RWTH AACHEN UNIVERSITY

Advanced Machine Learning
Lecture 15

Latent Factor Models

17.12.2012

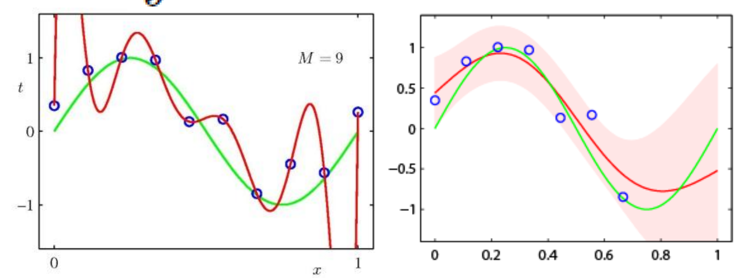
Bastian Leibe
RWTH Aachen
http://www.vision.rwth-aachen.de/

leibe@vision.rwth-aachen.de

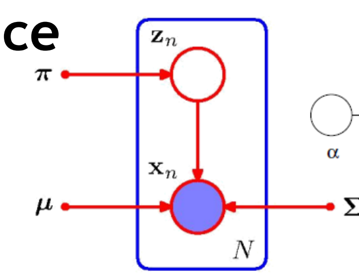
This Lecture: *Advanced Machine Learning*

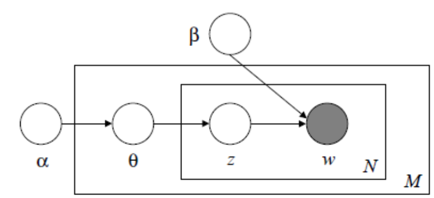
- Regression Approaches
  - Linear Regression
  - Regularization (Ridge, Lasso)
  - Kernels (Kernel Ridge Regression)
  - Gaussian Processes

$$f : \mathcal{X} \to \mathbb{R}$$



- Bayesian Estimation & Bayesian Non-Parametrics
  - Prob. Distributions, Approx. Inference
  - Mixture Models & EM
  - Dirichlet Processes
  - Latent Factor Models
  - Beta Processes





- SVMs and Structured Output Learning
  - SV Regression, SVDD
  - Large-margin Learning

$$f : \mathcal{X} \to \mathcal{Y}$$

Topics of This Lecture

- Latent Factor Models
  - Motivation
  - Example: PCA
  - Applications of PCA
  - Probabilistic PCA
  - Maximum Likelihood for PCA
  - Other Latent Factor Models: FA, ICA
- Towards Infinite Latent Factor Models
  - General formulation
  - Sparse latent factor models
  - Priors on binary matrices
  - Finite latent feature model

Mixture Models vs. Latent Factor Models

- **Mixture Models**
  - Assume that each observation was generated by exactly one of K components.
  - The uncertainty is just about which component is responsible.

- **Latent Factor Models**
  - Weaken this assumption.
  - Each observation is influenced by each of K components (factors or features) in a different way.
  - Sparse factor models: only a small subset of factors is active for each observation.

Latent Factor/Feature Models

- Most popular examples
  - Principal Component Analysis (PCA)
  - Factor Analysis (FA)
  - Independent Component Anlalysis (ICA)

- Properties
  - All of those assume that the number of factors K is known.
  - Usually, K is smaller than the dimensionality of the data: $K \ll D$

  $\Rightarrow$ Models provide dimensionality reduction.

- *Let's look at PCA and see how it fits into this framework...*

Principal Component Analysis

- **Goal**
  - Given a data set $X = \{\mathbf{x}_n\}$ in D dimensions, find the K-dimensional projection ($K < D$) that maximizes the variance of the projected data.
  - Intuition: preserve as much variance as possible.

- **One-dimensional example**
  - Project each data point $\mathbf{x}_n$ onto the unit vector $\mathbf{u}_1$

$$y_n = \mathbf{u}_1^T \mathbf{x}_n$$

  - What is the vector $\mathbf{u}_1$ that maximizes the variance of the projected data?

Principal Component Analysis

- **One-dimensional example (cont'd)**
  - Mean of the projected data

$$\bar{y} = \mathbf{u}_1^T \bar{\mathbf{x}} \qquad \bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n$$

  - Variance of the projected data

$$\frac{1}{N}\sum_{n=1}^{N}\{\mathbf{u}_1^T \mathbf{x}_n - \mathbf{u_1}^T \bar{\mathbf{x}}\}^2 = \mathbf{u_1}^T \mathbf{S} \mathbf{u}_1$$

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T$$

where S is the data covariance matrix.

Principal Component Analysis

- Optimization problem
  - Maximize the projected variance $\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1$ w.r.t. $\mathbf{u}_1$.
  - Problem: trivial solution is $\|\mathbf{u}_1\| \to \infty$.

  $\Rightarrow$ Need to enforce the normalization condition $\mathbf{u}_1^{\mathrm{T}} \mathbf{u}_1 = 1$.

  - Formulation with Lagrange multiplier

$$\arg\max_{\mathbf{u}_1} \quad \mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 + \lambda_1 (1 - \mathbf{u}_1^T \mathbf{u}_1)$$

  - Setting the derivative to zero

$$\mathbf{S}\mathbf{u}_1 = \lambda_1 \mathbf{u}_1$$

  $\Rightarrow$ Eigenvalue problem: $\mathbf{u}_1$ must be eigenvector of $\mathbf{S}$.

$$\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 = \mathbf{u}_1^T \lambda_1 \mathbf{u}_1 = \lambda_1 \mathbf{u}_1^T \mathbf{u}_1 = \lambda_1$$

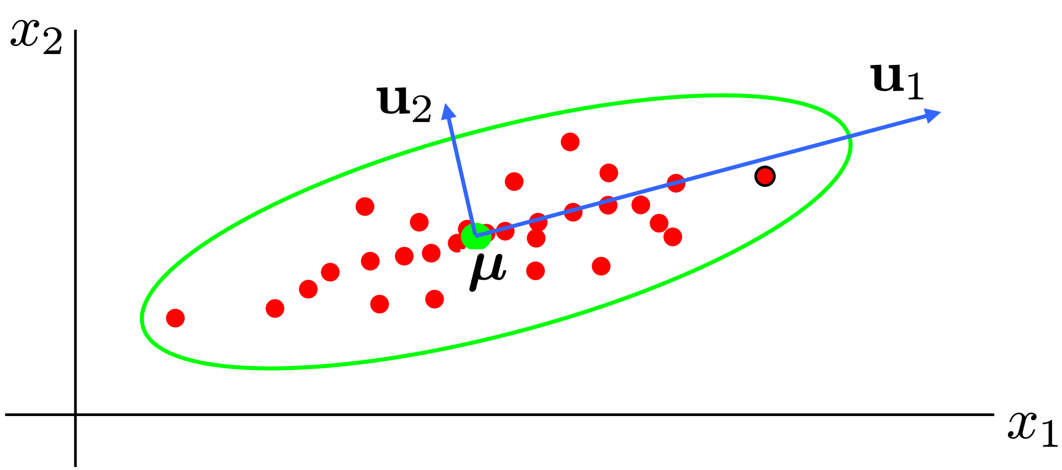
  $\Rightarrow$ Maximal variance if λ_1 is the largest eigenvalue of $\mathbf{S}$.

Principal Component Analysis

- **General case**
  - Inductively, we can show that the optimal linear projection into a K-dimensional space is given by the first K eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_K$ of S.

$$\mathbf{y}_n = \mathbf{U}_{1..K} \mathbf{x}_n$$

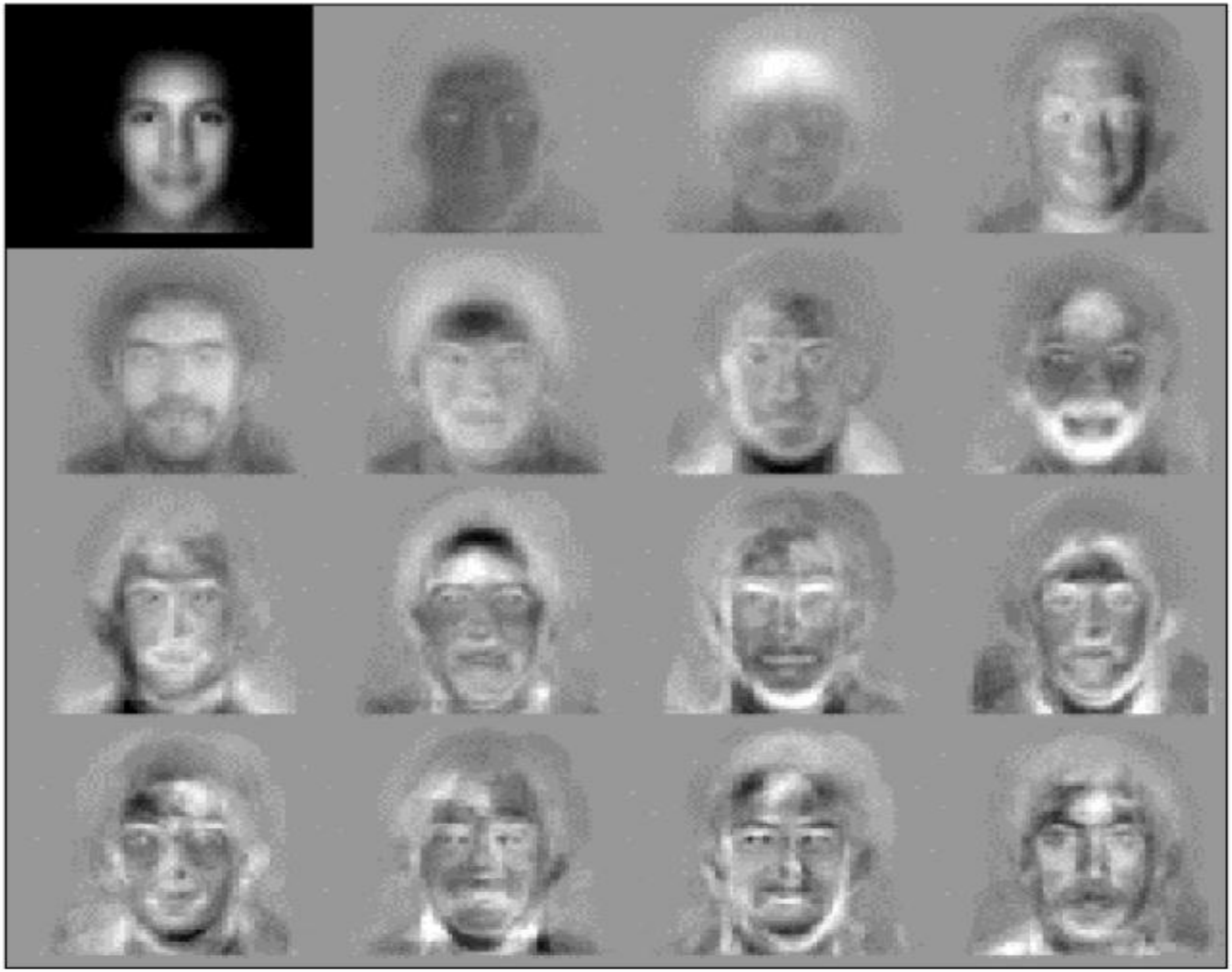
- **Graphical interpretation**



Uses of PCA

- **Dimensionality reduction**
  - Work in a subspace that contains only the K most important dimensions.
  - Advantages: faster processing, reduced memory footprint, robustness to noise.

Example: Eigenfaces [Turk & Pentland, 1993]



Slide credit: Peter Belhumeur

Uses of PCA

- Dimensionality reduction
  - Work in a subspace that contains only the K most important dimensions.
  - Advantages: faster processing, reduced memory footprint, robustness to noise.

- Data Preprocessing
  - Remove correlations between different dimensions of the data and bring them to a common scale.
  - Many classification or regression algorithms work better when the data is standardized, i.e., when each variable has zero mean and unit variance.
  - Using PCA, we can make a more substantial normalization of the data to give it zero mean and unit covariance. This is known as whitening.

PCA for Whitening

- Whitening procedure
  - Rewrite the eigenvector equation in matrix form

$$\mathbf{S}\mathbf{u}_1 = \lambda_1 \mathbf{u}_1 \qquad \Rightarrow \qquad \mathbf{S}\mathbf{U} = \mathbf{U}\mathbf{L}$$

where $\mathbf{L} = \mathrm{diag}\{\lambda_i\}, \ \mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_D]$.

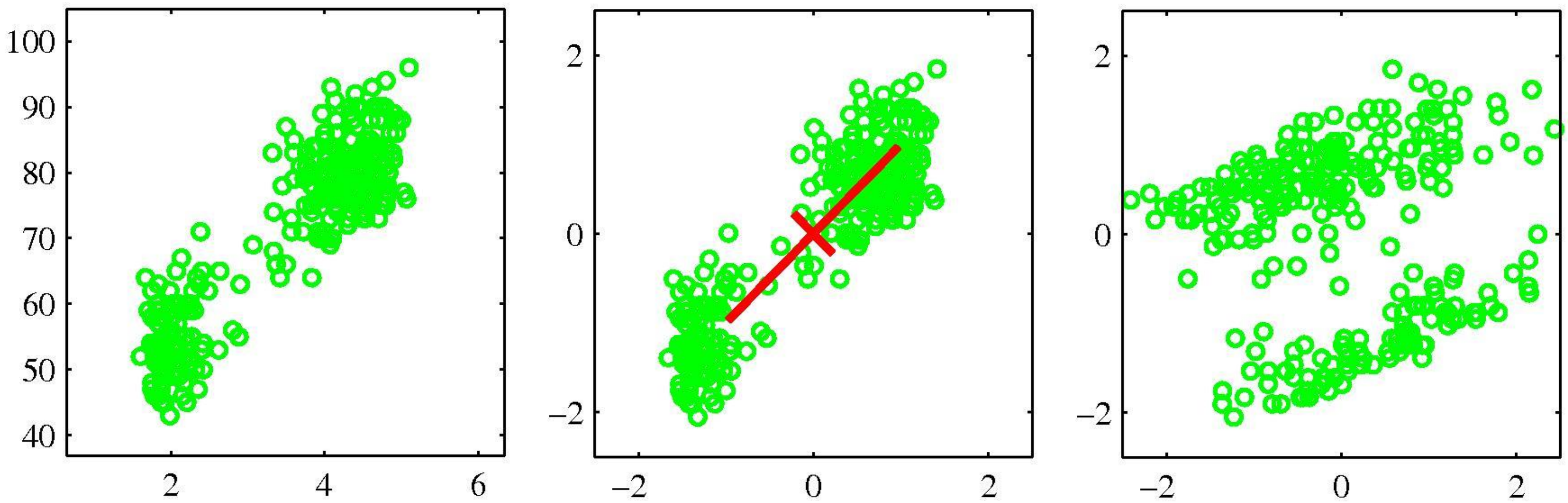
  - Define for each data point the transformed value as

$$\mathbf{y}_n = \mathbf{L}^{-1/2}\mathbf{U}^T(\mathbf{x}_n - \bar{\mathbf{x}})$$

$\Rightarrow$ The transformed set $\{\mathbf{y}_n\}$ has zero mean and unit covariance.

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_n\mathbf{y}_n^T &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{L}^{-1/2}\mathbf{U}^T(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}})^T\mathbf{U}\mathbf{L}^{-1/2} \\ &= \mathbf{L}^{-1/2}\mathbf{U}^T\mathbf{S}\mathbf{U}\mathbf{L}^{-1/2} = \mathbf{L}^{-1/2}\mathbf{L}\mathbf{L}^{-1/2} = \mathbf{I}\end{aligned}$$

Whitening Example



Original data

Principal axes

Whitened data

- Whitening result
  - Correlations are removed.
  - Distances are normalized to same value range.

Image source: C.M. Bishop

Probabilistic PCA

- **Discussion**
  - The formulation of PCA we have just seen was based on a linear projection of data into a lower-dim. subspace.
  - We now show that PCA can also be expressed as the ML solution of a probabilistic latent variable model.

- **Advantages of Probabilistic PCA**
  - We can derive an EM algorithm that is efficient in situations where only few leading eigenvectors are required.
  - Probabilistic model + EM makes it possible to deal with missing data values.
  - Basis for a Bayesian treatment of PCA in which the dimensionality of the principal subspace can be found automatically.

Topics of This Lecture

- Latent Factor Models
  - Motivation
  - Example: PCA
  - Applications of PCA
  - Probabilistic PCA
  - Maximum Likelihood for PCA
  - Other Latent Factor Models: FA, ICA
- Towards Infinite Latent Factor Models
  - General formulation
  - Sparse latent factor models
  - Priors on binary matrices
  - Finite latent feature model

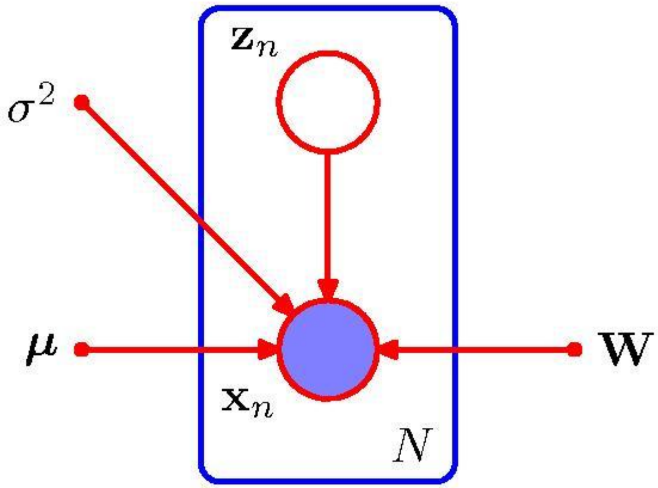
Probabilistic PCA

- Graphical Model
  - Introduce an explicit latent variable $\mathbf{z}$ corresponding to the principal component subspace.
  - Define a Gaussian prior distribution

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$$

  - Conditional distribution also Gaussian

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$$



$\Rightarrow$ Example of a Linear Gaussian framework: all of the marginal and conditional distributions are Gaussian

  - As we will see, the columns of $\mathbf{W}$ span an K-dimensional linear subspace within the data space that corresponds to the principal subspace.

Image source: C.M. Bishop

Probabilistic PCA

- **Generative interpretation**
  - D-Dimensional observed variable $\mathbf{x}$ is defined by a linear transformation of the K-dimensional latent variable $\mathbf{z}$, plus some added (isotropic Gaussian) noise.

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

  - Marginal distribution

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z}$$

  - Because of the linear-Gaussian model, this will again be Gaussian

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{C})$$

where

$$\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

Probabilistic PCA

- Properties
  - There is a rotational ambiguity in the parametrization.
  - Consider a rotation of the latent parameter space with orthonormal matrix $\mathbf{R}$ (orthogonality property: $\mathbf{R}\mathbf{R}^{\mathrm{T}} = \mathbf{I}$).

$$\widetilde{\mathbf{W}} = \mathbf{W}\mathbf{R}$$

$$\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^T = \mathbf{W}\mathbf{R}\mathbf{R}^T\mathbf{W}^T = \mathbf{W}\mathbf{W}^T$$

$\Rightarrow$ Thus, the covariance matrix $\mathbf{C}$ is independent of $\mathbf{R}$.

  - Efficiency trick: instead of evaluating $\mathbf{C}^{-1}$ directly, use the following equivalence ($\mathcal{O}(D^3) \rightarrow \mathcal{O}(K^3)$).

$$\mathbf{C}^{-1} = \sigma^{-2}\mathbf{I} - \sigma^{-2}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T$$

with $$\mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}$$

Probabilistic PCA

- **Posterior distribution**
  - Can again be derived from properties of linear Gaussian models

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z}|\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x}-\boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1}\right)$$

Topics of This Lecture

- Latent Factor Models
  - Motivation
  - Example: PCA
  - Applications of PCA
  - Probabilistic PCA
  - Maximum Likelihood for PCA
  - Other Latent Factor Models: FA, ICA
- Towards Infinite Latent Factor Models
  - General formulation
  - Sparse latent factor models
  - Priors on binary matrices
  - Finite latent feature model

Maximum Likelihood for PCA

- **Maximum Likelihood estimate**
  - **Log-likelihood function**

$$
\begin{aligned}
\log p(\mathbf{X}|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) &= \sum_{n=1}^{N} \log p(\mathbf{x}_n|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) \\
&= -\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) \\
&\quad -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\mathbf{C}|
\end{aligned}
$$

  - **Optimizing the parameters**

$$
\frac{\partial}{\partial \boldsymbol{\mu}} p(\mathbf{X}|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) \stackrel{!}{=} 0 \qquad \Rightarrow \qquad \boldsymbol{\mu} = \bar{\mathbf{x}}
$$

Maximum Likelihood for PCA

- **Maximum Likelihood estimate**
  - Plugging in the result for μ...

$$\log p(\mathbf{X}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2) = -\frac{N}{2}\left\{D\log(2\pi) + \log|\mathbf{C}| + \mathrm{Tr}(\mathbf{C}^{-1}\mathbf{S})\right\}$$

  - Maximizing w.r.t. $\mathbf{W}$ yields a closed-form solution:

$$\mathbf{W}_{\mathrm{ML}} = \mathbf{U}_K(\mathbf{L}_K - \sigma^2\mathbf{I})^{1/2}\mathbf{R}$$

  - where
    - $\mathbf{U}_K$ is a $D \times K$ matrix, whose columns are given by the K principal eigenvectors of the data covariance matrix $\mathbf{S}$,
    - $\mathbf{L}$ contains eigenvalues λ_i, and
    - $\mathbf{R}$ is an arbitrary $K \times K$ rotation matrix.

$\Rightarrow$ The columns of $\mathbf{W}$ define the principal subspace of standard PCA. For $\mathbf{R} = \mathbf{I}$, they correspond to the principal eigenvectors $[\mathbf{u}_1, \ldots, \mathbf{u}_K]$, scaled by the variance parameters λ_i - σ^2.

Maximum Likelihood for PCA

- **Maximum Likelihood estimate (cont'd)**
  - **Maximizing w.r.t. σ:**

$$\sigma^2_{\mathrm{ML}} = \frac{1}{D-K} \sum_{i=K+1}^{D} \lambda_i$$

$\Rightarrow$ ${\sigma^2}_{\mathrm{ML}}$ **is the average variance associated with the discarded dimensions.**

Interpretation of Probabilistic PCA

- **Putting all those results together...**
  - Consider again the covariance matrix

$$\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

where

$$\mathbf{W}_{\mathrm{ML}} = \mathbf{U}_K(\mathbf{L}_K - \sigma^2\mathbf{I})^{1/2}\mathbf{R}$$

$$\sigma^2_{\mathrm{ML}} = \frac{1}{D-K}\sum_{i=K+1}^{D}\lambda_i$$

$\Rightarrow$ The model correctly captures the variance of the data along the principal axes and approximates the variance in all remaining directions by σ^2, the average of the discarded eigenvalues.

  - To construct $\mathbf{C}$, we simply set $\mathbf{R} = \mathbf{I}$ and compute the principal eigenvalues and eigenvectors of the data covariance matrix $\mathbf{S}$.

$\Rightarrow$ If $\mathbf{C}$ is obtained in a different way, $\mathbf{R}$ may still be arbitrary.

Discussion: PCA vs. Probabilistic PCA

- **Comparison with standard PCA:**
  - PCA is generally formulated as a projection of points from the D-dimensional space onto a K-dimensional linear subspace.
  - Probabilistic PCA is more naturally expressed as a mapping from the latent space into the data space via

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

  - For applications such as visualization or data compression, we can reverse this mapping using Bayes' theorem.

$$\mathbb{E}[\mathbf{z}|\mathbf{x}] = \mathbf{M}^{-1}\mathbf{W}_{\mathrm{ML}}^{T}(\mathbf{x} - \bar{\mathbf{x}}) \quad \text{where} \quad \mathbf{M} = \mathbf{W}^{T}\mathbf{W} + \sigma^{2}\mathbf{I}$$

  - This projects to a point in data space given by

$$\mathbf{W}\mathbb{E}[\mathbf{z}|\mathbf{x}] + \boldsymbol{\mu}$$

  - In the limit $\sigma \to 0$, this reduces to the standard PCA model.

Topics of This Lecture

- **Latent Factor Models**
  - Motivation
  - Example: PCA
  - Applications of PCA
  - Probabilistic PCA
  - Maximum Likelihood for PCA
  - Other Latent Factor Models: FA, ICA
- **Towards Infinite Latent Factor Models**
  - General formulation
  - Sparse latent factor models
  - Priors on binary matrices
  - Finite latent feature model

Other Latent Factor Models

- Factor Analysis (FA)
  - Linear-Gaussian latent variable model, closely related to Probabilistic PCA.
  - Probabilistic PCA uses an isotropic covariance

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$$

  - Factor Analysis instead assumes a diagonal covariance

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \boldsymbol{\Psi}) \qquad \boldsymbol{\Psi} = \mathrm{diag}\{\psi_i\}$$

  - The FA model explains the observed covariance structure of the data by representing the independent variables associated with each coordinate by the matrix $\boldsymbol{\Psi}$ and capturing the covariance between variables in the matrix $\mathbf{W}$.
  - In the literature, the columns of $\mathbf{W}$ are called factor loadings, the diagonal elements ψ_i are called uniquenesses.

Other Latent Factor Models (2)

- Independent Component Analysis (ICA)
  - Model for which the observed variables are related linearly to the latent variables, but for which the latent distribution is non-Gaussian.
  - Consider a distribution over latent variables that factorizes

$$p(\mathbf{z}) = \prod_{j=1}^{K} p(z_j)$$

    i.e., the components z_j are independent.
  - This definition requires that the latent variables have a non-Gaussian distribution (as Gaussian models always have the rotational ambiguity $\mathbf{R}$ in latent space).
  - There is a large variety of ICA models and corresponding algorithms, differing mainly in the choice of latent-variable distribution.

Next Steps from Here...

- Discussion
  - We have now derived that the PCA result can be obtained as the ML estimate of the corresponding probabilistic model.
  - This result can directly be used to incorporate priors and derive a Bayesian extension of the model.
  - We can do similar things for FA and ICA...

- In the following, we will go into a different direction
  - What happens when we let $K \to \infty$?
  - Can we automatically determine K?

Topics of This Lecture

- Latent Factor Models
  - Motivation
  - Example: PCA
  - Applications of PCA
  - Probabilistic PCA
  - Maximum Likelihood for PCA
  - Other Latent Factor Models: FA, ICA
- Towards Infinite Latent Factor Models
  - General formulation
  - Sparse latent factor models
  - Priors on binary matrices
  - Finite latent feature model

General Latent Factor Models

- General formulation
  - Assume that the data are generated by noisy weighted combination of latent factors

$$\mathbf{x}_n = \mathbf{F}\mathbf{y}_n + \epsilon$$

  - E.g., in Factor Analysis, $\mathbf{F}$ would be a $D \times K$ ***factor loading matrix*** expressing how latent factor k influences observation dimension d. $\mathbf{y}_n$ would be a K-dimensional vector expressing the activity of each factor.

- Advantages of latent feature modeling
  - Each group of observations is associated with a subset of the possible latent features/factors.
  - Factorial power: There are 2^K combinations of K features, while accurate mixture modeling may require many more clusters.

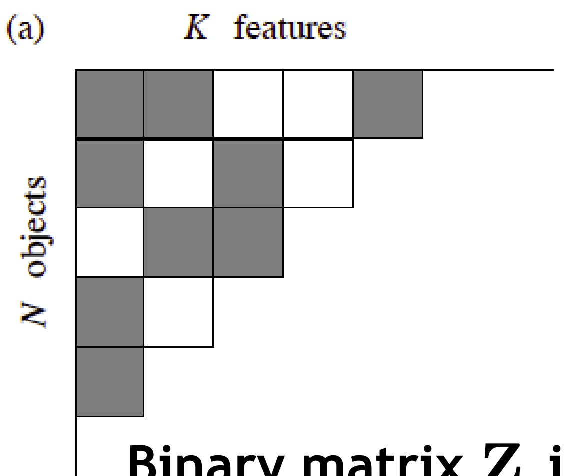
Sparse Latent Factor Models

- Goal: Infinite models
  - We would like to work with infinite-dimensional models ($K \rightarrow \infty$)
  - In order to do keep inference tractable, however, we have to restrict the model somehow.
  - Mixture Models: DPs enforce that the main part of the probability mass is concentrated on few cluster components.
  - Latent Factor Models: enforce that each object is represented by *a sparse subset* of an unbounded number of features.

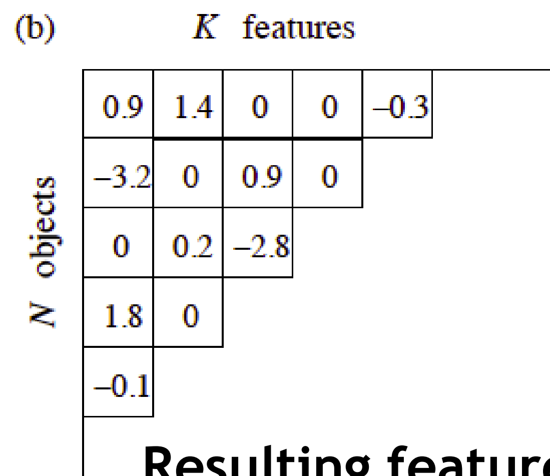
- Incorporating sparsity
  - Decompose $\mathbf{F}$ into the product of two components: $\mathbf{F} = \mathbf{Z} \otimes \mathbf{W}$, where $\otimes$ is the Hadamard product (element-wise product).
    - z_{mk} is a binary mask variable indicating whether factor k is "on".
    - w_{mk} is a continuous weight variable.

  $\Rightarrow$ Enforce sparsity by restricting the non-zero entries in $\mathbf{Z}$.

Sparse Latent Factor Models



Binary matrix $\mathbf{Z}$ indicating feature presence/absence



Resulting feature matrix $\mathbf{F}$ after multiplication with $\mathbf{W}$

- **Latent feature modeling**
  - In PCA (FA, ICA, etc.), objects have non-zero values on every feature and every entry of $\mathbf{Z}$ is 1.
  - In sparse latent feature models, only a sparse subset of features take non-zero values, and $\mathbf{Z}$ makes those subsets explicit.

Image source: Zoubin Gharahmani

Towards a Full Bayesian Treatment

- Inference in Latent Feature Models
  - Goal: Infer the latent factors, mask variables, and weights.
  - Classical approaches (PCA, FA, ICA) fit point estimates of the parameters through ML estimation.

- Bayesian approach
  - Specify a prior over latent features/factors $p(\mathbf{F})$ and a distribution over observed property distributions $p(\mathbf{X}|\mathbf{F})$.
  - Compute the posterior $p(\mathbf{F}, \mathbf{Z}, \mathbf{W}|\mathbf{X})$.
  - Our focus will be on $p(\mathbf{F}) = p(\mathbf{Z})p(\mathbf{W})$, showing how such a prior can be defined without placing an upper bound on the number of features/factors.

Priors on Binary Matrices

- Let's first go back to DPs/CRPs
  - Back there, we also had binary matrices due to 1-of-K coding.
  - *What is different here?*

- Binary matrices for clustering
  - We can think of CRPs as priors on infinite binary matrices, where...
  - ...each data point is assigned to one and only one cluster (class).
  - ...the rows sum to one.

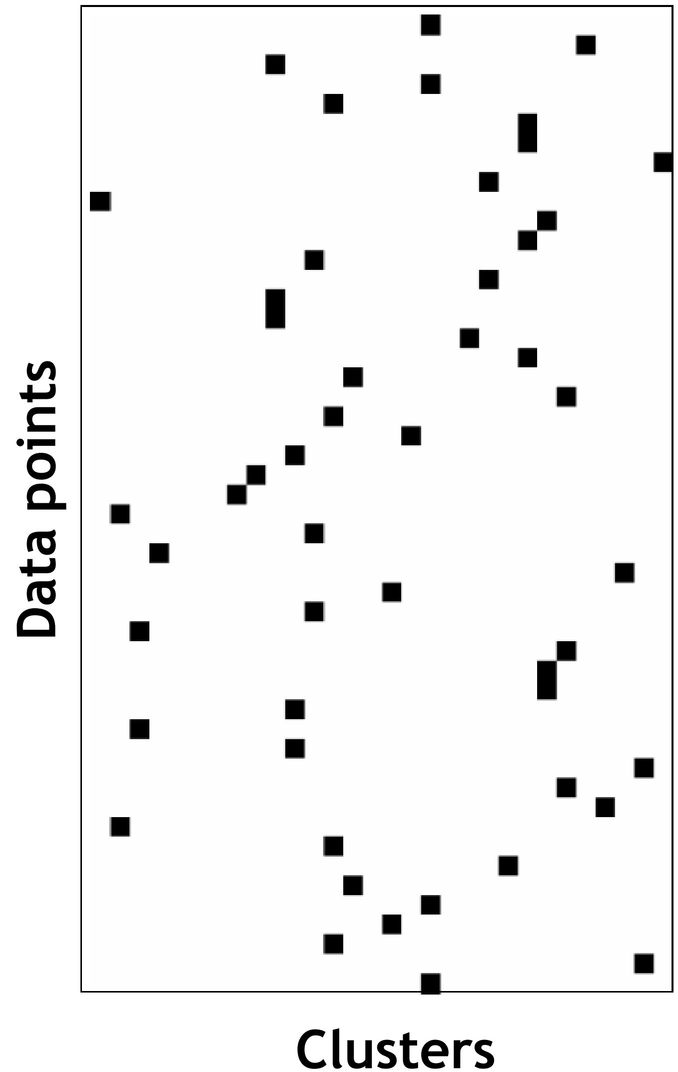


Image source: Zoubin Gharahmani

Priors on Binary Matrices

- Let's first go back to DPs/CRPs
  - Back there, we also had binary matrices due to 1-of-K coding.
  - *What is different here?*

- More general binary matrices
  - Each data point can now have multiple factors/features.
  - The rows sum to more than one.
  - $\Rightarrow$ *What is the corresponding prior on infinite binary matrices?*

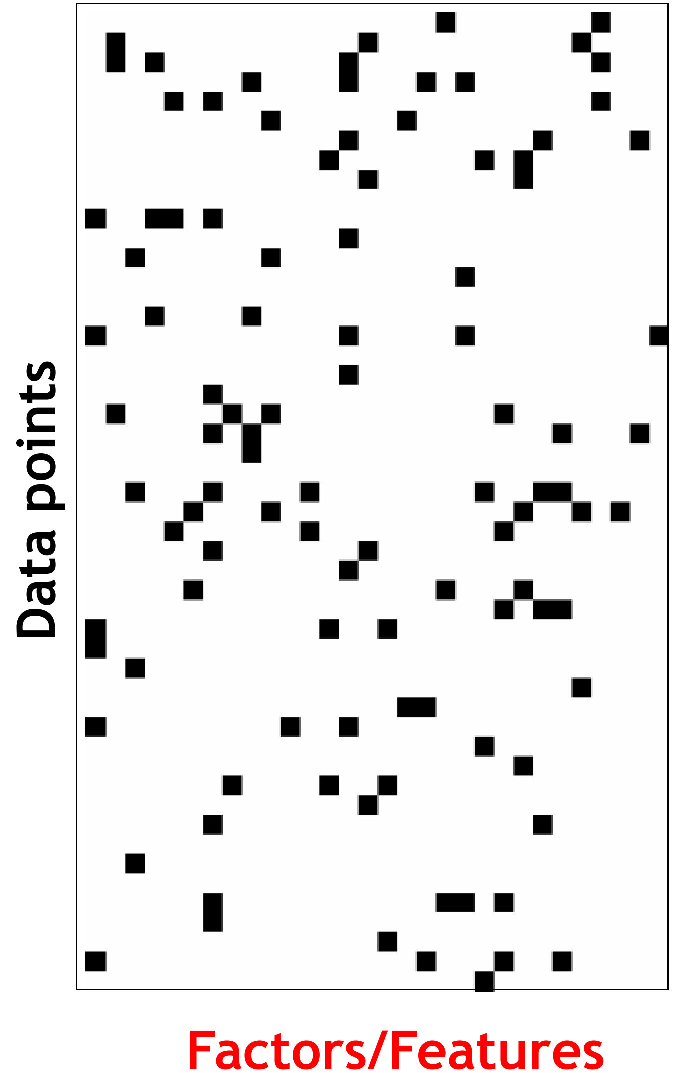


Image source: Zoubin Gharahmani

Priors on Latent Factor Models

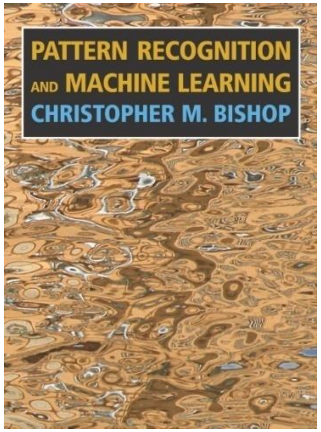
- Defining suitable priors
  - We will focus on defining a prior on $\mathbf{Z}$, since the effective dimensionality of the latent feature model is determined by $\mathbf{Z}$.
  - Assuming that $\mathbf{Z}$ is sparse, we can define a prior for infinite latent feature models by defining a distribution over infinite binary matrices.
- Desiderata for such a distribution
  - Objects should be exchangeable.
  - Inference should be tractable.
- Procedure
  - Start with a model that assumes a finite number of features and consider the limit as this number approaches infinity.

  $\Rightarrow$ *Next lecture...*

References and Further Reading

- More information on latent factor models and particularly PCA can be found in Chapter 12 of

Christopher M. Bishop
Pattern Recognition and Machine Learning
Springer, 2006



- Tutorial papers for infinite latent factor models
  - A good introduction to the topic
    - Z. Ghahramani, T.L. Griffiths, P. Sollich, "Bayesian Nonparametric Latent Feature Models", Bayesian Statistics, 2006.