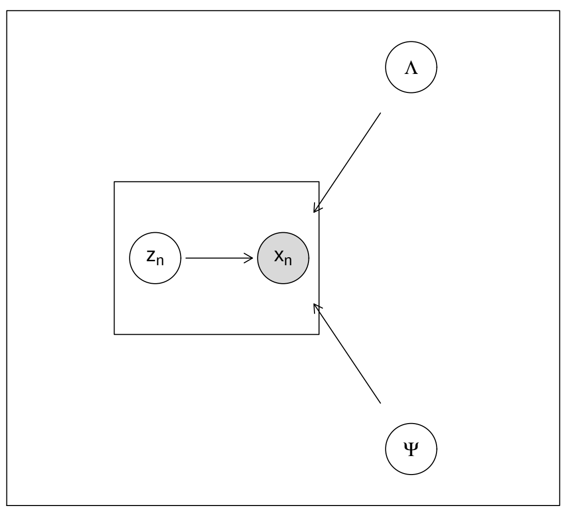
Expectation maximization, FA/PCA continued

Alex Acs

November 17, 2010

Figure 1 is the graphical model that motivates the following discussion of factor analysis (FA).

Figure 1:



In FA, the basic idea is to choose z from some distribution in q dimensions and project it onto a p-dimensional space and then choose x given the projection. To begin with, we define the variable distributions.

$$\begin{aligned}
&\langle z, x\rangle \text{ is a joint Gaussian}\\
&x \sim \mathcal{N}(0, \Lambda\Lambda^T + \Psi)\\
&z|x \sim \mathcal{N}(\Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}X, (I + \Lambda^T\Psi\Lambda^T)^{-1})
\end{aligned}$$

We want to get the MLE of Λ and Ψ, given that we have data, $\mathcal{D} = \{x_n\}_{n=1}^N$.

Notice that $x = \Lambda z + \epsilon$. If we know z, then this is a linear regression. But z is a hidden variable. So we are going to use the expectation maximization (EM) algorithm. Generally speaking, this is a way of solving maximization problems in the face of hidden variables. The EM algorithm for factor analysis is an iterative algorithm with 2 steps:

1. the E-step:
   - compute $p(z_n|x_n, \Lambda^{(t)}, \Psi^{(t)})$
   - the posterior $p(z|x)$ is defined above
2. The M-step:
   - $\Lambda^{(t=1)} = (\sum_n \mathrm{E}[z_n z_n^T|x_n, \Lambda^{(t)}, \Psi^{(t)}])^{-1}(\sum_n \mathrm{E}[z_n|x_n, \Lambda^{(t)}, \Psi^{(t)}]^T x_n)$
   - $\Psi^{(t+1)} =$ See book

EM is a way of finding approximate MLE's in latent variable models. We will be thinking about these in the rest of the course. Latent variable models posit hidden structure in observed data: clustering, subspace, trees, sequences etc.

One way to think about EM: in the E-step, we will fill in the hidden variables. In the M-step, we fit parameters to match the filled in variables (akin to taking the MLE estimate in a fully observed model). So, fill in z and then estimate the parameters. This gets us around having to integrate out the latent variables.

EM general setting

- $x_n = \{1....N\}$ observed data
- z_n hidden structure
- θ are the parameters we are interested in fitting.
- There is no particular graphical model.

What if z were observed? We could find the parameters by taking the max of the log-likelihood.

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \log p(x, z|\theta) \\ &= \arg\max_{\theta} \log p(x|z, \theta_x) + \log p(z|\theta_z)\end{aligned}$$

This function is called the complete log-likelihood. But z is hidden, so we are really after

$$
\begin{aligned}
\hat{\theta} &= \arg\max_{\theta} \log p(x|\theta) \\
&= \arg\max_{\theta} \log \sum_{z} p(x, z|\theta)
\end{aligned}
$$

Where the hidden variable has been factored out.

Note Jenson's inequality:
We will have a lower bound $\log p(x)$ on Jenson's inequality. If $\lambda \in (0,1)$ and φ is convex then:

$$
\lambda\varphi(x) + (1-\lambda)\varphi(y) \geq \varphi(\lambda(x) + (1-\lambda)(y))
$$

Which generalizes to expectations –

$$
\mathrm{E}[\varphi(x)] \geq \varphi(\mathrm{E}[x])
$$

And if φ is concave –

$$
\mathrm{E}[\varphi(x)] \leq \varphi(\mathrm{E}[x])
$$

Now back to EM:

$$
\begin{aligned}
\log p(x|\theta) &= \log \sum_{z} p(x, z|\theta) \\
&= \log \sum_{z} p(x, z|\theta) \frac{q(z)}{q(z)} \\
&= \log \mathrm{E}_q \left[\frac{p(x, z|\theta)}{q(z)} \right] \\
&\geq \mathrm{E}_q \left[\log \frac{p(x, z|\theta)}{q(z)} \right] \\
&= \mathrm{E}_q \log p(x, z|\theta) - \mathrm{E}_q \log q(z) \\
&\equiv \mathcal{Q}(\theta; q)
\end{aligned}
$$

This is the EM objective function. The EM algorithm will optimize the objective function. The EM is a coordinate ascent on $\mathcal{Q}$:

$$
\begin{aligned}
\mathrm{E} :& q^{(t+1)} = \arg\max_{q} \mathcal{Q}(\theta^{(t)}, q) \\
\mathrm{M} :& \theta^{(t+1)} = \arg\max_{\theta} \mathcal{Q}(\theta, q^{(t+1)})
\end{aligned}
$$

Holding θ fixed, the optimal $q(z)$ is $p(z|x, \theta^{(t)})$.

$$\begin{aligned}
&= \sum_z p(z|x) \log p(x, z) - \sum_z p(z|x) \log p(z|x) \\
&= \sum_z p(z|x) \log p(z|x) + \sum_z p(z|x) \log p(z|x) \log p(x) - \sum_z p(z|x) \log p(z|x) \\
&= \sum_z p(z|x) \log p(x) \\
&= \log p(x)
\end{aligned}$$

M-step:

$$\begin{aligned}
\theta^{(t+1)} &= \arg\max_\theta \mathrm{E}_q \log p(x, z|\theta) \\
&= \arg\max_\theta \mathrm{E}_q \log p(z|\theta) + \mathrm{E}_q \log p(x|z, \theta)
\end{aligned}$$

Which is the expected complete log-likelihood.

Mixture modeling

- E-step: estimate p(cluster|datapoint)
- M-step: reweight the data by p(|x) and do MLE.