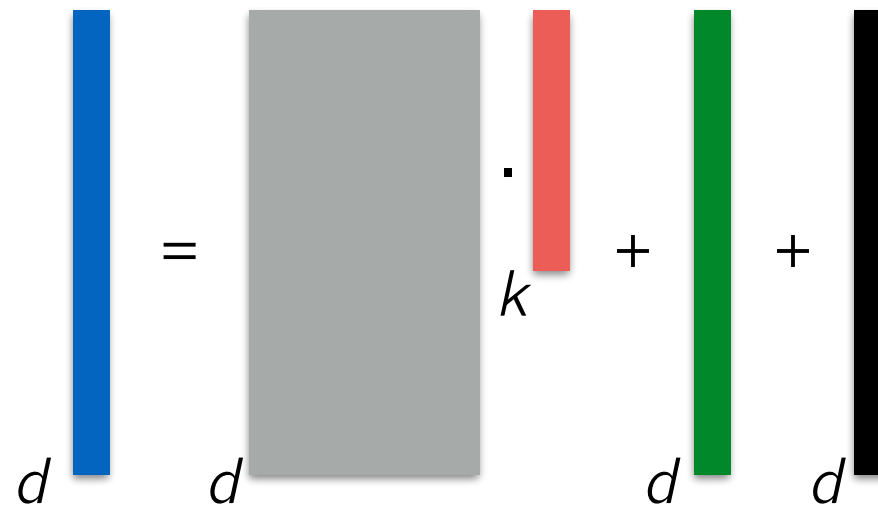
Model based on Mixture of Factor Analysers  
(MFA)

Note: the MFA model is used in unsupervised learning (e.g. model-based clustering, local dimensionality reduction)

# Model

Given an **unlabeled** training dataset  $D = \{\mathbf{y}_n\}_{n=1}^N$

$$\mathbf{y}_n = \Lambda \mathbf{x}_n + \boldsymbol{\mu} + \boldsymbol{\xi}$$

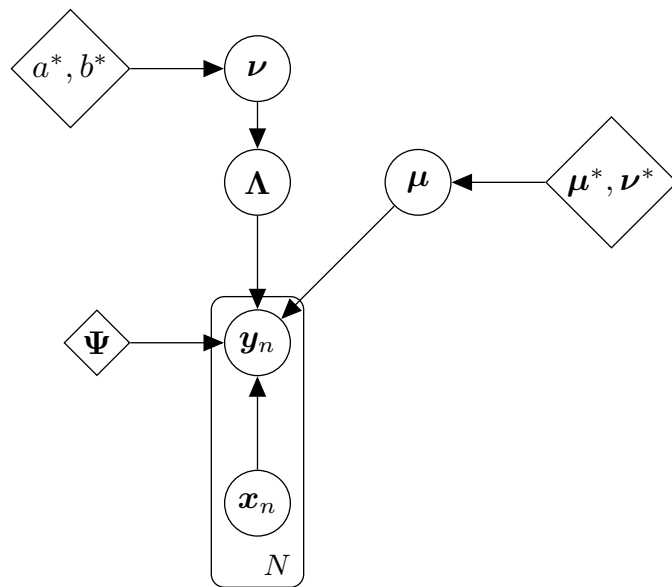


# Model

Given an **unlabeled** training dataset  $D = \{\mathbf{y}_n\}_{n=1}^N$

$$\mathbf{y}_n = \Lambda \mathbf{x}_n + \boldsymbol{\mu} + \boldsymbol{\xi}$$

$$\begin{matrix} d \\ \text{blue bar} \end{matrix} = \begin{matrix} d \\ \text{gray rectangle} \end{matrix} \cdot \begin{matrix} k \\ \text{red bar} \end{matrix} + \begin{matrix} d \\ \text{green bar} \end{matrix} + \begin{matrix} d \\ \text{black bar} \end{matrix}$$



$$\Lambda \sim \prod_{j=1}^k \mathcal{N}\left(\mathbf{0}, \frac{\mathbf{I}}{\nu(j)}\right)$$

$$\mathbf{x}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}^*, \text{diag}(\boldsymbol{\nu}^*)^{-1})$$

$$\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \Psi)$$

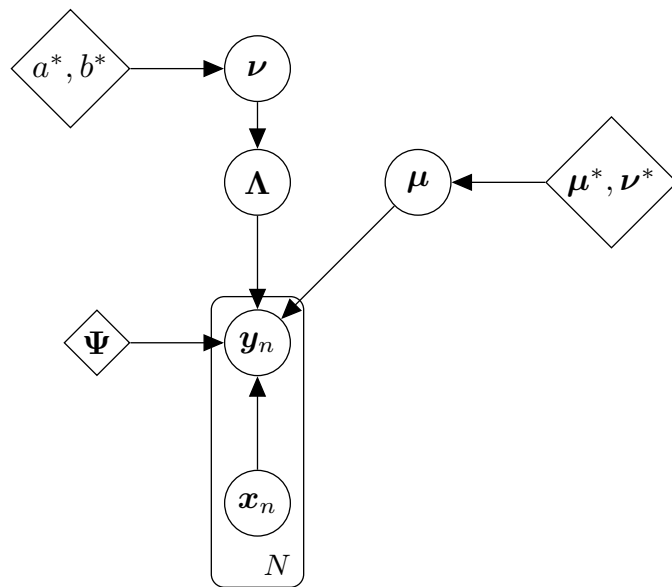
$$\boldsymbol{\nu} \sim \prod_{j=1}^k \text{Gamma}(\nu(j) | a^*, b^*)$$

# Model

Given an **unlabeled** training dataset  $D = \{\mathbf{y}_n\}_{n=1}^N$

$$\mathbf{y}_n = \Lambda \mathbf{x}_n + \boldsymbol{\mu} + \boldsymbol{\xi}$$

$$\begin{matrix} d \\ \text{blue bar} \end{matrix} = \begin{matrix} d \\ \text{grey rectangle} \end{matrix} \cdot \begin{matrix} k \\ \text{red bar} \end{matrix} + \begin{matrix} d \\ \text{green bar} \end{matrix} + \begin{matrix} d \\ \text{black bar} \end{matrix}$$



$$\begin{aligned} \Lambda &\sim \prod_{j=1}^k \mathcal{N}\left(\mathbf{0}, \frac{\mathbf{I}}{\nu(j)}\right) \\ \mathbf{x}_n &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ \boldsymbol{\mu} &\sim \mathcal{N}(\boldsymbol{\mu}^*, \text{diag}(\boldsymbol{\nu}^*)^{-1}) \\ \boldsymbol{\xi} &\sim \mathcal{N}(\mathbf{0}, \Psi) \\ \boldsymbol{\nu} &\sim \prod_{j=1}^k \text{Gamma}(\nu(j) | a^*, b^*) \end{aligned}$$



$$p(\mathbf{y}_n | \Lambda, \boldsymbol{\mu}, \Psi) \sim \mathcal{N}(\boldsymbol{\mu}, \Lambda \Lambda' + \Psi)$$

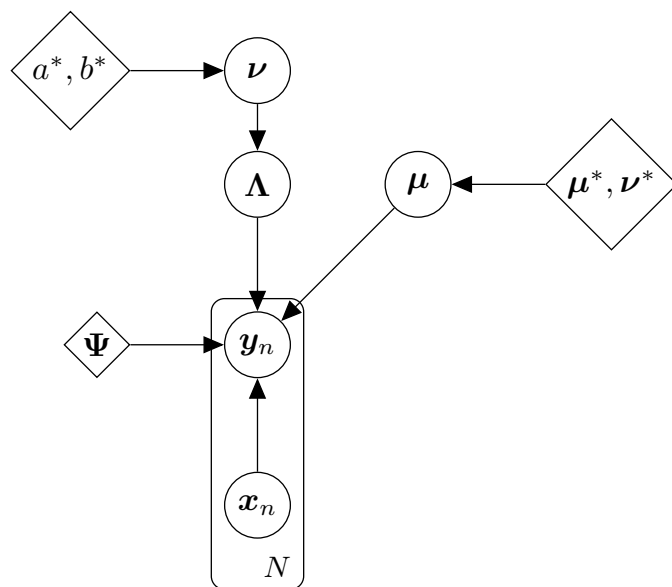
Analysers are described only by  $\Lambda, \boldsymbol{\mu}$

# Model

Given an **unlabeled** training dataset  $D = \{\mathbf{y}_n\}_{n=1}^N$

$$\mathbf{y}_n = \Lambda \mathbf{x}_n + \mu + \xi$$

$$\begin{matrix} d \\ \text{blue bar} \end{matrix} = \begin{matrix} d \\ \text{gray matrix} \end{matrix} \cdot \begin{matrix} k \\ \text{red bar} \end{matrix} + \begin{matrix} d \\ \text{green bar} \end{matrix} + \begin{matrix} d \\ \text{black bar} \end{matrix}$$

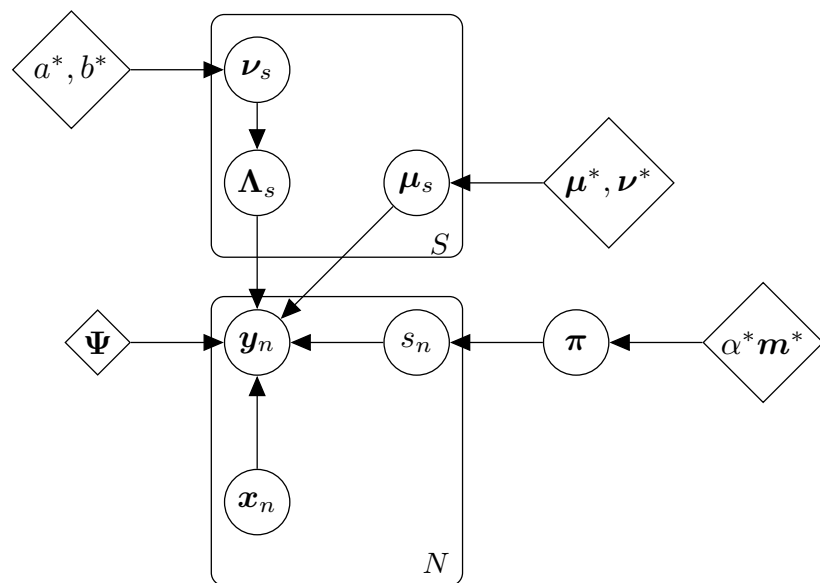


# Model

Given an **unlabeled** training dataset  $D = \{\mathbf{y}_n\}_{n=1}^N$

$$\mathbf{y}_n = \Lambda_{s_n} \mathbf{x}_n + \boldsymbol{\mu}_{s_n} + \boldsymbol{\xi}$$

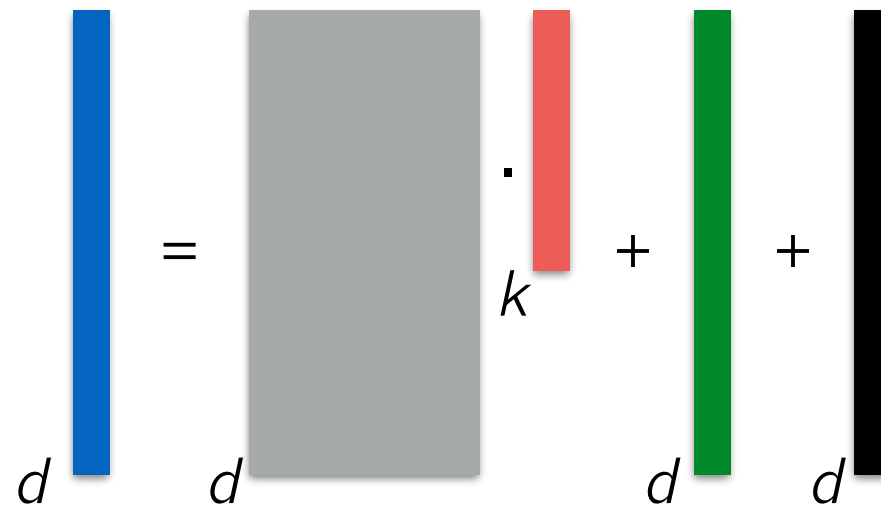
$$\begin{matrix} d \\ \text{blue bar} \end{matrix} = \begin{matrix} d \\ \text{gray bar} \end{matrix} \cdot \begin{matrix} k \\ \text{red bar} \end{matrix} + \begin{matrix} d \\ \text{green bar} \end{matrix} + \begin{matrix} d \\ \text{black bar} \end{matrix}$$



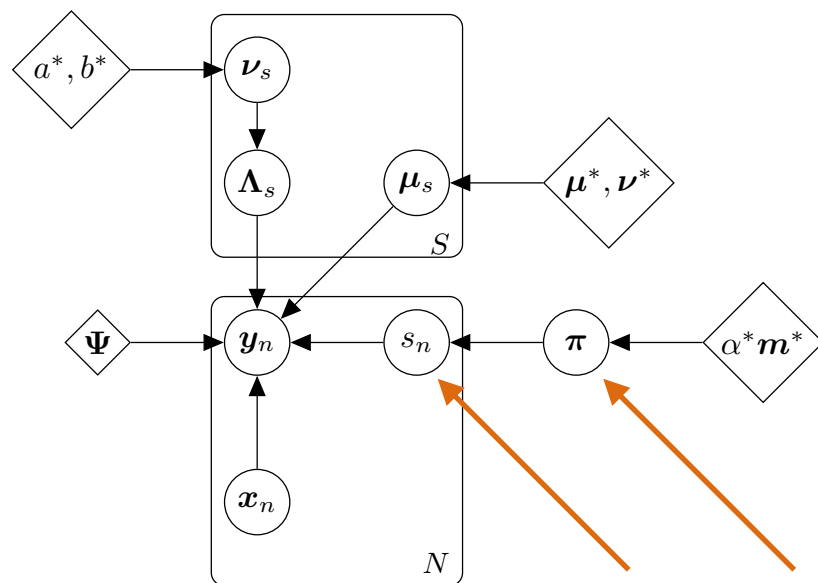
# Model

Given an **unlabeled** training dataset  $D = \{\mathbf{y}_n\}_{n=1}^N$

$$\mathbf{y}_n = \Lambda_{s_n} \mathbf{x}_n + \boldsymbol{\mu}_{s_n} + \boldsymbol{\xi}$$



A diagram illustrating the vector equation  $\mathbf{y}_n = \Lambda_{s_n} \mathbf{x}_n + \boldsymbol{\mu}_{s_n} + \boldsymbol{\xi}$ . It shows a blue vertical bar labeled  $d$  (representing  $\mathbf{y}_n$ ) equal to a gray vertical bar labeled  $d$  (representing  $\Lambda_{s_n}$ ) multiplied by a red vertical bar labeled  $k$  (representing  $\mathbf{x}_n$ ), plus a green vertical bar labeled  $d$  (representing  $\boldsymbol{\mu}_{s_n}$ ), plus a black vertical bar labeled  $d$  (representing  $\boldsymbol{\xi}$ ).



$$s_n \sim \prod_{s=1}^S \pi_s^{1_{s_n}(s)}$$

$$\boldsymbol{\pi} \sim \text{Dir}(\alpha^* \mathbf{m}^*)$$

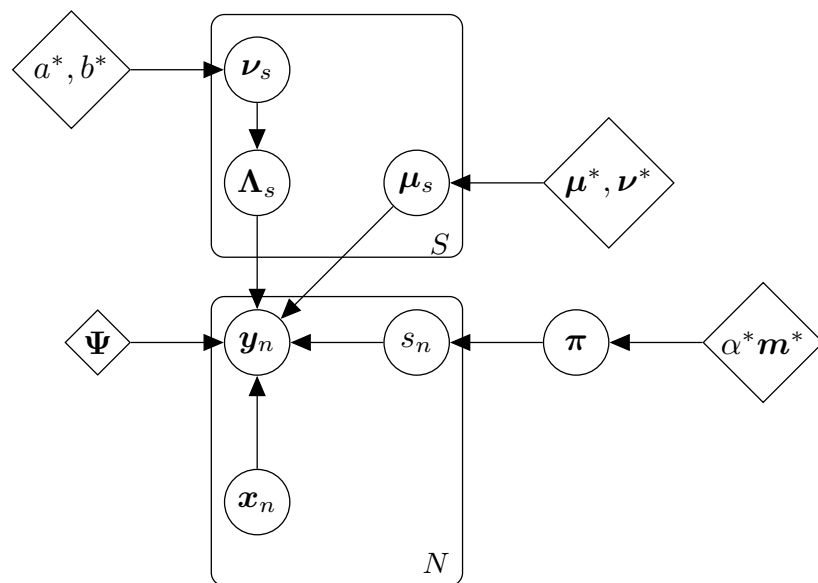
$$\mathbf{m}^* = [1/S, \dots, 1/S]$$

# Model

Given an **unlabeled** training dataset  $D = \{\mathbf{y}_n\}_{n=1}^N$

$$\mathbf{y}_n = \Lambda_{s_n} \mathbf{x}_n + \boldsymbol{\mu}_{s_n} + \boldsymbol{\xi}$$

$$\begin{matrix} d \\ \text{blue bar} \end{matrix} = \begin{matrix} d \\ \text{gray bar} \end{matrix} \cdot \begin{matrix} k \\ \text{red bar} \end{matrix} + \begin{matrix} d \\ \text{green bar} \end{matrix} + \begin{matrix} d \\ \text{black bar} \end{matrix}$$

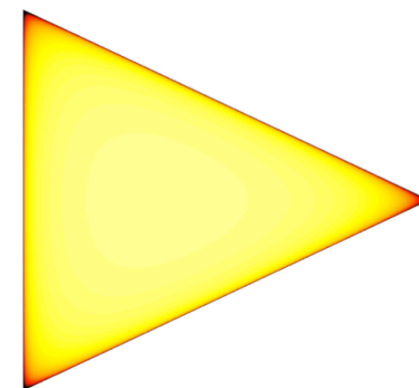


$$s_n \sim \prod_{s=1}^S \pi_s^{1_{s_n}(s)}$$

$$\boldsymbol{\pi} \sim \text{Dir}(\alpha^* \mathbf{m}^*)$$

$$\mathbf{m}^* = [1/S, \dots, 1/S]$$

Example with  
 $S = 3, \alpha^* = 2.1$



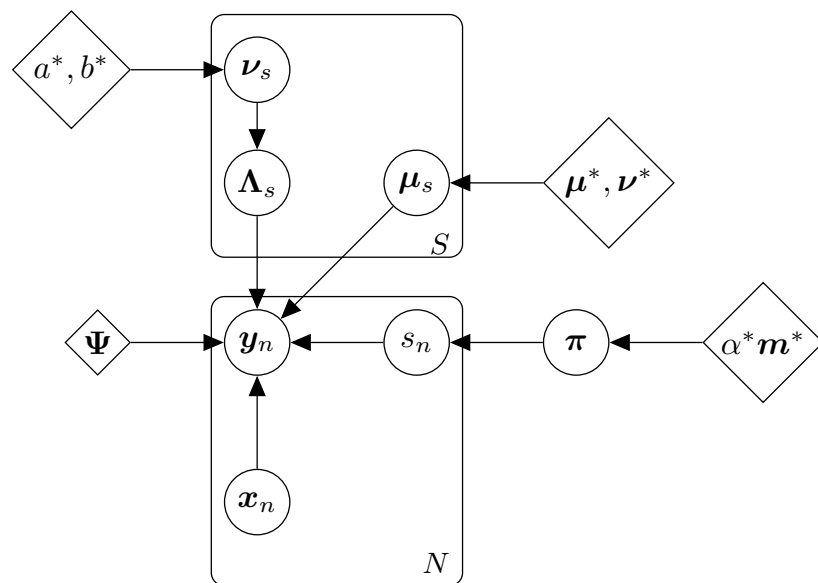


# Model

Given an **unlabeled** training dataset  $D = \{\mathbf{y}_n\}_{n=1}^N$

$$\mathbf{y}_n = \Lambda_{s_n} \mathbf{x}_n + \boldsymbol{\mu}_{s_n} + \boldsymbol{\xi}$$

$$\begin{matrix} d \\ \text{blue bar} \end{matrix} = \begin{matrix} d \\ \text{grey rectangle} \end{matrix} \cdot \begin{matrix} k \\ \text{red bar} \end{matrix} + \begin{matrix} d \\ \text{green bar} \end{matrix} + \begin{matrix} d \\ \text{black bar} \end{matrix}$$



$$s_n \sim \prod_{s=1}^S \pi_s^{1_{s_n}(s)}$$

$$\boldsymbol{\pi} \sim \text{Dir}(\alpha^* \mathbf{m}^*)$$

$$\mathbf{m}^* = [1/S, \dots, 1/S]$$

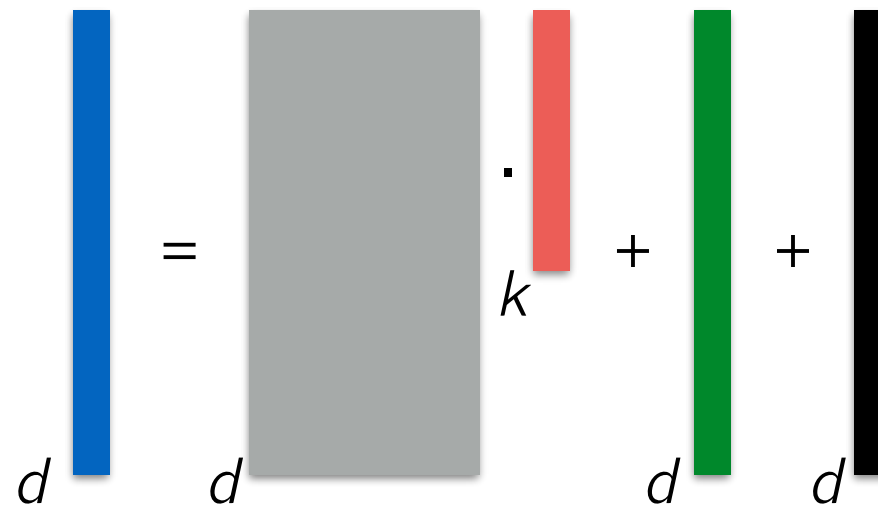


$$p(\mathbf{y}_n | \boldsymbol{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\Psi}) \sim \sum_{s_n=1}^S \pi_{s_n} \mathcal{N}(\boldsymbol{\mu}_{s_n}, \boldsymbol{\Lambda}_{s_n} \boldsymbol{\Lambda}_{s_n}^T + \boldsymbol{\Psi})$$

# Model

Given two sets: **labeled**  $D' = \{(\mathbf{y}_n, c_n)\}_{n=1}^N$  and **unlabeled**  $D'' = \{\mathbf{y}_n\}_{n=N+1}^M$

$$\mathbf{y}_n = \Lambda_{s_n} \mathbf{x}_n + \mu_{s_n} + \xi$$

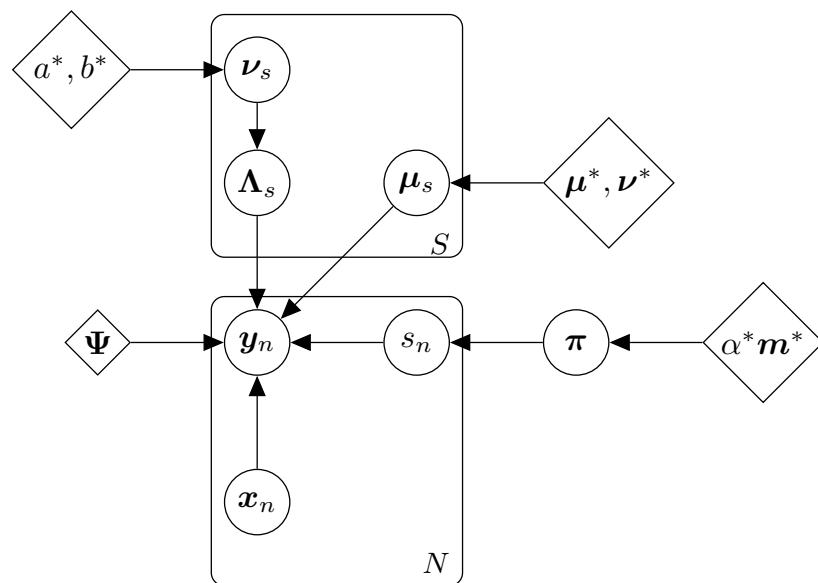


# Model

Given two sets: **labeled**  $D' = \{(\mathbf{y}_n, c_n)\}_{n=1}^N$  and **unlabeled**  $D'' = \{\mathbf{y}_n\}_{n=N+1}^M$

$$\mathbf{y}_n = \Lambda_{s_n} \mathbf{x}_n + \mu_{s_n} + \xi$$

$$\begin{matrix} d \\ \color{blue}{\text{bar}} \end{matrix} = \begin{matrix} d \\ \text{gray bar} \end{matrix} \cdot \begin{matrix} k \\ \color{red}{\text{bar}} \end{matrix} + \begin{matrix} d \\ \color{green}{\text{bar}} \end{matrix} + \begin{matrix} d \\ \text{black bar} \end{matrix}$$

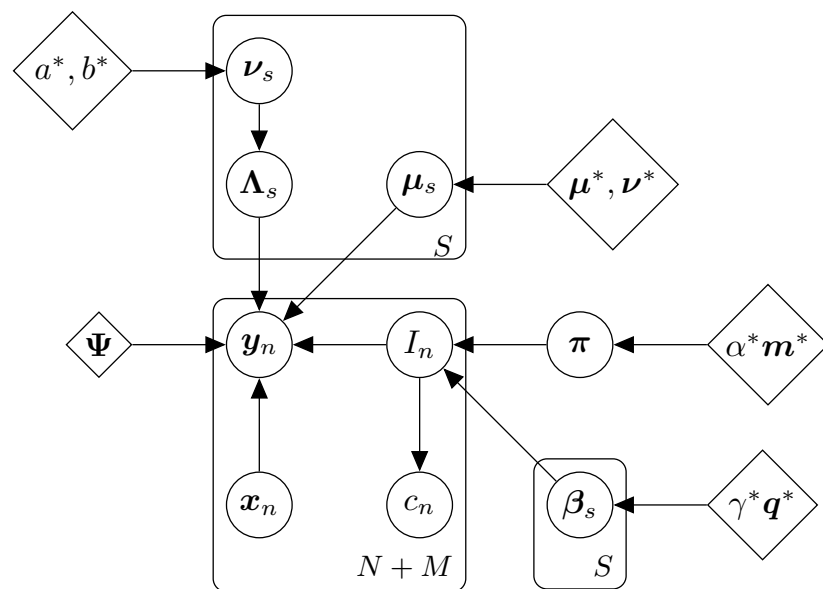


# Model

Given two sets: **labeled**  $D' = \{(\mathbf{y}_n, c_n)\}_{n=1}^N$  and **unlabeled**  $D'' = \{\mathbf{y}_n\}_{n=N+1}^M$

$$\mathbf{y}_n = \Lambda_{s_n} \mathbf{x}_n + \mu_{s_n} + \xi$$

$$\underset{d}{\text{blue bar}} = \underset{d}{\text{gray bar}} \cdot \underset{k}{\text{red bar}} + \underset{d}{\text{green bar}} + \underset{d}{\text{black bar}}$$

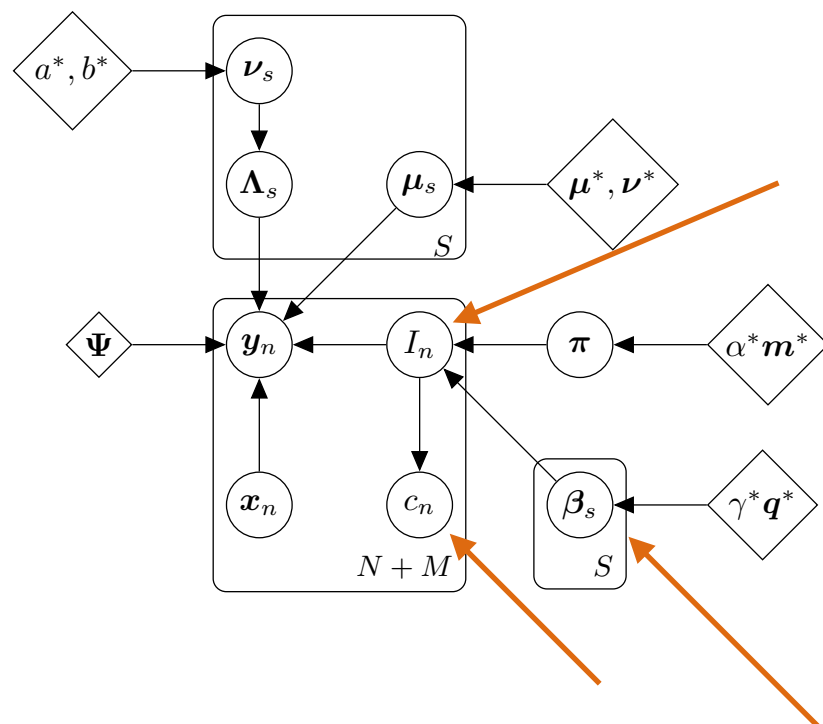


# Model

Given two sets: **labeled**  $D' = \{(\mathbf{y}_n, c_n)\}_{n=1}^N$  and **unlabeled**  $D'' = \{\mathbf{y}_n\}_{n=N+1}^M$

$$\mathbf{y}_n = \Lambda_{s_n} \mathbf{x}_n + \boldsymbol{\mu}_{s_n} + \boldsymbol{\xi}$$

$$\begin{matrix} d \\ d \end{matrix} \begin{matrix} \text{blue bar} \\ \text{gray bar} \end{matrix} = \begin{matrix} d \\ d \end{matrix} \begin{matrix} \text{gray bar} \\ \text{red bar} \end{matrix} \cdot \begin{matrix} k \\ k \end{matrix} + \begin{matrix} d \\ d \end{matrix} \begin{matrix} \text{green bar} \\ \text{black bar} \end{matrix} + \begin{matrix} d \\ d \end{matrix} \begin{matrix} \text{green bar} \\ \text{black bar} \end{matrix}$$



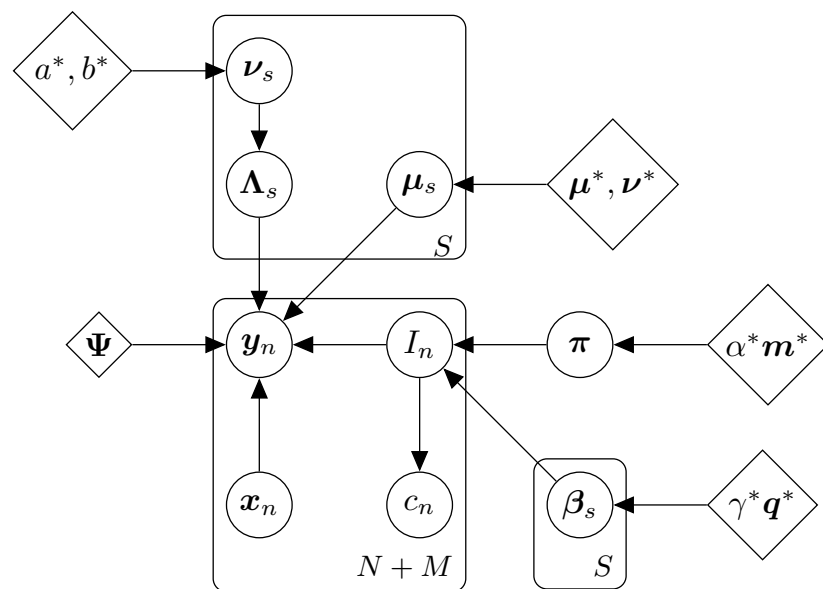
$$\begin{aligned} I_n &= (s_n, \ell_n) \\ \ell_n &\sim \prod_{\ell=1}^K \beta_{s_n}(\ell)^{1_{\ell_n}(\ell)} \\ c_n &\sim \delta(c_n - \ell_n) \\ \beta_s &\sim \text{Dir}(\gamma^* \mathbf{q}^*) \\ \mathbf{q}^* &= [1/K, \dots, 1/K] \end{aligned}$$

# Model

Given two sets: **labeled**  $D' = \{(\mathbf{y}_n, c_n)\}_{n=1}^N$  and **unlabeled**  $D'' = \{\mathbf{y}_n\}_{n=N+1}^M$

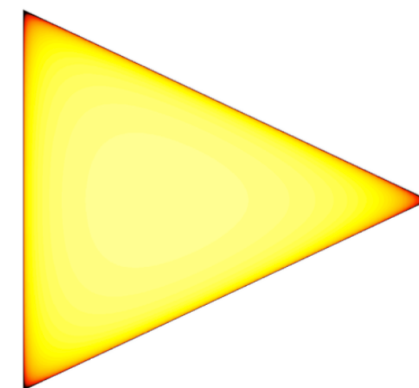
$$\mathbf{y}_n = \Lambda_{s_n} \mathbf{x}_n + \boldsymbol{\mu}_{s_n} + \boldsymbol{\xi}$$

$$\begin{matrix} d \\ \text{blue bar} \end{matrix} = \begin{matrix} d \\ \text{gray bar} \end{matrix} \cdot \begin{matrix} k \\ \text{red bar} \end{matrix} + \begin{matrix} d \\ \text{green bar} \end{matrix} + \begin{matrix} d \\ \text{black bar} \end{matrix}$$



$$\begin{aligned} I_n &= (s_n, \ell_n) \\ \ell_n &\sim \prod_{\ell=1}^K \beta_{s_n}(\ell)^{1_{\ell_n}(\ell)} \\ c_n &\sim \delta(c_n - \ell_n) \\ \beta_s &\sim \text{Dir}(\gamma^* \mathbf{q}^*) \\ \mathbf{q}^* &= [1/K, \dots, 1/K] \end{aligned}$$

Example with  
 $K = 3, \gamma^* = 2.1$

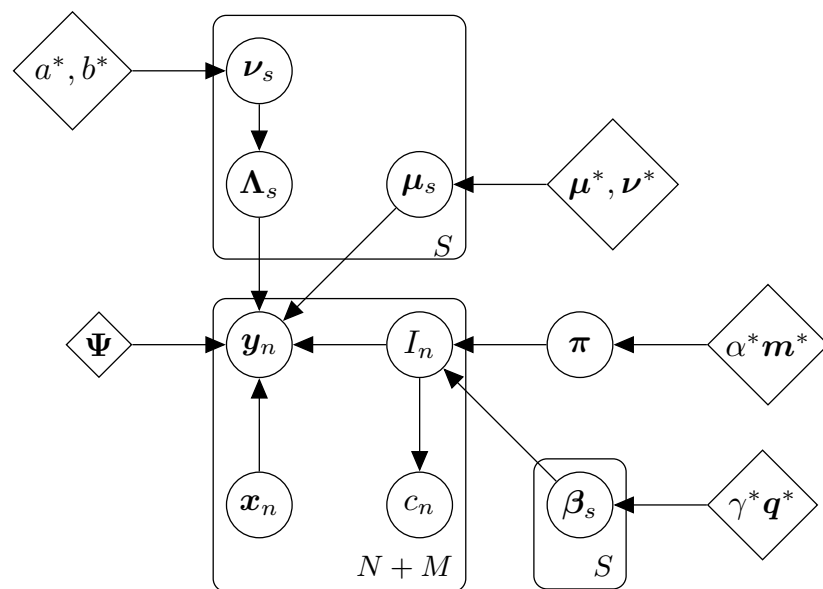


# Model

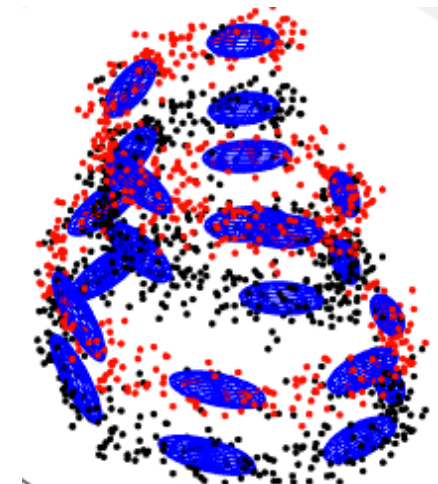
Given two sets: **labeled**  $D' = \{(\mathbf{y}_n, c_n)\}_{n=1}^N$  and **unlabeled**  $D'' = \{\mathbf{y}_n\}_{n=N+1}^M$

$$\mathbf{y}_n = \mathbf{\Lambda}_{s_n} \mathbf{x}_n + \boldsymbol{\mu}_{s_n} + \boldsymbol{\xi}$$

$$\begin{matrix} d \\ \text{blue bar} \end{matrix} = \begin{matrix} d \\ \text{gray bar} \end{matrix} \cdot \begin{matrix} k \\ \text{red bar} \end{matrix} + \begin{matrix} d \\ \text{green bar} \end{matrix} + \begin{matrix} d \\ \text{black bar} \end{matrix}$$



$$\begin{aligned} I_n &= (s_n, \ell_n) \\ \ell_n &\sim \prod_{\ell=1}^K \beta_{s_n}(\ell)^{1_{\ell_n}(\ell)} \\ c_n &\sim \delta(c_n - \ell_n) \\ \beta_s &\sim \text{Dir}(\gamma^* \mathbf{q}^*) \\ \mathbf{q}^* &= [1/K, \dots, 1/K] \end{aligned}$$

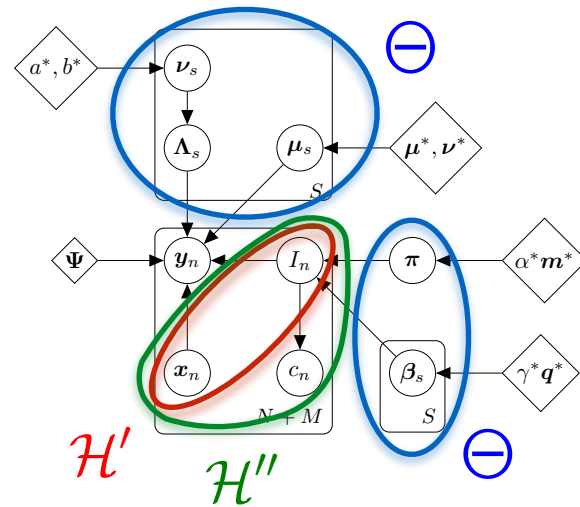


$$p(\mathbf{y}_n | \mathbf{\Lambda}, \boldsymbol{\mu}, \Psi) \sim \sum_{s_n=1}^S \pi_{s_n} \mathcal{N}(\boldsymbol{\mu}_{s_n}, \mathbf{\Lambda}_{s_n} \mathbf{\Lambda}_{s_n}^T + \Psi)$$

Information about clusters vs. classes

# Inference

Given two sets: **labeled**  $D' = \{(\mathbf{y}_n, c_n)\}_{n=1}^N$  and **unlabeled**  $D'' = \{\mathbf{y}_n\}_{n=N+1}^M$



$$\Theta = \{\pi\} \cup \{\beta_s, \Lambda_s, \mu_s, \nu_s\}_{s=1}^S$$

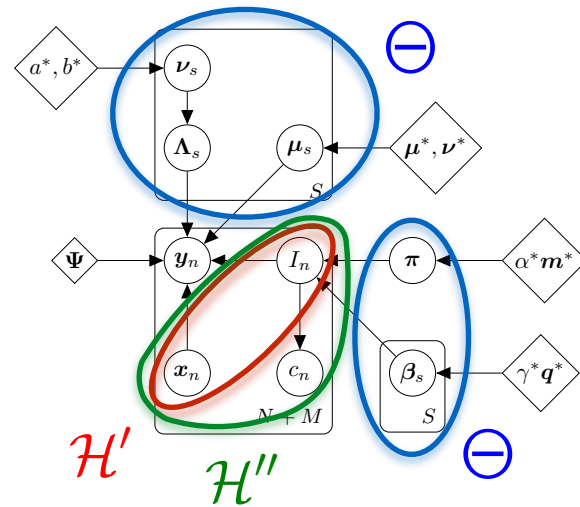
$$\mathcal{H}' = \{\mathbf{x}_n, s_n\}_{n=1}^N$$

$$\mathcal{H}'' = \{\mathbf{x}_n, (s_n, \ell_n), c_n\}_{n=N+1}^M$$



# Inference

Given two sets: **labeled**  $D' = \{(\mathbf{y}_n, c_n)\}_{n=1}^N$  and **unlabeled**  $D'' = \{\mathbf{y}_n\}_{n=N+1}^M$



$$\Theta = \{\pi\} \cup \{\beta_s, \Lambda_s, \mu_s, \nu_s\}_{s=1}^S$$

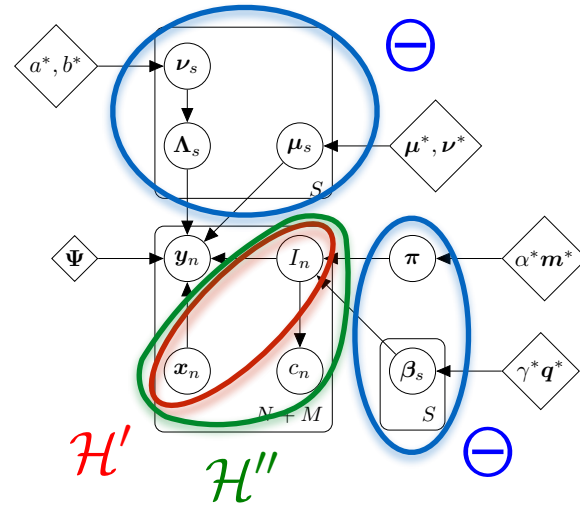
$$\mathcal{H}' = \{\mathbf{x}_n, s_n\}_{n=1}^N$$

$$\mathcal{H}'' = \{\mathbf{x}_n, (s_n, \ell_n), c_n\}_{n=N+1}^M$$

$$\begin{aligned} \log p(D', D'') &= \log \int_{\Theta, \mathcal{H}', \mathcal{H}''} p(D', D'', \Theta, \mathcal{H}', \mathcal{H}'') \\ &= \log \int_{\Theta, \mathcal{H}', \mathcal{H}''} q(\Theta, \mathcal{H}', \mathcal{H}'') \frac{p(D', D'', \Theta, \mathcal{H}', \mathcal{H}'')}{q(\Theta, \mathcal{H}', \mathcal{H}'')} \\ &\geq \int_{\Theta, \mathcal{H}', \mathcal{H}''} q(\Theta, \mathcal{H}', \mathcal{H}'') \log \frac{p(D', D'', \Theta, \mathcal{H}', \mathcal{H}'')}{q(\Theta, \mathcal{H}', \mathcal{H}'')} = \mathcal{F}(q(\cdot)) \end{aligned}$$

# Inference

Given two sets: **labeled**  $D' = \{(\mathbf{y}_n, c_n)\}_{n=1}^N$  and **unlabeled**  $D'' = \{\mathbf{y}_n\}_{n=N+1}^M$



$$\Theta = \{\pi\} \cup \{\beta_s, \Lambda_s, \mu_s, \nu_s\}_{s=1}^S$$

$$\mathcal{H}' = \{\mathbf{x}_n, s_n\}_{n=1}^N$$

$$\mathcal{H}'' = \{\mathbf{x}_n, (s_n, \ell_n), c_n\}_{n=N+1}^M$$

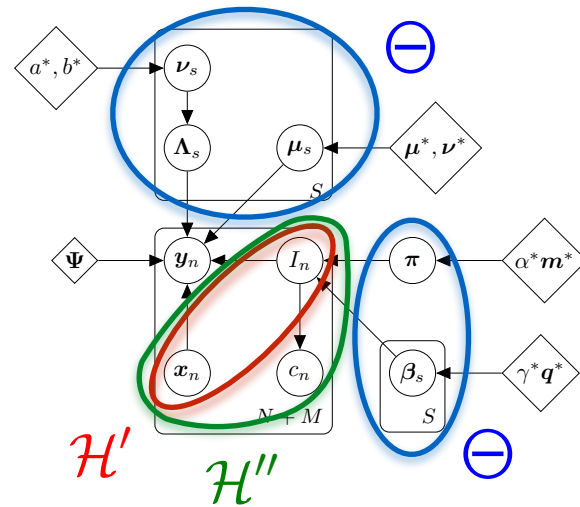
$$\begin{aligned} \log p(D', D'') &= \log \int_{\Theta, \mathcal{H}', \mathcal{H}''} p(D', D'', \Theta, \mathcal{H}', \mathcal{H}'') \\ &= \log \int_{\Theta, \mathcal{H}', \mathcal{H}''} q(\Theta, \mathcal{H}', \mathcal{H}'') \frac{p(D', D'', \Theta, \mathcal{H}', \mathcal{H}'')}{q(\Theta, \mathcal{H}', \mathcal{H}'')} \\ &\geq \int_{\Theta, \mathcal{H}', \mathcal{H}''} q(\Theta, \mathcal{H}', \mathcal{H}'') \log \frac{p(D', D'', \Theta, \mathcal{H}', \mathcal{H}'')}{q(\Theta, \mathcal{H}', \mathcal{H}'')} = \mathcal{F}(q(\cdot)) \end{aligned}$$

- Lower bound on the log-likelihood function
- Equality holds when  $q(\Theta, \mathcal{H}', \mathcal{H}'') = p(\Theta, \mathcal{H}', \mathcal{H}'' | D', D'')$
- Given conditional independence properties of graph  $q(\Theta, \mathcal{H}', \mathcal{H}'') = q(\Theta)q(\mathcal{H}'|\Theta)q(\mathcal{H}''|\Theta)$
- Strict inequality holds in general for:

$$q(\Theta) = q(\pi) \prod_{s=1}^S q(\beta_s) q(\nu_s) q(\Lambda_s, \mu_s) \quad q(\mathcal{H}'|\Theta) = \prod_{n=1}^N q(s_n) q(\mathbf{x}_n | s_n) \quad q(\mathcal{H}''|\Theta) = \prod_{n=1}^N q(l_n) q(\mathbf{x}_n | l_n)$$

# Inference

Given two sets: **labeled**  $D' = \{(\mathbf{y}_n, c_n)\}_{n=1}^N$  and **unlabeled**  $D'' = \{\mathbf{y}_n\}_{n=N+1}^M$



$$\Theta = \{\pi\} \cup \{\beta_s, \Lambda_s, \mu_s, \nu_s\}_{s=1}^S$$

$$\mathcal{H}' = \{\mathbf{x}_n, s_n\}_{n=1}^N$$

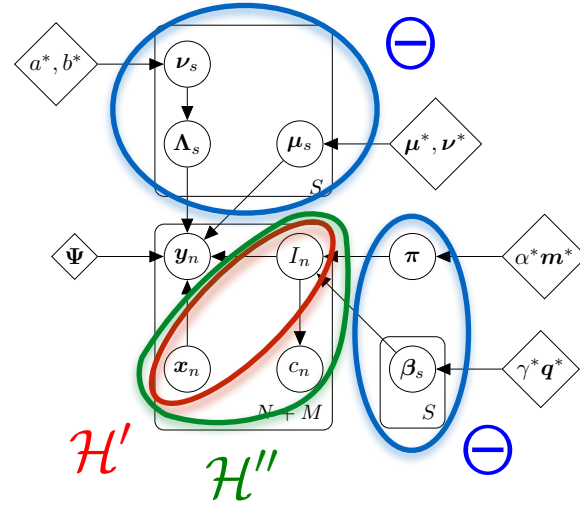
$$\mathcal{H}'' = \{\mathbf{x}_n, (s_n, \ell_n), c_n\}_{n=N+1}^M$$

$$\begin{aligned} \log p(D', D'') &= \log \int_{\Theta, \mathcal{H}', \mathcal{H}''} p(D', D'', \Theta, \mathcal{H}', \mathcal{H}'') \\ &= \log \int_{\Theta, \mathcal{H}', \mathcal{H}''} q(\Theta, \mathcal{H}', \mathcal{H}'') \frac{p(D', D'', \Theta, \mathcal{H}', \mathcal{H}'')}{q(\Theta, \mathcal{H}', \mathcal{H}'')} \\ &\geq \int_{\Theta, \mathcal{H}', \mathcal{H}''} q(\Theta, \mathcal{H}', \mathcal{H}'') \log \frac{p(D', D'', \Theta, \mathcal{H}', \mathcal{H}'')}{q(\Theta, \mathcal{H}', \mathcal{H}'')} = \mathcal{F}(q(\cdot)) \end{aligned}$$

$$q(\Theta, \mathcal{H}', \mathcal{H}'') = q(\Theta)q(\mathcal{H}'|\Theta)q(\mathcal{H}''|\Theta)$$

# Inference

Given two sets: **labeled**  $D' = \{(\mathbf{y}_n, c_n)\}_{n=1}^N$  and **unlabeled**  $D'' = \{\mathbf{y}_n\}_{n=N+1}^M$



$$\Theta = \{\pi\} \cup \{\beta_s, \Lambda_s, \mu_s, \nu_s\}_{s=1}^S$$

$$\mathcal{H}' = \{\mathbf{x}_n, s_n\}_{n=1}^N$$

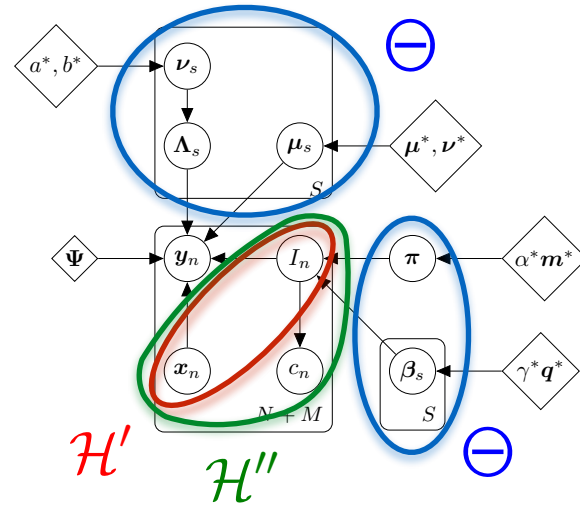
$$\mathcal{H}'' = \{\mathbf{x}_n, (s_n, \ell_n), c_n\}_{n=N+1}^M$$

$$\begin{aligned} \mathcal{F}(q(\cdot)) &= \int_{\Theta, \mathcal{H}', \mathcal{H}''} q(\Theta, \mathcal{H}', \mathcal{H}'') \log \frac{p(D', D'', \Theta, \mathcal{H}', \mathcal{H}'')}{q(\Theta, \mathcal{H}', \mathcal{H}'')} \\ &= \int_{\Theta, \mathcal{H}', \mathcal{H}''} q(\Theta) q(\mathcal{H}'|\Theta) q(\mathcal{H}''|\Theta) \log \frac{p(D', D'', \Theta, \mathcal{H}', \mathcal{H}'')}{q(\Theta) q(\mathcal{H}'|\Theta) q(\mathcal{H}''|\Theta)} \\ &= \int_{\Theta, \mathcal{H}', \mathcal{H}''} q(\Theta) q(\mathcal{H}'|\Theta) q(\mathcal{H}''|\Theta) \log \frac{p(D'|\Theta, \mathcal{H}') p(D''|\Theta, \mathcal{H}'') p(\mathcal{H}'|\Theta) p(\mathcal{H}''|\Theta) p(\Theta)}{q(\Theta) q(\mathcal{H}'|\Theta) q(\mathcal{H}''|\Theta)} \\ &= \int_{\Theta, \mathcal{H}', \mathcal{H}''} q(\Theta) q(\mathcal{H}'|\Theta) q(\mathcal{H}''|\Theta) \left[ \log \frac{p(\Theta)}{q(\Theta)} + \log \frac{p(D'|\Theta, \mathcal{H}') p(\mathcal{H}'|\Theta)}{q(\mathcal{H}'|\Theta)} + \log \frac{p(D''|\Theta, \mathcal{H}'') p(\mathcal{H}''|\Theta)}{q(\mathcal{H}''|\Theta)} \right] \\ &= \int_{\Theta} q(\Theta) \left[ \log \frac{p(\Theta)}{q(\Theta)} + \int_{\mathcal{H}'} q(\mathcal{H}'|\Theta) \log \frac{p(D'|\Theta, \mathcal{H}') p(\mathcal{H}'|\Theta)}{q(\mathcal{H}'|\Theta)} + \int_{\mathcal{H}''} q(\mathcal{H}''|\Theta) \log \frac{p(D''|\Theta, \mathcal{H}'') p(\mathcal{H}''|\Theta)}{q(\mathcal{H}''|\Theta)} \right] \end{aligned}$$

$q(\Theta, \mathcal{H}', \mathcal{H}'') = q(\Theta) q(\mathcal{H}'|\Theta) q(\mathcal{H}''|\Theta)$

# Inference

Given two sets: **labeled**  $D' = \{(\mathbf{y}_n, c_n)\}_{n=1}^N$  and **unlabeled**  $D'' = \{\mathbf{y}_n\}_{n=N+1}^M$



$$\Theta = \{\pi\} \cup \{\beta_s, \Lambda_s, \mu_s, \nu_s\}_{s=1}^S$$

$$\mathcal{H}' = \{\mathbf{x}_n, s_n\}_{n=1}^N$$

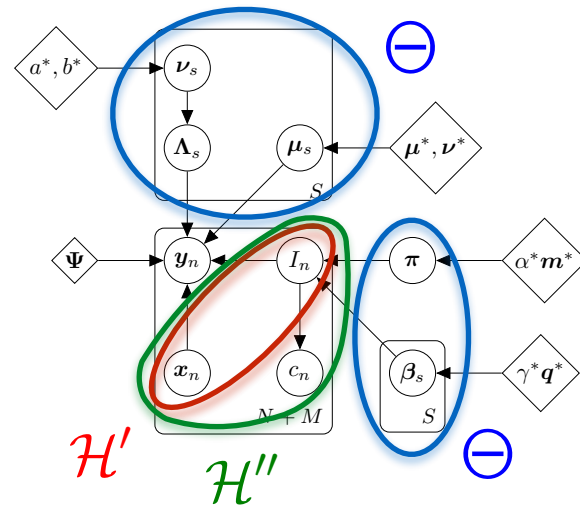
$$\mathcal{H}'' = \{\mathbf{x}_n, (s_n, \ell_n), c_n\}_{n=N+1}^M$$

$$\begin{aligned} \mathcal{F}(q(\cdot)) &= \int_{\Theta, \mathcal{H}', \mathcal{H}''} q(\Theta, \mathcal{H}', \mathcal{H}'') \log \frac{p(D', D'', \Theta, \mathcal{H}', \mathcal{H}'')}{q(\Theta, \mathcal{H}', \mathcal{H}'')} \\ &= \int_{\Theta, \mathcal{H}', \mathcal{H}''} q(\Theta) q(\mathcal{H}'|\Theta) q(\mathcal{H}''|\Theta) \log \frac{p(D', D'', \Theta, \mathcal{H}', \mathcal{H}'')}{q(\Theta) q(\mathcal{H}'|\Theta) q(\mathcal{H}''|\Theta)} \\ &= \int_{\Theta, \mathcal{H}', \mathcal{H}''} q(\Theta) q(\mathcal{H}'|\Theta) q(\mathcal{H}''|\Theta) \log \frac{p(D'|\Theta, \mathcal{H}') p(D''|\Theta, \mathcal{H}'') p(\mathcal{H}'|\Theta) p(\mathcal{H}''|\Theta) p(\Theta)}{q(\Theta) q(\mathcal{H}'|\Theta) q(\mathcal{H}''|\Theta)} \\ &= \int_{\Theta, \mathcal{H}', \mathcal{H}''} q(\Theta) q(\mathcal{H}'|\Theta) q(\mathcal{H}''|\Theta) \left[ \log \frac{p(\Theta)}{q(\Theta)} + \log \frac{p(D'|\Theta, \mathcal{H}') p(\mathcal{H}'|\Theta)}{q(\mathcal{H}'|\Theta)} + \log \frac{p(D''|\Theta, \mathcal{H}'') p(\mathcal{H}''|\Theta)}{q(\mathcal{H}''|\Theta)} \right] \\ &= \int_{\Theta} q(\Theta) \left[ \log \frac{p(\Theta)}{q(\Theta)} + \int_{\mathcal{H}'} q(\mathcal{H}'|\Theta) \log \frac{p(D'|\Theta, \mathcal{H}') p(\mathcal{H}'|\Theta)}{q(\mathcal{H}'|\Theta)} + \int_{\mathcal{H}''} q(\mathcal{H}''|\Theta) \log \frac{p(D''|\Theta, \mathcal{H}'') p(\mathcal{H}''|\Theta)}{q(\mathcal{H}''|\Theta)} \right] \end{aligned}$$

Compute functional derivatives with respect to  $q(\Theta)$ ,  $q(\mathcal{H}'|\Theta)$ ,  $q(\mathcal{H}''|\Theta)$  and equate them to 0.

# Prediction

Given two sets: **labeled**  $D' = \{(\mathbf{y}_n, c_n)\}_{n=1}^N$  and **unlabeled**  $D'' = \{\mathbf{y}_n\}_{n=N+1}^M$



$$\Theta = \{\pi\} \cup \{\beta_s, \Lambda_s, \mu_s, \nu_s\}_{s=1}^S$$

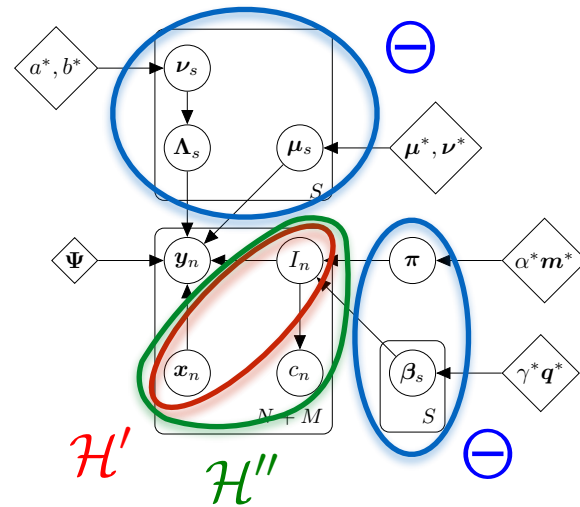
$$\mathcal{H}' = \{\mathbf{x}_n, s_n\}_{n=1}^N$$

$$\mathcal{H}'' = \{\mathbf{x}_n, (s_n, \ell_n), c_n\}_{n=N+1}^M$$

$$q(\Theta), q(\mathcal{H}'|\Theta), q(\mathcal{H}''|\Theta)$$

# Prediction

Given two sets: **labeled**  $D' = \{(\mathbf{y}_n, c_n)\}_{n=1}^N$  and **unlabeled**  $D'' = \{\mathbf{y}_n\}_{n=N+1}^M$



$$\Theta = \{\pi\} \cup \{\beta_s, \Lambda_s, \mu_s, \nu_s\}_{s=1}^S$$

$$\mathcal{H}' = \{\mathbf{x}_n, s_n\}_{n=1}^N$$

$$\mathcal{H}'' = \{\mathbf{x}_n, (s_n, \ell_n), c_n\}_{n=N+1}^M$$

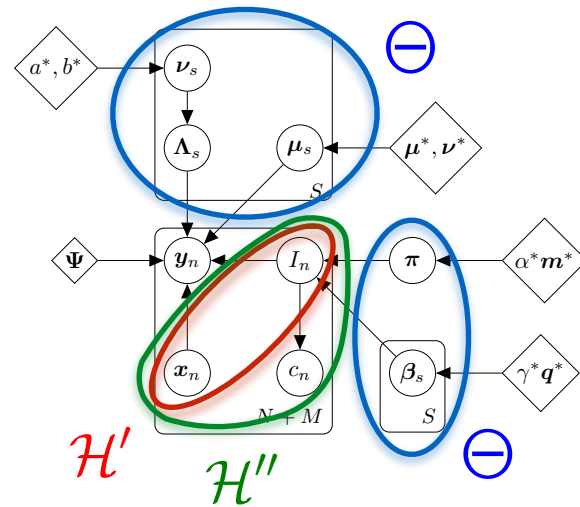
$$q(\Theta), q(\mathcal{H}'|\Theta), q(\mathcal{H}''|\Theta)$$



$$\begin{aligned} \log p(\mathbf{y}_t | D', D'') &= \log \int_{\Theta, \{\mathbf{x}_t, l_t\}} p(\mathbf{y}_t, \mathbf{x}_t, l_t, \Theta | D', D'') \\ &= \log \int_{\Theta} p(\Theta | D', D'') \left[ \int_{\{\mathbf{x}_t, l_t\}} p(\mathbf{y}_t, \mathbf{x}_t, l_t, \Theta | D', D'') \right] \\ &= \log \int_{\Theta} p(\Theta | D', D'') \left[ \int_{\{\mathbf{x}_t, l_t\}} q(\mathbf{x}_t, l_t) \frac{p(\mathbf{y}_t, \mathbf{x}_t, l_t, \Theta | D', D'')}{q(\mathbf{x}_t, l_t)} \right] \\ &= \log \int_{\Theta} p(\Theta | D', D'') \left[ \int_{\{\mathbf{x}_t, l_t\}} q(\mathbf{x}_t, l_t) \frac{p(\mathbf{y}_t | \mathbf{x}_t, l_t, \Theta) p(\mathbf{x}_t, l_t | \Theta)}{q(\mathbf{x}_t, l_t)} \right] \\ &\geq \int_{\Theta} p(\Theta | D', D'') \left[ \int_{\{\mathbf{x}_t, l_t\}} q(\mathbf{x}_t, l_t) \log \frac{p(\mathbf{y}_t | \mathbf{x}_t, l_t, \Theta) p(\mathbf{x}_t, l_t | \Theta)}{q(\mathbf{x}_t, l_t)} \right] \\ &\approx \int_{\Theta} q(\Theta) \left[ \int_{\{\mathbf{x}_t, l_t\}} q(\mathbf{x}_t, l_t) \log \frac{p(\mathbf{y}_t | \mathbf{x}_t, l_t, \Theta) p(\mathbf{x}_t, l_t | \Theta)}{q(\mathbf{x}_t, l_t)} \right] \end{aligned}$$

# Prediction

Given two sets: **labeled**  $D' = \{(\mathbf{y}_n, c_n)\}_{n=1}^N$  and **unlabeled**  $D'' = \{\mathbf{y}_n\}_{n=N+1}^M$



$$\Theta = \{\pi\} \cup \{\beta_s, \Lambda_s, \mu_s, \nu_s\}_{s=1}^S$$

$$\mathcal{H}' = \{\mathbf{x}_n, s_n\}_{n=1}^N$$

$$\mathcal{H}'' = \{\mathbf{x}_n, (s_n, \ell_n), c_n\}_{n=N+1}^M$$

$$q(\Theta), q(\mathcal{H}'|\Theta), q(\mathcal{H}''|\Theta)$$



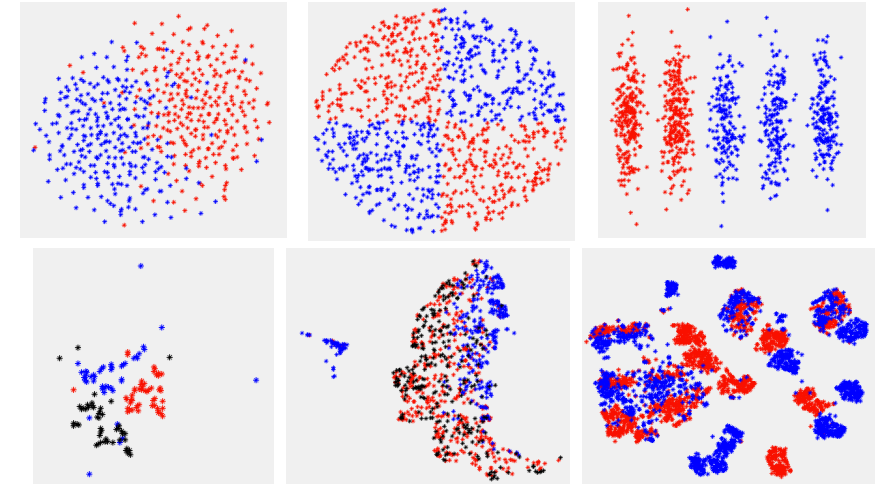
$$\begin{aligned} \log p(\mathbf{y}_t | D', D'') &= \log \int_{\Theta, \{\mathbf{x}_t, l_t\}} p(\mathbf{y}_t, \mathbf{x}_t, l_t, \Theta | D', D'') \\ &= \log \int_{\Theta} p(\Theta | D', D'') \left[ \int_{\{\mathbf{x}_t, l_t\}} p(\mathbf{y}_t, \mathbf{x}_t, l_t, \Theta | D', D'') \right] \\ &= \log \int_{\Theta} p(\Theta | D', D'') \left[ \int_{\{\mathbf{x}_t, l_t\}} q(\mathbf{x}_t, l_t) \frac{p(\mathbf{y}_t, \mathbf{x}_t, l_t, \Theta | D', D'')}{q(\mathbf{x}_t, l_t)} \right] \\ &= \log \int_{\Theta} p(\Theta | D', D'') \left[ \int_{\{\mathbf{x}_t, l_t\}} q(\mathbf{x}_t, l_t) \frac{p(\mathbf{y}_t | \mathbf{x}_t, l_t, \Theta) p(\mathbf{x}_t, l_t | \Theta)}{q(\mathbf{x}_t, l_t)} \right] \\ &\geq \int_{\Theta} p(\Theta | D', D'') \left[ \int_{\{\mathbf{x}_t, l_t\}} q(\mathbf{x}_t, l_t) \log \frac{p(\mathbf{y}_t | \mathbf{x}_t, l_t, \Theta) p(\mathbf{x}_t, l_t | \Theta)}{q(\mathbf{x}_t, l_t)} \right] \\ &\approx \int_{\Theta} q(\Theta) \left[ \int_{\{\mathbf{x}_t, l_t\}} q(\mathbf{x}_t, l_t) \log \frac{p(\mathbf{y}_t | \mathbf{x}_t, l_t, \Theta) p(\mathbf{x}_t, l_t | \Theta)}{q(\mathbf{x}_t, l_t)} \right] \end{aligned}$$

Compute  $q(\mathbf{x}_t, l_t)$  for a test sample



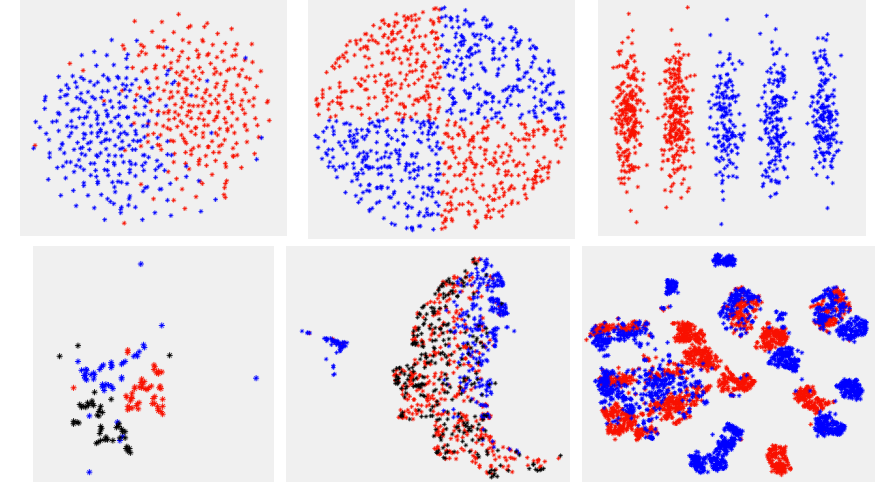
# Experiments

| Data sets | Classes | Features | Instances |
|-----------|---------|----------|-----------|
| G50C      | 2       | 50       | 550       |
| CAKE      | 2       | 2        | 1000      |
| TOES      | 2       | 2        | 1000      |
| IRIS      | 3       | 4        | 150       |
| USPS      | 3       | 256      | 1918      |
| ISOLET    | 2       | 617      | 3119      |

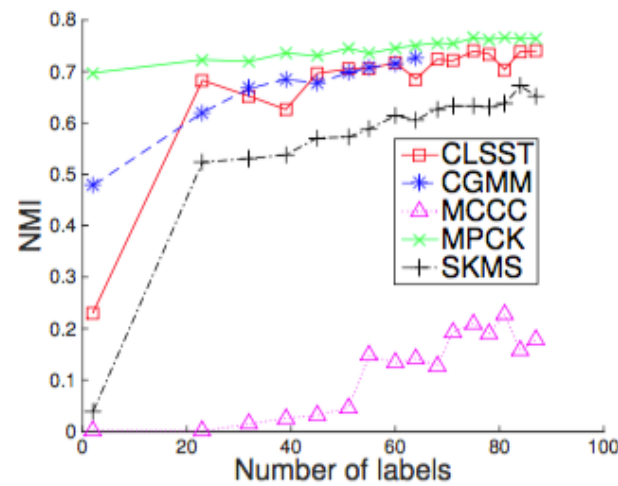


# Experiments

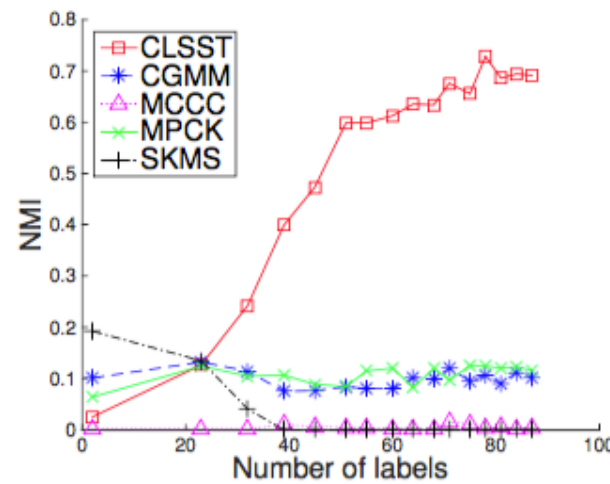
| Data sets | Classes | Features | Instances |
|-----------|---------|----------|-----------|
| G50C      | 2       | 50       | 550       |
| CAKE      | 2       | 2        | 1000      |
| TOES      | 2       | 2        | 1000      |
| IRIS      | 3       | 4        | 150       |
| USPS      | 3       | 256      | 1918      |
| ISOLET    | 2       | 617      | 3119      |



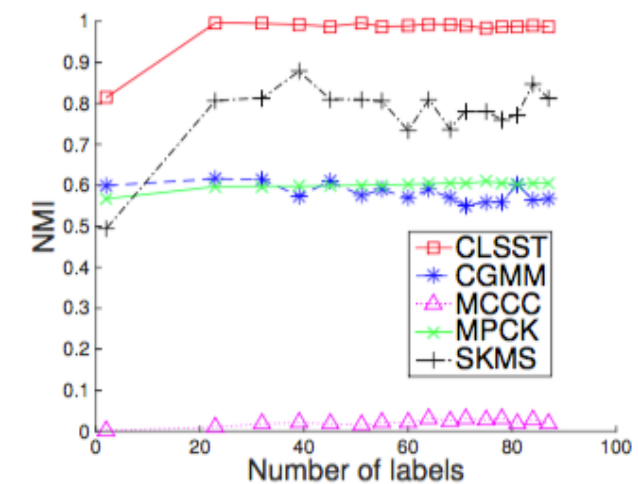
## Clustering



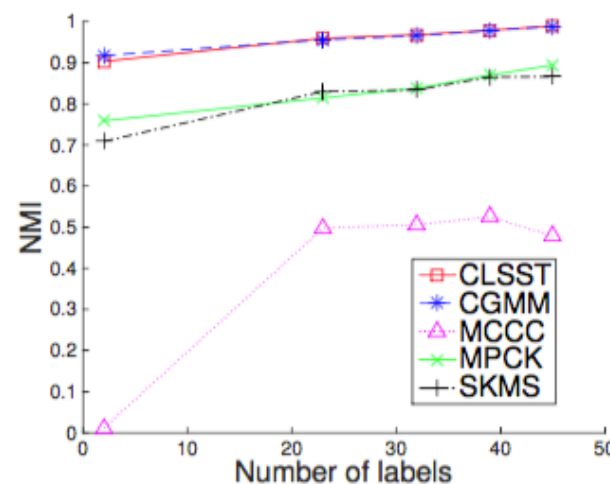
(a) G50C



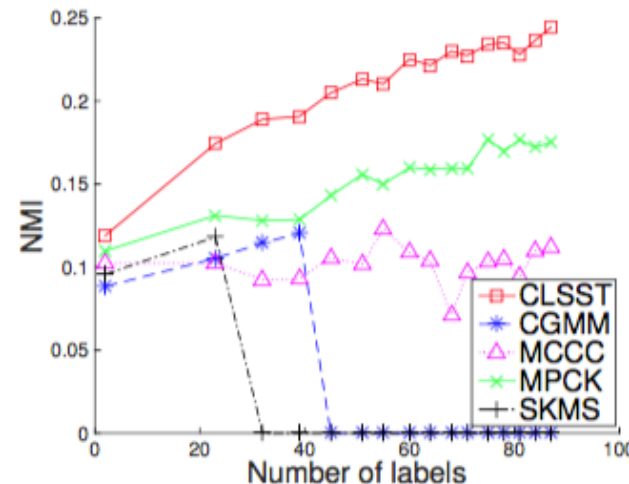
(b) CAKE



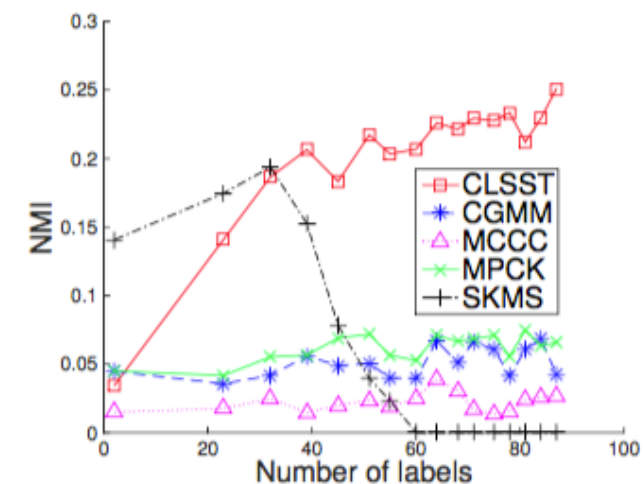
(c) TOES



(d) IRIS



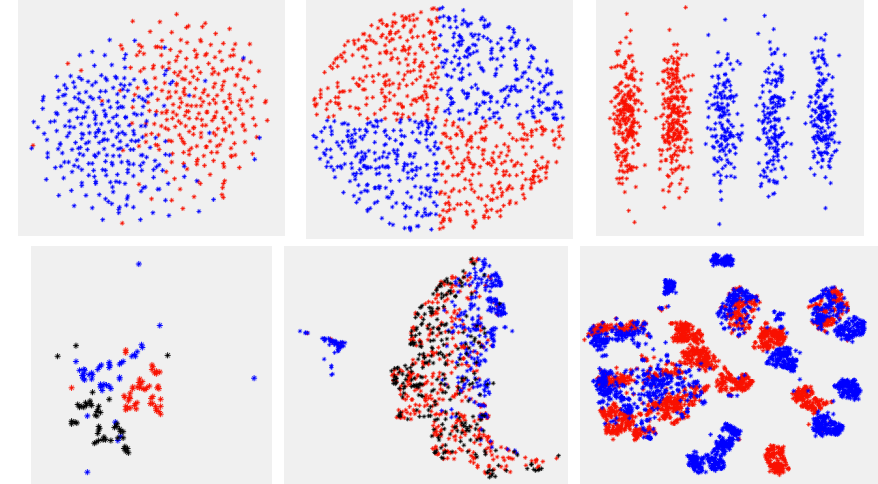
(e) USPS



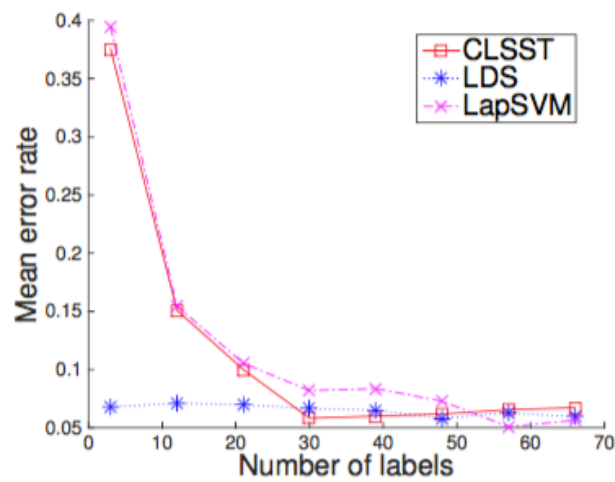
(f) ISOLET

# Experiments

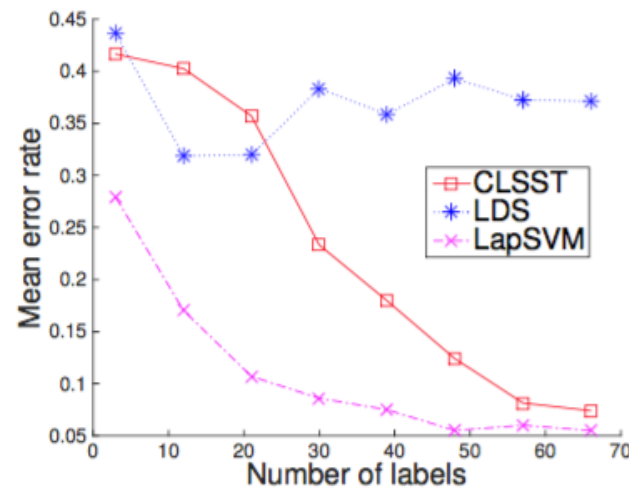
| Data sets | Classes | Features | Instances |
|-----------|---------|----------|-----------|
| G50C      | 2       | 50       | 550       |
| CAKE      | 2       | 2        | 1000      |
| TOES      | 2       | 2        | 1000      |
| IRIS      | 3       | 4        | 150       |
| USPS      | 3       | 256      | 1918      |
| ISOLET    | 2       | 617      | 3119      |



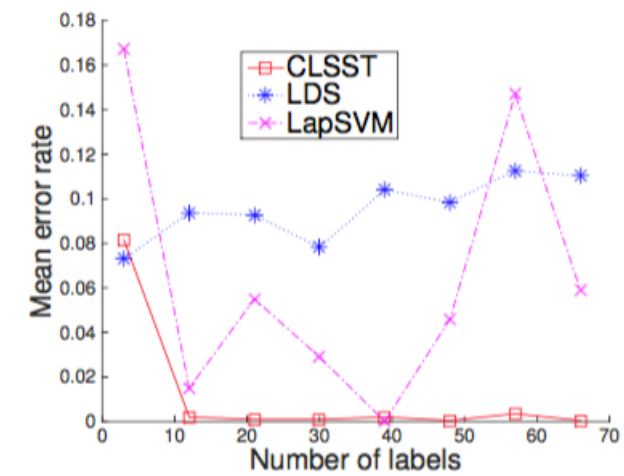
## Classification



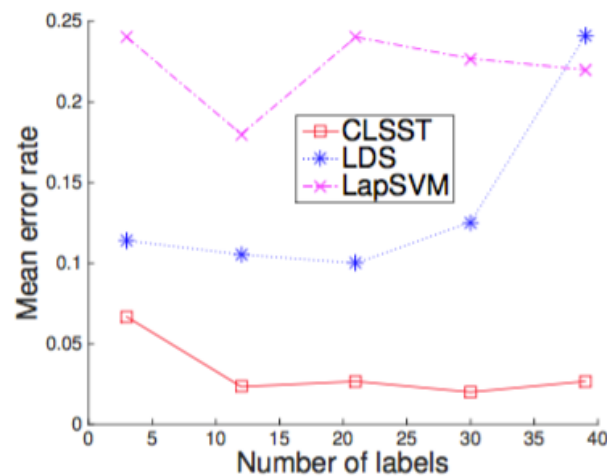
(a) G50C



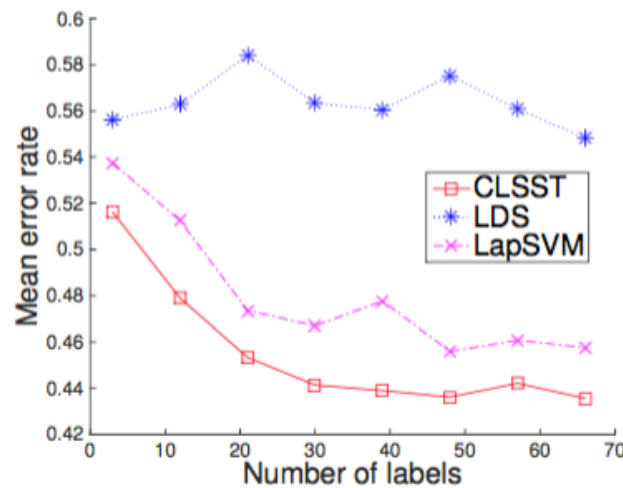
(b) CAKE



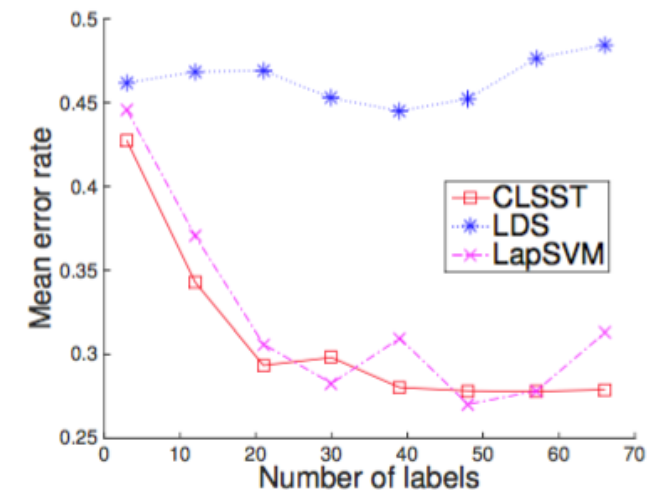
(c) TOES



(d) IRIS



(e) USPS



(f) ISOLET

# Experiments

| <b>Dataset</b>                | <b>Classes</b> | <b>Features</b> | <b>Instances</b> |
|-------------------------------|----------------|-----------------|------------------|
| Breast cancer<br>(discovery)  | 5              | 754             | 997              |
| Breast cancer<br>(validation) | 5              | 754             | 995              |

# Experiments

| Dataset                    | Classes | Features | Instances |
|----------------------------|---------|----------|-----------|
| Breast cancer (discovery)  | 5       | 754      | 997       |
| Breast cancer (validation) | 5       | 754      | 995       |

| Cluster     | [13]   | CLSST (fixed $S$ ) | CLSST (variable $S$ ) |
|-------------|--------|--------------------|-----------------------|
| 1           | 0.8235 | 0.9266             | 0.9117                |
| 2           | 0.8099 | 0.8639             | 0.8377                |
| 3           | 0.7281 | 0.7899             | 0.7931                |
| 4           | 0.7091 | 0.6867             | 0.7730                |
| 5           | 0.6866 | 0.6842             | 0.7624                |
| 6           | 0.6455 | 0.6794             | 0.5833                |
| 7           | 0.6015 | 0.6780             | 0.5745                |
| 8           | 0.5818 | 0.6000             | -                     |
| 9           | 0.5072 | 0.5965             | -                     |
| 10          | 0.4481 | 0.5574             | -                     |
| <b>Avg.</b> | 0.654  | <b>0.706</b>       | <b>0.748</b>          |
| <b>Min.</b> | 0.448  | <b>0.557</b>       | <b>0.575</b>          |
| <b>Max.</b> | 0.824  | <b>0.927</b>       | <b>0.912</b>          |

IGP is increased at least of 5%! But further analysis is required to prove the biological relevance.

## Conclusions & Future work

- Proposed model based on MFA for SSL (clustering/classification)
- Clustering: handling multi-groups per class + problem of cluster assumption
- Classification: discovered clusters help classification (comparison with discriminative approaches)
- Real-world problem: promising results (future research)

**Thank You**