Self-adjusting Models for Semi-supervised Learning in Partially-observed Settings

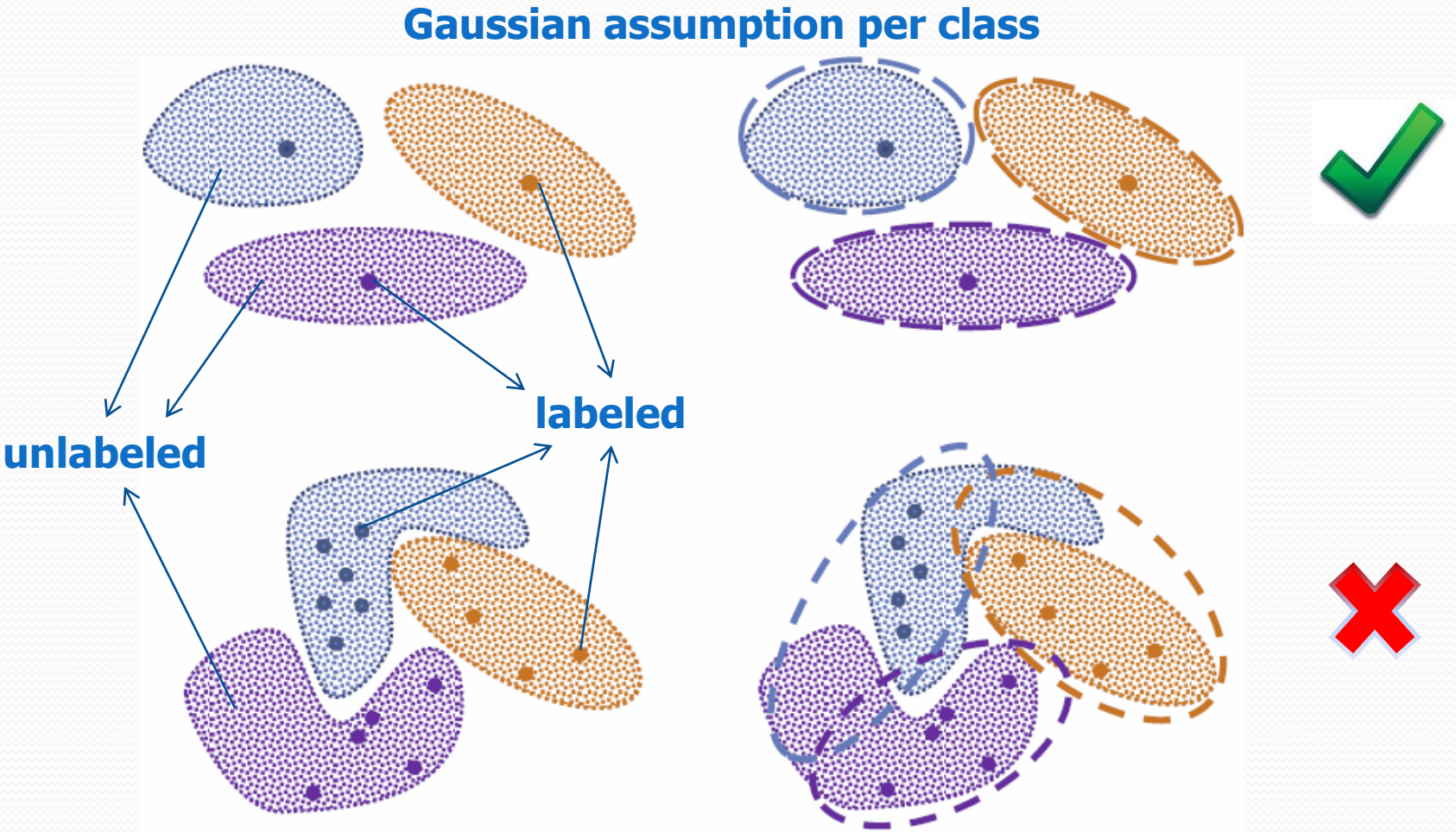
Ferit Akova[a,b]

in collaboration with
Yuan Qi[b], Bartek Rajwa[c] and Murat Dundar[a]

[a] Computer & Information Science Department, Indiana University – Purdue University, Indianapolis (IUPUI)
[b] Computer Science Department, Purdue University, West Lafayette, IN
[c] Discovery Park, Purdue University, West Lafayette, IN

Overview

Semi-supervised learning and the fixed model assumption



Overview

A new direction for Semi-supervised learning

- utilizes unlabeled data to improve learning even when labeled data is partially-observed
- uses self-adjusting generative models instead of fixed ones
- discovers new classes and new components of existing classes

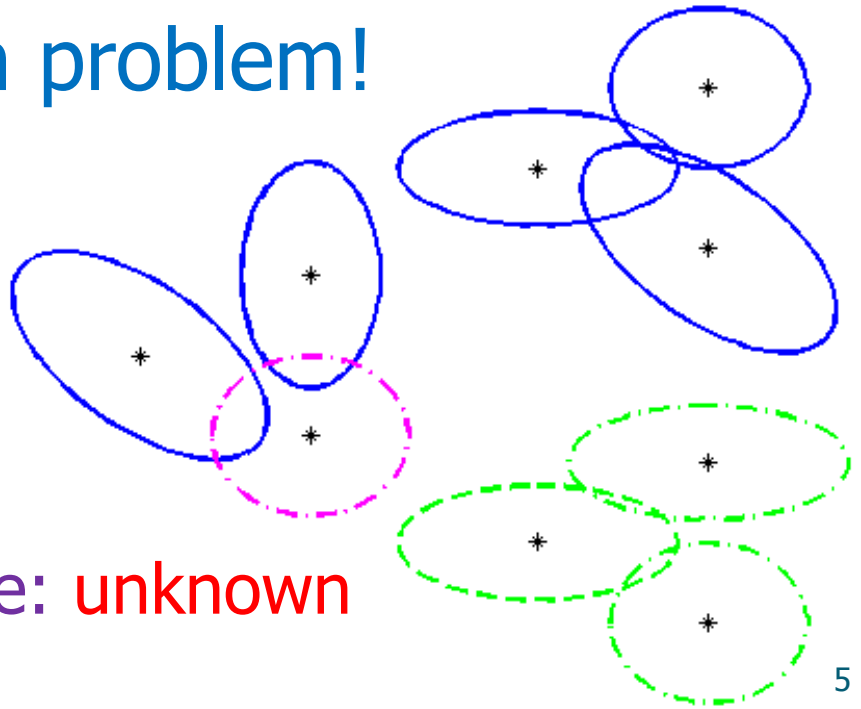
Outline

1. Learning in Non-exhaustive Settings
2. Motivating Problems
3. Overview of the Proposed Approach
4. Partially-observed Hierarchical Dirichlet Processes
5. Illustration and Experiments
6. Conclusion and Future Work

Non-exhaustive Setting

- Training dataset is unrepresentative if the list of classes is incomplete, i.e., non-exhaustive
- Future samples of unknown classes will be misclassified (into one of the existing classes) with a probability one

ill-defined classification problem!



blue: known

green & purple: unknown

What may lead to non-exhaustiveness?

- Some classes may not be in existence
- Classes may exist but may not be known
- Classes may be known but samples are unobtainable

Exhaustive training data not realistic for many problems

Some Application Domains

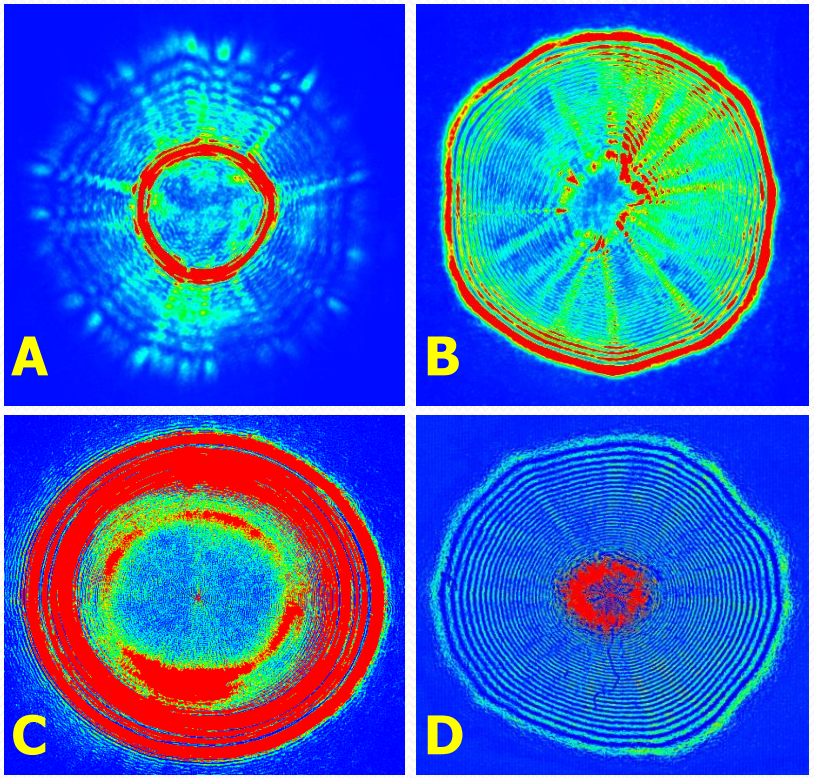
- Classification of documents by topics
  - research articles, web pages, news articles
- Image annotation
- Object categorization
- Bio-detection
- Hyperspectral image analysis

Biodetection

Food Pathogens

- Acquired samples are from most prevalent classes
- High mutation rate, new classes can emerge anytime
- An exhaustive training library simply impractical

Inherently non-exhaustive setting



(A) Listeria monocytogenes 7644,
(B) E. coli ETEC O25,
(C) Staphylococcus aureus P103,
(D) Vibrio cholerae O1E

Hyperspectral Data Analysis

- Military projects, GIS, urban planning, ...
- Physically inaccessible or dynamically changing areas
  - Enemy territories, special military bases
  - urban fields, construction areas

Impractical to obtain an exhaustive training data

Semi-supervised Learning (SSL)

- Traditional approaches
  - 1. self-training, 2. co-training, 3. transductive methods, 4. graph-based methods, 5. generative mixture models
- Unlabeled data improves classification under certain conditions, but primarily:
  - model assumption matches the model generating the data
- Limited labeled data not only scarce, but usually data distribution not fully represented or maybe evolving

SSL in Non-exhaustive Settings

A new framework for semi-supervised learning

- replaces the (brute-force fitting of a) fixed data model
- dynamically includes new classes/components
- classifies incoming samples more accurately

A self-adjusting model to better accommodate unlabeled data

Our Approach in a Nutshell

- Classes as Gaussian mixture model (GMM) with unknown number of components
- Extension of HDP to dynamically model new components/classes
- Parameter sharing across inter- & intra-class components
- Collapsed Gibbs sampler for inference

Our Notation

- J: number of known classes/groups
- n_j: number of samples in class j
- $x_{ji} \in \Re^d$, $i = \{1, \ldots, n_j\}$, $j = \{1, \ldots, J\}$
- $\mathbf{t} = \left\{ \{t_{ji}\}_{i=1}^{n_j} \right\}_{j=1}^{J}$, $\mathbf{k} = \left\{ \{k_{jt}\}_{t=1}^{m_{j.}} \right\}_{j=1}^{J}$
- $\tilde{J}$: current number of new classes
- n^u: total number of unlabeled samples

$$\tilde{x} = \{\tilde{x}_i\}_{i=1}^{n_u} \text{ and } \tilde{t} = \{\tilde{t}_i\}_{i=1}^{n_u}$$

$$\tilde{y} = \{\tilde{y}_i\}_{i=1}^{n_u} \qquad \tilde{y}_i \in \{1, \ldots, J + \tilde{J}\}$$

DP, HDP Briefly…

- Dirichlet Process (DP): a nonparametric prior over the number of mixture components with base distribution G_0 and parameter α
- Hierarchical DP: models each group/class as a DP mixture and couples the G_j's through a higher level DP

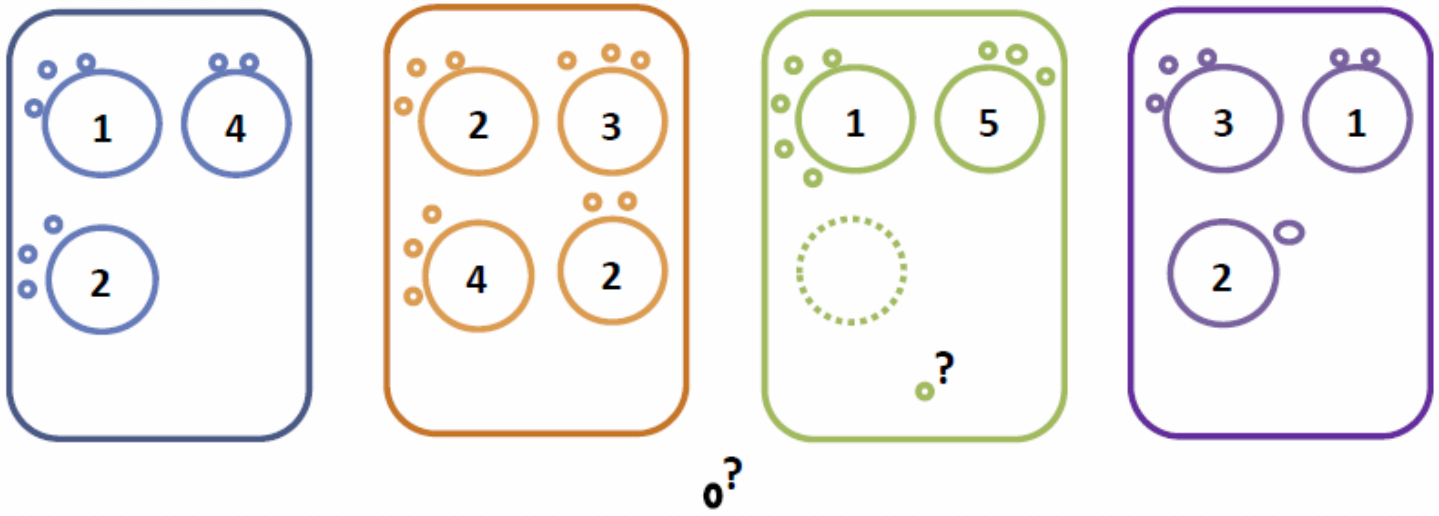
$$x_{ji}|\theta_{ji} \sim p(\cdot|\theta_{ji}) \quad for\ each\ j,\ i$$
$$\theta_{ji}|G_j \sim G_j \quad for\ each\ j,\ i$$

$$G_j|G_0,\alpha \sim DP(G_0,\alpha) \quad for\ each\ j$$
$$G_0|H,\gamma \sim DP(H,\gamma)$$

- α controls the prior probability of a new component

Modeling with HDP

- Chinese Restaurant Franchise (CRF) analogy
  - Restaurants correspond to classes, tables to mixture components and dishes in the "global menu" to unique parameters
  - First customer at a table orders a dish
  - Popular dishes more likely to be chosen
  - Role of γ in picking a new dish from the menu



Conditional Priors in CRF

- Seating customers and assigning dishes to tables
- t_{ji} – index of the table for customer i in restaurant j
- k_{jt} – index of the dish served at table t in restaurant j

$$t_{ji}|t_{j1},\dots,t_{j,i-1},\alpha \sim \frac{\alpha}{n_j+\alpha}\delta_t{}^{new} + \sum_{t=1}^{m_{j\cdot}} \frac{n_{jt}}{n_j+\alpha}\delta_t$$

$$k_{jt}|k_{j1},\dots,k_{j,t-1},\gamma \sim \frac{\gamma}{m_{\cdot\cdot}+\gamma}\delta_k{}^{new} + \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot}+\gamma}\delta_k$$

Inference in HDP

- Gibbs sampler to iteratively sample the indicator variables for tables and dishes given the state of all others

$$\mathbf{t} = \left\{\{t_{ji}\}_{i=1}^{n_j}\right\}_{j=1}^{J}, \mathbf{k} = \left\{\{k_{jt}\}_{t=1}^{m_{j.}}\right\}_{j=1}^{J}, \phi = \{\phi_k\}_{k=1}^{K}$$

- Conjugate pair of H and P(.|ϕ) allows for integrating out ϕ to obtain a collapsed version

- α and γ also sampled in each sweep based on number of tables and dishes, respectively. (Escobar & West, 1994)

Gibbs Sampler for **t** and **k**

- Conditional weighted by number of samples

$$p(t_{ji} = t | \boldsymbol{t} \backslash t_{ji}, \boldsymbol{k}, \boldsymbol{\phi}, \boldsymbol{x}) \propto \begin{cases} \alpha p(x_{ji}) & \text{for } t = m_{j\cdot} + 1 \\ n_{jt}^{-i} p(x_{ji} | \phi_{k_{jt}}) & \text{for } t \in \{1, \ldots, m_{j\cdot}\} \end{cases}$$

- Joint probability weighted by number of components

$$p(k_{jt} = k | \boldsymbol{t}, \boldsymbol{k} \backslash k_{jt}, \boldsymbol{\phi}, \boldsymbol{x}) \propto \begin{cases} \gamma \prod_{i:t_{ji}=t} p(x_{ji}) & \text{for } k = K + 1 \\ m_{\cdot k}^{-jt} \prod_{i:t_{ji}=t} p(x_{ji} | \phi_k) & \text{for } k \in \{1, \ldots, K\} \end{cases}$$

Defining Partially-observed Setting

- Observed classes/subclasses: Those initially available in the training library.
- Unobserved classes/subclasses: Those not represented in the training library
- New classes: classes discovered online, verified offline
  - limited to a single component until manual verification

HDP in a Partially-observed Setting

- Two tasks:
  1. Inferring component membership of labeled samples
  2. Inferring both the group and component membership of unlabeled samples
- Unlabeled samples evaluated for **all existing components**

Inference in Partially-observed HDP

- Updated Gibbs sampling inference for $\boldsymbol{t_{ji}}$

$$p(t_{ji} = t | \boldsymbol{t} \backslash t_{ji}, \boldsymbol{k}, \boldsymbol{\phi}, \boldsymbol{x}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \tilde{\boldsymbol{t}}) \propto$$
$$\begin{cases} \alpha\, p(x_{ji}) & \text{for } t = m_{j\cdot} + 1 \\ (n_{jt}^{-i} + \tilde{n}_{jt}) p(x_{ji} | \phi_{k_{jt}}) & \text{for } t \in \{1, \ldots, m_{j\cdot}\} \end{cases}$$

$$p(\tilde{t}_i = t | \boldsymbol{t}, \boldsymbol{k}, \boldsymbol{\phi}, \boldsymbol{x}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{t}} \backslash \tilde{t}_i, \tilde{\boldsymbol{y}}, \tilde{\boldsymbol{k}}) \propto$$
$$\begin{cases} \alpha\, p(\tilde{x}_i) & \text{for } t = 1 \\ & \underline{j = J + \bar{J} + 1} \\ (n_{jt} + \tilde{n}_{jt}^{-i}) p(\tilde{x}_i | \phi_{k_{jt}}) & \text{for } t \in \{1, \ldots, m_{j\cdot}\} \\ & \underline{j \in \{1, \ldots, J + \bar{J}\}} \end{cases}$$

Inference in Partially-observed HDP

- Updated inference for $\boldsymbol{k_{jt}}$ for existing and new classes

$$p(k_{jt} = k | \boldsymbol{t}, \boldsymbol{k} \backslash k_{jt}, \boldsymbol{\phi}, \boldsymbol{x}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \tilde{\boldsymbol{t}}, \tilde{\boldsymbol{k}}) \propto$$
$$\begin{cases} \gamma \prod_{i:t_{ji}=t} p(x_{ji}) \prod_{i:\tilde{t}_i=t \wedge \tilde{y}_i=j} p(\tilde{x}_i) & \text{for } k = K+1 \\ m_{\cdot k}^{-jt} \prod_{i:t_{ji}=t} p(x_{ji}|\phi_k) \prod_{i:\tilde{t}_i=t \wedge \tilde{y}_i=j} p(\tilde{x}_i|\phi_k) & \text{for } k \in \{1,\ldots,K\} \end{cases}$$

$$p(\tilde{k}_j = k | \boldsymbol{t}, \boldsymbol{k}, \boldsymbol{\phi}, \boldsymbol{x}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \tilde{\boldsymbol{t}}, \tilde{\boldsymbol{k}} \backslash \tilde{k}_j) \propto$$
$$\begin{cases} \gamma \prod_{i:\tilde{y}_i=j} p(\tilde{x}_i) & \text{for } k = K+1 \\ m_{\cdot k}^{-j} \prod_{i:\tilde{y}_i=j} p(\tilde{x}_i|\phi_k) & \text{for } k \in \{1,\ldots,K\} \end{cases}$$

Gaussian Mixture Model Data

$$p(x|\phi_j) = N(\mu_j, \Sigma_j), \quad \phi_j = \{\mu_j, \Sigma_j\}$$

$$H = p(\mu, \Sigma) = \underbrace{\mathcal{N}\left(\mu|\mu_0, \frac{\Sigma}{\kappa}\right)}_{p(\mu|\Sigma)} \times \underbrace{W^{-1}(\Sigma|\Sigma_0, m)}_{p(\Sigma)}$$

$\Sigma_0, m, \mu_0, \kappa$ estimated from labeled data by Empirical Bayes

Inference from GMM Data

- Evaluating Gibbs sampler requires $P(x|\phi_{jt})$ and $P(x)$
- True ϕ_{jt} unknown; integrating out …

$p(x|\phi_{jt}) \sim p(x|D_{jt})$, where $D_{jt} \equiv \{\bar{x}_{jt}, S_{jt}\}$, sufficient statistics

$$p(x|D_{jt}) = \int p\left(x\middle|\phi_{jt}\right) p\left(\phi_{jt}\middle|D_{jt}\right) \partial\phi_{jt}, \text{ where } \phi = \{\mu, \Sigma\}$$

$$p(x|D_{jt}) \sim \textbf{Student-t}\left(\hat{\mu}_{jt}, \hat{\Sigma}_{jt}, v_{jt}\right)$$

$$\mu_{jt} = \frac{n_{jt}\bar{x}_{jt} + \kappa\mu_0}{n_{jt} + \kappa}$$

Parameter Sharing in a GMM

- $P(x/D_{jt})$ replaced by $P(x/D_{.k})$, $D_{.k}$ samples from all components sharing ϕ_k
- Sharing both μ and Σ would make components unidentifiable
- $\boldsymbol{p(x|D_{.k}) = \int p(x|\phi_k)p(\phi_k|D_{.k})\partial\phi_k, \sim \text{Student–t}(\hat{\mu}_{jt}, \hat{\Sigma}_k, v_k)}$

$$\hat{\Sigma}_{jt} = \frac{\Sigma_0 + (n_{jt}-1)S_{jt} + \frac{n_{jt}\kappa}{n_{jt}+\kappa}(\bar{x}_{jt}-\mu_0)(\bar{x}_{jt}-\mu_0)^T}{\frac{(\kappa+n_{jt})\,v}{(\kappa+n_{jt}+1)}}$$

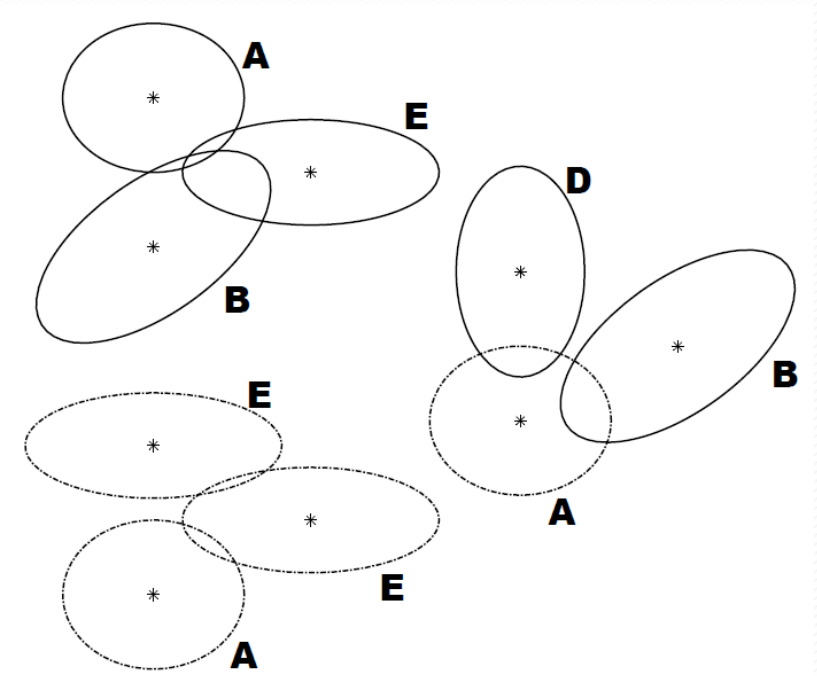
$$\hat{\Sigma}_{jt} = \frac{\Sigma_0 + \sum_{jt:k_{jt}=k}(n_{jt}-1)S_{jt} + \frac{n_{jt}\kappa}{n_{jt}+\kappa}(\bar{x}_{jt}-\mu_0)(\bar{x}_{jt}-\mu_0)^T}{\frac{(\kappa+n_{jt})\,v}{(\kappa+n_{jt}+1)}}$$

$$v_{jt} = m + n_{jt} - d + 1$$

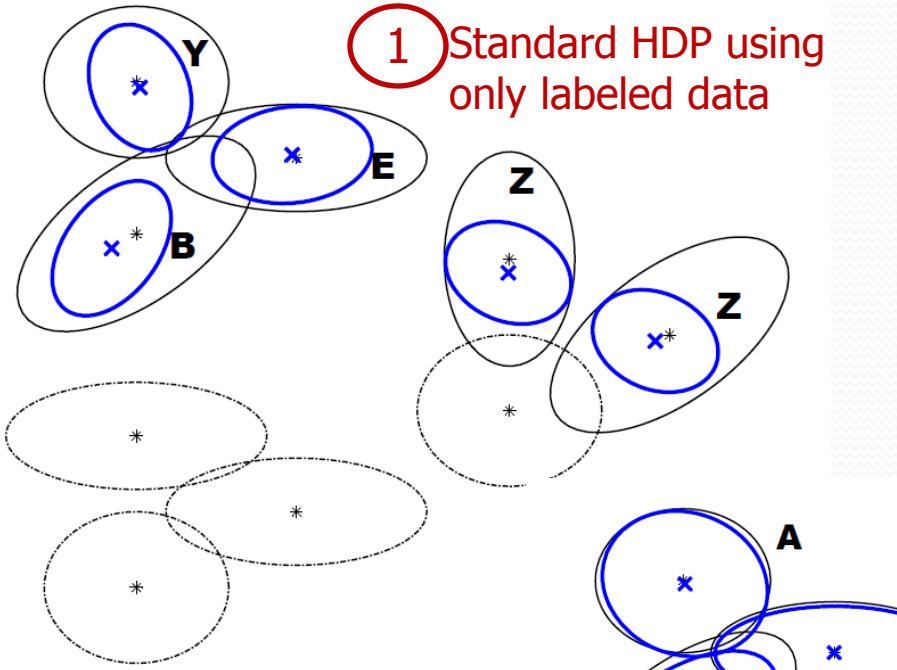
$$v_{jt} = m + \sum_{jt:k_{jt}=k}(n_{jt}-1) - d + 2$$

Illustrative Example

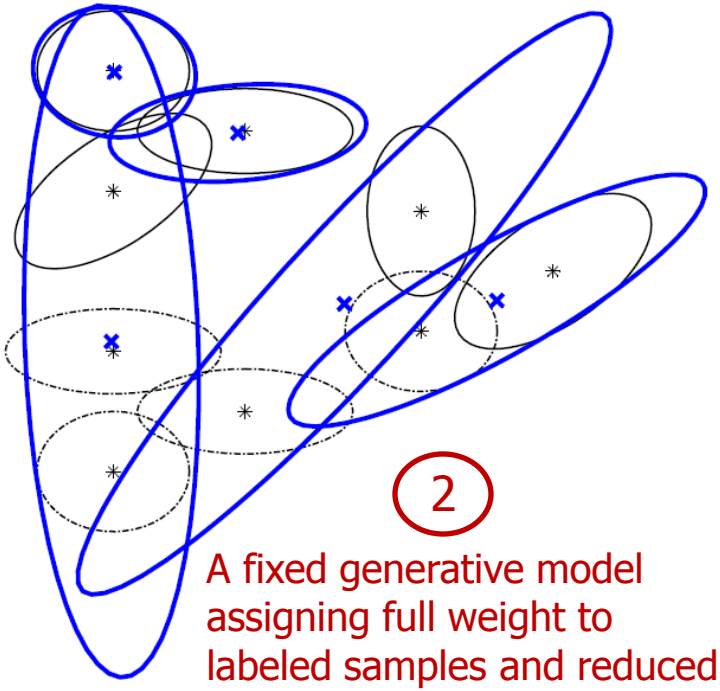
- 3 classes as a mixture of 3 components
- 110 samples in each component, 10 randomly selected as labeled 100 considered as unlabeled
- Covariance matrices from a set of 5 templates



Illustrative Example

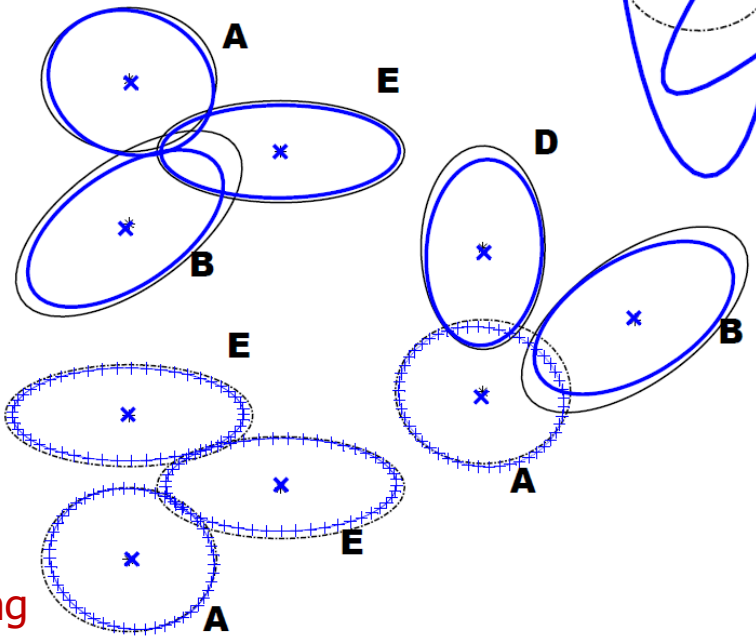


1 Standard HDP using only labeled data



2

A fixed generative model assigning full weight to labeled samples and reduced weight to unlabeled ones.



3

SA-SSL using labeled and unlabeled data with parameter sharing

Experiments - Evaluated Classifiers

- Baseline supervised learning methods using only labeled data
  - Naïve-Bayes (SL-NB), Maximum likelihood (SL-ML), expectation maximization (SL-EM)

- Benchmark semi-supervised learning methods
  - Self-training with base learners ML and NB (SELF)
  - Co-training with base learners ML and NB (CO-TR)
  - SSL-EM: Standard generative model approach
  - SSL-MOD: EM based approach with unobserved class modeling
  - SA-SSL: Proposed Self-adjusting SSL approach

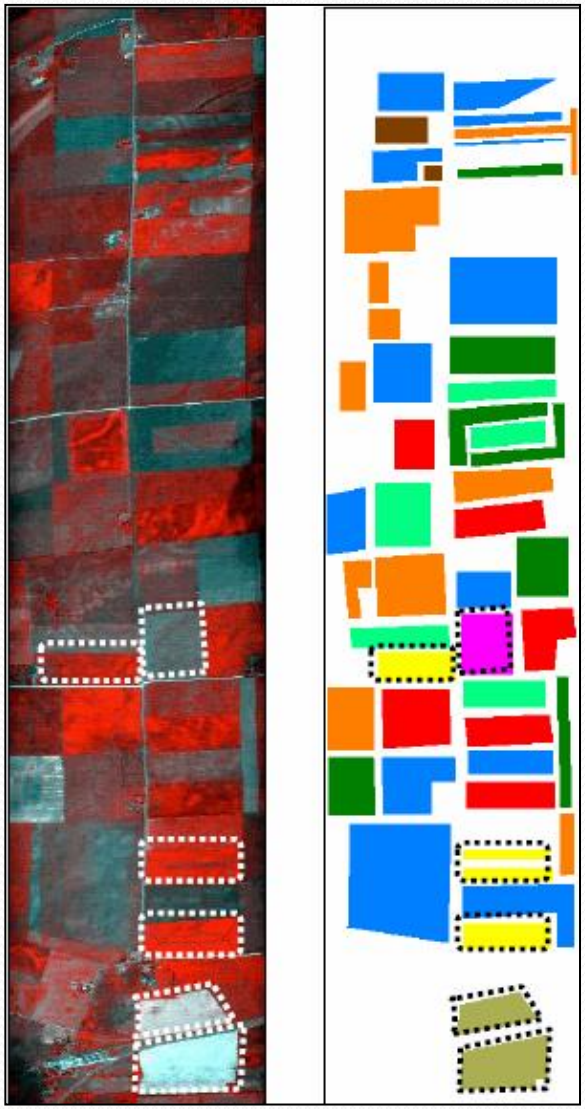
Experiments – Classifier Design

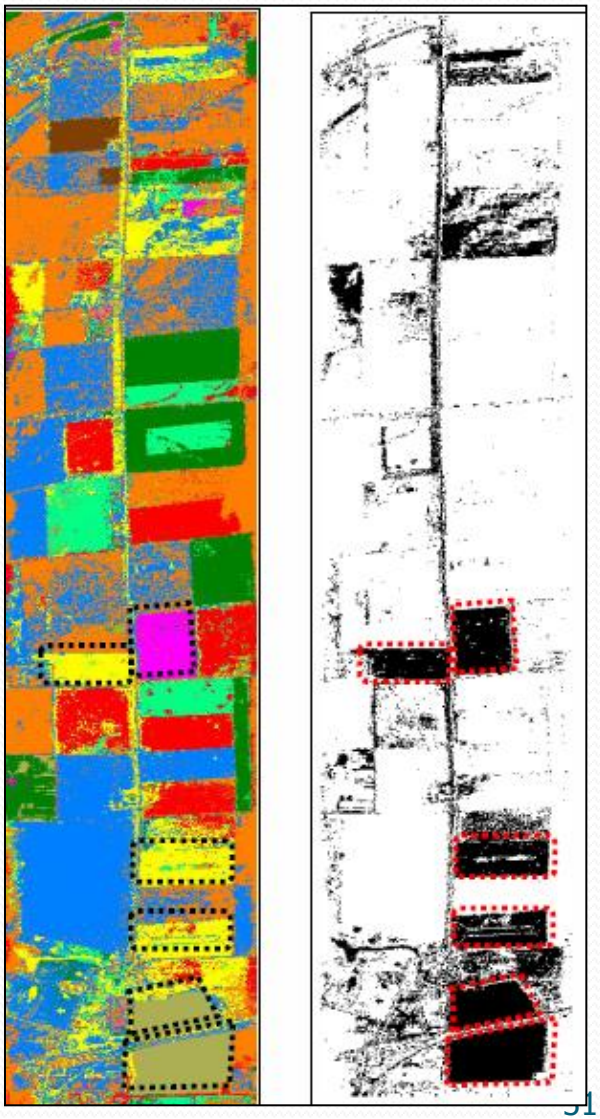
- Split available labeled data into **train**, **unlabeled** and **test**
- Stratified sampling to represent each class proportionally
- Consider some classes "**unobserved**" moving their samples from training set to unlabeled set
- **Non-exhaustive training** set, exhaustive unlabeled and test sets

Experiments – Evaluation

- Overall classification accuracy
- Average accuracies on observed and unobserved classes
- Newly created components associated with unobserved classes according to majority of samples
- Repeated with 10 random test/train/unlabeled splits

Remote Sensing





Remote Sensing Results

- 20 components and 10 unique covariance matrices in total
- Two to three components per each of the 8 classes
- Half of the components shares covariance matrices

Method	Acc	Acc-O	Acc-U
SA-SSL	**0.83**	**0.83**	**0.81**
SSL-SVM	0.62	0.81	0
SSL-EM	0.63	0.81	0
SSL-MOD	0.60	0.78	0.01
SELF	0.61	0.79	0
CO-TR	0.63	0.82	0
SL-ML	0.63	0.82	0
SL-NB	0.56	0.73	0
SL-EM	0.63	0.82	0

Pathogen Detection Experiment

- Total of 2054 samples from 28 bacteria classes
- Each class contains between 40 to 100 samples
- 22 feature samples
- 4 classes made unobserved, 24 classes remains observed
- 30% as **test**, 20% as **train** and remaining 50% as **unlabeled**

- Totally 180 components, 150 unique covariance matrices
- Five to six components per each class
- One sixth of the components shared parameter with others

Method	Acc	Acc-O	Acc-U
SA-SSL	**0.81**	**0.80**	**0.84**
SSL-EM	0.64	0.75	0
SSL-MOD	0.67	0.74	0.26
SELF	0.59	0.70	0
CO-TR	0.60	0.72	0
SL-ML	0.62	0.73	0
SL-NB	0.52	0.62	0
SL-EM	0.30	0.35	0

Recap of the Contributions

- A new approach to learning with a non-exhaustively defined labeled data set
- A unique framework to utilize unlabeled samples in partially-observed semi-supervised settings

1) Extension of HDP model to entertain unlabeled data and to discover & recover new classes
2) Fully Bayesian treatment of mixture components to allow parameter sharing across different components
   a) addresses the curse of dimensionality
   b) connects observed classes with unobserved ones

Future Work

- Replace Gibbs sampler with more scalable approximate inference methods
- Speed-up for real-time analysis of sequential data via a sequential MCMC sampler
- Extend the framework to hierarchically-structured datasets to associate discovered classes with higher level groups of classes