Probabilistic Modelling and Reasoning
Factor Analysis and Beyond

School of Informatics, University of Edinburgh

Instructor: Dr Chris Williams

November 2007

We consider in turn Principal Components Analysis, Factor Analysis, Independent Components Analysis and Non-linear Factor Analysis.

1 Principal Components Analysis

Principal Components Analysis (PCA) is a well-established linear technique for dimensionality reduction. We consider reducing dimensionality from the data space of dimensionality d with data vector $\mathbf{x}$ to a space of dimensionality m. Let the sample data have mean $\boldsymbol{\mu}$ and covariance matrix S, and eigenvalues/vectors such that $S\mathbf{w}_j = \lambda_j \mathbf{w}_j$. We order the eigenvalues so that $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_d > 0$.

The first derivation of PCA concerns the idea of a "bottleneck" network architecture. We input a vector $\mathbf{x}$ which is transformed into a code vector $\mathbf{z}$ of lower dimensionality, and we then expand back up to the original dimensionality. With a linear architecture, this is achieved by a $m \times d$ matrix A and a $d \times m$ matrix B so that the reconstructed output $\hat{\mathbf{x}}$ is given by $\hat{\mathbf{x}} = \boldsymbol{\mu} + BA(\mathbf{x} - \boldsymbol{\mu})$. The optimal choice for A is to project into the subspace spanned by the first m principal components of S. The optimal orthogonal projection is given by the matrix $A = W^T$, where $W = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m)^T$, and $B = W$.

PCA can also be derived by choosing projections of the data which maximize the variance in the projected space. For example, we first look for a vector $\mathbf{a}$ (with $\mathbf{a}.\mathbf{a} = 1$) such that the projection of the original data $\mathbf{a}.(\mathbf{x} - \boldsymbol{\mu})$ has maximal variance. This turns out to be given by $\mathbf{a} = \mathbf{w}_1$. We then look for a second direction, orthogonal to the first, along which the variance is maximized, and then a third, and so on. The eigenvectors taken in order are the directions we are looking for.

Although PCA is very useful for dimensionality reduction, one problem is that it does not define a probability density model of the data. For that we turn to Factor Analysis.

2 Factor Analysis

The factor analysis (FA) model is a latent-variable model with

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

where $\mathbf{z} \sim N(0, I_m)$ and W is a general $d \times m$ matrix called the factor loadings matrix. The noise term $\mathbf{e}$ is independent of $\mathbf{z}$ and has the form $\mathbf{e} \sim N(0, \Psi)$ with Ψ diagonal. $\boldsymbol{\mu}$ is a constant vector whose maximum likelihood estimator is the mean of the data. Given this formulation, the model for $\mathbf{x}$ is is Gaussian $N(\boldsymbol{\mu}, C)$ with $C = WW^T + \Psi$, as $E[\mathbf{x}] = \boldsymbol{\mu}$ and $E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = E[(W\mathbf{z} + \mathbf{e})(W\mathbf{z} + \mathbf{e})^T] = WW^T + \Psi$. Note that the observed variables $\mathbf{x}$ are conditionally independent given $\mathbf{z}$ (because Ψ is diagonal). The aim of the factor analysis model is to explain the covariance between variables rather than the variance (cf PCA). In factor analysis the columns of W will not in general correspond to the principal subspace of the data.

The solution for the factor loadings matrix is not unique. If W is a solution, then so is WR where R is a $m \times m$ rotation matrix ($RR^T = I$), as then $(WR)(WR)^T = WRR^TW^T = WW^T$. This causes problems if one wishes to interpret the factors (as in social sciences research). A unique solution can be imposed by various conditions, for example by making $W^T\Psi^{-1}W$ to be diagonal.

Factor analysis attempts to find a model for the covariance matrix S. As such, it will make sense if the number of parameters in the model (dm for the factor loadings and d for Ψ) are fewer than the $d(d+1)/2$ in the original matrix S. (Note that the condition $W^T\Psi^{-1}W$ further reduces the number of free parameters in the FA model by $m(m-1)/2$.)

Factor analysis can be used for data visualization, as can PCA. The posterior distribution $p(\mathbf{z}|\mathbf{x}) \propto p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ is Gaussian.

There is no analytical solution for the maximum likelihood estimators of W and Ψ, and these parameters are determined by an iterative procedure. One possibility is the EM algorithm for factor analysis, due to [8]. If one considers the Probabilistic PCA model of [2] where $\Psi = \sigma^2 I$, then W_{ML} spans the subspace defined by the first m eigenvectors of S.

An interesting example of the application of FA is for handwritten digit recognition [4]. Here the authors effected digit recognition by building class-conditional density models of 8×8 images. They used 10-dimensional FA models which they fitted to the data. In fact they used a mixture of FA models for each class as this improved performance.

2.1 FA Example 1: Exam Scores

This example is taken from Mardia, Kent and Bibby *Multivariate Analysis*, (Academic Press, 1979). 88 students take five different examinations in mechanics, vectors, algebra, analysis and statistics. Each exam is marked out of 100%. The correlation matrix derived from this data is

$$\begin{pmatrix} 1 & 0.553 & 0.547 & 0.410 & 0.389 \\ & 1 & 0.610 & 0.485 & 0.437 \\ & & 1 & 0.711 & 0.665 \\ & & & 1 & 0.607 \\ & & & & 1 \end{pmatrix}$$

The correlation matrix is obtained from the covariance matrix C by dividing every entry c_{ij} by $\sqrt{c_{ii}c_{jj}}$. This is the same effect as would be obtained by rescaling each of the 5 columns in the data matrix to have unit variance before computing the covariance matrix. The entries below the diagonal need not be filled in as this is a symmetric matrix.

Maximum likelihood factor analysis was performed on this matrix (and uniqueness was obtained by imposing that $W^T\Psi^{-1}W$ be diagonal). We consider only models with $m = 1$ or $m = 2$, otherwise there would be more free parameters in the factor analysis model than entries in the correlation matrix. The results obtained are:

	m = 1	m = 2	(not rotated)	m = 2	(rotated)
Variable	$\mathbf{w}_1$	$\mathbf{w}_1$	$\mathbf{w}_2$	$\tilde{\mathbf{w}}_1$	$\tilde{\mathbf{w}}_2$
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

For the $m = 1$ solution the factor loadings are all positive and roughly equal, suggesting something like the idea that smarter people do better on all exams. For the $m = 2$ solution the first factor is similar, and the second factor represents a contrast across the range of examinations. The final two columns are a rotation of the factors imposing a different constraint to the one given above.

2.2 FA Example 2: Geometrical transformations with noise

Suppose we have k points in the plane with coordinates $(x_1, y_1), \ldots, (x_k, y_k)$. These are taken to be important points on a shape outline, so that by joining-the-dots we obtain a polygonal approximation of the desired shape, say a hand outline. However, we do not always see the shape in the same position; say that it can be scaled and rotated relative to the canonical shape. If we scale by a factor s and rotate by an angle θ and then add some Gaussian nose, then the relationship of the transformed point $(\tilde{x}_i, \tilde{y}_i)$ to (x_i, y_i) is given by

$$\begin{pmatrix} \tilde{x}_i \\ \tilde{y}_i \end{pmatrix} = \begin{pmatrix} s\cos\theta & s\sin\theta \\ -s\sin\theta & s\cos\theta \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix} + \begin{pmatrix} e_i^x \\ e_i^y \end{pmatrix}$$

We then let $z_1 = s\cos\theta$ and $z_2 = s\sin\theta$ be chosen from independent $N(0,1)$ Gaussian distributions. (In fact this implies that θ is chosen uniformly in $[0, 2\pi)$ and that $s^2 \sim \chi_2^2$, but this detail is not really important.) This relationship can be then rewritten as

$$\begin{pmatrix} \tilde{x}_i \\ \tilde{y}_i \end{pmatrix} = \begin{pmatrix} x_1 & y_1 \\ y_1 & -x_1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} + \begin{pmatrix} e_i^x \\ e_i^y \end{pmatrix}$$

and if we consider the vector $\mathbf{x} = (\tilde{x}_1, \tilde{y}_1, \tilde{x}_2, \tilde{y}_2, \ldots, \tilde{x}_k, \tilde{y}_k)^T$ we see that it it is generated by a factor analysis model

$$\mathbf{x} = W\mathbf{z} + \mathbf{e}.$$

Exercise: Write down the $2k \times 2$ matrix W in terms of the x_is and y_is.

If we are given data generated from this model for many different $\mathbf{z}$'s then we can fit a two-factor FA model to the data and recover the x_is and y_is. [In fact from the construction above we know of certain equalities in the factor loadings matrix which we could take into account during learning if we so wished.] Note that rotation of the factors here corresponds to a rotation of the canonical figure in the x, y plane.

Exercise: extend this model to add in translations as well as scaling and rotation.

3 Independent Components Analysis

Independent Components Analysis (ICA) is a recent development in latent variable modelling. We consider the same linear model as for FA, i.e.

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

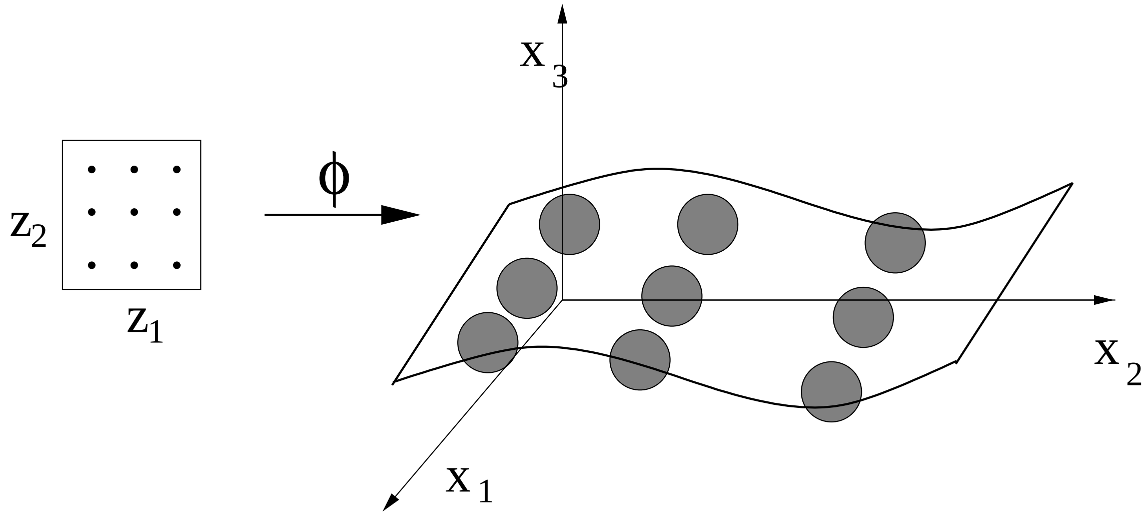

Figure 1: Graphical view of the Generative Topographic Mapping set-up. The function $\boldsymbol{\phi}$ maps from the 2-d latent space to the data space.

with independent z's so that $p(\mathbf{z}) = \prod_i p(z_i)$, but the key difference is that the latent variables now have a non-Gaussian distribution. For example we may use $p(z_i) \propto e^{-|z_i|}$. The model for $p(\mathbf{x})$ is now non-Gaussian and it is not simply the covariance structure of the data that is needed to fit the model. There has been a considerable amount of work on ICA in recent years—see, for example, the review article by [5] and the web site `http://www.cnl.salk.edu/~tony/ica.html` .

4 Non-linear Factor Analysis

In both Factor Analysis and ICA we have considered a linear transformation from the latent to the data space. However, it is clear that this can be generalized to a non-linear transformation $\boldsymbol{\phi}(\mathbf{z})$, where $\boldsymbol{\phi}(\mathbf{z})$ is a vector valued function made up of components $\phi_1(\mathbf{z}), \phi_2(\mathbf{z}), \ldots, \phi_d(\mathbf{z})$. We then have $p(\mathbf{x}|\mathbf{z}) \sim N(\boldsymbol{\phi}(\mathbf{z}), \sigma^2 I)$. As before

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

but this integral cannot now be done analytically. One approach is to approximate it using samples from $p(\mathbf{z})$, so that

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

This follows the general recipe for approximating $\int f(\mathbf{z}) p(\mathbf{z}) d\mathbf{z}$ with $\frac{1}{K} \sum_{k=1}^{K} f(\mathbf{z}_k)$, where the $\mathbf{z}_k$'s are drawn from $p(\mathbf{z})$.

A specific example of non-linear factor analysis (NLFA) is the Generative Topographic Mapping (GTM) model of [1]. Here $p(\mathbf{z})$ is taken to be uniform in $[-1, 1]^m$ with samples taken on a grid in this latent space, and $\phi_i(\mathbf{z})$ being modelled as $\phi_i(\mathbf{z}) = \sum_j w_{ij} \psi_j(\mathbf{z})$ for a set of basis functions $\psi_j(\mathbf{z})$. The set-up for GTM is illustrated in Figure 1. As the number of sample points used scales exponentially with $\dim(\mathbf{z})$ only 1-d or 2-d latent spaces are usually considered with the GTM model.

Maximum likelihood estimation of the model parameters $\{w_{ij}\}$ and σ^2 can be achieved using the EM algorithm. We are actually fitting a *constrained* mixture of Gaussians to the data. This algorithm is quite like the Self-Organizing Map of Kohonen [6] but it is more principled as there is an objective function. Data visualization can be done by looking at the posterior distribution $p(\mathbf{z}|\mathbf{x})$ in latent space, although note that is is now possible for the posterior distribution to be multimodal, so that the posterior mean may not be a good summary of this distribution.

There are other similar NLFA models in the literature, for example density nets [7] and principal curves/surfaces [3, 9].

References

[1] C. M. Bishop, M. Svensen, and C. K. I. Williams. GTM: The Generative Topographic Mapping. *Neural Computation*, 10(1):215–234, 1998.

[2] Tipping. M. E. and Bishop. C. M. Probabilistic principal components analysis. *J. Roy. Statistical Society B*, 61(3):611–622, 1999.

[3] T. Hastie and W. Stuetzle. Principal Curves. *J. American Statistical Association*, 84:502–516, 1989.

[4] G. E. Hinton, P. Dayan, and M. Revow. Modelling the Manifolds of Images of Handwritten Digits. *IEEE Trans. on Neural Networks*, 8(1):65–74, 1997.

[5] A. Hyvärinen. Survey on Independent Component Analysis. *Neural Computing Surveys*, 2:94–128, 1999.

[6] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 1982.

[7] D. J. C. MacKay. Bayesian neural networks and density networks. *Nuclear Instruments and Methods in Physics Research, A*, 354(1):73–80, 1995.

[8] D. B. Rubin and D. T. Thayer. EM Algorithms for ML Factor Analysis. *Psychometrika*, 47(1):69–76, 1982.

[9] R. Tibshrani. Principal curves revisited. *Statistics and Computing*, 2:183–190, 1992.