Classtering: Joint Classification and Clustering with Mixture of Factor Analysers

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

UNIVERSITY OF TRENTO - Italy

Slides and code @ http://emsansone.github.io/publications/

Motivation

Semi-supervised learning (SSL)

Classification

Clustering

- Problem of label propagation
- Cluster assumption

Discriminative

$$f : y \rightarrow c$$

Generative

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

Motivation

Semi-supervised learning (SSL)

Classification

Clustering

- Problem of label propagation
- Cluster assumption

Desired:
- No wrong label propagation
- Relaxing the cluster assumption

Discriminative

$$f : y \rightarrow c$$

Generative

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

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

Motivation

Semi-supervised learning (SSL)

Classification

Discriminative

$$f : y \rightarrow c$$

Generative

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

Desired:
- Model inter- and intra-class variabilities
- Achieve possibly "good performance"

Clustering

- Problem of label propagation
- Cluster assumption

Desired:
- No wrong label propagation
- Relaxing the cluster assumption

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

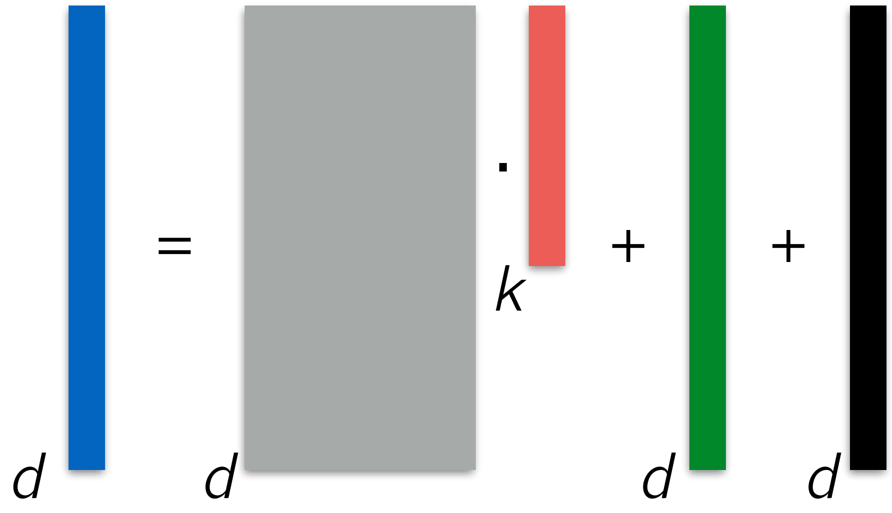
Model based on Mixture of Factor Analysers (MFA)

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

Model

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

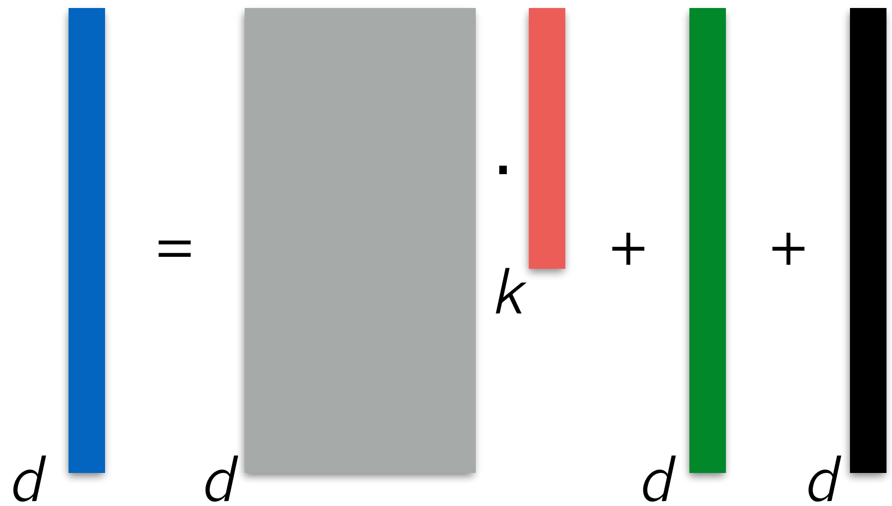
$$\boldsymbol{y}_n = \Lambda \boldsymbol{x}_n + \boldsymbol{\mu} + \boldsymbol{\xi}$$

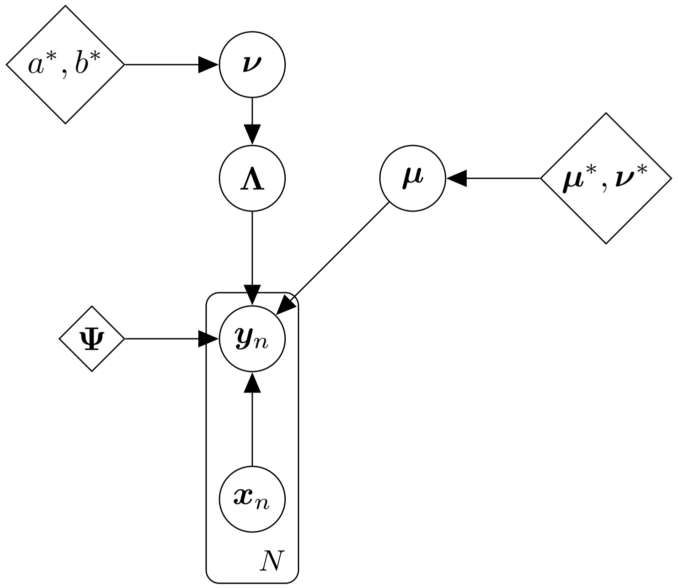


Model

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

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





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

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

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

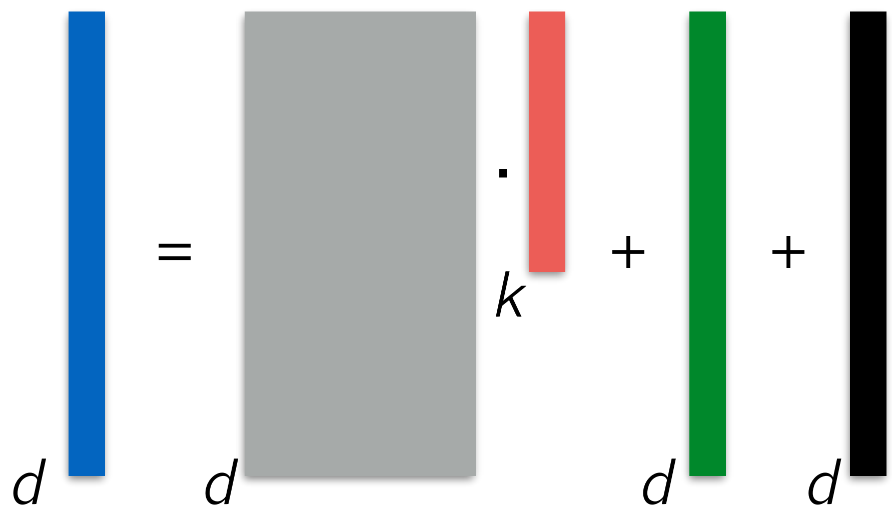
$$\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$$

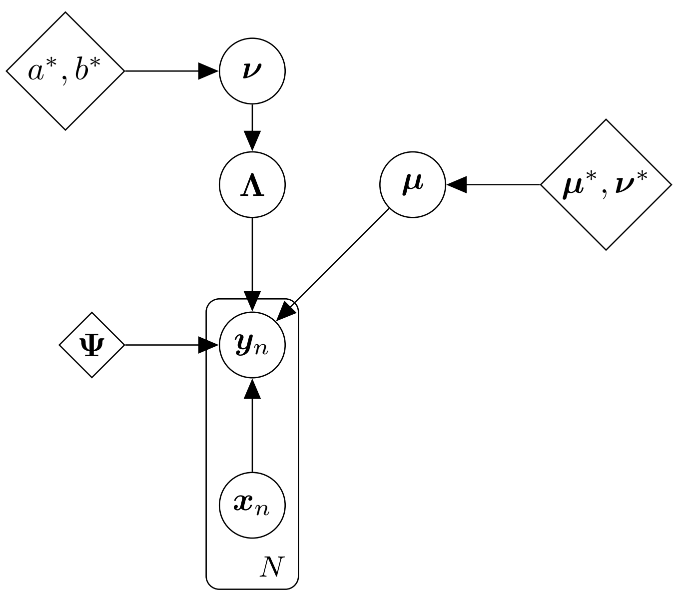
$$\boldsymbol{\nu} \sim \prod_{j=1}^{k} \text{Gamma}(\boldsymbol{\nu}(j) | a^*, b^*)$$

Model

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

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





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

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

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

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

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

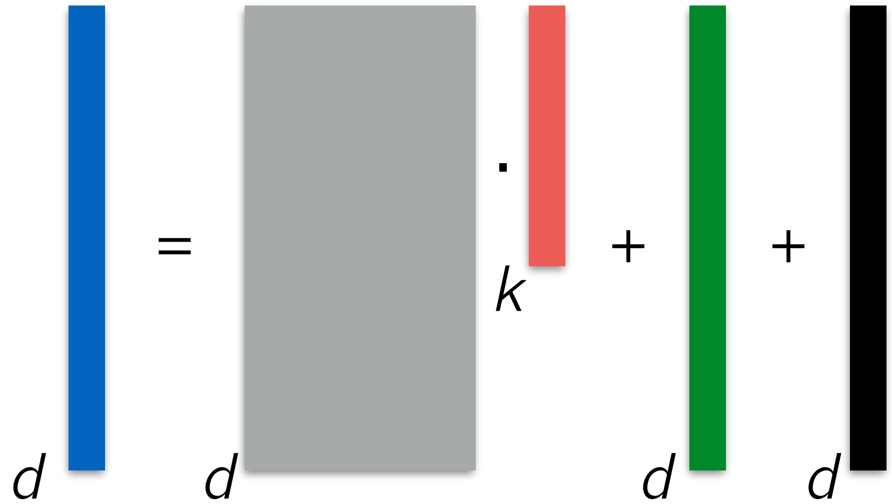
$$p(\boldsymbol{y}_n | \boldsymbol{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\Psi}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})$$

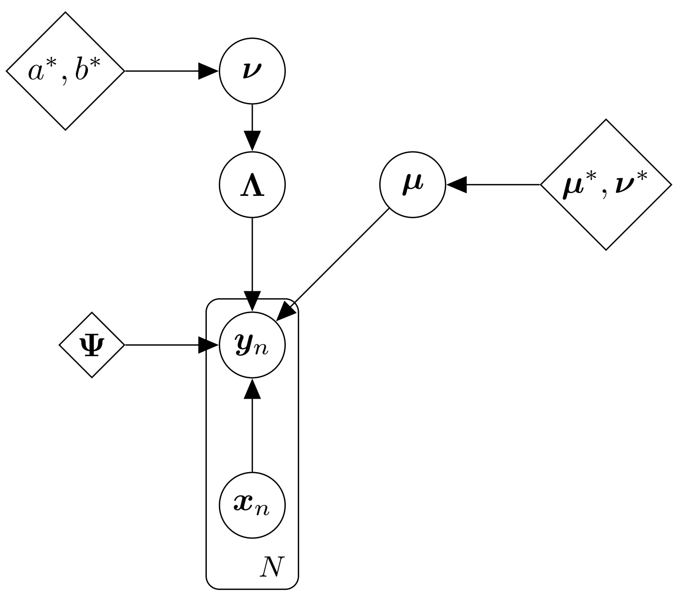
Analyser is described only by $\boldsymbol{\Lambda}, \boldsymbol{\mu}$

Model

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

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

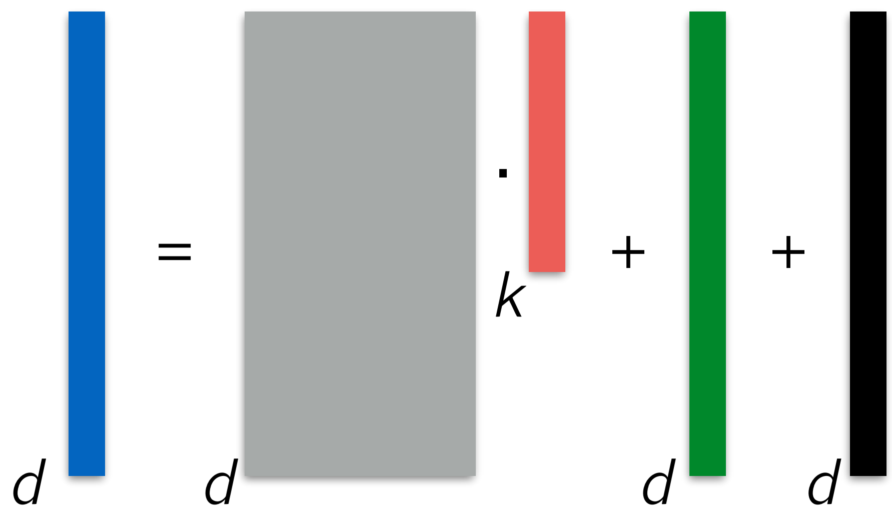


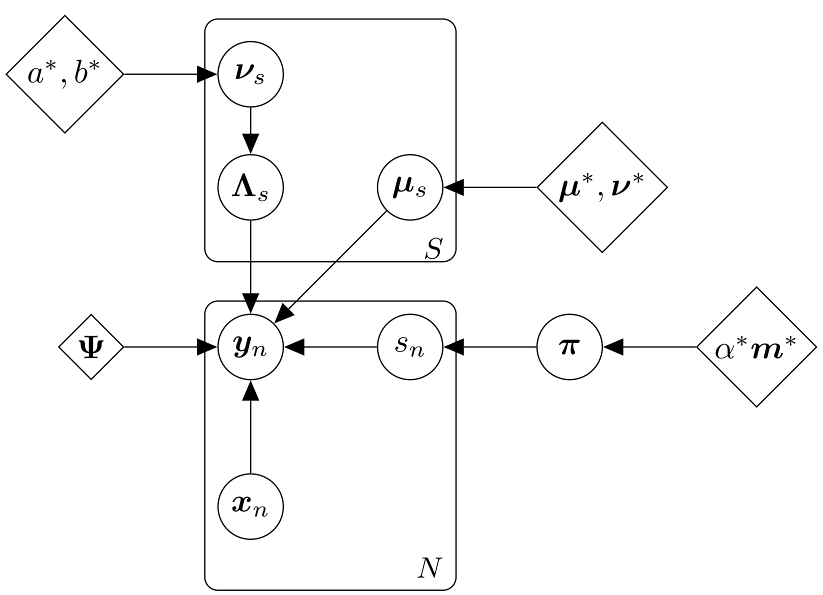


Model

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

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

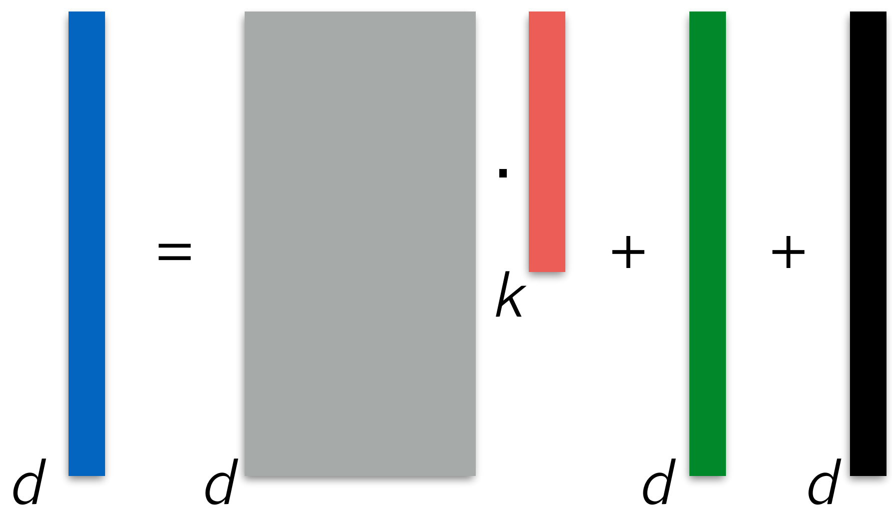


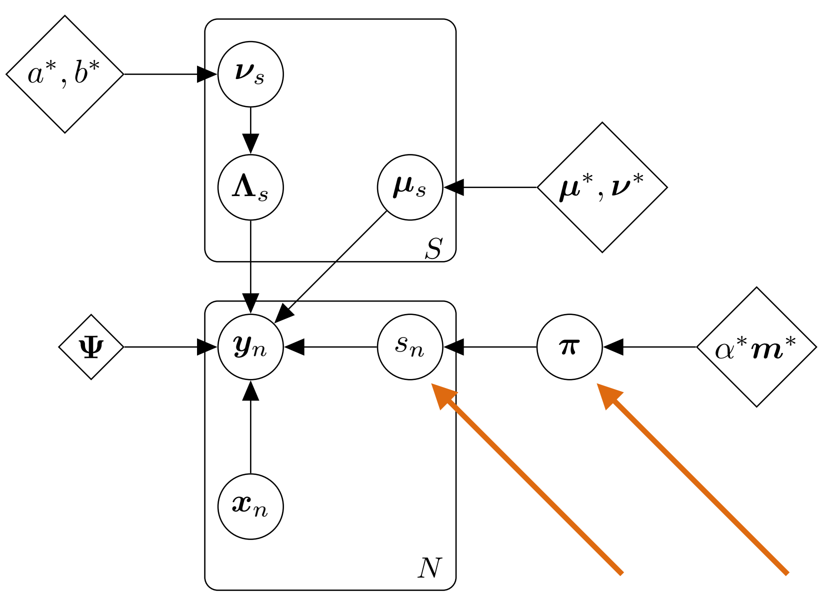


Model

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

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





$$s_n \sim \prod_{s=1}^{S} \boldsymbol{\pi}_s^{1_{s_n}(s)}$$

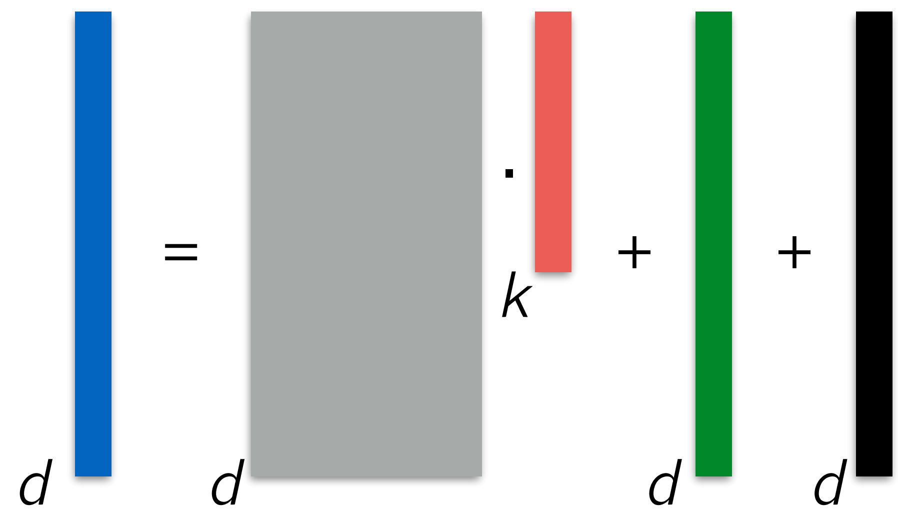
$$\boldsymbol{\pi} \sim \text{Dir}(\alpha^* \boldsymbol{m}^*)$$

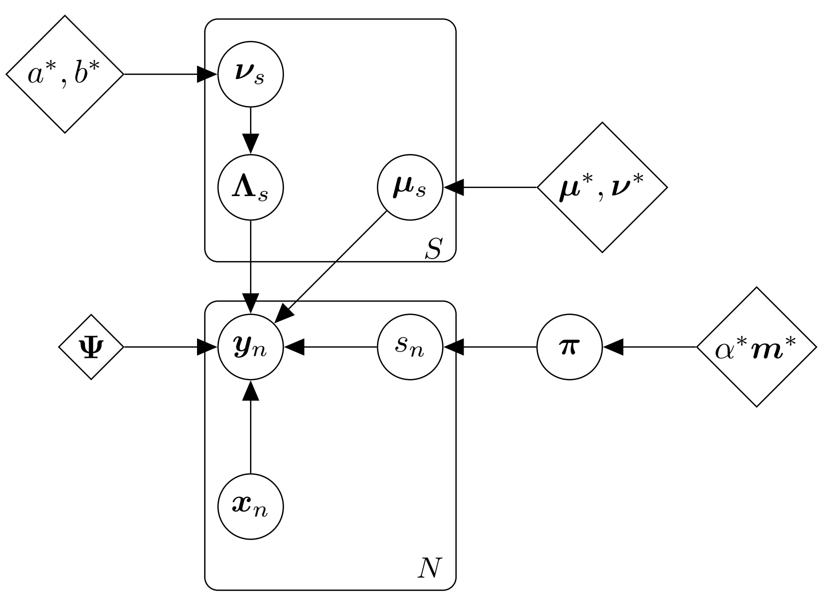
$$\boldsymbol{m}^* = [1/S, \ldots, 1/S]$$

Model

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

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





$$s_n \sim \prod_{s=1}^{S} \boldsymbol{\pi}_s^{1_{s_n}(s)}$$

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

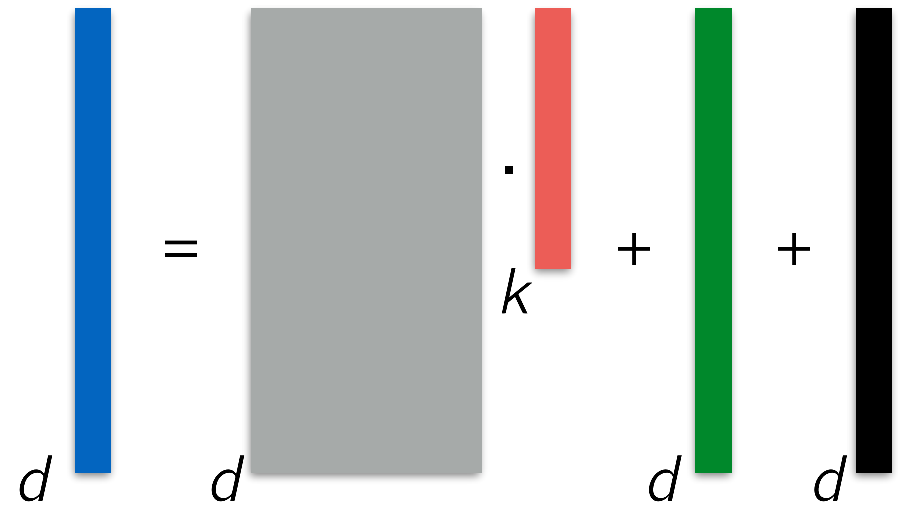
$$\boldsymbol{m}^* = [1/S, \ldots, 1/S]$$

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

Model

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

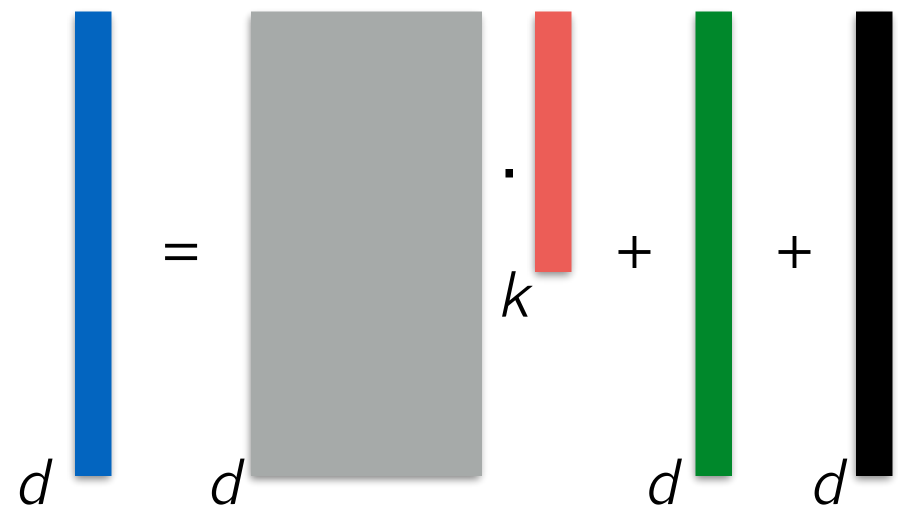
$$\boldsymbol{y}_n = \boldsymbol{\Lambda}_{s_n} \boldsymbol{x}_n + \boldsymbol{\mu}_{s_n} + \boldsymbol{\xi}$$

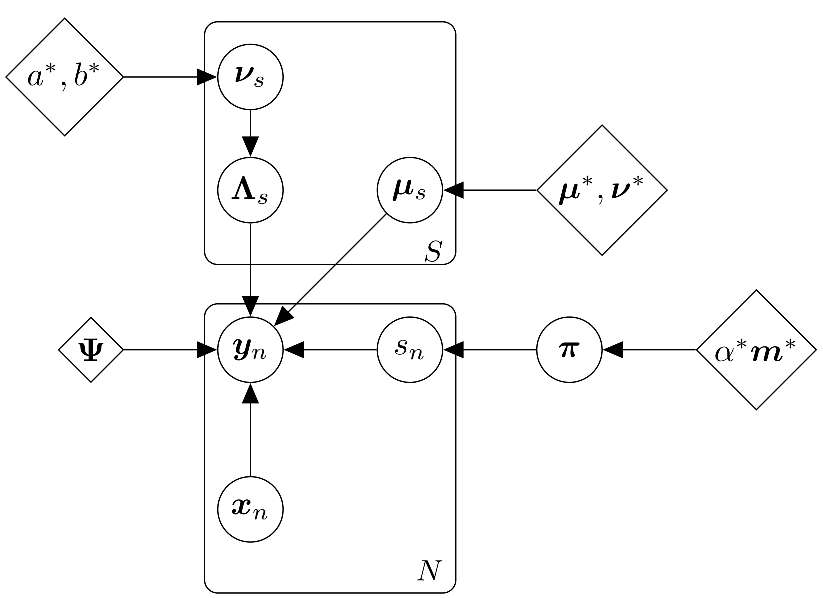


Model

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

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

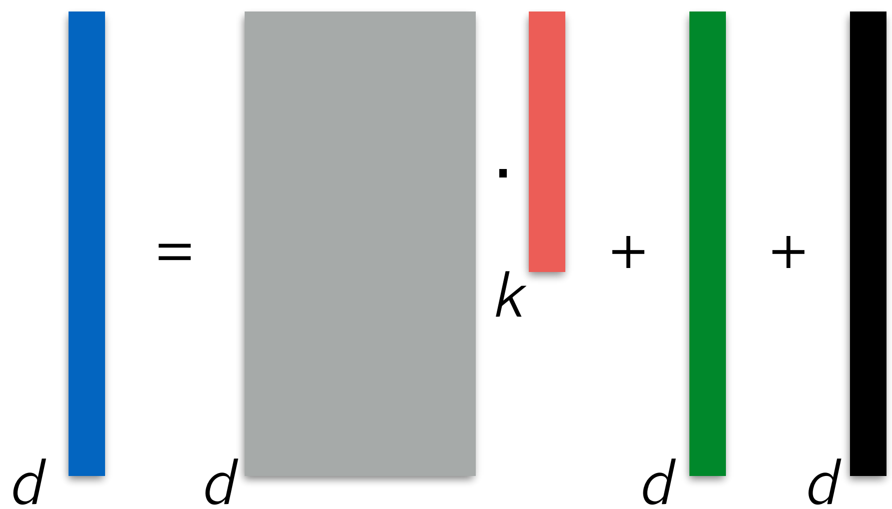


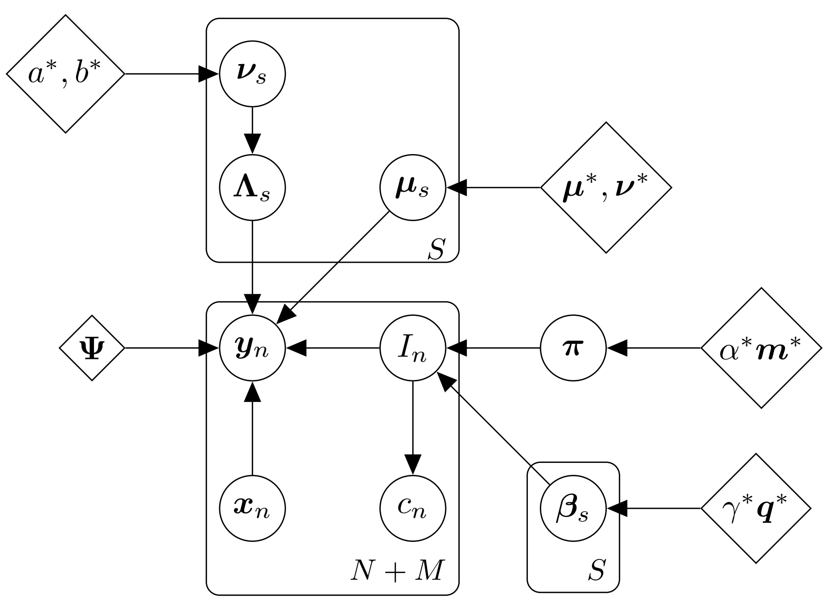


Model

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

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

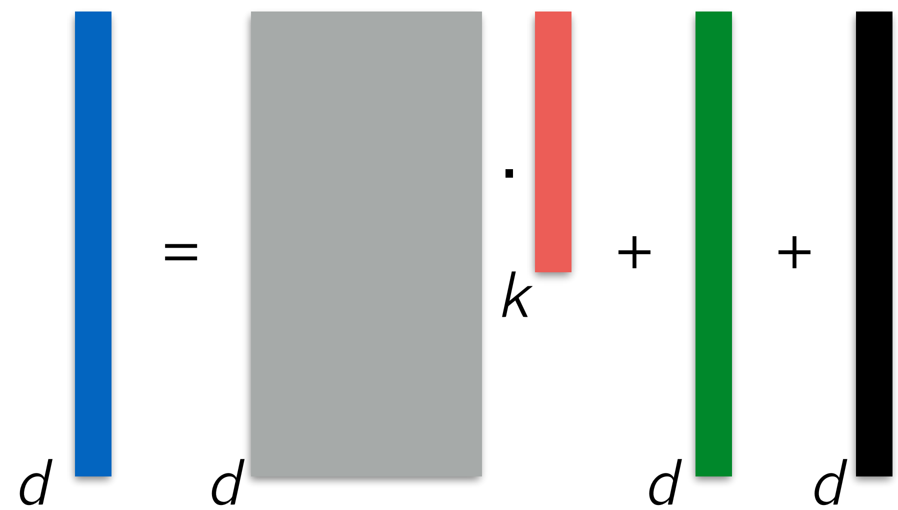


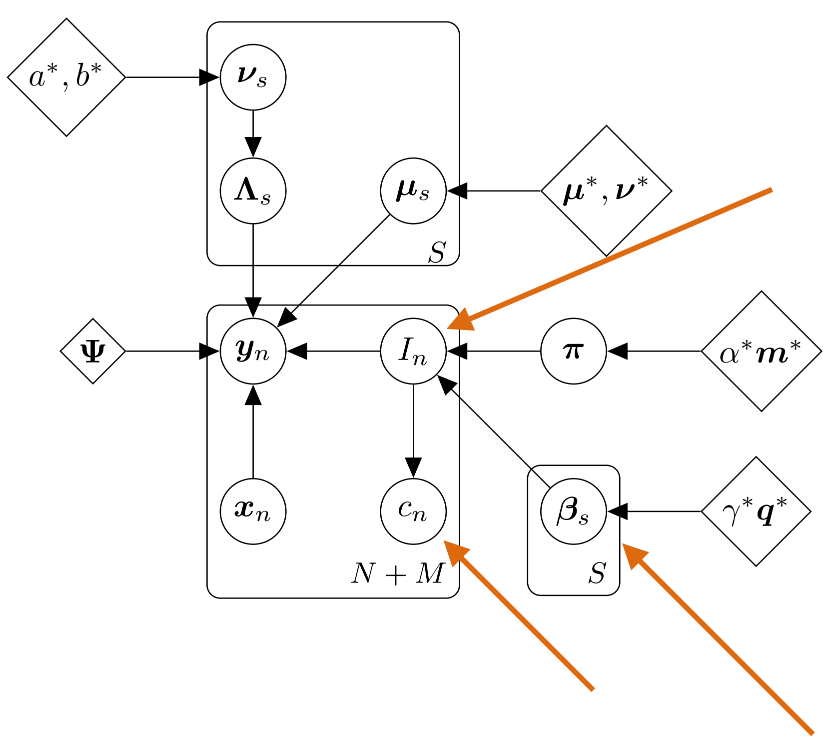


Model

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

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





$$I_n = (s_n, \ell_n)$$

$$\ell_n \sim \prod_{\ell=1}^{K} \boldsymbol{\beta}_{s_n}(\ell)^{1_{\ell_n}(\ell)}$$

$$c_n \sim \delta(c_n - \ell_n)$$

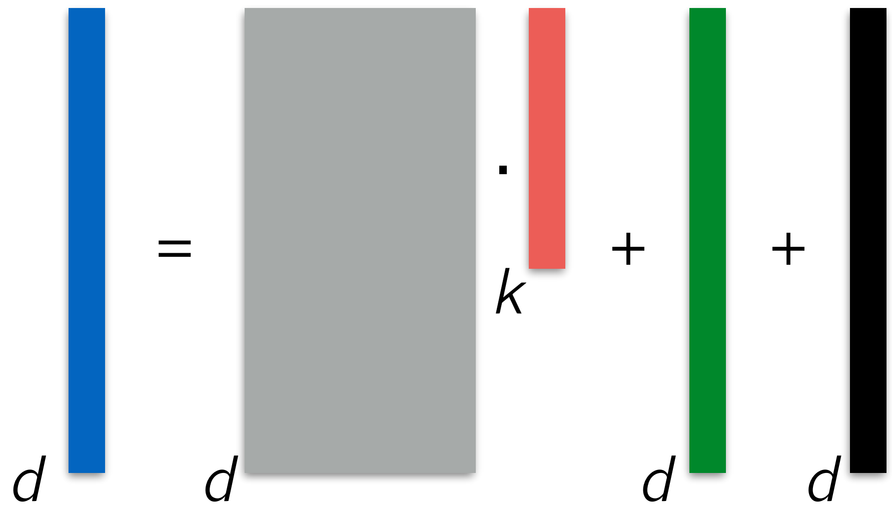
$$\boldsymbol{\beta}_s \sim \text{Dir}(\gamma^* \boldsymbol{q}^*)$$

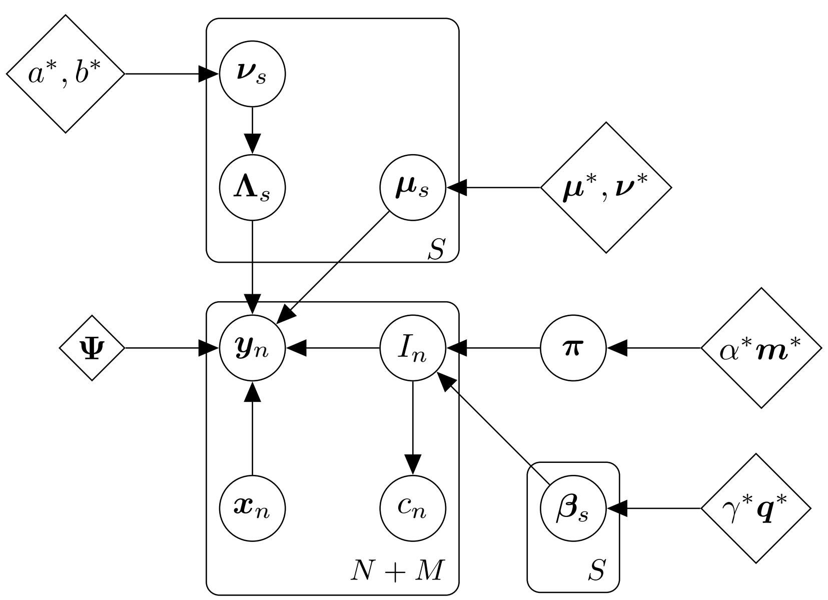
$$\boldsymbol{q}^* = [1/K, \ldots, 1/K]$$

Model

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

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





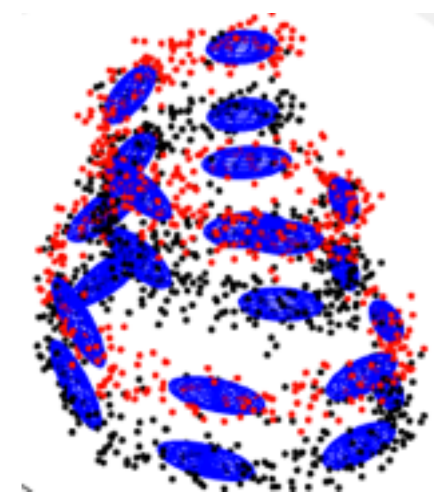
$$I_n = (s_n, \ell_n)$$

$$\ell_n \sim \prod_{\ell=1}^{K} \boldsymbol{\beta}_{s_n}(\ell)^{1_{\ell_n}(\ell)}$$

$$c_n \sim \delta(c_n - \ell_n)$$

$$\boldsymbol{\beta}_s \sim \mathrm{Dir}(\gamma^* \boldsymbol{q}^*)$$

$$\boldsymbol{q}^* = [1/K, \ldots, 1/K]$$

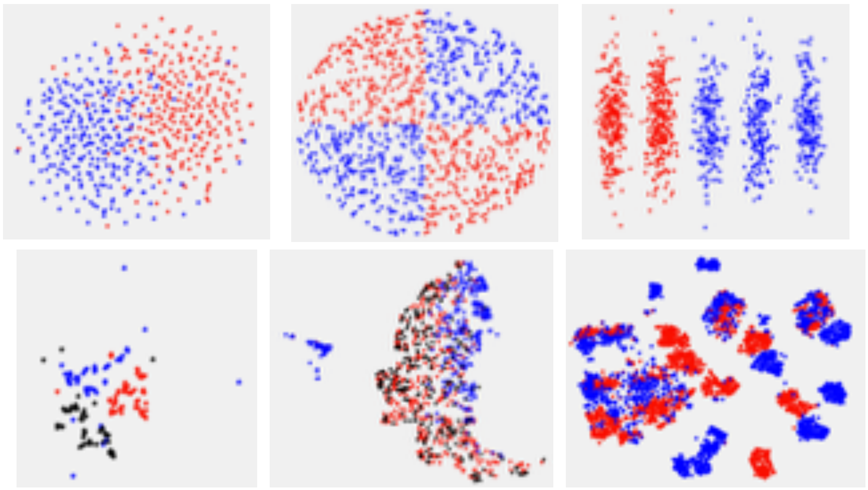


$$p(\boldsymbol{y}_n | \boldsymbol{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\Psi}) \sim \sum_{s_n=1}^{S} \boldsymbol{\pi}_{s_n} \mathcal{N}(\boldsymbol{\mu}_{s_n}, \boldsymbol{\Lambda}_{s_n}\boldsymbol{\Lambda}_{s_n}^T + \boldsymbol{\Psi})$$

Information about clusters vs. classes

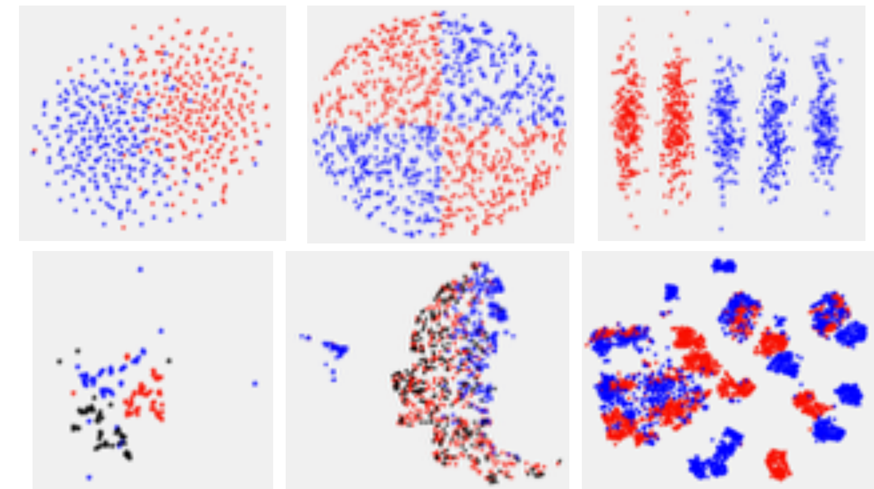
Experiments

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

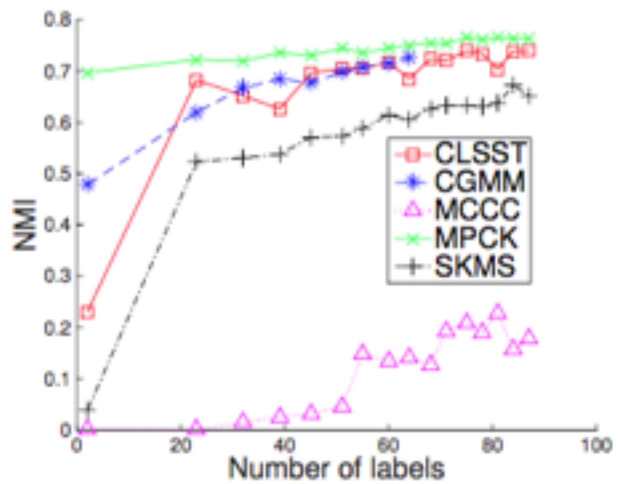


Experiments

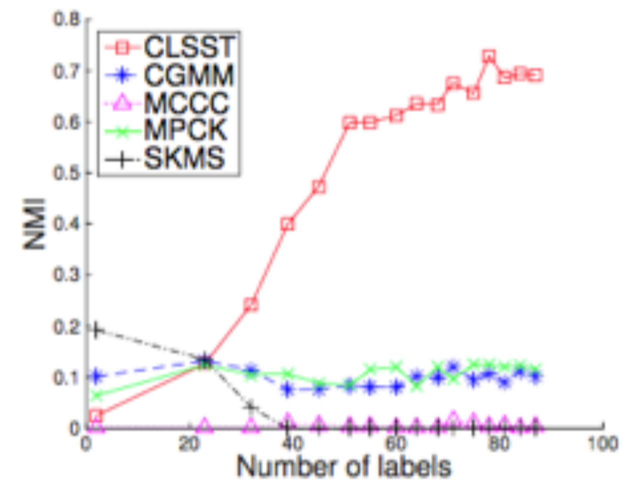
Data sets	Classes	Features	Instances
G50C	2	50	550
CAKE	2	2	1000
TOES	2	2	1000
IRIS	3	4	150
USPS	3	256	1918
ISOLET	2	617	3119



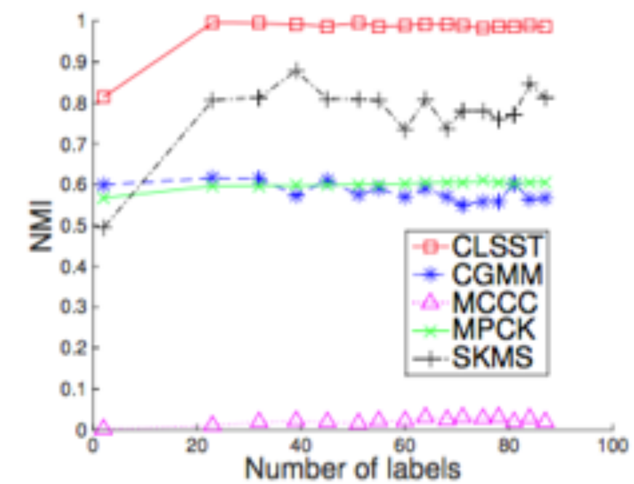
Clustering



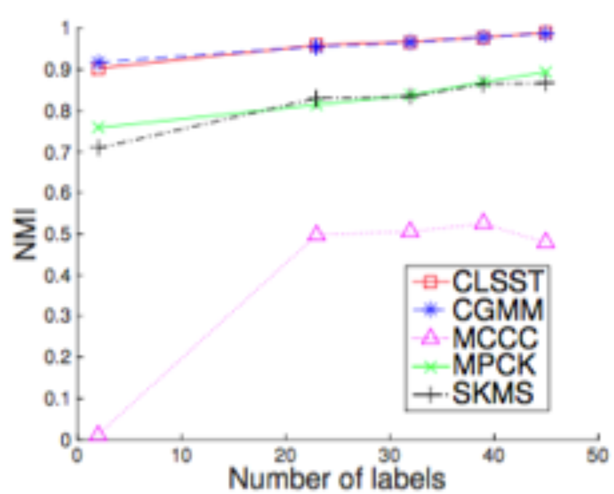
(a) G50C



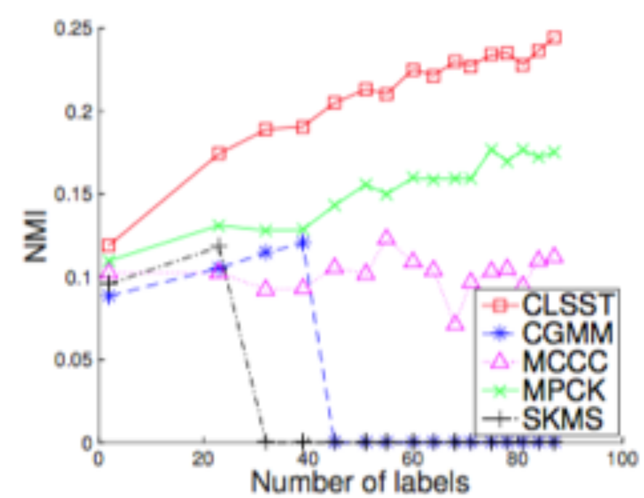
(b) CAKE



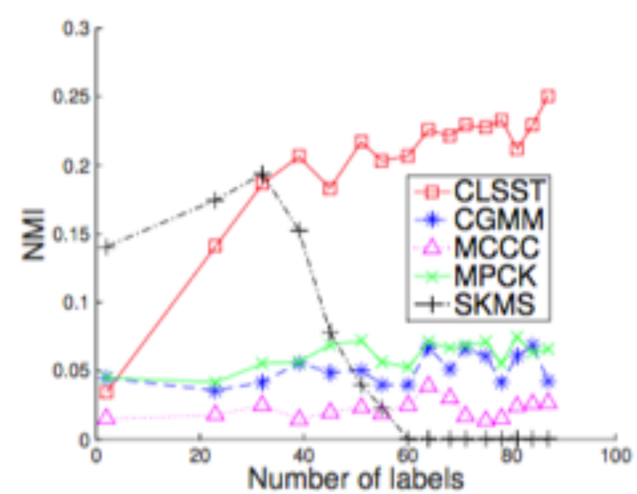
(c) TOES



(d) IRIS



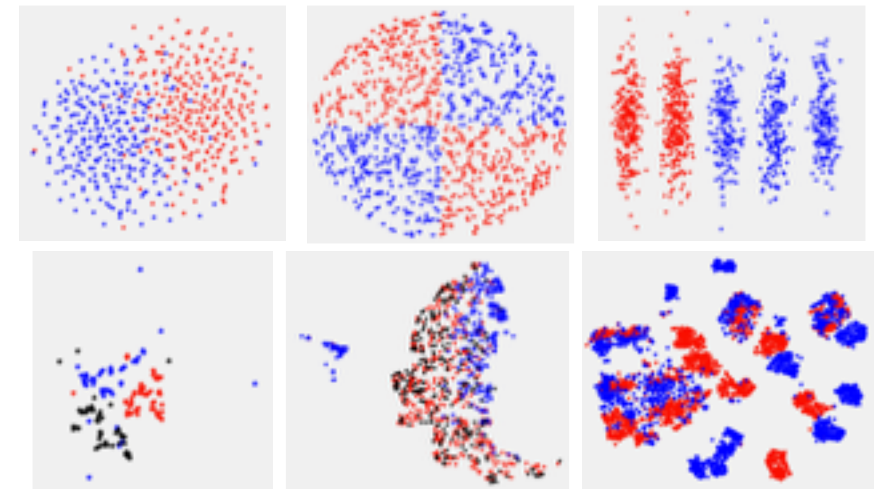
(e) USPS



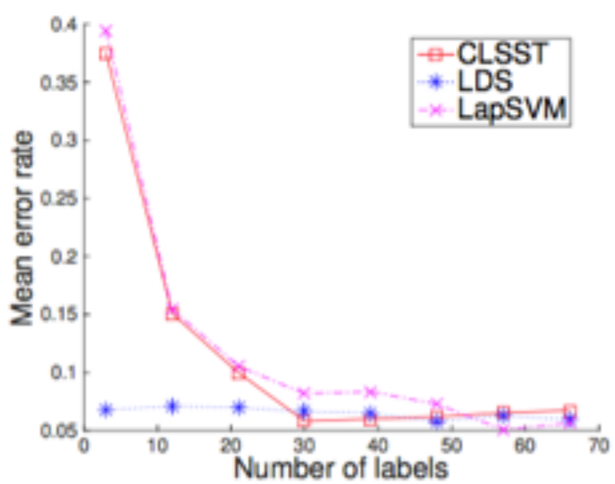
(f) ISOLET

Experiments

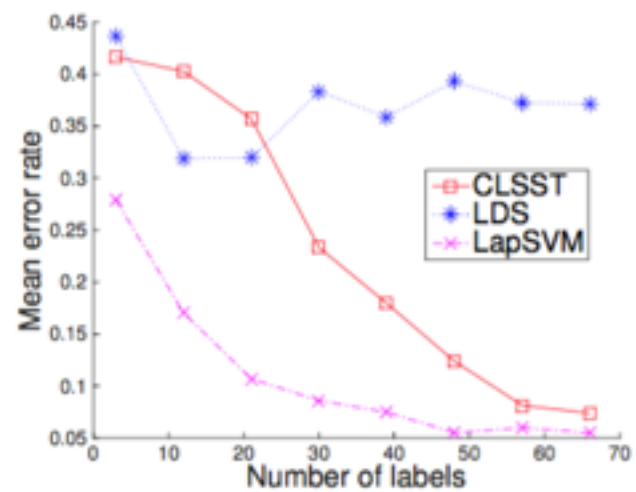
Data sets	Classes	Features	Instances
G50C	2	50	550
CAKE	2	2	1000
TOES	2	2	1000
IRIS	3	4	150
USPS	3	256	1918
ISOLET	2	617	3119



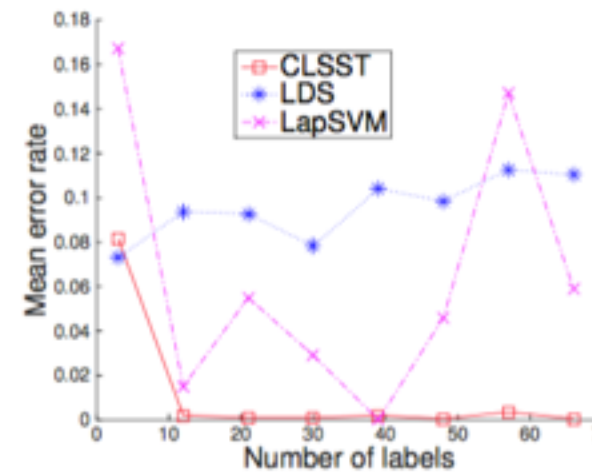
Classification



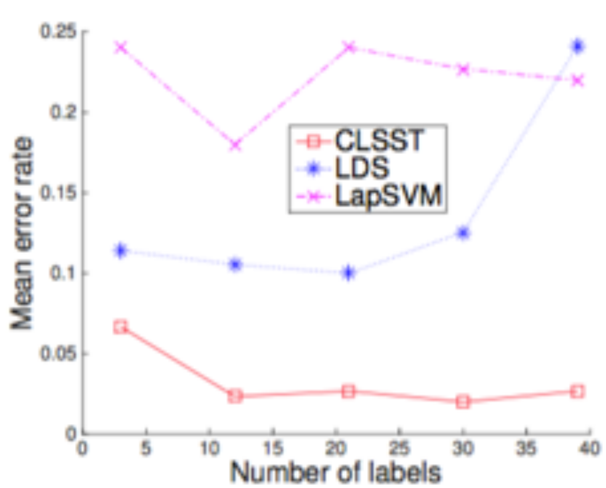
(a) G50C



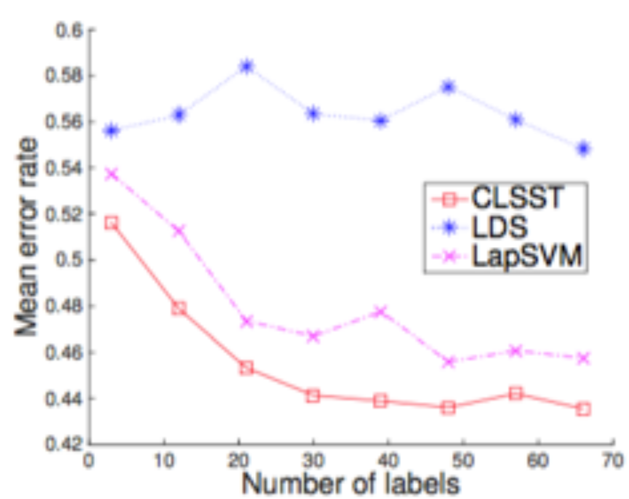
(b) CAKE



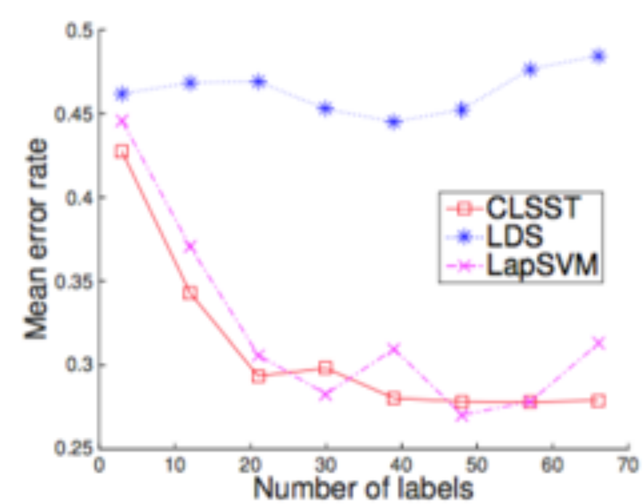
(c) TOES



(d) IRIS



(e) USPS



(f) ISOLET

Experiments

Dataset	Classes	Features	Instances
Breast cancer (discovery)	5	754	997
Breast cancer (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	-
Avg.	0.654	**0.706**	**0.748**
Min.	0.448	**0.557**	**0.575**
Max.	0.824	**0.927**	**0.912**

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

Conclusions & Future work

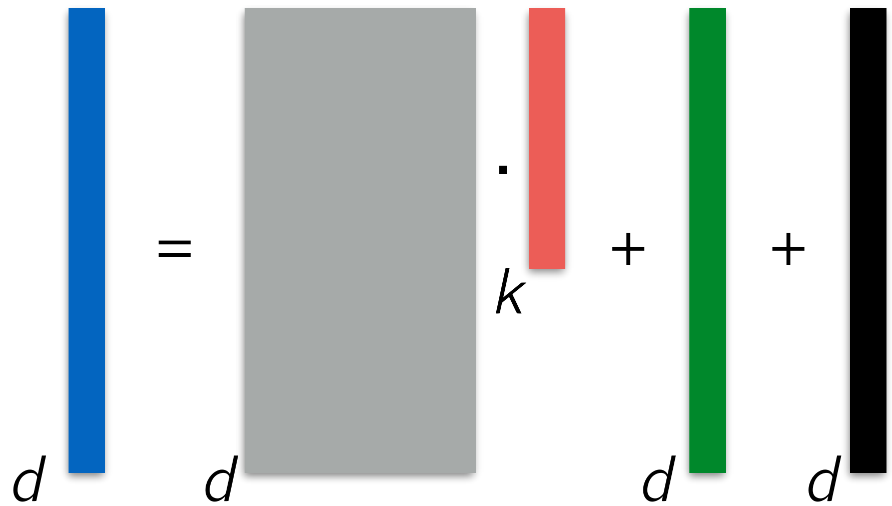
- Proposed model based on MFA for SSL (clustering/classification)
- Clustering: handling multi-groups per class + problem of cluster assumption
- Classification: discovered clusters help classification (comparison with discriminative approaches)
- Real-world problem: promising results (future research)

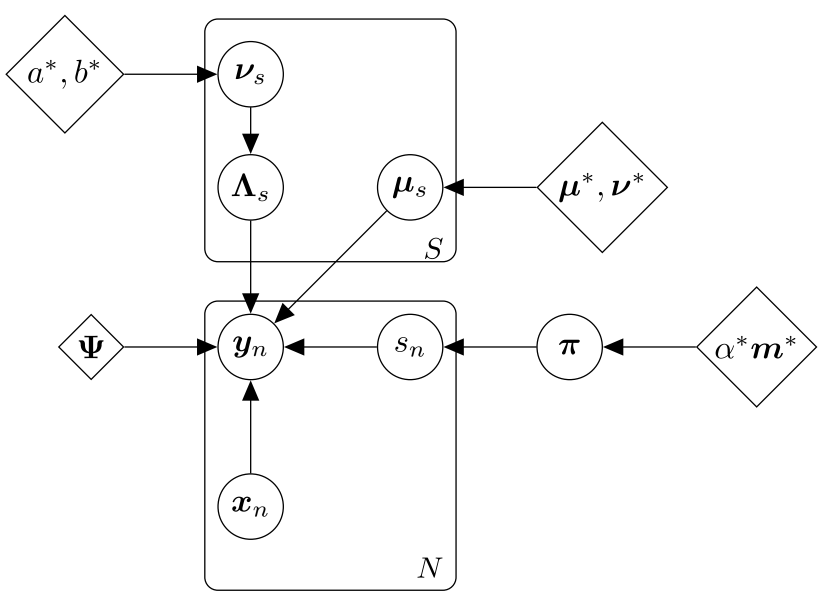
Thank You

Model

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

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



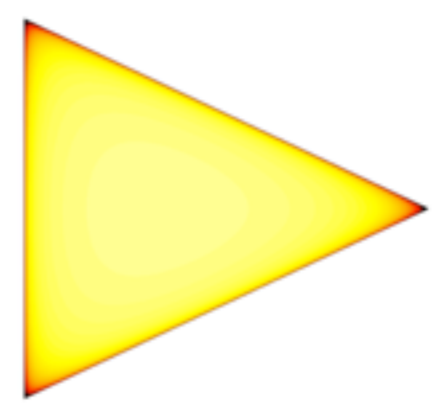


$$s_n \sim \prod_{s=1}^{S} \boldsymbol{\pi}_s^{1_{s_n}(s)}$$

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

$$\boldsymbol{m}^* = [1/S, \ldots, 1/S]$$

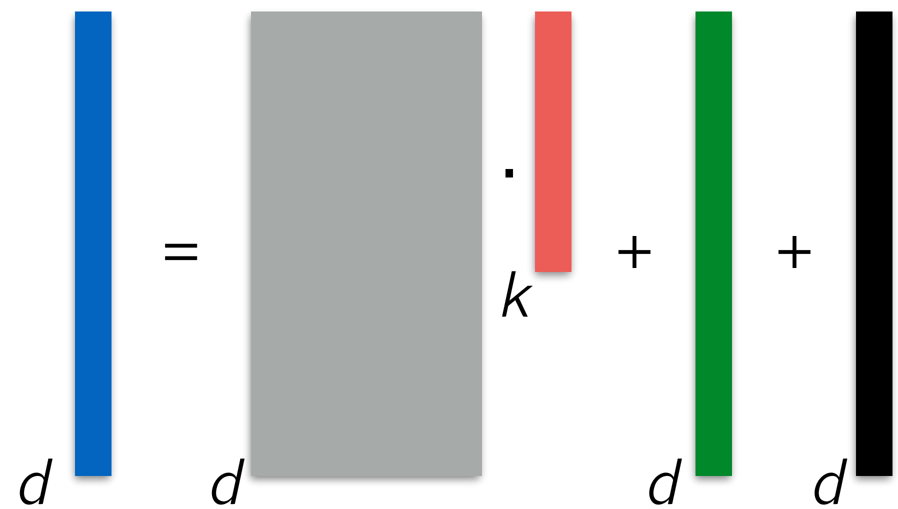
Example with
$S = 3, \alpha^* = 2.1$

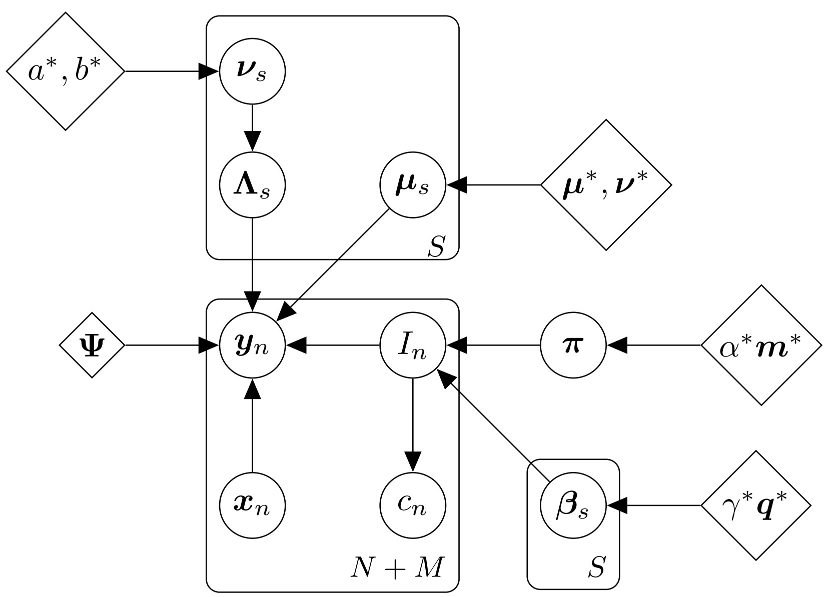


Model

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

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





$$I_n = (s_n, \ell_n)$$

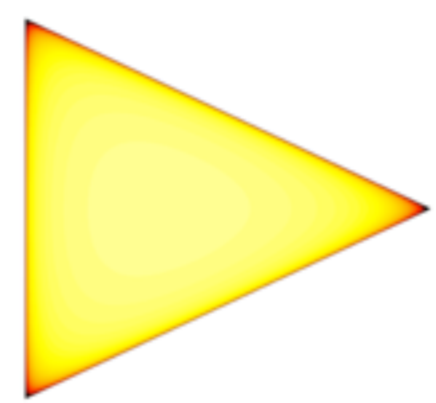
$$\ell_n \sim \prod_{\ell=1}^{K} \boldsymbol{\beta}_{s_n}(\ell)^{1_{\ell_n}(\ell)}$$

$$c_n \sim \delta(c_n - \ell_n)$$

$$\boldsymbol{\beta}_s \sim \text{Dir}(\gamma^* \boldsymbol{q}^*)$$

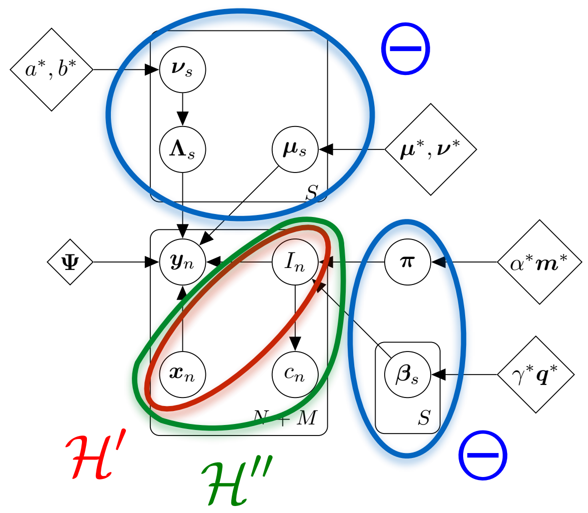
$$\boldsymbol{q}^* = [1/K, \ldots, 1/K]$$

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



Inference

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



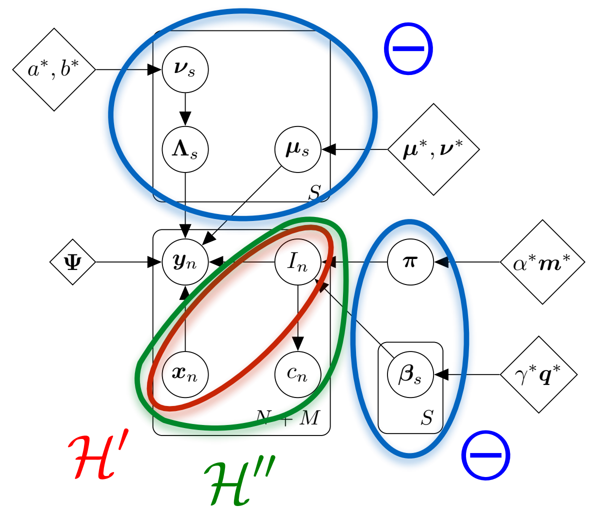
$\Theta = \{\boldsymbol{\pi}\} \cup \{\boldsymbol{\beta}_s, \boldsymbol{\Lambda}_s, \boldsymbol{\mu}_s, \boldsymbol{\nu}_s\}_{s=1}^{S}$

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

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

Inference

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



$\Theta = \{\boldsymbol{\pi}\} \cup \{\boldsymbol{\beta}_s, \boldsymbol{\Lambda}_s, \boldsymbol{\mu}_s, \boldsymbol{\nu}_s\}_{s=1}^{S}$

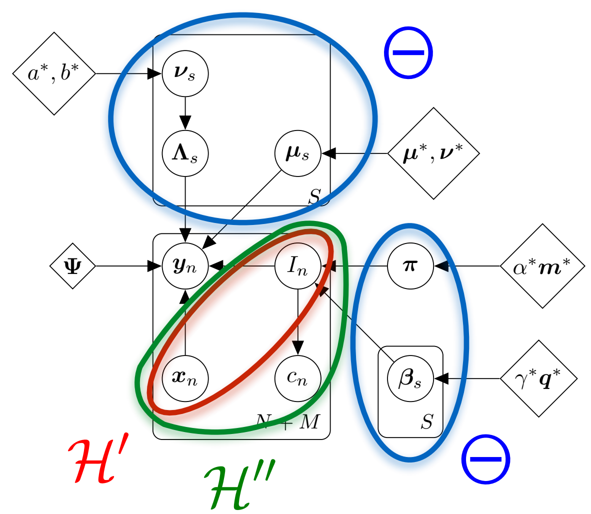
$\mathcal{H}' = \{\boldsymbol{x}_n, s_n\}_{n=1}^{N}$

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

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

Inference

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



$\Theta = \{\boldsymbol{\pi}\} \cup \{\boldsymbol{\beta}_s, \boldsymbol{\Lambda}_s, \boldsymbol{\mu}_s, \boldsymbol{\nu}_s\}_{s=1}^{S}$

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

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

$$\log p(D', D'') = \log \int_{\Theta, \mathcal{H}', \mathcal{H}''} p(D', D'', \Theta, \mathcal{H}', \mathcal{H}'')$$

$$= \log \int_{\Theta, \mathcal{H}', \mathcal{H}''} q(\Theta, \mathcal{H}', \mathcal{H}'') \frac{p(D', D'', \Theta, \mathcal{H}', \mathcal{H}'')}{q(\Theta, \mathcal{H}', \mathcal{H}'')}$$

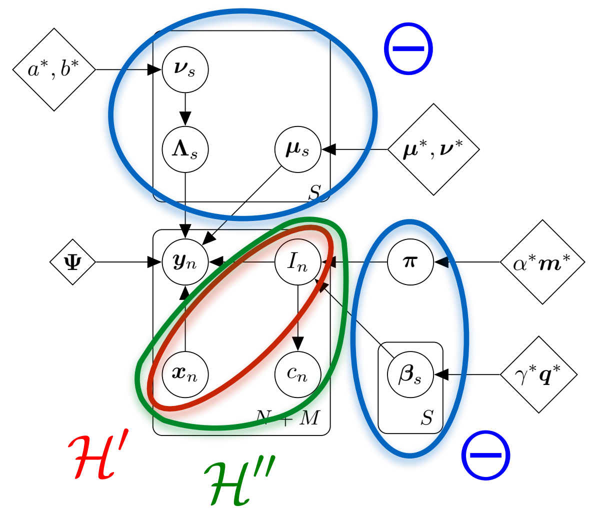
$$\geq \int_{\Theta, \mathcal{H}', \mathcal{H}''} q(\Theta, \mathcal{H}', \mathcal{H}'') \log \frac{p(D', D'', \Theta, \mathcal{H}', \mathcal{H}'')}{q(\Theta, \mathcal{H}', \mathcal{H}'')} = \mathcal{F}(q(\cdot))$$

- 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:

$$q(\Theta) = q(\pi) \prod_{s=1}^{S} q(\boldsymbol{\beta}_s) q(\boldsymbol{\nu}_s) q(\boldsymbol{\Lambda}_s, \boldsymbol{\mu}_s) \qquad q(\mathcal{H}' | \Theta) = \prod_{n=1}^{N} q(s_n) q(\boldsymbol{x}_n | s_n) \qquad q(\mathcal{H}'' | \Theta) = \prod_{n=1}^{N} q(I_n) q(\boldsymbol{x}_n | I_n)$$

Inference

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



$\Theta = \{\boldsymbol{\pi}\} \cup \{\boldsymbol{\beta}_s, \boldsymbol{\Lambda}_s, \boldsymbol{\mu}_s, \boldsymbol{\nu}_s\}_{s=1}^{S}$

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

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

$$\log p(D', D'') = \log \int_{\Theta, \mathcal{H}', \mathcal{H}''} p(D', D'', \Theta, \mathcal{H}', \mathcal{H}'')$$

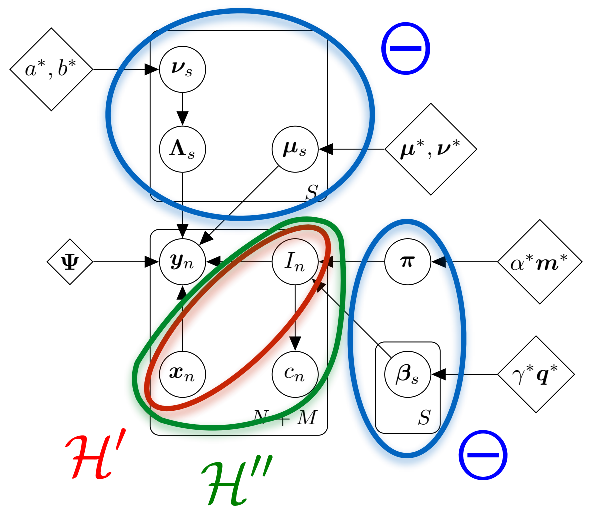
$$= \log \int_{\Theta, \mathcal{H}', \mathcal{H}''} q(\Theta, \mathcal{H}', \mathcal{H}'') \frac{p(D', D'', \Theta, \mathcal{H}', \mathcal{H}'')}{q(\Theta, \mathcal{H}', \mathcal{H}'')}$$

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

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

Inference

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



$\Theta = \{\boldsymbol{\pi}\} \cup \{\boldsymbol{\beta}_s, \boldsymbol{\Lambda}_s, \boldsymbol{\mu}_s, \boldsymbol{\nu}_s\}_{s=1}^{S}$

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

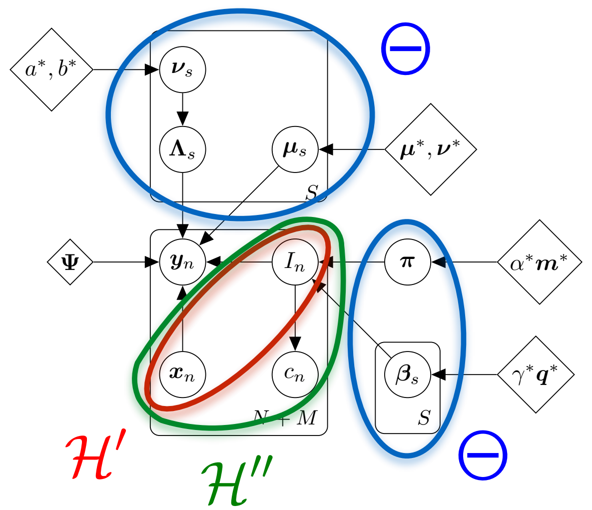
$\mathcal{H}'' = \{\boldsymbol{x}_n, (s_n, \ell_n), c_n\}_{n=N+1}^{M}$

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

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

Inference

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



$$\Theta = \{\boldsymbol{\pi}\} \cup \{\boldsymbol{\beta}_s, \boldsymbol{\Lambda}_s, \boldsymbol{\mu}_s, \boldsymbol{\nu}_s\}_{s=1}^{S}$$
$$\mathcal{H}' = \{\boldsymbol{x}_n, s_n\}_{n=1}^{N}$$
$$\mathcal{H}'' = \{\boldsymbol{x}_n, (s_n, \ell_n), c_n\}_{n=N+1}^{M}$$

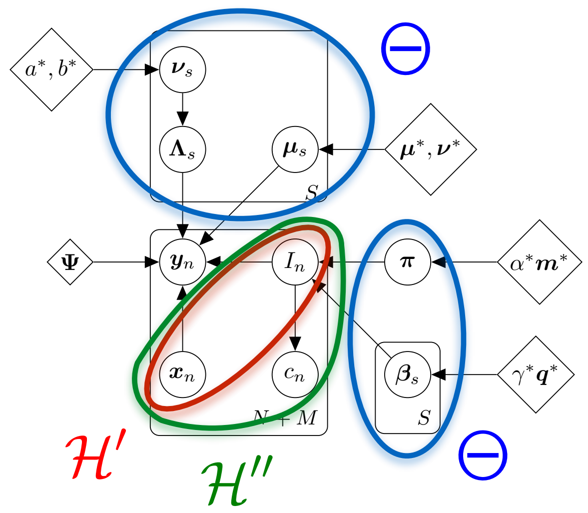
$$q(\Theta, \mathcal{H}', \mathcal{H}'') = q(\Theta) q(\mathcal{H}'|\Theta) q(\mathcal{H}''|\Theta)$$

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

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** $D' = \{(\boldsymbol{y}_n, c_n)\}_{n=1}^{N}$ and **unlabeled** $D'' = \{\boldsymbol{y}_n\}_{n=N+1}^{M}$



$\Theta = \{\boldsymbol{\pi}\} \cup \{\boldsymbol{\beta}_s, \boldsymbol{\Lambda}_s, \boldsymbol{\mu}_s, \boldsymbol{\nu}_s\}_{s=1}^{S}$

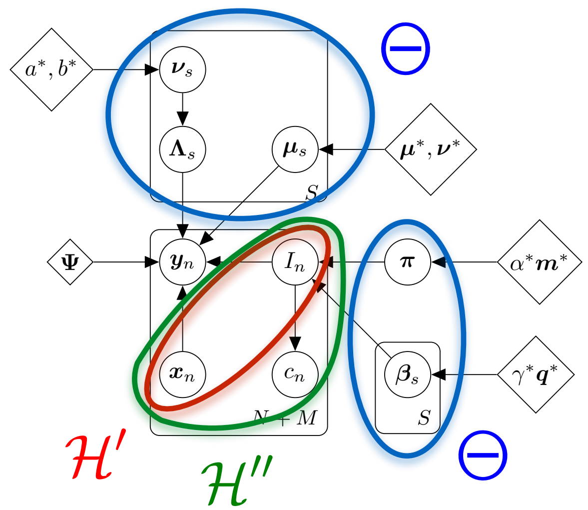
$\mathcal{H}' = \{\boldsymbol{x}_n, s_n\}_{n=1}^{N}$

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

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

Prediction

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



$\Theta = \{\boldsymbol{\pi}\} \cup \{\boldsymbol{\beta}_s, \boldsymbol{\Lambda}_s, \boldsymbol{\mu}_s, \boldsymbol{\nu}_s\}_{s=1}^{S}$

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

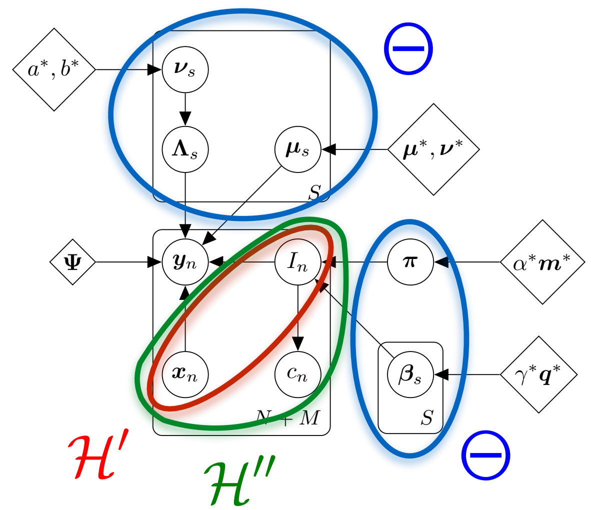
$\mathcal{H}'' = \{\boldsymbol{x}_n, (s_n, \ell_n), c_n\}_{n=N+1}^{M}$

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

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

Prediction

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



$\Theta = \{\boldsymbol{\pi}\} \cup \{\boldsymbol{\beta}_s, \boldsymbol{\Lambda}_s, \boldsymbol{\mu}_s, \boldsymbol{\nu}_s\}_{s=1}^{S}$

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

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

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

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

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