Overview

Factor Analysis and Beyond

Chris Williams
School of Informatics, University of Edinburgh
October 2011

Covariance matrix

- Let \( \langle \cdot \rangle \) 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 \( \Sigma \)
- The population covariance matrix is denoted \( \Sigma \), 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 $\langle x \rangle = 0$
- $y = w^T x$
- Choose $w$ to maximize $\langle y^2 \rangle$, subject to $w^T w = 1$
- Solution: $w$ is the eigenvector corresponding to the largest eigenvalue of $\Sigma = \langle xx^T \rangle$

Generalize this to consider projection from $d$ dimensions down to $m$
- $\Sigma$ has eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_d \geq 0$
- The directions to choose are the first $m$ eigenvectors of $\Sigma$ corresponding to $\lambda_1, \ldots, \lambda_m$
- $w_i^T w_j = 0 \quad i \neq j$
- 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: $x = (x_1, \ldots, x_d)$,
- Latent variables: $z = (z_1, \ldots, z_m)$, $z \sim N(0, I_m)$
- Noise variables: $e = (e_1, \ldots, e_d)$, $e \sim N(0, \Psi)$, where $\Psi = \text{diag}(\psi_1, \ldots, \psi_d)$.

Assume $x = \mu + Wz + e$

Then covariance structure of $x$ is $C = WW^T + \Psi$

$W$ is called the factor loadings matrix
\( p(x) \) is like a multivariate Gaussian pancake
\[
p(x|z) \sim N(Wz + \mu, \Psi)
\]
\[
p(x) = \int p(x|z)p(z)dz
\]
\[
p(x) \sim N(\mu, WW^T + \Psi)
\]

Rotation of solution: if \( W \) is a solution, so is \( WR \) where \( RRT = I_m \) 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. \( \Psi \) 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]

<table>
<thead>
<tr>
<th>Variable</th>
<th>( w_1 )</th>
<th>( w_1 )</th>
<th>( w_2 )</th>
<th>( \tilde{w}_1 )</th>
<th>( \tilde{w}_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>mechanics</td>
<td>1</td>
<td>0.553</td>
<td>0.547</td>
<td>0.410</td>
<td>0.389</td>
</tr>
<tr>
<td>vectors</td>
<td>1</td>
<td>0.610</td>
<td>0.485</td>
<td>0.437</td>
<td></td>
</tr>
<tr>
<td>algebra</td>
<td>1</td>
<td>0.711</td>
<td>0.665</td>
<td></td>
<td></td>
</tr>
<tr>
<td>analysis</td>
<td>1</td>
<td>0.607</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>statistics</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

1-factor and first factor of the 2-factor solutions differ (cf PCA). Problem of interpretation due to rotation of factors.
FA for visualization

\[ p(z|x) \propto p(z)p(x|z) \]

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

Learning \( W, \Psi \)

- Maximum likelihood solution available (Lawley/Jöreskog).
- EM algorithm for ML solution (Rubin and Thayer, 1982)
  - E-step: for each \( x_i \), infer \( p(z|x_i) \)
  - M-step: do linear regression from \( z \) to \( x \) 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 \( \Psi \to 0 \).
Example Application: Handwritten Digits Recognition

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

- Do digit recognition with class-conditional densities
- $8 \times 8$ 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!

Independent Components Analysis

- A non-Gaussian latent variable model, plus linear transformation, e.g.
  \[
  p(z) \propto \prod_{i=1}^{m} e^{-|z_i|} \\
  x = Wz + \mu + e
  \]
- Rotational symmetry in $z$-space is now broken
- $p(x)$ is non-Gaussian, go beyond second-order statistics of data for fitting model
- Can be used with $\text{dim}(z) = \text{dim}(x)$ for blind source separation
- http://www.cnl.salk.edu/~tony/ica.html
- Blind source separation demo: Te-Won Lee

Useful Texts on PCA and FA

A General View of Latent Variable Models

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

Non-linear Factor Analysis

\[
p(x) = \int p(x|z)p(z)dz
\]

For PPCA
\[
p(x|z) \sim N(Wz + \mu, \sigma^2 I)
\]

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

\[
p(x) \simeq \frac{1}{K} \sum_{k=1}^{K} p(x|z_k)
\]

where the samples $z_k$ are drawn from the density $p(z)$. Note that the approximation to $p(x)$ is a mixture of Gaussians.

Fitting the Model to Data

- Adjust the parameters of $\phi$ and $\sigma^2$ to maximize the log likelihood of the data.
- For a simple form of mapping $\phi(z) = \sum_i w_i \psi_i(z)$ we can obtain EM updates for the weights $\{w_i\}$ and the variance $\sigma^2$.
- 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)