Factor Analysis

Prof. Dr. H. H. Takada

Quantitative Research – Itaú Asset Management
Institute of Mathematics and Statistics – University of São Paulo

Factor Analysis

FA is essentially a probabilistic extension of Principal Components Analysis. It is very widely used in practice and one of the central tools in statistical analysis.
$\mathbf{v}$ is 'visible' data vector. The dataset is then given by a set of vectors,

$$\mathcal{V} = \left\{\mathbf{v}^1, \ldots, \mathbf{v}^N\right\}$$

where $\dim(\mathbf{v}) = D$. Our interest is to find a lower H-dimensional probabilistic description of this data.

$$\mathbf{v} = \mathbf{F}\mathbf{h} + \mathbf{c} + \boldsymbol{\epsilon}$$

where the noise $\boldsymbol{\epsilon}$ is Gaussian distributed,

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \boldsymbol{\Psi})$$

Probabilistic PCA

$$\boldsymbol{\Psi} = \sigma^2\mathbf{I}$$

Factor Analysis

$$\boldsymbol{\Psi} = \mathrm{diag}(\psi_1, \ldots, \psi_D)$$

A probabilistic description

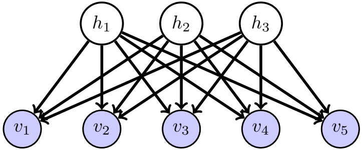
$$p(\mathbf{v}|\mathbf{h}) = \mathcal{N}(\mathbf{v}|\mathbf{F}\mathbf{h}+\mathbf{c},\boldsymbol{\Psi}) \propto e^{-\frac{1}{2}(\mathbf{v}-\mathbf{F}\mathbf{h}-\mathbf{c})^{\mathsf{T}}\boldsymbol{\Psi}^{-1}(\mathbf{v}-\mathbf{F}\mathbf{h}-\mathbf{c})}$$

To complete the model, we need to specify the hidden distribution $p(\mathbf{h})$. A convenient choice is a Gaussian

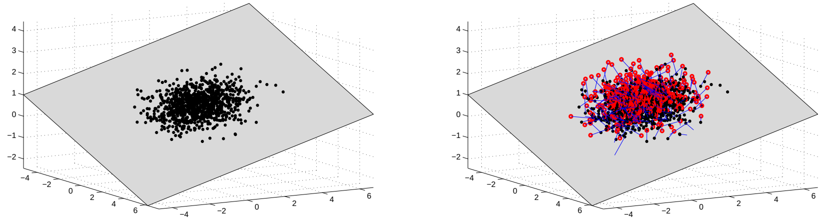
$$p(\mathbf{h}) = \mathcal{N}(\mathbf{h}|\mathbf{0},\mathbf{I}) \propto e^{-\mathbf{h}^{\mathsf{T}}\mathbf{h}/2}$$

$$p(\mathbf{v}) = \int p(\mathbf{v}|\mathbf{h})\,p(\mathbf{h})\,d\mathbf{h} = \mathcal{N}\left(\mathbf{v}\middle|\mathbf{c},\mathbf{F}\mathbf{F}^{\mathsf{T}}+\boldsymbol{\Psi}\right)$$

The coordinates $\mathbf{h}$ will be preferentially concentrated around values close to $\mathbf{0}$. If we sample a $\mathbf{h}$ from $p(\mathbf{h})$ and then draw a value for $\mathbf{v}$ using $p(\mathbf{v}|\mathbf{h})$, the sampled $\mathbf{v}$ vectors would produce a saucer or 'pancake' of points in the $\mathbf{v}$ space. Using a correlated Gaussian prior $p(\mathbf{h}) = \mathcal{N}(\mathbf{h}|\mathbf{0},\boldsymbol{\Sigma}_H)$ has no effect on the complexity of the model since $\boldsymbol{\Sigma}_H$ can be absorbed into $\mathbf{F}$.



Pancakes



(a) (b)

Figure: Factor Analysis: 1000 points generated from the model. **(a)**: 1000 latent two-dimensional points $\mathbf{h}^n$ sampled from $\mathcal{N}(\mathbf{h}|\mathbf{0},\mathbf{I})$. These are transformed to a point on the three-dimensional plane by $\mathbf{x}_0^n = \mathbf{c} + \mathbf{F}\mathbf{h}^n$. The covariance of $\mathbf{x}_0$ is degenerate, with covariance matrix $\mathbf{F}\mathbf{F}^\mathsf{T}$. **(b)**: For each point $\mathbf{x}_0^n$ on the plane a random noise vector is drawn from $\mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0},\boldsymbol{\Psi})$ and added to the in-plane vector to form a sample $\mathbf{x}^n$, plotted in red. The distribution of points forms a 'pancake' in space. Points 'underneath' the plane are not shown.

Maximum Likelihood

For a set of data $\mathcal{V}$ and using the usual i.i.d. assumption, the log likelihood is

$$\log p(\mathcal{V}) = \sum_{n=1}^{N} \log p(\mathbf{v}^n) = -\frac{1}{2} \sum_{n=1}^{N} (\mathbf{v}^n - \mathbf{c})^{\mathsf{T}} \boldsymbol{\Sigma}_D^{-1} (\mathbf{v}^n - \mathbf{c}) - \frac{N}{2} \log\det(2\pi\boldsymbol{\Sigma}_D)$$

where

$$\boldsymbol{\Sigma}_D \equiv \mathbf{F}\mathbf{F}^{\mathsf{T}} + \boldsymbol{\Psi}$$

Differentiating $\log p(\mathcal{V})$ with respect to $\mathbf{c}$ and equating to zero,

$$\mathbf{c} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{v}^n \equiv \bar{\mathbf{v}}$$

With this setting

$$\log p(\mathcal{V}) = -\frac{N}{2} \left(\mathrm{trace}\left(\boldsymbol{\Sigma}_D^{-1} \mathbf{S} \right) + \log\det(2\pi\boldsymbol{\Sigma}_D) \right)$$

where $\mathbf{S}$ is the sample covariance matrix

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

Maximum Likelihood

First, let's fix $\mathbf{\Psi}$ and try to find the optimal $\mathbf{F}$. Define

$$\tilde{\mathbf{S}} = \mathbf{\Psi}^{-\frac{1}{2}}\mathbf{S}\mathbf{\Psi}^{-\frac{1}{2}}$$

and consider the eigen-decomposition of $\tilde{\mathbf{S}}$

$$\tilde{\mathbf{S}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathsf{T}}$$

Then take the eigenvectors corresponding to the largest eigenvalues, to give the non-square matrix $\mathbf{U}_H$. Then one may show that, optimally

$$\mathbf{F} = \mathbf{\Psi}^{\frac{1}{2}}\mathbf{U}_H\left(\mathbf{\Lambda}_H - \mathbf{I}_H\right)^{\frac{1}{2}}\mathbf{R}$$

where

$$\mathbf{\Lambda}_H \equiv \mathrm{diag}\left(\lambda_1, \ldots, \lambda_H\right)$$

are the H largest eigenvalues of $\tilde{\mathbf{S}}$, and $\mathbf{R}$ is an arbitrary orthogonal matrix.

Maximum Likelihood

Now, let's fix $\mathbf{F}$ and find the optimal $\mathbf{\Psi}$. One can show that this is given by

$$\mathbf{\Psi} = \mathrm{diag}\left(\mathbf{S} - \mathbf{F}\mathbf{F}^{\mathsf{T}}\right)$$

Optimising $\mathbf{F}$ and $\mathbf{\Psi}$

A simple procedure to find the optimal $\mathbf{F}$ and $\mathbf{\Psi}$ is:

1. Update $\mathbf{F}$ for fixed $\mathbf{\Psi}$.
2. Update $\mathbf{\Psi}$ for fixed $\mathbf{F}$.

and iterate these steps until convergence.

Probabilistic PCA

In the special case that

$$\mathbf{\Psi} = \sigma^2 \mathbf{I}$$

one may equivalently write

$$\mathbf{F} = \mathbf{U}_H \left(\mathbf{\Lambda}_H - \sigma^2 \mathbf{I}_H\right)^{\frac{1}{2}} \mathbf{R}$$

where $\mathbf{R}$ is an arbitrary orthogonal matrix with $\mathbf{R}^{\mathsf{T}}\mathbf{R} = \mathbf{I}$ and $\mathbf{U}_H$, $\mathbf{\Lambda}_H$ are the eigenvectors and corresponding eigenvalues of the sample covariance $\mathbf{S}$. Classical PCA is recovered in the limit $\sigma^2 \to 0$. Note that for a full correspondence with PCA, one needs to set $\mathbf{R} = \mathbf{I}$, which points $\mathbf{F}$ along the principal directions.

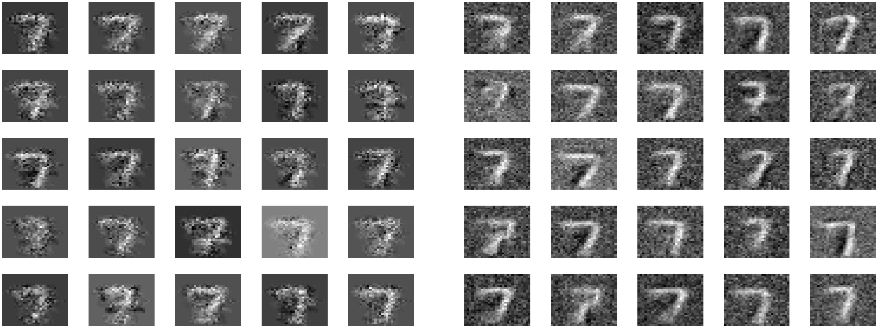
Optimal σ^2

The Maximum Likelihood optimal setting for σ^2 is

$$\sigma^2 = \frac{1}{D - H} \sum_{j=H+1}^{D} \lambda_j$$

The single-shot training nature of PPCA makes it an attractive algorithm and also gives a useful initialisation for Factor Analysis.

pPCA versus FA



(a) Factor Analysis (b) PPCA

Figure: **(a)**: 25 samples from the learned FA model of a dataset of handwritten '7s'. Note how the noise variance depends on the pixel, being zero for pixels on the boundary of the image. **(b)**: 25 samples from the learned PPCA model.

Canonical Correlation Analysis

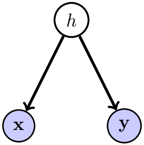
CCA is a classical method to associate data from two different spaces x and y (x might be a speech signal and y the corresponding video). It is straightforward to show that CCA is a limiting zero-noise case of a constrained FA model:

$$p(\mathbf{x}, \mathbf{y}) = \int p(\mathbf{x}|h)p(\mathbf{y}|h)p(h)dh$$

where

$$p(\mathbf{x}|h) = \mathcal{N}\left(\mathbf{x}|h\mathbf{a}, \mathbf{\Psi}_x\right), \quad p(\mathbf{y}|h) = \mathcal{N}\left(\mathbf{y}|h\mathbf{b}, \mathbf{\Psi}_y\right), \quad p(h) = \mathcal{N}\left(h|0, 1\right)$$

$\mathbf{a}$, $\mathbf{b}$ can be set by Maximum Likelihood.



Canonical Correlation Analysis. CCA corresponds to the latent variable model in which a common latent variable generates both the observed x and y variables. This is therefore a formed of constrained Factor Analysis.

The extension to vector $\mathbf{h}$ is clear.