Lecture 11:

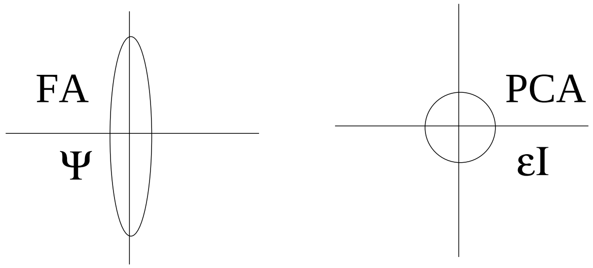
Principal Component Analysis

Sam Roweis

February 11, 2002

Gaussians are Footballs in High-D

- Recall the intuition that Gaussians are hyperellipsoids.
- Mean == centre of football
  Eigenvectors of covariance matrix == axes of football
  Eigenvalues == lengths of axes
- In FA our football is an axis aligned cigar.
  In PCA our football is a sphere of radius σ^2.



Principal Component Analysis

- In Factor Analysis, we can write the marginal density explicitly:

$$p(\mathbf{y}|\theta) = \int_{\mathbf{x}} p(\mathbf{x})p(\mathbf{y}|\mathbf{x},\theta)d\mathbf{x} = \mathcal{N}(\mathbf{x}|\mu\ ,\ \Lambda\Lambda^\top + \Psi)$$

- Noise Ψ mut be restricted for model to be interesting. (Why?)
- In Factor Analysis the restriction is that Ψ is *diagonal* (axis-aligned).
- What if we further restrict $\Psi = \sigma^2 I$ (ie *spherical*)?
- We get the Principal Component Analysis (PCA) model:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|0, I)$$
$$p(\mathbf{y}|\mathbf{x},\theta) = \mathcal{N}(\mathbf{y}|\mu + \Lambda\mathbf{x}, \sigma^2 I)$$

where μ is the mean vector,
columns of Λ are the *principal components* (usually orthogonal),
and σ^2 is the *global sensor noise*.

Likelihood Functions

- For both FA and PCA, the data model is Gaussian.
  Thus, the likelihood function is simple:

$$\begin{aligned}\ell(\theta;\mathcal{D}) &= -\frac{M}{2}\log|\Lambda\Lambda^\top + \Psi| - \frac{1}{2}\sum_n (\mathbf{y}^n - \mu)^\top(\Lambda\Lambda^\top + \Psi)^{-1}(\mathbf{y}^n - \mu) \\ &= -\frac{M}{2}\log|\mathbf{V}| - \frac{1}{2}\mathrm{trace}\left[\mathbf{V}^{-1}\sum_n(\mathbf{y}^n - \mu)(\mathbf{y}^n - \mu)^\top\right] \\ &= -\frac{M}{2}\log|\mathbf{V}| - \frac{1}{2}\mathrm{trace}\left[\mathbf{V}^{-1}\mathbf{S}\right]\end{aligned}$$

  $\mathbf{V}$ is model covariance; $\mathbf{S}$ is sample data covariance.
- In other words, we are trying to make the constrained model covariance as close as possible to the observed covariance, where "close" means the trace of the ratio.
- Thus, the sufficient statistics are the same as for the Gaussian: mean $\sum_n \mathbf{y}^n$ and covariance $\sum_n(\mathbf{y}^n - \mu)(\mathbf{y}^n - \mu)^\top$.

Fitting the PCA model

- The standard EM algorithm applies to PCA also:
  E-step: $q^{t+1} = p(\mathbf{x}^n|\mathbf{y}^n, \theta^t)$
  M-step: $\theta^{t+1} = \mathrm{argmax}_\theta \sum_n \int_\mathbf{x} q^{t+1}(\mathbf{x}^n|\mathbf{y}^n) \log p(\mathbf{y}^n, \mathbf{x}^n|\theta) d\mathbf{x}^n$
- For this we need the conditional distribution (inference) and the expected log of the complete data. Results:

$$\begin{aligned}
\mathbf{E-step}: q^{t+1} &= p(\mathbf{x}|\mathbf{y}, \theta^t) = \mathcal{N}(\mathbf{x}^n|\mathbf{m}^n, \mathbf{V}^n) \\
\mathbf{V}^n &= (I + \sigma^{-2}\Lambda^\top\Lambda)^{-1} \\
\mathbf{m}^n &= \sigma^{-2}\mathbf{V}^n\Lambda^\top(\mathbf{y} - \mu)
\end{aligned}$$

$$\begin{aligned}
\mathbf{M-step}: \Lambda^{t+1} &= \left(\sum_n \mathbf{y}^n \mathbf{m}^{n\top}\right)\left(\sum_n \mathbf{V}^n\right)^{-1} \\
\sigma^{2^{t+1}} &= \frac{1}{MN}\sum_i \left[\sum_n \mathbf{y}^n\mathbf{y}^{n\top} + \Lambda^{t+1}\sum_n \mathbf{m}^n\mathbf{y}^{n\top}\right]_{ii}
\end{aligned}$$

Inference is Linear

- Recall the inference formulas for FA:

$$\begin{aligned}
p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{V}) \\
\mathbf{V} &= I - \Lambda^\top(\Lambda\Lambda^\top + \Psi)^{-1}\Lambda \\
&= (I + \Lambda^\top\Psi^{-1}\Lambda)^{-1} \\
\mathbf{m} &= \Lambda^\top(\Lambda\Lambda^\top + \Psi)^{-1}(\mathbf{y} - \mu) \\
&= \mathbf{V}\Lambda^\top\Psi^{-1}(\mathbf{y} - \mu)
\end{aligned}$$

- Note: inference of the posterior mean is just a linear operation!

$$\mathbf{m} = \beta(\mathbf{y} - \mathbf{m})$$

  where β can be computed beforehand given the model parameters.
- Also: posterior covariance does not depend on observed data!

$$\mathrm{cov}[\mathbf{x}|\mathbf{y}] = \mathbf{V} = (I + \Lambda^\top\Psi^{-1}\Lambda)^{-1}$$

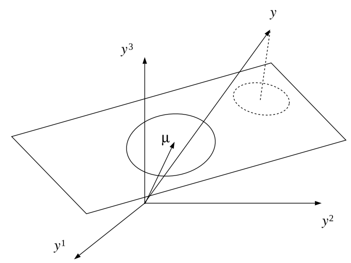
Direct Fitting

- For FA the parameters are coupled in a way that makes it impossible to solve for the ML params directly.
  We must use EM or other nonlinear optimization techniques.
- But for PCA, the ML params can be solved for directly:
  The k^{th} column of Λ is the k^{th} largest eigenvalue of the sample covariance $\mathbf{S}$ times the associated eigenvector.
- The global sensor noise σ^2 is the sum of all the eigenvalues smaller than the k^{th} one.
- This technique is good for initializing FA also.

Zero Noise Limit

- The traditional PCA model is actually a limit as $\sigma^2 \to 0$.
  The model we saw is actually called "probabilistic PCA".
- However, the ML parameters Λ^* are the same.
  The only difference is the global sensor noise σ^2.
- In the zero noise limit inference is easier: orthogonal projection.

$$\lim_{\sigma^2 \to 0} \Lambda^\top(\Lambda\Lambda^\top + \sigma^2 I)^{-1} = (\Lambda^\top\Lambda)^{-1}\Lambda^\top$$

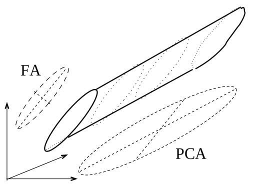


Scale Invariance in Factor Analysis

- In FA the *scale* of the data is unimportant: we can multiply $\mathbf{y}_i$ by α_i without changing anything:

$$\begin{aligned} \mu_i &\leftarrow \alpha_i \mu_i \\ \Lambda_{ij} &\leftarrow \alpha_i \Lambda_{ij} \quad \forall j \\ \Psi_i &\leftarrow \alpha_i^2 \Psi_i \end{aligned}$$

- However, the *rotation* of the data *is* important.
- FA looks for directions of large correlation in the data, so it is not fooled by large variance noise.



Model Invariance and Identfiability

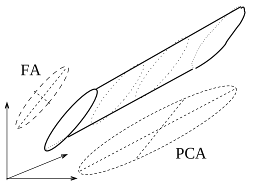
- There is *degeneracy* in the FA model.
- Since Λ only appears as outer product $\Lambda\Lambda^\top$, the model is invariant to rotation and axis flips of the latent space.
- We can replace Λ with $\Lambda\mathbf{Q}$ for any unitary matrix $\mathbf{Q}$ and the model remains the same: $(\Lambda\mathbf{Q})(\Lambda\mathbf{Q})^\top = \Lambda(\mathbf{Q}\mathbf{Q}^\top)\Lambda^\top = \Lambda\Lambda^\top$.
- This means that there is no "one best" setting of the parameters. An infinite number of parameters all give the ML score!
- Such models are called un-identifiable since two people both fitting ML params to the identical data will not be guaranteed to identify the same parameters.

Rotational Invariance in PCA

- In PCA the *rotation* of the data is unimportant: we can multiply the data $\mathbf{y}$ by and rotation $\mathbf{Q}$ without changing anything:

$$\begin{aligned} \mu &\leftarrow \mathbf{Q}\mu \\ \Lambda &\leftarrow \mathbf{Q}\Lambda \\ \Psi &\leftarrow \text{unchanged} \end{aligned}$$

- However, the *scale* of the data *is* important.
- PCA looks for directions of large variance, so it will chase big noise directions.



Latent Covariance in Factor Analysis and PCA

- What if we allow the latent variable $\mathbf{x}$ to have a covariance matrix of its own: $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|0, \mathbf{P})$?
- We can still compute the marginal probability:

$$p(\mathbf{y}|\theta) = \int_{\mathbf{x}} p(\mathbf{x})p(\mathbf{y}|\mathbf{x},\theta)d\mathbf{x} = \mathcal{N}(\mathbf{x}|\mu\,,\ \Lambda\mathbf{P}\Lambda^\top + \Psi)$$

- We can always absorb $\mathbf{P}$ into the loading matrix Λ by diagonalizing it: $P = EDE^\top$ and setting $\Lambda = \Lambda ED^{1/2}$.
- Thus, there is another degeneracy in FA, between $\mathbf{P}$ and Λ: we can set $\mathbf{P}$ to be the identity, to be diagonal, whatever we want.
- Traditionally we break this degeneracy by either:
  - set the covariance $\mathbf{P}$ of the latent variable to be I (FA) or
  - fore the columns of Λ to be orthonormal (PCA)