"Classical" Factor Analysis and PCA

Francis J. DiTraglia

January 21, 2014

These notes draw on material from Chapters 11–12 of Murphy's *Machine Learning: A Probabilistic Perspective*, and Andrew Ng's lecture notes for CS229 at Stanford.

1 EM Algorithm

1.1 The Idea behind the EM Algorithm

For simplicity, we'll consider an iid setup for now although the EM can be used in situations with dependence. We'll also suppose that the latent variable is continuous. If it's discrete the idea is exactly the same but the integral is replaced by a sum.

$$\ell(\theta) = \sum_{t=1}^{T} \log p(\mathbf{x}_t; \theta) = \sum_{t=1}^{T} \log \left(\int p(\mathbf{x}_t, \mathbf{z}_t; \theta) \, d\mathbf{z} \right)$$

where $\mathbf{x}_t$ is observed and $\mathbf{z}_t$ is unobserved. In many interesting models there is no explicit formula for the MLE in terms of the marginal density $p(\mathbf{x}_t; \theta)$ but there *is* an explicit formula in terms of the *joint* density $p(\mathbf{x}_t, \mathbf{z}_t; \theta)$. This is exactly the setting in which the EM algorithm is useful. Rather than directly maximizing $\ell(\theta)$, the EM algorithm proceeds *iteratively* over the following two steps:

(E-step) Construct a *lower bound* for $\ell(\theta)$

(M-step) Optimize the lower bound over θ

Roughly speaking, the EM algorithm converts a single complicated optimization problem into a sequence of simple optimization problems. The trick is to ensure that the resulting sequence of estimators converges to the MLE. Jensen's Inequality is the key so I'll briefly remind you of a few importnat facts before proceeding.

Figure 1: Picture of the EM Algorithm

1.2 Jensen's Inequality

Recall that a function is called *convex* if its Hessian matrix is positive semi-definite and *strictly convex* if its Hessian matrix is positive definite. For functions of a single variable the condition is $f''(x) \geq 0 \quad \forall x \in \mathbb{R}$ for *convex* and $f''(x) > 0 \quad \forall x \in \mathbb{R}$ for *strictly convex*. In statistics, one of the most useful results concerning convex functions is *Jensen's Inequality*

Proposition 1.1 (Jensen's Inequality)**.** *Let f be a convex function and X be a random variable. Then $E[f(X)] \geq f(E[X])$. If f is strictly convex then the inequality is strict unless $P(X = E[X]) = 1$, i.e. X is a constant. For the equivalent results for concave functions, simply reverse the inequality.*

1.3 A Lower Bound for the Likelihood

Let $f_t(\mathbf{z}_t)$ be *some arbitrary* density function over the support of $\mathbf{z}_t$, that is any function satisfying $f_t(\mathbf{z_t}) \geq 0$ and

$$\int f_t(\mathbf{z}_t)\, d\mathbf{z}_t = 1$$

We have

$$\begin{aligned} \ell(\theta) = \sum_{t=1}^{T} \log p(\mathbf{x}_t; \theta) &= \sum_{t=1}^{T} \log \left(\int p(\mathbf{x}_t, \mathbf{z}_t; \theta)\, d\mathbf{z}_t \right) \\ &= \sum_{t=1}^{T} \log \left(\int f_t(\mathbf{z}_t) \left[\frac{p(\mathbf{x}_t, \mathbf{z}_t; \theta)}{f_t(\mathbf{z}_t)} \right] d\mathbf{z}_t \right) \end{aligned}$$

Now we use Jensen's inequality and the fact that log is a concave function over its domain to find that

$$\log \left(\int f_t(\mathbf{z}_t) \left[\frac{p(\mathbf{x}_t, \mathbf{z}_t; \theta)}{f_t(\mathbf{z}_t)} \right] d\mathbf{z}_t \right) \geq \int f_t(\mathbf{z}_t) \log \left[\frac{p(\mathbf{x}_t, \mathbf{z}_t; \theta)}{f_t(\mathbf{z}_t)} \right] d\mathbf{z}_t$$

What's going on here? Since f_t is a *density* the integral inside the parentheses is *an expectation* of a particular function of the argument of integration $\mathbf{z}_t$. The parameter θ and the observed vector of realizations $\mathbf{x}_t$ are constants with

respect to the integration. Substituting the preceding inequality into the sum, we have established that

$$\ell(\theta) \geq \sum_{t=1}^{T} \left(\int f_t(\mathbf{z}_t) \log \left[\frac{p(\mathbf{x}_t, \mathbf{z}_t; \theta)}{f_t(\mathbf{z}_t)} \right] d\mathbf{z}_t \right)$$

for *any* density function f_t. This is the *lower bound* for the likelihood that we will use in the E-step. The question is, how should we choose f_t?

The key idea is to turn the *inequality* into an *equality* at a particular value of θ. Intuitively, we want to ensure that, in a given iteration of the algorithm, the actual likelihood and the lower bound *agree* at the value of θ that emerged from the *preceding* iteration. In this way, our sequence of approximating functions will "trace out a path" along the true likelihood, ultimately ensuring that the EM algorithm will converge to the MLE. Since log is in fact *strictly* concave, the only way for Jensen's inequality to hold with equality is if

$$\frac{p(\mathbf{x}_t, \mathbf{z}_t; \theta)}{f_t(\mathbf{z}_t)} = c$$

for some constant c that *does not depend* on $\mathbf{z}_t$. The question is, how should we choose f_t to achieve this? Rearranging, integrating, and using the fact that f_t is a density,

$$\begin{aligned} c f_t(\mathbf{z}_t) &= p(\mathbf{x}_t, \mathbf{z}_t; \theta) \\ c \int f_t(\mathbf{z}_t)\, d\mathbf{z}_t &= \int p(\mathbf{x}_t, \mathbf{z}_t; \theta)\, d\mathbf{z}_t \\ c &= p(\mathbf{x}_t; \theta) \end{aligned}$$

Substituting for c, solving for f_t and using the definition of a conditional density we have

$$f_t(\mathbf{z}_t) = \frac{p(\mathbf{x}_t, \mathbf{z}_t; \theta)}{p(\mathbf{x}_t; \theta)} = p(\mathbf{z}_t | \mathbf{x}_t; \theta)$$

In other words, to make the lower bound hold with equality at a particular value of θ, say θ^*, it suffices to set f_t equal to the *conditional* density of $\mathbf{z}_t$ *given* $\mathbf{x}_t$ *evaluated* at θ^*. Crucially this is a both a probability density and a function of $\mathbf{z}_t$ *only* since we plug in the observed value of $\mathbf{x}_t$.

1.4 The Algorithm

In the previous subsection we showed that if we set $f_t(\mathbf{z}_t) = p(\mathbf{z}_t | \mathbf{x}_t; \theta^*)$ then

$$\ell(\theta^*) = \sum_{t=1}^{T} \left(\int f_t(\mathbf{z}_t) \log \left[\frac{p(\mathbf{x}_t, \mathbf{z}_t; \theta^*)}{f_t(\mathbf{z}_t)} \right] d\mathbf{z}_t \right)$$

and, more generally for *any* value of θ

$$\ell(\theta) \geq \sum_{t=1}^{T} \left(\int f_t(\mathbf{z}_t) \log \left[\frac{p(\mathbf{x}_t, \mathbf{z}_t; \theta)}{f_t(\mathbf{z}_t)} \right] d\mathbf{z}_t \right)$$

by Jensen's Inequality. Now we are ready to state the EM algorithm:

Algorithm 1.1 (EM Algorithm)**.** First select a starting value $\theta^{(1)}$. Then repeat the following two steps repeatedly until convergence

(E-step) For each t set $f_t^{(j-1)}(\mathbf{z}_t) = p(\mathbf{z}_t|\mathbf{x}_t; \theta^{(j-1)})$ where $\theta^{(j-1)}$ is the solution from the M-step of the *preceding* iteration.

(M-step) $\theta^{(j)} = \arg\max_{\theta \in \Theta} \sum_{t=1}^{T} \left(\int f_t^{(j-1)}(\mathbf{z}_t) \log \left[\frac{p(\mathbf{x}_t, \mathbf{z}_t; \theta)}{f_t^{(j-1)}(\mathbf{z}_t)} \right] d\mathbf{z}_t \right)$

If $j = 2$ then $\theta^{(j-1)}$ is simply the starting value $\theta^{(1)}$.

Note that in the M-step the argument θ over which we maximize *only* enters the expression $p(\mathbf{x}_t, \mathbf{z}_t; \theta)$. The density $f^{(j-1)}(\mathbf{z}_t)$ does *not* depend on θ, it depends on the *constant* $\theta^{(j-1)}$ that solved the M-step of the *previous iteration*. The amazing thing about the EM algorithm is that it is *guaranteed* to converge to a local maximum of the likelihood function: each successive iteration *monotonically* improves the likelihood as we will see below. This fact along the the way we constructed our lower bound to hold with equality at the value of θ from the *previous* M-step gives us an excellent tool for debugging our code: simply plot

$$\ell(\theta^{(j)}) = \sum_{t=1}^{T} \left(\int f_t^{(j)}(\mathbf{z}_t) \log \left[\frac{p(\mathbf{x}_t, \mathbf{z}_t; \theta^{(j)})}{f_t^{(j)}(\mathbf{z}_t)} \right] d\mathbf{z}_t \right)$$

against j. The preceding expression is the *objective function* from the $(j+1)$th M-step evaulated at the *solution* from the jth M-step. By construction, this If the is equal to the likelihood evaulated at $\theta^{(j)}$. If the plot is *not* increasing monotonically in j, then there must be a bug in your code.

1.5 Why Does the EM Algorithm Converge?

Let $\theta^{(j)}$ and $\theta^{(j+1)}$ be two succesive solutions to the M-step of the EM algorithm. We will now show that $\ell(\theta^{(j)}) \leq \theta^{(j+1)}$. In other words, the EM algorithm *monotonically* improves the likelihood in each iteration. Since $\{\theta^{(j)}\}$

is a monotonic sequence, it converges as long as it is bounded (Rudin Theorem 3.14). Since $\ell(\theta^{(1)})$ is a lower bound, if follows that the EM algorithm is *guaranteed* to converge to a local maximum of the likelihood function provided that the likelihood function is bounded above. All that remains is to actually demonstrate that $\ell(\theta^{(j)}) \leq \theta^{(j+1)}$.

By definition,

$$\theta^{(j+1)} = \arg\max_{\theta \in \Theta} \sum_{t=1}^{T} \left(\int f_t^{(j)}(\mathbf{z}_t) \log \left[\frac{p(\mathbf{x}_t, \mathbf{z}_t; \theta)}{f_t^{(j)}(\mathbf{z}_t)} \right] d\mathbf{z}_t \right)$$

Now let $\tilde{\theta}$ be some arbitrary value of θ. Since $\theta^{(j+1)}$ is the arg max, evaluating the objective function at $\tilde{\theta}$ cannot yield a greater value than evaluating it at $\theta^{(j+1)}$. Since this holds for *any* $\tilde{\theta}$ it holds in particular for $\theta^{(j)}$. Hence,

$$\begin{aligned} \sum_{t=1}^{T} \left(\int f_t^{(j)}(\mathbf{z}_t) \log \left[\frac{p(\mathbf{x}_t, \mathbf{z}_t; \theta^{(j+1)})}{f_t^{(j)}(\mathbf{z}_t)} \right] d\mathbf{z}_t \right) &\geq \sum_{t=1}^{T} \left(\int f_t^{(j)}(\mathbf{z}_t) \log \left[\frac{p(\mathbf{x}_t, \mathbf{z}_t; \theta^{(j)})}{f_t^{(j)}(\mathbf{z}_t)} \right] d\mathbf{z}_t \right) \\ &= \ell(\theta^{(j)}) \end{aligned}$$

since we chose $f_t^{(j)}(\mathbf{z}_t)$ to make Jensen's Inequality strict at $\theta^{(j)}$. Now, recall from above that for *any density* $f_t(\mathbf{z}_t)$ and *any* value of θ,

$$\ell(\theta) \geq \sum_{t=1}^{T} \left(\int f_t(\mathbf{z}_t) \log \left[\frac{p(\mathbf{x}_t, \mathbf{z}_t; \theta)}{f_t(\mathbf{z}_t)} \right] d\mathbf{z}_t \right)$$

by Jensen's Inequality. Since this holds in general, it also holds in particular for $\theta = \theta^{(j+1)}$ and $f_t(\mathbf{z}_t) = f_t^{(j)}(\mathbf{z}_t)$. Hence,

$$\ell(\theta^{(j+1)}) \geq \sum_{t=1}^{T} \left(\int f_t^{(j)}(\mathbf{z}_t) \log \left[\frac{p(\mathbf{x}_t, \mathbf{z}_t; \theta^{(j+1)})}{f_t^{(j)}(\mathbf{z}_t)} \right] d\mathbf{z}_t \right)$$

Combining the two inequalities gives $\ell(\theta^{(j+1)}) \geq \ell(\theta^{(j)})$ as claimed.

2 Factor Analysis

Before we proceed, I'll just remind you of some key facts about normal distributions and we'll need below.

2.1 Facts about the Multivariate Normal Distribution

2.1.1 Linear Combinations

Suppose that $X \sim N(\mu, \Sigma)$ and $Y = a + BX$ where a is a vector and B a matrix of constants. Then $Y \sim (a + B\mu, B\Sigma B')$.

2.1.2 Marginals and Conditionals

Let X_1 and X_2 be random vectors such that $(X_1', X_2') \sim N(\mu, \Sigma)$ where

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Then,

$$\begin{aligned} X_1 &\sim N(\mu_1, \Sigma_{11}) \\ X_2 &\sim N(\mu_2, \Sigma_{22}) \\ X_1|X_2 &\sim N(\mu_{1|2}, \Sigma_{1|2}) \end{aligned}$$

where,

$$\begin{aligned} \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned}$$

2.2 The Factor Analysis Model

Classical Factor Analysis specifies a joint distribution on the observable random p-vector X and an unobserved or "latent" random k-vector Z, as follows

$$\begin{aligned} Z &\sim N_k(0_k, I_k) \\ \epsilon &\sim N_p(0_p, \Psi) \\ Z &\perp \epsilon \\ X &= \mu + \Lambda Z + \epsilon \end{aligned}$$

where μ is a $p \times 1$ vector of parameters, Λ is a $p \times k$ matrix of parameters called the *factor loading matrix*, and Ψ is a $p \times p$ *diagonal* matrix of parameters. Factor Analysis can be viewed as a "low rank parameterization" of a multivariate normal distribution. The idea ois that, while X is a random p-vector, its realizations lie *close* to a k-dimensional affine subspace: Λ maps Z from $\mathbb{R}^k$ to a linear subspace of $\mathbb{R}^p$, μ shifts this subspace away from the origin,

Figure 2: Picture of Factor Analysis

and ϵ adds axis-aligned Gaussian noise. Hence it makes sense to require that k is strictly less than both p, the dimension of X, and T, the sample size.

The intution is as follows: Factor Analysis "forces" Z to "explain" the correlation structure of X. This is why Ψ is required to be diagonal. The diagonal elements of Ψ are sometimes called the *idiosyncratic variance terms*, since each corresponds to a *single* component of X.

The factor analysis model implies that the joint distribution of Z and X is normal. Specifically,

$$\begin{aligned}
\begin{bmatrix} Z \\ X \end{bmatrix} &= \begin{bmatrix} 0_k \\ \mu \end{bmatrix} + \begin{bmatrix} I_k & 0_{k\times p} \\ \Lambda & I_p \end{bmatrix} \begin{bmatrix} Z \\ \epsilon \end{bmatrix} \\
&= \begin{bmatrix} 0_k \\ \mu \end{bmatrix} + \begin{bmatrix} I_k & 0_{k\times p} \\ \Lambda & I_p \end{bmatrix} N\left(\begin{bmatrix} 0_k \\ 0_p \end{bmatrix}, \begin{bmatrix} I_k & 0_{k\times p} \\ 0_{p\times k} & \Psi \end{bmatrix} \right) \\
&\sim N\left(\begin{bmatrix} 0_k \\ \mu \end{bmatrix}, \begin{bmatrix} I & \Lambda' \\ \Lambda & \Lambda\Lambda' + \Psi \end{bmatrix} \right)
\end{aligned}$$

The algebra for the variance matrix calculation is as follows:

$$\begin{aligned}
V &= \begin{bmatrix} I_k & 0_{k\times p} \\ \Lambda & I_p \end{bmatrix} \begin{bmatrix} I_k & 0_{k\times p} \\ 0_{p\times k} & \Psi \end{bmatrix} \begin{bmatrix} I_k & 0_{k\times p} \\ \Lambda & I_p \end{bmatrix}' \\
&= \begin{bmatrix} I_k & 0_{k\times p} \\ \Lambda & \Psi \end{bmatrix} \begin{bmatrix} I_k & \Lambda' \\ 0_{p\times k} & I_p \end{bmatrix} \\
&= \begin{bmatrix} I_k & \Lambda' \\ \Lambda & \Lambda\Lambda' + \Psi \end{bmatrix}
\end{aligned}$$

2.3 The Factor Analysis Model is Not Identified

Suppose we want to estimated the parameters μ, Λ, Ψ of the factor analysis model. The first natural question is whether this model is even identified. Unfortunately the answer is no. To see why, suppose that R is an orthogonal matrix, i.e. $RR' = R'R = I$. Geometrically, R is a rotation: it leaves the length of any vector v unchanged since

$$||Rv|| = \sqrt{(Rv)'(Rv)} = \sqrt{v'R'Rv} = \sqrt{v'v} = ||v||$$

From the joint distribution for X and Z that we derived above it follows that the marginal distribution of X is $N(\mu, \Lambda\Lambda' + \Psi)$. Thus if we observe realizations

$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$ of a sequence of iid random vectors $X_1, X_2, \ldots, X_T$ generated from the Factor Analysis model the log-likelihood is given by

$$\ell(\mu, \Lambda, \Psi) = \log \left[\prod_{t=1}^{T} \frac{\exp\left\{-\frac{1}{2}(\mathbf{x}_t - \mu)'(\Lambda\Lambda' + \Psi)^{-1}(\mathbf{x}_t - \mu)\right\}}{(2\pi)^{T/2}|\Lambda\Lambda' + \Psi|^{1/2}} \right]$$

Now suppose that we evaluate the log-likelihood at $\widetilde{\Lambda}R$ rather than Λ. Since Λ only enters through the outer product $\Lambda\Lambda'$ the likelihood is *unchanged*:

$$\widetilde{\Lambda}\widetilde{\Lambda}' = (\Lambda R)(\Lambda R)' = \Lambda R R' \Lambda' = \Lambda\Lambda'$$

We have shown that the matrix of factor loadings is *only identified up to a rotation*. Another way to think about this is in terms of the latent variable Z, Since $X = \mu + \Lambda Z + \epsilon$, post-multiplying Λ by R is the same as *pre-multiplying* Z by R. As explained above, this constitutes a *rotation* of the vector Z. But since Z is a *spherical* normal distribution, rotating it cannot change the likelihood.

If we merely plan to use Factor Analysis for *prediction* this lack of identification is irrelevant: it does not affect the predictive performance of the model in any way. If we ultimately hope to *interpret* the latent factors, however the lack of identification becomes problematic. There are various ways to get a unique solution for the factor loadings Λ that involve making various restrictions on the matrix of factor loadings Λ. The first question is: how many restrictions do we need?

Since the lack of identication comes from rotational invariance, the first step is to count the number of free parameters in a $k \times k$ rotation matrix. Start with the first column: it has $k - 1$ free parameters since the only constraint is that it have length one. The second column must also have length one, but it has the further restriction that it must be orthogonal to the first column. Hence it has $k - 2$ free parameters. Continuing in this way, we see that there are $(k-1) + (k-2) + \ldots + (k-k+1) = k(k-2)/2$ free parameters in a general $k \times k$ rotation matrix.

The mean vector μ doesn't provide any problems for identification since we can always demean X before proceeding. Excluding μ, the factor analysis model has $k(p+1)$ free parameters: Λ is a $p \times k$ matrix and Ψ is a *diagonal* $p \times p$ matrix.

2.4 The Latent Factors

The unobserved random variables $Z_1, \ldots, Z_T$ that generate $X_1, \ldots, X_T$ under the Factor Analysis Model are called the *latent factors* or the *latent scores*. In some settings the factor scores are given a particular interpretation