Lecture 10:

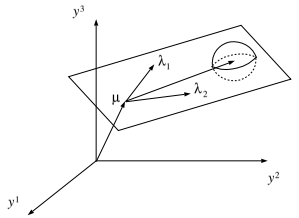
Factor Analysis

Sam Roweis

February 6, 2002

Factor Analysis

- When we assume that the subspace is *linear* and that the underlying latent variable has a Gaussian distribution we get a model known as *factor analysis*:
  — data $\mathbf{y}$ (p-dim);
  — latent variable $\mathbf{x}$ (k-dim)



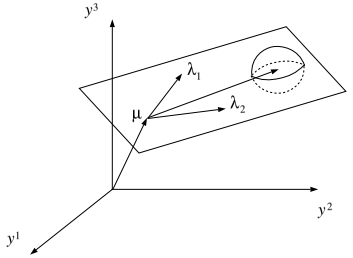
$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|0, I)$$
$$p(\mathbf{y}|\mathbf{x}, \theta) = \mathcal{N}(\mathbf{y}|\mu + \Lambda\mathbf{x}, \Psi)$$

where μ is the mean vector, Λ is the p by k *factor loading matrix*, and Ψ is the *sensor noise covariance* (ususally diagonal).

- Important: since the product of Gaussians is still Gaussian, the joint distribution $p(\mathbf{x}, \mathbf{y})$, the other marginal $p(\mathbf{y})$ and the conditional $p(\mathbf{x}|\mathbf{y})$ are also Gaussian.

Continuous Latent Variables

- In many models there are some *underlying causes* of the data.
- Mixture models use a discrete class variable: clustering.
- Sometimes, it is more appropriate to think in terms of continuous *factors* which control the data we observe. Geometrically, this is equivalent to thinking of a data *manifold* or subspace.



- To generate data, first generate a point within the manifold then add noise. Coordinates of point are components of latent variable.

Marginal Data Distribution

- Just as with discrete latent variables, we can compute the marginal density $p(\mathbf{y}|\theta)$ by summing out $\mathbf{x}$. But now the sum is an integral:

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

  which can be done by completing the square in the exponent.

- However, since the marginal is Gaussian, we can also just compute its mean and covariance. (Assume noise uncorrelated with data.)

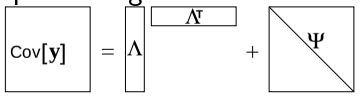
$$\begin{aligned} E[\mathbf{y}] &= E[\mu + \Lambda\mathbf{x} + \text{noise}] = \mu + \Lambda E[\mathbf{x}] + E[\text{noise}] \\ &= \mu + \Lambda \cdot 0 + 0 = \mu \\ \text{Cov}[\mathbf{y}] &= E[(\mathbf{y} - \mu)(\mathbf{y} - \mu)^\top] \\ &= E[(\mu + \Lambda\mathbf{x} + \text{noise} - \mu)(\mu + \Lambda\mathbf{x} + \text{noise} - \mu)^\top] \\ &= E[(\Lambda\mathbf{x} + n)(\Lambda\mathbf{x} + n)^\top] = \Lambda E(\mathbf{x}\mathbf{x}^\top)\Lambda^\top + E(nn^\top) \\ &= \Lambda\Lambda^\top + \Psi \end{aligned}$$

Constrained Covariance

- Marginal density for factor analysis ($\mathbf{y}$ is p-dim, $\mathbf{x}$ is k-dim):

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

- So the effective covariance is the low-rank outer product of two long skinny matrices plus a diagonal matrix:



- In other words, factor analysis is just a constrained Gaussian model. (If Ψ were not diagonal then we could model any Gaussian and it would be pointless.)
- It is easy to find μ: just take the mean of the data.
  From now on assume we have done this and re-centred $\mathbf{y}$.

Complete Data Likelihood

- Write down the joint distribtion of $\mathbf{x}$ and $\mathbf{y}$:

$$p(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}) = \mathcal{N}(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} | \begin{bmatrix}0\\ \mu\end{bmatrix}, \begin{bmatrix}I & \Lambda^\top\\ \Lambda & \Lambda\Lambda^\top + \Psi\end{bmatrix})$$

  where the corner elements $\Lambda^\top, \Lambda$ come from $\mathrm{Cov}[\mathbf{x}, \mathbf{y}]$:

$$\begin{aligned}\mathrm{Cov}[\mathbf{x}, \mathbf{y}] &= E[(\mathbf{x} - 0)(\mathbf{y} - \mu)^\top] = E[\mathbf{x}(\mu + \Lambda\mathbf{x} + n - \mu)^\top]\\ &= E[\mathbf{x}(\Lambda\mathbf{x} + n)^\top] = \Lambda^\top\end{aligned}$$

- This gives the complete likelihood (ignoring mean):

$$\begin{aligned}\ell_c(\Lambda, \Psi) &= -\frac{M}{2}\log|\Psi| - \frac{1}{2}\sum_n \mathbf{x}^\top\mathbf{x} - \frac{1}{2}\sum_n (\mathbf{y}^n - \Lambda\mathbf{x}^n)^\top\Psi^{-1}(\mathbf{y}^n - \Lambda\mathbf{x}^n\\ &= -\frac{M}{2}\log|\Psi| - \frac{M}{2}\mathrm{trace}[S\Psi^{-1}]\\ S &= \frac{1}{M}\sum_n (\mathbf{y}^n - \Lambda\mathbf{x}^n)(\mathbf{y}^n - \Lambda\mathbf{x}^n)^\top\end{aligned}$$

EM for Factor Analysis

- We will do maximum likelihood learning using (surprise, surprise) the EM algorithm.
  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:

$$\begin{aligned}\mathbf{E-step}: q^{t+1} &= p(\mathbf{x}|\mathbf{y}, \theta^t) = \mathcal{N}(\mathbf{x}^n|\mathbf{m}^n, \mathbf{V}^n)\\ \mathbf{V}^n &= (I + \Lambda^\top\Psi^{-1}\Lambda)^{-1}\\ \mathbf{m}^n &= \mathbf{V}^n\Lambda^\top\Psi^{-1}(\mathbf{y} - \mu)\\ \mathbf{M-step}: \Lambda^{t+1} &= \left(\sum_n \mathbf{y}^n\mathbf{m}^{n\top}\right)\left(\sum_n \mathbf{V}^n\right)^{-1}\\ \Psi^{t+1} &= \frac{1}{M}\mathrm{diag}\left[\sum_n \mathbf{y}^n\mathbf{y}^{n\top} + \Lambda^{t+1}\sum_n \mathbf{m}^n\mathbf{y}^{n\top}\right]\end{aligned}$$

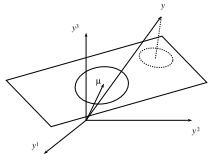
Inference in Factor Analysis

- Apply the Gaussian conditioning formulas to the joint distribution we derived above. This gives:

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

- Now apply the matrix inversion lemma to get:

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



Gaussian Conditioning

- Remember the formulas for condtional Gaussian distributions:

$$p(\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix}) = \mathcal{N}(\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix} \mid \begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix})$$
$$p(\mathbf{x}_1|\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_1|\mathbf{m}_{1|2}, \mathbf{V}_{1|2})$$
$$\mathbf{m}_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2)$$
$$\mathbf{V}_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Derivatives

- You need these tricks to compute the M-step derivatives:

$$\frac{\partial}{\partial A}\log|A| = A^{-\top}$$
$$\frac{\partial}{\partial A}\text{trace}[B^\top A] = B$$
$$\frac{\partial}{\partial A}\text{trace}[BA^\top CA] = 2CAB$$

Matrix Inversion Lemma

- There is a good trick for inverting matrices when they can be decomposed into the sum of an easily inverted matrix (D) and a low rank outer product. It is called the *matrix inversion lemma*.

$$(D - AB^{-1}A^\top)^{-1} = D^{-1} + D^{-1}A(B - A^\top D^{-1}A)^{-1}A^\top D^{-1}$$

Means, Variances and Covariances

- Remember the definition of the mean and covariance of a vector random variable:

$$E[x] = \int_{\mathbf{x}} \mathbf{x}p(\mathbf{x})d\mathbf{x} = \mathbf{m}$$
$$\text{Cov}[x] = E[(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^\top] = \int_x (\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^\top p(\mathbf{x})d\mathbf{x} = \mathbf{V}$$

  which is the expected value of the outer product of the variable with itself, after subtracting the mean.

- Also, the covariance between two variables:

$$\text{Cov}[\mathbf{x},\mathbf{y}] = E[(\mathbf{x}-\mathbf{m}_\mathbf{x})(\mathbf{y}-\mathbf{m}_\mathbf{y})^\top]$$
$$= \int_{\mathbf{xy}} (\mathbf{x}-\mathbf{m}_\mathbf{x})(\mathbf{y}-\mathbf{m}_\mathbf{y})^\top p(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y} = \mathbf{C}$$

  which is the expected value of the outer product of one variable with another, after subtracting their means.
  Note: $\mathbf{C}$ is not symmetric.