Factor Analysis — A Probabilistic Model Related to PCA

PCA doesn't provide a probabilistic model of the data. If we use $m = 10$ principal components for data with $p = 1000$ variables, it's not clear what we're saying about the distribution of this data.

A latent variable model called *factor analysis* is similar, and does treat the data probabilistically.

We assume that each data item, $x = (x_1, \ldots, x_p)$ is generated using m latent variables $z_1, \ldots, z_m$. the relationship of x to z is assumed to be linear.

The z_i are independent of each other. They all have Gaussian distributions with mean 0 and variance 1. (This is just a convention — any mean and variance would do as well.)

The observed data, x, are obtained by

$$x \;=\; \mu \;+\; \Lambda z \;+\; \epsilon$$

where μ is a vector of means for the p components of x, Λ is a $p \times m$ matrix, and ϵ is a vector of p "residuals", assumed to be independent, and to come from Gaussian distributions with mean zero. The variance of ϵ_j is σ_j^2.

The Distribution Defined by a Factor Analysis Model

Since the factor analysis model expresses x as a linear combination of independent Gaussian variables, the distribution of x will be multivariate Gaussian. The mean vector will be μ. The covariance matrix will be

$$E\Big((x-\mu)(x-\mu)^T\Big) \;=\; E\Big((\Lambda z)(\Lambda z)^T \;+\; \epsilon\epsilon^T \;+\; (\Lambda z)\epsilon^T \;+\; \epsilon(\Lambda x)^T\Big)$$

Because ϵ and z are independent, and have means of zero, the last two terms have expectation zero, so the covariance is

$$E\Big((\Lambda z)(\Lambda z)^T \;+\; \epsilon\epsilon^T\Big) \;=\; \Lambda E(zz^T)\Lambda^T \;+\; E(\epsilon\epsilon^T) \;=\; \Lambda\Lambda^T + \Sigma$$

where Σ is the diagonal matrix containign the residual variances, σ_j^2.

This form of covariance matrix has $mp + p$ free parameters, as opposed to $p(p+1)/2$ for a unrestricted covariance matrix. So when m is small, factor analysis is a restricted Gaussian model.

Fitting Factor Analysis Models

We can estimate the parameters of a factor analysis model (Λ and the σ_j) by maximum likelihood.

This is a moderately difficult optimization problem. There are local maxima, so trying multiple initial values may be a good idea.

When there is more than one latent factor ($m > 1$), the result is non-unique, since the latent space can be rotated (with a corresponding change to Λ) without affecting the probability distribution of the observed data.

Sometimes, one or more of the σ_j are estimated to be zero. This is maybe not too realistic.

Factor Analysis in R

The `factanal` procedure in R does maximum likelihood factor analysis. An example with simulated data, using $m = 1$:

```
> n = 1000              # number of training cases
> z = rnorm(n)          # simulate values for the latent factor
> x = cbind (           # simulate observed data
+   4+3*z+rnorm(n,0,0.1),
+   1-2*z+rnorm(n,0,0.3),
+   4*z+rnorm(n,0,1))
>
> f = factanal(x,1)  # find maximum likelihood estimate
>
> f$loadings *          # look at lambda, correcting for factanal
+  apply(x,2,sd)        #   having standardized variables

Loadings:
     Factor1
[1,]  3.036
[2,] -2.031
[3,]  4.080

               Factor1
SS loadings     29.994
Proportion Var   9.998
>
> sqrt(f$uniquenesses * # look at noise standard deviations
+  apply(x,2,var))
[1] 0.2152241 0.2874030 0.9887391
```

Factor Analysis and PCA

If we constrain all the σ_j to be equal, the results of maximum likelihood factor analysis are essential the same as PCA. The mapping $x = \Lambda z$ defines an embedding of an m-dimensional manifold in p-dimensional space, which corresponds to the hyperplane spanned by the first m principal components.

But if the σ_j can be different, factor analysis can produce much different results from PCA:

- Unlike PCA, maximum likelihood factor analysis is not sensitive to the units used, or other scaling of the variables.
- Lots of noise in a variable (unrelated to anything else) will not affect the result of factor analysis except to increase σ_j for that variable. In contrast, a noisy variable may dominate the first principle component (at least if the variable is not rescaled to make the noise smaller).
- In general, the first m principal components are chosen to capture as much *variance* as possible, but the m latent variables in a factor analysis model are chosen to explain as much *covariance* as possible.