A Unified Framework for Probabilistic Component analysis

Debabrota Basu

School of Computing
National University of Singapore

Reference Paper

Nicolaou, Mihalis A., Stefanos Zafeiriou, and Maja Pantic. "A unified framework for probabilistic component analysis." *Machine Learning and Knowledge Discovery in Databases*. Springer Berlin Heidelberg, 2014. 469-484.

Roadmap

- **Introduction**
- **Overview of CA techniques**
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - Locality Preserving Projections (LPP)
  - Slow Feature Analysis (SFA)
- **Steps to Unification**
- **Unified Maximum Likelihood framework**
  - Defining priors and Markov random fields
  - Maximum likelihood solution
- **Unified Expectation Minimization framework**
  - Generalizing the prior
  - Expectation step
  - Minimization step
- **Experiments**
- **Discussions**

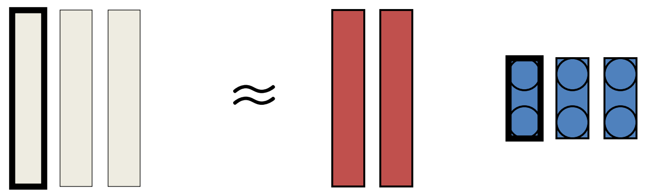
Roadmap

- **Introduction**
- **Overview of CA techniques**
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - Locality Preserving Projections (LPP)
  - Slow Feature Analysis (SFA)
- **Steps to Unification**
- **Unified Maximum Likelihood framework**
  - Defining priors and Markov random fields
  - Maximum likelihood solution
- **Unified Expectation Minimization framework**
  - Generalizing the prior
  - Expectation step
  - Minimization step
- **Experiments**
- **Discussions**

What is Component Analysis?

- Component analysis is a method of projecting data to subspace
- Subspace is a “manifold” (surface) embedded in a higher dimensional vector space
  - Data (e.g. images) are represented as points in a high dimensional vector space
  - Constraints in the natural world and the extraction process causes the points to “live” in a lower dimensional subspace
- Dimensionality reduction
  - Achieved by extracting ‘important’ features from the dataset
    → Learning
  - Desirable to avoid the “curse of dimensionality” in pattern recognition
    → Classification
- Examples- PCA, LDA, ICA, LPP, SFA, Kernel methods….

Projection to Subspaces



X_{FxT}

Sample data set
or
observation space

W_{FxN}

Projection
matrix

Y_{NxT}

Component set
or
latent space

$$\underline{\mathbf{x}}_i \approx \sum_{b=1..N} w_{bi}\, \underline{\mathbf{y}}_b$$

- Selection of W
  - Orthonormal bases
    - Y is simply projection of X onto W: Y = W^T X
  - General independent bases
    - If *N*=*F*, Q is obtained by solving linear system
    - If *N*<F, have to do some optimization (e.g., least squares)

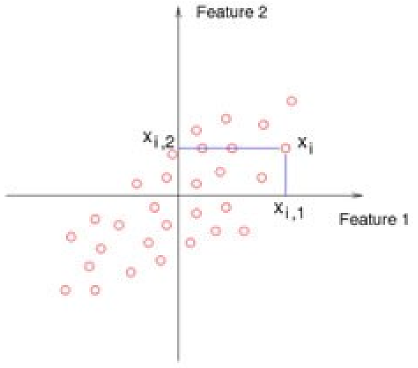
- *Different criterion for selecting W leads to different subspace methods*

-**Motivation for unification**

Roadmap

- **Introduction**
- **Overview of CA techniques**
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - Locality Preserving Projections (LPP)
  - Slow Feature Analysis (SFA)
- **Steps to Unification**
- **Unified Maximum Likelihood framework**
  - Defining priors and Markov random fields
  - Maximum likelihood solution
- **Unified Expectation Minimization framework**
  - Generalizing the prior
  - Expectation step
  - Minimization step
- **Experiments**
- **Discussions**

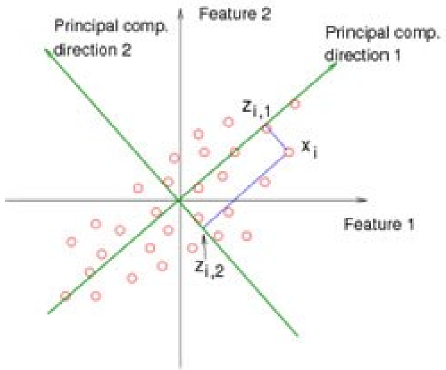
Principal Component Analysis



Algebra:
Orthogonal Transform

$\xrightarrow{\text{PCA}}$

Geometry:
Axis Rotation



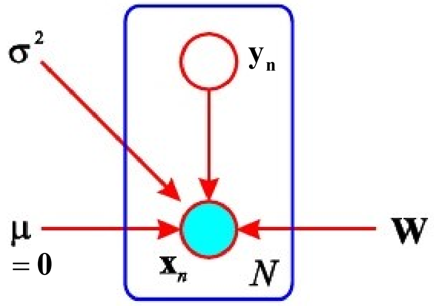
- Equivalent optimization problem

$$\mathbf{W}_0 = \underset{\mathbf{W}}{\arg\max}\ tr\left[\mathbf{W}^{\mathrm{T}}\mathbf{S}\mathbf{W}\right] \qquad ,\text{s.t. } \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}$$

- Probabilistic PCA

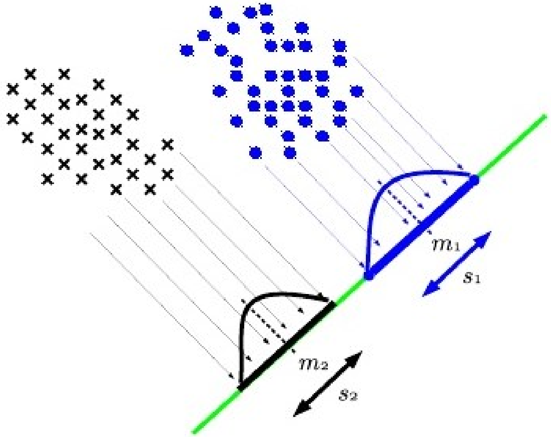
$$\mathbf{x}_\mathbf{i} = \mathbf{W}\mathbf{y}_\mathbf{i} + \boldsymbol{\varepsilon}_\mathbf{i}$$

$$s.t.\quad \mathbf{y}_\mathbf{i} \sim \qquad\qquad\qquad\qquad \sim$$



- **Motivation-**
  If *N* < *F*, the latent variables will offer a more parsimonious representation.

Linear Discriminant Analysis



- **Motivation-**
  Minimizing within-class variance i.e, $s_1 + s_2$ and maximizing between-class variance i.e, $(m_1 - m_2)^2$

- This is equivalent to finding a projection

$$\mathbf{W}_0 = \arg\max_{\mathbf{W}} \frac{\mathbf{W}^{\mathrm{T}}\mathbf{S}_b\mathbf{W}}{\mathbf{W}^{\mathrm{T}}\mathbf{S}_w\mathbf{W}}$$

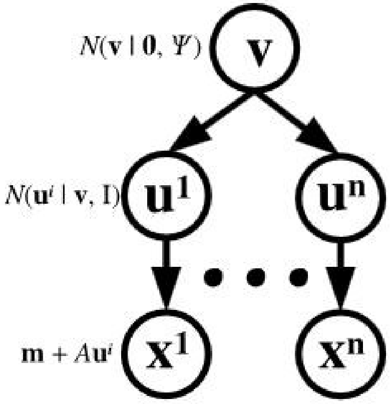
- This can be adopted as

$$\boxed{\mathbf{W}_0 = \arg\min_{\mathbf{W}} tr[\mathbf{W}^{\mathrm{T}}\mathbf{S}_w\mathbf{W}] \quad , s.t.\ \mathbf{W}^{\mathrm{T}}\mathbf{S}_b\mathbf{W} = \mathbf{I}}$$

- Probabilistic LDA is given as a generative model

$$P(\mathbf{y}) = \aleph(\mathbf{m}, \mathbf{A\Psi A}^{\mathrm{T}})$$

$$P(\mathbf{x}\,|\,\mathbf{y}) = \aleph(\mathbf{m}, \mathbf{AA}^{\mathrm{T}})$$


$N(\mathbf{v} \mid \mathbf{0}, \Psi)$
v
$N(\mathbf{u}^i \mid \mathbf{v}, \mathrm{I})$
$\mathbf{u}^1$
$\mathbf{u}^n$
$\mathbf{m} + A\mathbf{u}^i$
$\mathbf{x}^1$
$\mathbf{x}^n$

- Achilles' heel of PLDA:

Every class has to have same number of data points.

-Unrealistic!!!

Locality Preserving Projections

- **Motivation-**
  Finding a projection **W** such that locality of original samples is preserved in latent space.

- This is equivalent to

$$\boxed{\mathbf{W}_0 = \arg\min_{\mathbf{W}} \, tr[\mathbf{W}^{\mathbf{T}}\mathbf{X}\mathbf{L}\mathbf{X}^{\mathbf{T}}\mathbf{W}] \quad , s.t. \; \mathbf{W}^{\mathbf{T}}\mathbf{X}\mathbf{D}\mathbf{X}^{\mathbf{T}}\mathbf{W} = \mathbf{I}}$$

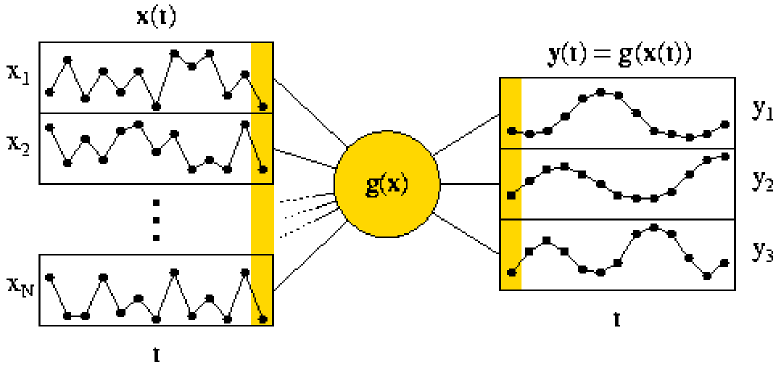
$$\mathbf{U} = [u_{ij}] = \left[\exp\left\{ -\frac{\left\| x_i - x_j \right\|^2}{\gamma} \right\} \right] \qquad \mathbf{D} = diag(\mathbf{U1}) \qquad \mathbf{L} = \mathbf{D} \text{-} \mathbf{U}$$

- Here, **U** represents the Heat kernel. This is used to represent locality.

- w_{ij} results a heavy penalty if the data points are mapped far apart.

- **No probabilistic version was proposed.**

Slow Feature Analysis



- **Motivation-**
  Finding a projection **W** such that features of the output signal varies slowest with time.

- This is equivalent to

$$\mathbf{W}_0 = \arg\min_{\mathbf{W}} tr[\mathbf{W}^{\mathrm{T}} \dot{\mathbf{X}} \dot{\mathbf{X}}^{T} \mathbf{W}] \quad , s.t.\ \mathbf{W}^{\mathrm{T}}\mathbf{S}\mathbf{W} = \mathbf{I}$$

$$\dot{\mathbf{X}} = [\dot{\mathbf{x}}_{\mathrm{j}}] = \left[\mathbf{x}_{\mathrm{j}} - \mathbf{x}_{\mathrm{j-1}}\right]$$

⇨ **First time derivative Matrix**

- **The generative model is an one-step linear Gaussian system**

$$P\left(\mathbf{x}_{\mathrm{t}} \mid \mathbf{y}_{\mathrm{t}}, \mathbf{W}, \sigma_X\right) = \aleph\left(\mathbf{W}^{-1}\mathbf{y}_{\mathrm{t}}, \sigma_X^2 \mathbf{I}\right)$$

$$p(\mathbf{y}_t|\mathbf{y}_{t-1}, \lambda_{1:N}, \sigma_{1:N}^2) = \prod_{n=1}^{N} p(y_{n,t}|y_{n,t-1}, \lambda_n, \sigma_n^2)$$

Roadmap

- **Introduction**
- **Overview of CA techniques**
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - Locality Preserving Projections (LPP)
  - Slow Feature Analysis (SFA)
- **Steps to Unification**
- **Unified Maximum Likelihood framework**
  - Defining priors and Markov random fields
  - Maximum likelihood solution
- **Unified Expectation Minimization framework**
  - Generalizing the prior
  - Expectation step
  - Minimization step
- **Experiments**
- **Discussions**

Steps to Unification

- Unified Maximum Likelihood Framework
  - A linear generative model of observation is assumed with white Gaussian noise over latent space
  - Use Markov Random Fields to calculate the prior
    - MRF encapsulates connectivity of latent variables in CA's
  - Projection directions (**W**) for CA's are engendered by ML estimation of joint PDF $P(\mathbf{X} \mid \boldsymbol{\Psi})$

- Unified Expectation Minimization framework
  - Generalize the prior for arbitrary number of MRFs
  - Using mean-field approximation calculate the marginal distribution
  - Execute the expectation and maximization steps of EM algorithm respectively

Steps to Unification

- Unified Maximum Likelihood Framework
  - A linear generative model of observation is assumed with white Gaussian noise over latent space
  - Use Markov Random Fields to calculate the prior
    - MRF encapsulates connectivity of latent variables in CA's
  - Projection directions (**W**) for CA's are engendered by ML estimation of joint PDF $P(\mathbf{X}|\boldsymbol{\Psi})$

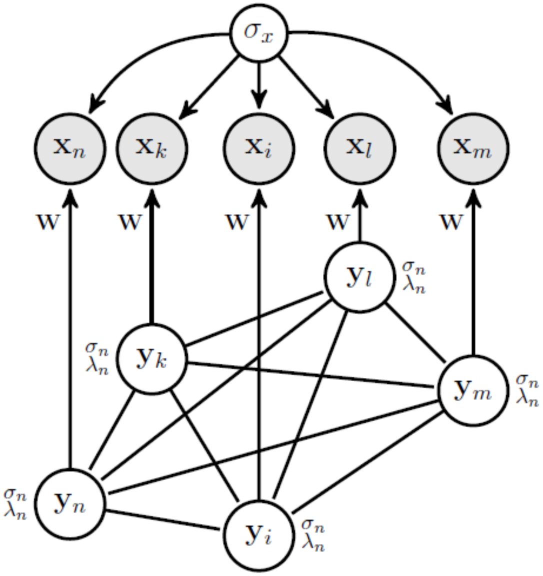
- Unified Expectation Minimization framework
  - Generalize the prior for arbitrary number of MRFs
  - Using mean-field approximation calculate the marginal distribution
  - Execute the expectation and maximization steps of EM algorithm respectively

Roadmap

- **Introduction**
- **Overview of CA techniques**
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - Locality Preserving Projections (LPP)
  - Slow Feature Analysis (SFA)
- **Steps to Unification**
- **Unified Maximum Likelihood framework**
  - Calculation of priors
  - Maximum likelihood solution
- **Unified Expectation Minimization framework**
  - Generalizing the prior
  - Expectation step
  - Minimization step
- **Experiments**
- **Discussions**

MRFs and Latent Connectivity

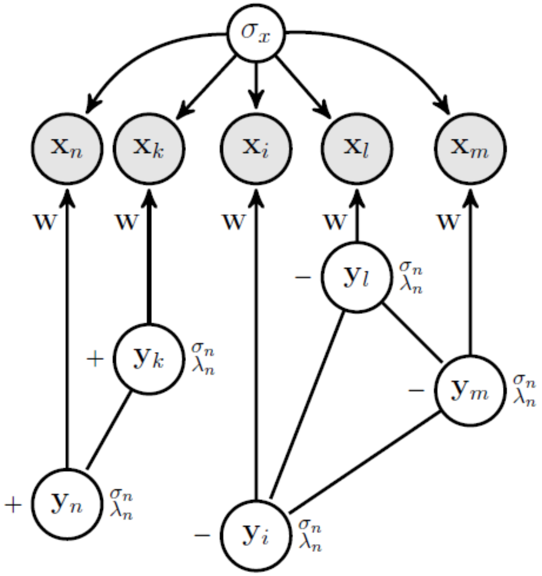
EM-PCA



(a)

Fully Connected MRF

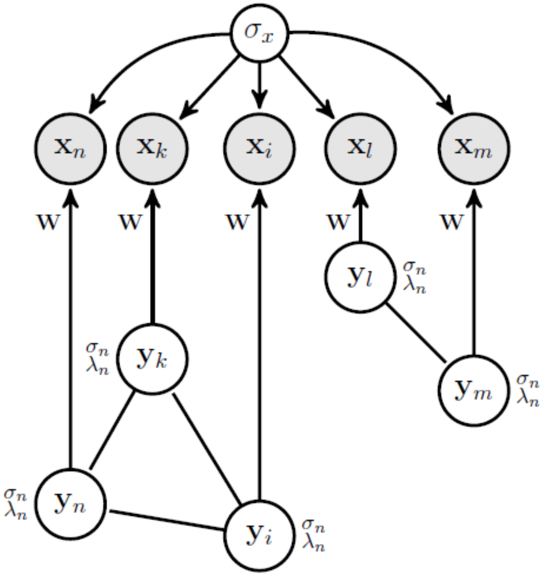
EM-LDA



(b)

Within-class Connected MRF

EM-LPP



(c)

Locally Connected MRF

Calculation of priors

- The unified formula for the prior of component analysis methods is of the form

$$P(\mathbf{Y}\,|\,\beta) \propto \exp\left\{-\frac{1}{2}\left(tr\left[\mathbf{\Lambda}^{(1)}\mathbf{Y}\mathbf{B}^{(1)}\mathbf{Y}^{\mathrm{T}}\right]+tr\left[\mathbf{\Lambda}^{(2)}\mathbf{Y}\mathbf{B}^{(2)}\mathbf{Y}^{\mathrm{T}}\right]\right)\right\}$$

- $\mathbf{B}^{(1)}$and $\mathbf{B}^{(2)}$ are functional form of potentials which encapsulate the latent covariance connectivity of neighborhoods.
- $\mathbf{\Lambda}^{(1)}$ and $\mathbf{\Lambda}^{(2)}$ are functions of parameters of MRF $\beta=\left\{\lambda_{1:N},\sigma_{1:N}^{2}\right\}$

	PCA	LDA	LPP	SFA
$\mathbf{B}^{(1)}$	$\mathbf{I}$	$\mathbf{M}_c = \mathbf{I} - diag\left[\mathbf{C}_c\right]$	$\mathbf{L}=\mathbf{D}^{-1}\mathbf{L}$	$\mathbf{K}_1 = \mathbf{P}_1\mathbf{P}_1^{\mathrm{T}}$
$\mathbf{B}^{(2)}$	$\mathbf{M}=-\frac{1}{T}\mathbf{1}\mathbf{1}^{\mathrm{T}}$	$\mathbf{M}_t = \mathbf{I}+\mathbf{M}$	$\mathbf{D}=\mathbf{I}$	$\mathbf{I}$

Maximum Likelihood (ML) solution

- If we consider the linear generative model,

$$\mathbf{x_i} = \mathbf{W}^{-1}\mathbf{y_i} + \boldsymbol{\varepsilon}_\mathbf{i} \qquad , s.t. \quad \boldsymbol{\varepsilon}_\mathbf{i} \sim \qquad\qquad \mathbf{I}\Big)$$

$$\Rightarrow \;\; P\left(\mathbf{x_t} \,|\, \mathbf{y_t}, \mathbf{W}, \sigma_x^2\right) = \aleph\left(\mathbf{W}^{-1}\mathbf{y}_t, \sigma_x^2\right)$$

- Thus, the likelihood will be

$$P(\mathbf{X} \,|\, \boldsymbol{\Psi}) = \int \prod_{t=1}^{T} P\left(\mathbf{x_t} \,|\, \mathbf{y_t}, \mathbf{W}, \sigma_x^2\right) P(\mathbf{Y} \,|\, \beta)\; d\mathbf{Y}$$

- Maximum likelihood solution for our model gives

$$\mathbf{I} = \boldsymbol{\Lambda}^{(1)}\mathbf{W}\mathbf{X}\mathbf{B}^{(1)}\mathbf{X}^\mathrm{T}\mathbf{W}^\mathrm{T} + \boldsymbol{\Lambda}^{(2)}\mathbf{W}\mathbf{X}\mathbf{B}^{(2)}\mathbf{X}^\mathrm{T}\mathbf{W}^\mathrm{T}$$

- **W** simultaneously diagonalises $\mathbf{X}\mathbf{B}^{(1)}\mathbf{X}^\mathbf{T}$ and $\mathbf{X}\mathbf{B}^{(2)}\mathbf{X}^\mathbf{T}$.

Properties of ML solution

- **W** is independent of setting of λ_n , if they are all different.

- If $0 < \lambda_n < 1$, then larger values of λ_n corresponds to
  - More expressive PCA
  - More discriminant LDA
  - More local LPP
  - Slower latent variables in SFA

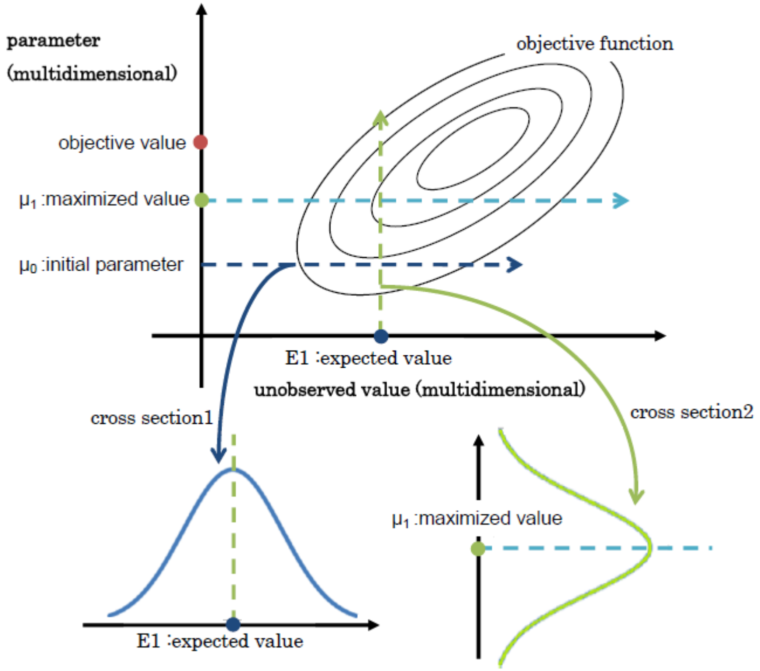
- To get the exact equivalence, we moreover need **scaling.**
  - Assuming, $\sigma_n^2 = 1 - \lambda_n^2$ scales LDA, SFA and LPP.
  - In PCA, σ_n should be kept analogous to eigenvalues of covariance matrix.

Roadmap

- **Introduction**
- **Overview of CA techniques**
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - Locality Preserving Projections (LPP)
  - Slow Feature Analysis (SFA)
- **Steps to Unification**
- **Unified Maximum Likelihood framework**
  - Defining priors and Markov random fields
  - Maximum likelihood solution
- **Unified Expectation Minimization framework**
  - Generalizing the prior
  - Expectation step
  - Minimization step
- **Experiments**
- **Discussions**

Expectation-Maximization

- Iterative method for parameter (θ) estimation where you have missing data ($\mathbf{Y}$).



- Starting from an initial guess, each iteration consists
  - An Expectation (E) step

  where it computes expectation of log likelihood over pre estimated parameters and available data

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{\mathbf{i}}) \triangleq \qquad P\left(\mathbf{X},\mathbf{Y}\,|\,\boldsymbol{\theta}\right)|\,\mathbf{X},\boldsymbol{\theta}^{\mathbf{i}}\,\big]$$

  - A Maximization (M) step

  where parameters are updated

$$\boldsymbol{\theta}^{\mathbf{i+1}} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta},\boldsymbol{\theta}^{\mathbf{i}})$$

Generalizing prior

- The prior is defined as product of $\mathcal{M}$ MRFs as

$$P(\mathbf{Y}|\beta) = \prod_{\mu \in \mathcal{M}} \frac{1}{Z^{\mu}} \exp\{Q^{\mu}\}$$

$$Q^{\mu} = -\sum_{n=1}^{N} \frac{f_{\mu}(\lambda_n)}{2\sigma_n^2} \frac{1}{c} \sum_{i \in \omega_i} \frac{1}{c_j^{\mu}} \sum_{j \in \omega_j^{\mu}} (y_{n,i} - \phi_{\mu}(\lambda_n) y_{n,j})^2$$

- If the linear generative model is assumed, using mean-field approximation we can write

$$P(\mathbf{Y}|\beta) \approx \prod_{i=1}^{T} P(\mathbf{y}_i | \mathbf{m}_i^{\mathcal{M}}, \beta^{\mathcal{M}}) = \prod_{i=1}^{T} \aleph\left(\mathbf{m}_i^{M}, \Sigma^{M}\right)$$

  - $\mathbf{m}_i^{M}$ depends on model specific connectivity and depends on $\mathrm{E}[\mathbf{y}_\mathrm{i}]$
  - Σ^{M} depends on $\beta = \left\{\lambda_{1:N}, \sigma_{1:N}^2\right\}$

- Linear generative model is assumed.

$$\mathbf{x}_i = \mathbf{W}\mathbf{y}_i + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma_x^2)$$

Expectation Step

- Compute the first order moment on the latent posterior which returns a Gaussian distribution.

$$P(\mathbf{y}_i|\mathbf{x}_i, \mathbf{m}_i^{\mathcal{M}}, \Psi^{\mathcal{M}}) = \mathcal{N}(\mathbf{y}_i|(\mathbf{W}^T\mathbf{x}_i + \boldsymbol{\Sigma}^{\mathcal{M}^{-1}}\mathbf{m}_i^{\mathcal{M}})\mathbf{A}, \sigma_x^{\mathcal{M}^2}\mathbf{A})$$

- It in turn gives us, expectation terms for missing data

$$\mathbb{E}^{\mathcal{M}}[\mathbf{y}_i] = \underbrace{\mathbf{y}_i|(\mathbf{W}^T\mathbf{x}_i + \boldsymbol{\Sigma}^{\mathcal{M}^{-1}}\mathbf{m}_i^{\mathcal{M}})\mathbf{A}}_{\text{mean}}$$

$$\mathbb{E}^{\mathcal{M}}[\mathbf{y}_i\mathbf{y}_i^T] = \underbrace{\sigma_x^{\mathcal{M}^2}\mathbf{A}}_{\text{covariance}} + \mathbb{E}[\mathbf{y}_i]\mathbb{E}[\mathbf{y}_i]^T$$

Maximization Step

- By applying mean-field approximation the data-likelihood can be factorized as,

$$P(\mathbf{Y}, \mathbf{X} | \Psi^{\mathcal{M}}) \approx \prod_{i=1}^{T} P(\mathbf{x}_i | \mathbf{y}_i, \theta^{\mathcal{M}}) P(\mathbf{y}_i | \mathbf{m}_i^{\mathcal{M}}, \beta^{\mathcal{M}})$$

- Thus, the maximization term becomes

$$\theta^{\mathcal{M}} = \arg\max \left\{ \textstyle\sum_{i=1}^{T} \int_{\mathbf{y}_i} P(\mathbf{y}_i | \mathbf{x}_i, \mathbf{m}_i^{\mathcal{M}}, \Psi^{\mathcal{M}}) \log P(\mathbf{x}_i | \mathbf{y}_i, \theta^{\mathcal{M}}) d\mathbf{y}_i \right\}$$

$$\beta^{\mathcal{M}} = \arg\max \left\{ \textstyle\sum_{i=1}^{T} \int_{\mathbf{y}_i} P(\mathbf{y}_i | \mathbf{x}_i, \mathbf{m}_i^{\mathcal{M}}, \Psi^{\mathcal{M}}) \log P(\mathbf{y}_i | \mathbf{m}_i^{\mathcal{M}}, \beta^{\mathcal{M}}) d\mathbf{y}_i \right\}$$

- This gives us closed form update rules for model parameters.

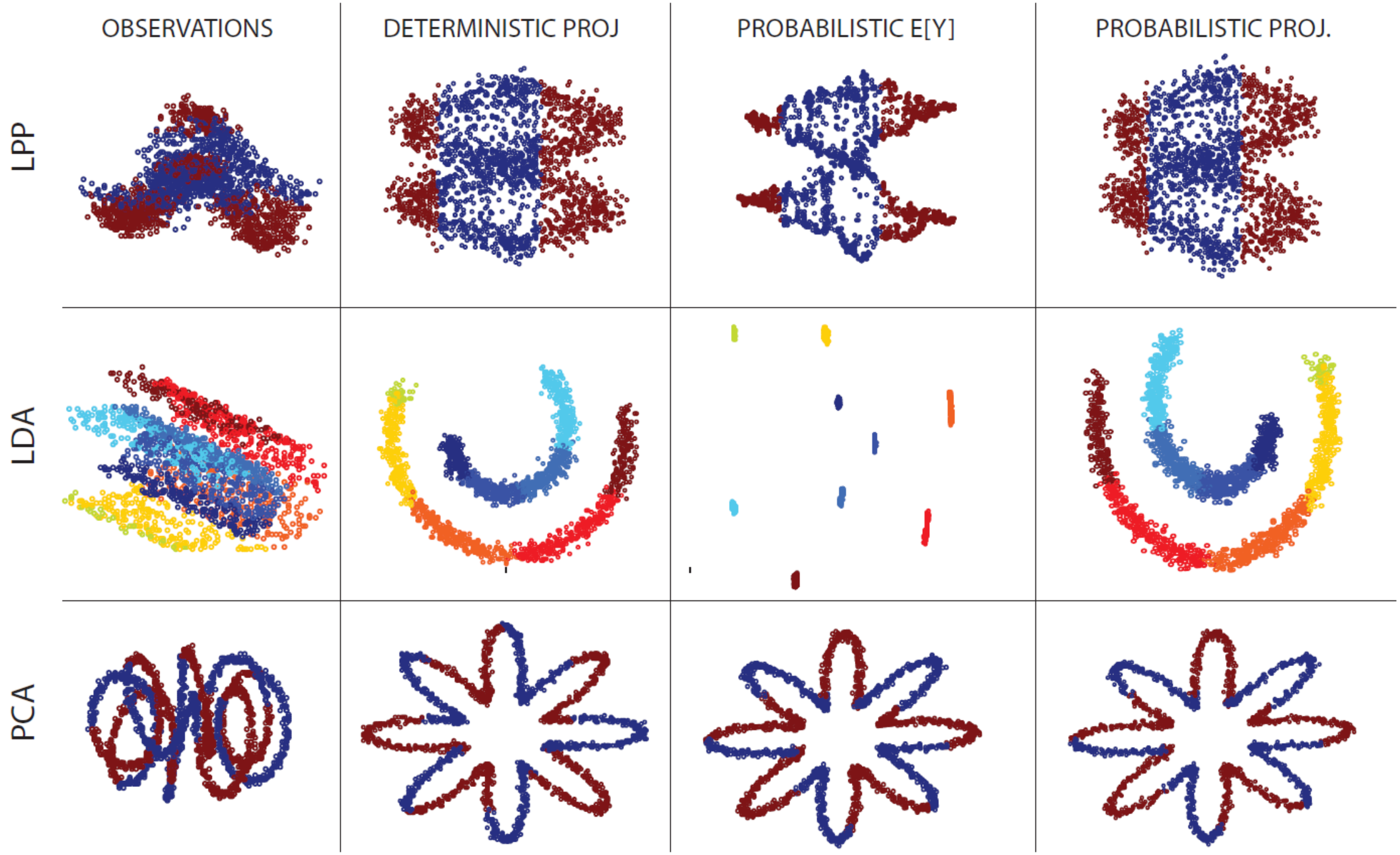
Features of EM solutions

- EM-PCA
  - Equivalent to PPCA when $\lambda_n = 0 \ \ and \ \ \sigma_n = 1$
  - Generally shifted by a mean field
  - Models per dimension variance, that PCA cannot
  - Complexity is $O(TNF)$, unlike $O(T^3)$ for deterministic PCA (*F*,*N*<<*T*)
- EM for SFA
  - Undirected MRF interpretation
    - Autoregressive SFA
    - Can learn bi-directional latent dependencies
  - Directed Dynamic Bayesian Network interpretation
    - A direction specific model of our EM model with directed MRF prior
- Probabilistic LDA
  - Only need to estimate likelihood of each test datum in each class
  - Probabilistic nature can be exploited to infer the most likely class assignment of unseen data

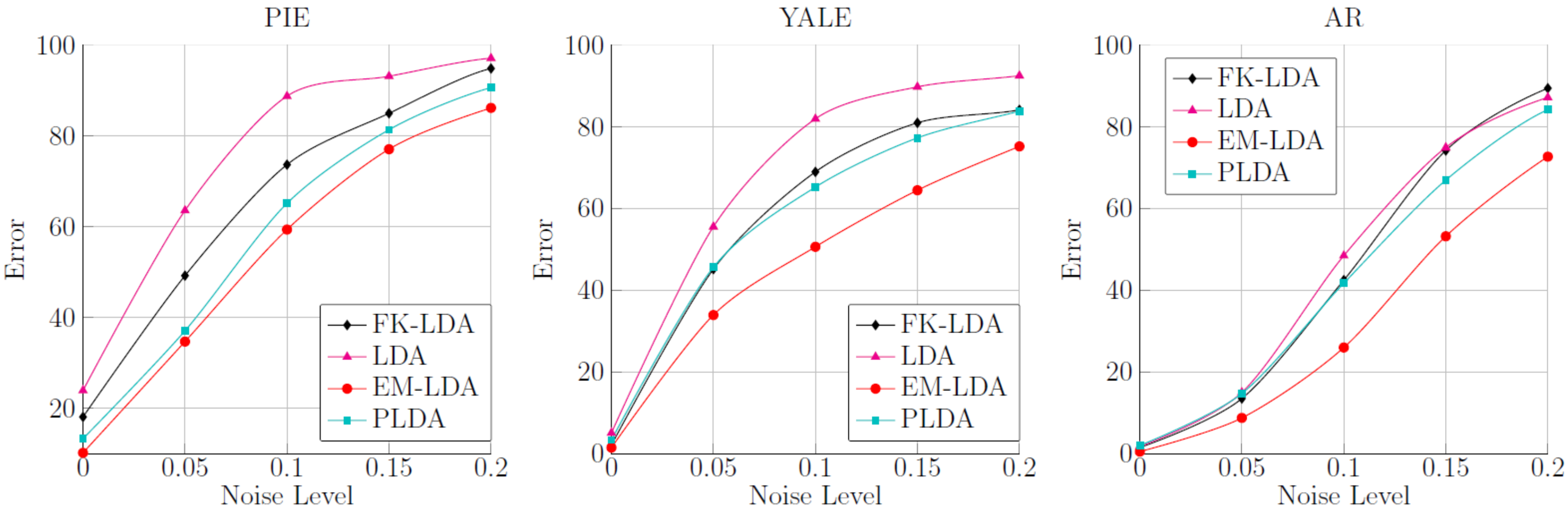
Roadmap

- **Introduction**
- **Overview of CA techniques**
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - Locality Preserving Projections (LPP)
  - Slow Feature Analysis (SFA)
- **Steps to Unification**
- **Unified Maximum Likelihood framework**
  - Defining priors and Markov random fields
  - Maximum likelihood solution
- **Unified Expectation Minimization framework**
  - Generalizing the prior
  - Expectation step
  - Minimization step
- **Experiments**
- **Discussions**

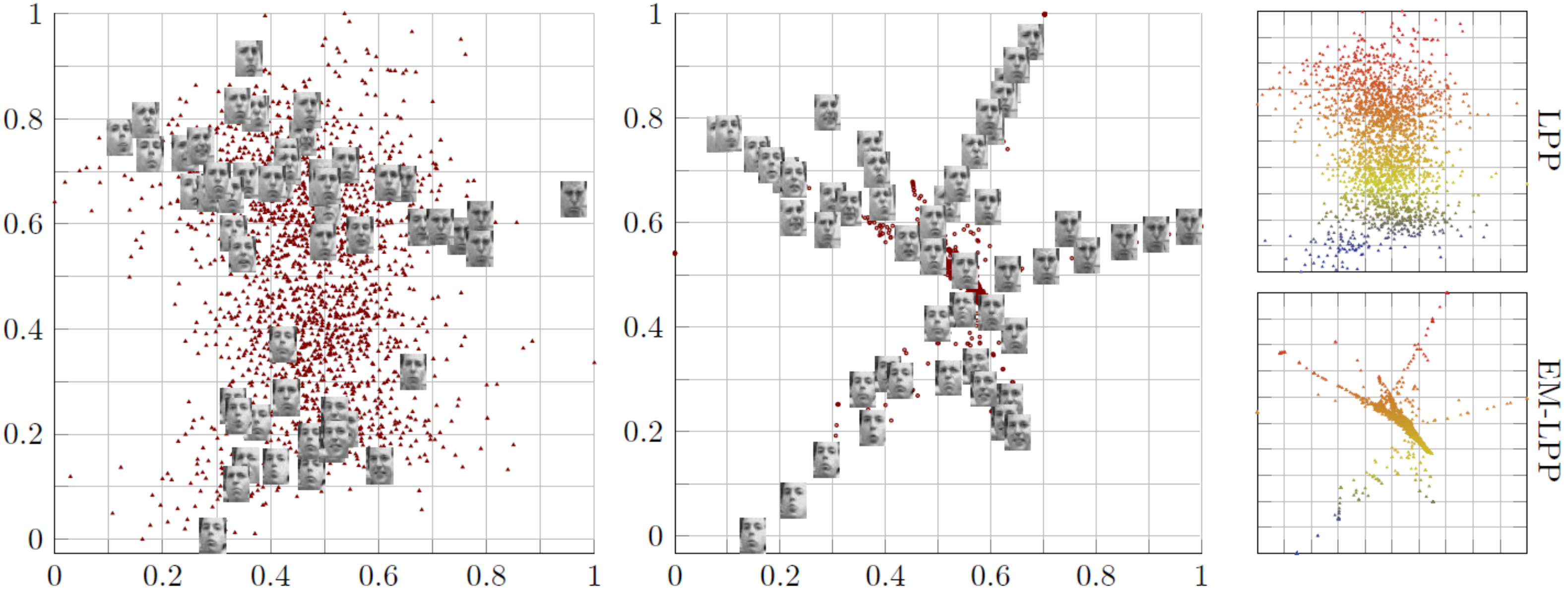
Proof of equivalence



Face recognition: EM-LDA



Face Visualization: EM-LPP



(a) LPP (b) EM-LPP (c)

Roadmap

- **Introduction**
- **Overview of CA techniques**
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - Locality Preserving Projections (LPP)
  - Slow Feature Analysis (SFA)
- **Steps to Unification**
- **Unified Maximum Likelihood framework**
  - Defining priors and Markov random fields
  - Maximum likelihood solution
- **Unified Expectation Minimization framework**
  - Generalizing the prior
  - Expectation step
  - Minimization step
- **Experiments**
- **Discussions**

Discussions(1)

- All component analysis methods are constraint based subspace projection
- Subspace methods can be modeled probabilistically
  - By defining a prior as product of MRFs having different latent neighborhood connectivity
  - Estimating maximum likelihood depending on a linear model with white Gaussian noise
- An EM algorithm for each of the subspace method can be proposed
  - Use of mean field approximation and MRF priors give us the updates

Discussions(2)

- EM variants of these algorithms are compatible with state-of-art
- Most variants are less computationally complex
- This method models variance per dimension
- Efficient CA's can be generated just by varying prior MRF connectivity
- Experiments show the EM variants are more immune to noise in data and also more efficient

Questions?

Thank you...