Factor Analysis and Beyond

Chris Williams

School of Informatics, University of Edinburgh

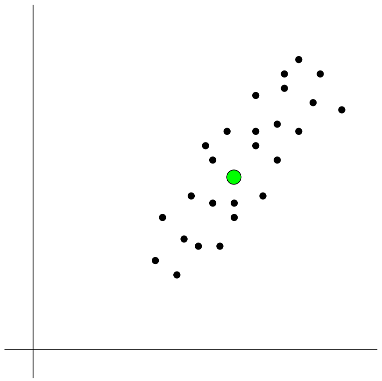
October 2011

Overview

- Principal Components Analysis
- Factor Analysis
- Independent Components Analysis
- Non-linear Factor Analysis
- Reading: Handout on "Factor Analysis and Beyond", Bishop §12.1, 12.2 (but not 12.2.1, 12.2.2, 12.2.3), 12.4 (but not 12.4.2)

Covariance matrix

- Let $\langle \quad \rangle$ denote an average
- Suppose we have a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_d)^T$
- $\langle \mathbf{X} \rangle$ denotes the mean of $\mathbf{X}$, $(\mu_1, \mu_2, \ldots \mu_d)^T$
- $\sigma_{ii} = \langle (X_i - \mu_i)^2 \rangle$ is the variance of component i (gives a measure of the "spread" of component i)
- $\sigma_{ij} = \langle (X_i - \mu_i)(X_j - \mu_j) \rangle$ is the covariance between components i and j



- In d-dimensions there are d variances and $d(d-1)/2$ covariances which can be arranged into a *covariance matrix* Σ
- The *population* covariance matrix is denoted Σ, the *sample* covariance matrix is denoted S

Principal Components Analysis

If you want to use a single number to describe a whole vector drawn from a known distribution, pick the projection of the vector onto the direction of maximum variation (variance)

- Assume $\langle \mathbf{x} \rangle = \mathbf{0}$
- $y = \mathbf{w}.\mathbf{x}$
- Choose $\mathbf{w}$ to maximize $\langle y^2 \rangle$, subject to $\mathbf{w}.\mathbf{w} = 1$
- Solution: $\mathbf{w}$ is the eigenvector corresponding to the largest eigenvalue of $\Sigma = \langle \mathbf{x}\mathbf{x}^T \rangle$

- Generalize this to consider projection from d dimensions down to m
- Σ has eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_d \geq 0$
- The directions to choose are the first m eigenvectors of Σ corresponding to $\lambda_1, \ldots, \lambda_m$
- $\mathbf{w}_i.\mathbf{w}_j = 0 \qquad i \neq j$
- Fraction of total variation explained by using m principal components is
$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{d} \lambda_i}$$
- PCA is basically a rotation of the axes in the data space

Factor Analysis

- A latent variable model; can the observations be explained in terms of a small number of unobserved latent variables ?
- FA is a proper statistical model of the data; it explains covariance between variables rather than variance (*cf* PCA)
- FA has a controversial rôle in social sciences

- visible variables : $\mathbf{x} = (x_1, \ldots, x_d)$,
- latent variables: $\mathbf{z} = (z_1, \ldots, z_m)$, $\mathbf{z} \sim N(0, I_m)$
- noise variables: $\mathbf{e} = (e_1, \ldots, e_d)$, $\mathbf{e} \sim N(0, \Psi)$, where $\Psi = \mathrm{diag}(\psi_1, \ldots, \psi_d)$.

Assume
$$\mathbf{x} = \boldsymbol{\mu} + W\mathbf{z} + \mathbf{e}$$
then covariance structure of $\mathbf{x}$ is
$$C = WW^T + \Psi$$
W is called the factor loadings matrix

$p(\mathbf{x})$ is like a multivariate Gaussian pancake

$$p(\mathbf{x}|\mathbf{z}) \sim N(W\mathbf{z} + \boldsymbol{\mu}, \Psi)$$
$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$$
$$p(\mathbf{x}) \sim N(\boldsymbol{\mu}, WW^T + \Psi)$$

- Rotation of solution: if W is a solution, so is WR where $RR^T = I_m$ as $(WR)(WR)^T = WW^T$. Causes a problem if we want to interpret factors. Unique solution can be imposed by various conditions, e.g. that $W^T\Psi^{-1}W$ is diagonal.
- Is the FA model a simplification of the covariance structure? S has $d(d+1)/2$ independent entries. Ψ and W together have $d + dm$ free parameters (and uniqueness condition above can reduce this). FA model makes sense if number of free parameters is less than $d(d+1)/2$.

FA example

[from Mardia, Kent & Bibby, table 9.4.1]

- Correlation matrix

$$\begin{array}{l} \text{mechanics} \\ \text{vectors} \\ \text{algebra} \\ \text{analysis} \\ \text{statstics} \end{array} \begin{pmatrix} 1 & 0.553 & 0.547 & 0.410 & 0.389 \\ & 1 & 0.610 & 0.485 & 0.437 \\ & & 1 & 0.711 & 0.665 \\ & & & 1 & 0.607 \\ & & & & 1 \end{pmatrix}$$

- Maximum likelihood FA (impose that $W^T\Psi^{-1}W$ is diagonal). Require $m \leq 2$ otherwise more free parameters than entries in S.

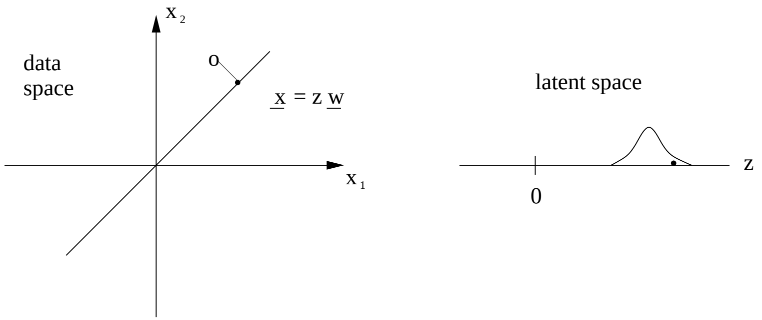
Variable	m = 1 $\mathbf{w}_1$	m = 2 $\mathbf{w}_1$	(not rotated) $\mathbf{w}_2$	m = 2 $\tilde{\mathbf{w}}_1$	(rotated) $\tilde{\mathbf{w}}_2$
1	0.600	0.628	0.372	0.270	0.678
2	0.667	0.696	0.313	0.360	0.673
3	0.917	0.899	-0.050	0.743	0.510
4	0.772	0.779	-0.201	0.740	0.317
5	0.724	0.728	-0.200	0.698	0.286

- 1-factor and first factor of the 2-factor solutions differ (cf PCA)
- problem of interpretation due to rotation of factors

FA for visualization

$$p(\mathbf{z}|\mathbf{x}) \propto p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$$

Posterior is a Gaussian. If $\mathbf{z}$ is low dimensional. Can be used for visualization (as with PCA)



Learning W, Ψ

- Maximum likelihood solution available (Lawley/Jöreskog).
- EM algorithm for ML solution (Rubin and Thayer, 1982)
  - E-step: for each $\mathbf{x}_i$, infer $p(\mathbf{z}|\mathbf{x}_i)$
  - M-step: do linear regression from $\mathbf{z}$ to $\mathbf{x}$ to get W
- Choice of m difficult (see Bayesian methods later).

Comparing FA and PCA

- Both are linear methods and model second-order structure S
- FA is invariant to changes in scaling on the axes, but not rotation invariant (cf PCA).
- FA models *covariance*, PCA models *variance*

Probabilistic PCA

Tipping and Bishop (1997), see Bishop §12.2

Let $\Psi = \sigma^2 I$.

- In this case W_{ML} spans the space defined by the first m eigenvectors of S
- PCA and FA give same results as $\Psi \rightarrow 0$.

Example Application: Handwritten Digits Recognition

Hinton, Dayan and Revow, IEEE Trans Neural Networks 8(1), 1997

- Do digit recognition with class-conditional densities
- 8×8 images $\Rightarrow 64 \cdot 65/2$ entries in the covariance matrix.
- 10-dimensional latent space used
- Visualization of W matrix. Each hidden unit gives rise to a weight image ...
- In practice use a mixture of FAs!

Useful Texts

on PCA and FA

- B. S. Everitt and G. Dunn "Applied Multivariate Data Analysis" Edward Arnold, 1991.
- C. Chatfield and A. J. Collins "Introduction to Multivariate Analysis", Chapman and Hall, 1980.
- K. V. Mardia, J. T. Kent and J. M. Bibby "Multivariate Analysis", Academic Press, 1979.

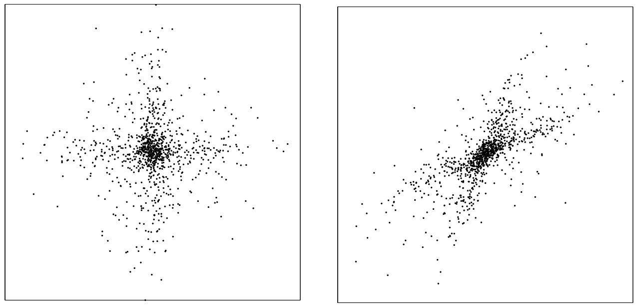
Independent Components Analysis

- A non-Gaussian latent variable model, plus linear transformation, e.g.

$$p(\mathbf{z}) \propto \prod_{i=1}^{m} e^{-|z_i|}$$

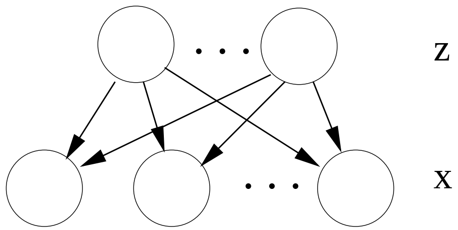
$$\mathbf{x} = W\mathbf{z} + \boldsymbol{\mu} + \mathbf{e}$$

- Rotational symmetry in **z**-space is now broken
- $p(\mathbf{x})$ is non-Gaussian, go beyond second-order statistics of data for fitting model
- Can be used with $\dim(\mathbf{z}) = \dim(\mathbf{x})$ for blind source separation
- http://www.cnl.salk.edu/~tony/ica.html
- Blind source separation demo: Te-Won Lee



unmixed mixed

A General View of Latent Variable Models



- Clustering: $\mathbf{z}$ is one-on-in-m encoding
- Factor analysis: $\mathbf{z} \sim N(0, I_m)$
- ICA: $p(\mathbf{z}) = \prod_i p(z_i)$, and each $p(z_i)$ is non-Gaussian
- Latent Dirichlet Allocation: $\mathbf{z} \sim \mathrm{Dir}(\boldsymbol{\alpha})$ (Blei et al, 2003). Used especially for "topic modelling" of documents

Non-linear Factor Analysis

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

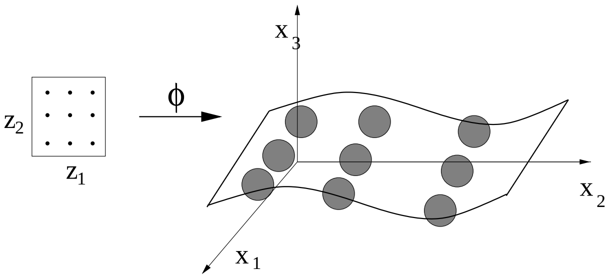
For PPCA

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

If we make the prediction of the mean a non-linear function of $\mathbf{z}$, we get non-linear factor analysis, with $p(\mathbf{x}|\mathbf{z}) \sim N(\phi(\mathbf{z}), \sigma^2 I)$ and $\phi(\mathbf{z}) = (\phi_1(\mathbf{z}), \phi_2(\mathbf{z}), \ldots, \phi_d(\mathbf{z}))^T$. However, there is a problem— we can't do the integral analytically, so we need to approximate it.

$$p(\mathbf{x}) \simeq \frac{1}{K}\sum_{k=1}^{K} p(\mathbf{x}|\mathbf{z}_k)$$

where the samples $\mathbf{z}_k$ are drawn from the density $p(\mathbf{z})$. Note that the approximation to $p(\mathbf{x})$ is a mixture of Gaussians.



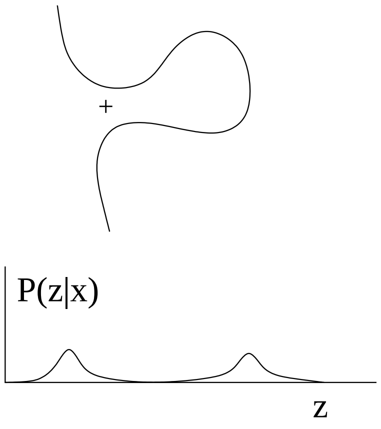
- Generative Topographic Mapping (Bishop, Svensen and Williams, 1997/8)
- Do GTM demo

Fitting the Model to Data

- Adjust the parameters of ϕ and σ^2 to maximize the log likelihood of the data.
- For a simple form of mapping $\phi(\mathbf{z}) = \sum_i \mathbf{w}_i \psi_i(\mathbf{z})$ we can obtain EM updates for the weights $\{\mathbf{w}_i\}$ and the variance σ^2.
- We are fitting a *constrained* mixture of Gaussians to the data. The algorithm works quite like Kohonen's self-organizing map (SOM), but is more principled as there is an objective function.

Visualization

- The mean may be a bad summary of the posterior distribution.



Manifold Learning

- A manifold is a topological space that is locally Euclidean
- We are particularly interested in the case of non-linear dimensionality reduction, where a low-dimensional nonlinear manifold is embedded in a high-dimensional space
- As well as GTM, there are other methods for non-linear dimensionality reduction. Some recent methods based on eigendecomposition include:
  - Isomap (Renenbaum et al, 2000)
  - Local linear embedding (Roweis and Saul, 2000)
  - Lapacian eigenmaps (Belkin and Niyogi, 2001)