CS229 Lecture notes

Andrew Ng

Part X
Factor analysis

When we have data $x^{(i)} \in \mathbb{R}^d$ that comes from a mixture of several Gaussians, the EM algorithm can be applied to fit a mixture model. In this setting, we usually imagine problems where we have sufficient data to be able to discern the multiple-Gaussian structure in the data. For instance, this would be the case if our training set size n was significantly larger than the dimension d of the data.

Now, consider a setting in which $d \gg n$. In such a problem, it might be difficult to model the data even with a single Gaussian, much less a mixture of Gaussian. Specifically, since the n data points span only a low-dimensional subspace of $\mathbb{R}^d$, if we model the data as Gaussian, and estimate the mean and covariance using the usual maximum likelihood estimators,

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$$

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)} - \mu)(x^{(i)} - \mu)^T,$$

we would find that the matrix Σ is singular. This means that Σ^{-1} does not exist, and $1/|\Sigma|^{1/2} = 1/0$. But both of these terms are needed in computing the usual density of a multivariate Gaussian distribution. Another way of stating this difficulty is that maximum likelihood estimates of the parameters result in a Gaussian that places all of its probability in the affine space spanned by the data,[1] and this corresponds to a singular covariance matrix.

---

[1] This is the set of points x satisfying $x = \sum_{i=1}^{n} \alpha_i x^{(i)}$, for some α_i's so that $\sum_{i=1}^{n} \alpha_1 = 1$.

More generally, unless n exceeds d by some reasonable amount, the maximum likelihood estimates of the mean and covariance may be quite poor. Nonetheless, we would still like to be able to fit a reasonable Gaussian model to the data, and perhaps capture some interesting covariance structure in the data. How can we do this?

In the next section, we begin by reviewing two possible restrictions on Σ that allow us to fit Σ with small amounts of data but neither will give a satisfactory solution to our problem. We next discuss some properties of Gaussians that will be needed later; specifically, how to find marginal and conditonal distributions of Gaussians. Finally, we present the factor analysis model, and EM for it.

1 Restrictions of Σ

If we do not have sufficient data to fit a full covariance matrix, we may place some restrictions on the space of matrices Σ that we will consider. For instance, we may choose to fit a covariance matrix Σ that is diagonal. In this setting, the reader may easily verify that the maximum likelihood estimate of the covariance matrix is given by the diagonal matrix Σ satisfying

$$\Sigma_{jj} = \frac{1}{n} \sum_{i=1}^{n} (x_j^{(i)} - \mu_j)^2.$$

Thus, Σ_{jj} is just the empirical estimate of the variance of the j-th coordinate of the data.

Recall that the contours of a Gaussian density are ellipses. A diagonal Σ corresponds to a Gaussian where the major axes of these ellipses are axis-aligned.

Sometimes, we may place a further restriction on the covariance matrix that not only must it be diagonal, but its diagonal entries must all be equal. In this setting, we have $\Sigma = \sigma^2 I$, where σ^2 is the parameter under our control. The maximum likelihood estimate of σ^2 can be found to be:

$$\sigma^2 = \frac{1}{nd} \sum_{j=1}^{d} \sum_{i=1}^{n} (x_j^{(i)} - \mu_j)^2.$$

This model corresponds to using Gaussians whose densities have contours that are circles (in 2 dimensions; or spheres/hyperspheres in higher dimensions).

If we are fitting a full, unconstrained, covariance matrix Σ to data, it is necessary that $n \geq d + 1$ in order for the maximum likelihood estimate of Σ not to be singular. Under either of the two restrictions above, we may obtain non-singular Σ when $n \geq 2$.

However, restricting Σ to be diagonal also means modeling the different coordinates x_i, x_j of the data as being uncorrelated and independent. Often, it would be nice to be able to capture some interesting correlation structure in the data. If we were to use either of the restrictions on Σ described above, we would therefore fail to do so. In this set of notes, we will describe the factor analysis model, which uses more parameters than the diagonal Σ and captures some correlations in the data, but also without having to fit a full covariance matrix.

2 Marginals and conditionals of Gaussians

Before describing factor analysis, we digress to talk about how to find conditional and marginal distributions of random variables with a joint multivariate Gaussian distribution.

Suppose we have a vector-valued random variable

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

where $x_1 \in \mathbb{R}^r$, $x_2 \in \mathbb{R}^s$, and $x \in \mathbb{R}^{r+s}$. Suppose $x \sim \mathcal{N}(\mu, \Sigma)$, where

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

Here, $\mu_1 \in \mathbb{R}^r$, $\mu_2 \in \mathbb{R}^s$, $\Sigma_{11} \in \mathbb{R}^{r \times r}$, $\Sigma_{12} \in \mathbb{R}^{r \times s}$, and so on. Note that since covariance matrices are symmetric, $\Sigma_{12} = \Sigma_{21}^T$.

Under our assumptions, x_1 and x_2 are jointly multivariate Gaussian. What is the marginal distribution of x_1? It is not hard to see that $\mathrm{E}[x_1] = \mu_1$, and that $\mathrm{Cov}(x_1) = \mathrm{E}[(x_1 - \mu_1)(x_1 - \mu_1)] = \Sigma_{11}$. To see that the latter is true, note that by definition of the joint covariance of x_1 and x_2, we have

that

$$
\begin{aligned}
\mathrm{Cov}(x) &= \Sigma \\
&= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \\
&= \mathrm{E}[(x-\mu)(x-\mu)^T] \\
&= \mathrm{E}\left[\begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T \right] \\
&= \mathrm{E}\left[\begin{matrix} (x_1-\mu_1)(x_1-\mu_1)^T & (x_1-\mu_1)(x_2-\mu_2)^T \\ (x_2-\mu_2)(x_1-\mu_1)^T & (x_2-\mu_2)(x_2-\mu_2)^T \end{matrix} \right].
\end{aligned}
$$

Matching the upper-left subblocks in the matrices in the second and the last lines above gives the result.

Since marginal distributions of Gaussians are themselves Gaussian, we therefore have that the marginal distribution of x_1 is given by $x_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$.

Also, we can ask, what is the conditional distribution of x_1 given x_2? By referring to the definition of the multivariate Gaussian distribution, it can be shown that $x_1|x_2 \sim \mathcal{N}(\mu_{1|2}, \Sigma_{1|2})$, where

$$
\begin{aligned}
\mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \tag{1} \\
\Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{2}
\end{aligned}
$$

When we work with the factor analysis model in the next section, these formulas for finding conditional and marginal distributions of Gaussians will be very useful.

3   The Factor analysis model

In the factor analysis model, we posit a joint distribution on (x, z) as follows, where $z \in \mathbb{R}^k$ is a latent random variable:

$$
\begin{aligned}
z &\sim \mathcal{N}(0, I) \\
x|z &\sim \mathcal{N}(\mu + \Lambda z, \Psi).
\end{aligned}
$$

Here, the parameters of our model are the vector $\mu \in \mathbb{R}^d$, the matrix $\Lambda \in \mathbb{R}^{d \times k}$, and the diagonal matrix $\Psi \in \mathbb{R}^{d \times d}$. The value of k is usually chosen to be smaller than d.

Thus, we imagine that each datapoint $x^{(i)}$ is generated by sampling a k dimension multivariate Gaussian $z^{(i)}$. Then, it is mapped to a d-dimensional affine space of $\mathbb{R}^d$ by computing $\mu + \Lambda z^{(i)}$. Lastly, $x^{(i)}$ is generated by adding covariance Ψ noise to $\mu + \Lambda z^{(i)}$.

Equivalently (convince yourself that this is the case), we can therefore also define the factor analysis model according to

$$
\begin{aligned}
z &\sim \mathcal{N}(0, I) \\
\epsilon &\sim \mathcal{N}(0, \Psi) \\
x &= \mu + \Lambda z + \epsilon
\end{aligned}
$$

where ϵ and z are independent.

Let's work out exactly what distribution our model defines. Our random variables z and x have a joint Gaussian distribution

$$
\begin{bmatrix} z \\ x \end{bmatrix} \sim \mathcal{N}(\mu_{zx}, \Sigma).
$$

We will now find μ_{zx} and Σ.

We know that $\mathrm{E}[z] = 0$, from the fact that $z \sim \mathcal{N}(0, I)$. Also, we have that

$$
\begin{aligned}
\mathrm{E}[x] &= \mathrm{E}[\mu + \Lambda z + \epsilon] \\
&= \mu + \Lambda \mathrm{E}[z] + \mathrm{E}[\epsilon] \\
&= \mu.
\end{aligned}
$$

Putting these together, we obtain

$$
\mu_{zx} = \begin{bmatrix} \vec{0} \\ \mu \end{bmatrix}
$$

Next, to find Σ, we need to calculate $\Sigma_{zz} = \mathrm{E}[(z - \mathrm{E}[z])(z - \mathrm{E}[z])^T]$ (the upper-left block of Σ), $\Sigma_{zx} = \mathrm{E}[(z - \mathrm{E}[z])(x - \mathrm{E}[x])^T]$ (upper-right block), and $\Sigma_{xx} = \mathrm{E}[(x - \mathrm{E}[x])(x - \mathrm{E}[x])^T]$ (lower-right block).

Now, since $z \sim \mathcal{N}(0, I)$, we easily find that $\Sigma_{zz} = \mathrm{Cov}(z) = I$. Also,

$$
\begin{aligned}
\mathrm{E}[(z - \mathrm{E}[z])(x - \mathrm{E}[x])^T] &= \mathrm{E}[z(\mu + \Lambda z + \epsilon - \mu)^T] \\
&= \mathrm{E}[zz^T]\Lambda^T + \mathrm{E}[z\epsilon^T] \\
&= \Lambda^T.
\end{aligned}
$$

In the last step, we used the fact that $\mathrm{E}[zz^T] = \mathrm{Cov}(z)$ (since z has zero mean), and $\mathrm{E}[z\epsilon^T] = \mathrm{E}[z]\mathrm{E}[\epsilon^T] = 0$ (since z and ϵ are independent, and

hence the expectation of their product is the product of their expectations). Similarly, we can find Σ_{xx} as follows:

$$
\begin{aligned}
\mathrm{E}[(x - \mathrm{E}[x])(x - \mathrm{E}[x])^T] &= \mathrm{E}[(\mu + \Lambda z + \epsilon - \mu)(\mu + \Lambda z + \epsilon - \mu)^T] \\
&= \mathrm{E}[\Lambda z z^T \Lambda^T + \epsilon z^T \Lambda^T + \Lambda z \epsilon^T + \epsilon \epsilon^T] \\
&= \Lambda \mathrm{E}[z z^T] \Lambda^T + \mathrm{E}[\epsilon \epsilon^T] \\
&= \Lambda \Lambda^T + \Psi.
\end{aligned}
$$

Putting everything together, we therefore have that

$$
\begin{bmatrix} z \\ x \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \vec{0} \\ \mu \end{bmatrix}, \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix} \right). \tag{3}
$$

Hence, we also see that the marginal distribution of x is given by $x \sim \mathcal{N}(\mu, \Lambda\Lambda^T + \Psi)$. Thus, given a training set $\{x^{(i)}; i = 1, \ldots, n\}$, we can write down the log likelihood of the parameters:

$$
\ell(\mu, \Lambda, \Psi) = \log \prod_{i=1}^{n} \frac{1}{(2\pi)^{n/2} |\Lambda\Lambda^T + \Psi|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu)^T (\Lambda\Lambda^T + \Psi)^{-1} (x^{(i)} - \mu) \right).
$$

To perform maximum likelihood estimation, we would like to maximize this quantity with respect to the parameters. But maximizing this formula explicitly is hard (try it yourself), and we are aware of no algorithm that does so in closed-form. So, we will instead use to the EM algorithm. In the next section, we derive EM for factor analysis.

4 EM for factor analysis

The derivation for the E-step is easy. We need to compute $Q_i(z^{(i)}) = p(z^{(i)}|x^{(i)}; \mu, \Lambda, \Psi)$. By substituting the distribution given in Equation (3) into the formulas (1-2) used for finding the conditional distribution of a Gaussian, we find that $z^{(i)}|x^{(i)}; \mu, \Lambda, \Psi \sim \mathcal{N}(\mu_{z^{(i)}|x^{(i)}}, \Sigma_{z^{(i)}|x^{(i)}})$, where

$$
\begin{aligned}
\mu_{z^{(i)}|x^{(i)}} &= \Lambda^T (\Lambda\Lambda^T + \Psi)^{-1} (x^{(i)} - \mu), \\
\Sigma_{z^{(i)}|x^{(i)}} &= I - \Lambda^T (\Lambda\Lambda^T + \Psi)^{-1} \Lambda.
\end{aligned}
$$

So, using these definitions for $\mu_{z^{(i)}|x^{(i)}}$ and $\Sigma_{z^{(i)}|x^{(i)}}$, we have

$$
Q_i(z^{(i)}) = \frac{1}{(2\pi)^{k/2} |\Sigma_{z^{(i)}|x^{(i)}}|^{1/2}} \exp\left(-\frac{1}{2}(z^{(i)} - \mu_{z^{(i)}|x^{(i)}})^T \Sigma_{z^{(i)}|x^{(i)}}^{-1} (z^{(i)} - \mu_{z^{(i)}|x^{(i)}}) \right).
$$

Let's now work out the M-step. Here, we need to maximize

$$\sum_{i=1}^{n} \int_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \mu, \Lambda, \Psi)}{Q_i(z^{(i)})} dz^{(i)} \tag{4}$$

with respect to the parameters μ, Λ, Ψ. We will work out only the optimization with respect to Λ, and leave the derivations of the updates for μ and Ψ as an exercise to the reader.

We can simplify Equation (4) as follows:

$$\sum_{i=1}^{n} \int_{z^{(i)}} Q_i(z^{(i)}) \left[\log p(x^{(i)}|z^{(i)}; \mu, \Lambda, \Psi) + \log p(z^{(i)}) - \log Q_i(z^{(i)}) \right] dz^{(i)} \tag{5}$$

$$= \sum_{i=1}^{n} \mathrm{E}_{z^{(i)} \sim Q_i} \left[\log p(x^{(i)}|z^{(i)}; \mu, \Lambda, \Psi) + \log p(z^{(i)}) - \log Q_i(z^{(i)}) \right] \tag{6}$$

Here the $z^{(i)} \sim Q_i$ subscript indicates that the expectation is with respect to $z^{(i)}$ drawn from Q_i. In the subsequent development, we will omit this subscript when there is no risk of ambiguity. Dropping terms that do not depend on the parameters, we find that we need to maximize:

$$\sum_{i=1}^{n} \mathrm{E} \left[\log p(x^{(i)}|z^{(i)}; \mu, \Lambda, \Psi) \right]$$

$$= \sum_{i=1}^{n} \mathrm{E} \left[\log \frac{1}{(2\pi)^{d/2} |\Psi|^{1/2}} \exp\left(-\frac{1}{2} (x^{(i)} - \mu - \Lambda z^{(i)})^T \Psi^{-1} (x^{(i)} - \mu - \Lambda z^{(i)}) \right) \right]$$

$$= \sum_{i=1}^{n} \mathrm{E} \left[-\frac{1}{2} \log |\Psi| - \frac{n}{2} \log(2\pi) - \frac{1}{2} (x^{(i)} - \mu - \Lambda z^{(i)})^T \Psi^{-1} (x^{(i)} - \mu - \Lambda z^{(i)}) \right]$$

Let's maximize this with respect to Λ. Only the last term above depends on Λ. Taking derivatives, and using the facts that $\mathrm{tr}\, a = a$ (for $a \in \mathbb{R}$), $\mathrm{tr} AB = \mathrm{tr} BA$, and $\nabla_A \mathrm{tr} ABA^T C = CAB + C^T AB^T$, we get:

$$\nabla_\Lambda \sum_{i=1}^{n} -\mathrm{E} \left[\frac{1}{2} (x^{(i)} - \mu - \Lambda z^{(i)})^T \Psi^{-1} (x^{(i)} - \mu - \Lambda z^{(i)}) \right]$$

$$= \sum_{i=1}^{n} \nabla_\Lambda \mathrm{E} \left[-\mathrm{tr} \frac{1}{2} z^{(i)^T} \Lambda^T \Psi^{-1} \Lambda z^{(i)} + \mathrm{tr} z^{(i)^T} \Lambda^T \Psi^{-1} (x^{(i)} - \mu) \right]$$

$$= \sum_{i=1}^{n} \nabla_\Lambda \mathrm{E} \left[-\mathrm{tr} \frac{1}{2} \Lambda^T \Psi^{-1} \Lambda z^{(i)} z^{(i)^T} + \mathrm{tr} \Lambda^T \Psi^{-1} (x^{(i)} - \mu) z^{(i)^T} \right]$$

$$= \sum_{i=1}^{n} \mathrm{E} \left[-\Psi^{-1} \Lambda z^{(i)} z^{(i)^T} + \Psi^{-1} (x^{(i)} - \mu) z^{(i)^T} \right]$$

Setting this to zero and simplifying, we get:

$$\sum_{i=1}^{n} \Lambda \mathrm{E}_{z^{(i)}\sim Q_i}\left[z^{(i)}z^{(i)T}\right] = \sum_{i=1}^{n}(x^{(i)} - \mu)\mathrm{E}_{z^{(i)}\sim Q_i}\left[z^{(i)T}\right].$$

Hence, solving for Λ, we obtain

$$\Lambda = \left(\sum_{i=1}^{n}(x^{(i)} - \mu)\mathrm{E}_{z^{(i)}\sim Q_i}\left[z^{(i)T}\right]\right)\left(\sum_{i=1}^{n}\mathrm{E}_{z^{(i)}\sim Q_i}\left[z^{(i)}z^{(i)T}\right]\right)^{-1}. \quad (7)$$

It is interesting to note the close relationship between this equation and the normal equation that we'd derived for least squares regression,

$$\text{``}\theta^T = (y^T X)(X^T X)^{-1}.\text{''}$$

The analogy is that here, the x's are a linear function of the z's (plus noise). Given the "guesses" for z that the E-step has found, we will now try to estimate the unknown linearity Λ relating the x's and z's. It is therefore no surprise that we obtain something similar to the normal equation. There is, however, one important difference between this and an algorithm that performs least squares using just the "best guesses" of the z's; we will see this difference shortly.

To complete our M-step update, let's work out the values of the expectations in Equation (7). From our definition of Q_i being Gaussian with mean $\mu_{z^{(i)}|x^{(i)}}$ and covariance $\Sigma_{z^{(i)}|x^{(i)}}$, we easily find

$$\begin{aligned}
\mathrm{E}_{z^{(i)}\sim Q_i}\left[z^{(i)T}\right] &= \mu_{z^{(i)}|x^{(i)}}^T \\
\mathrm{E}_{z^{(i)}\sim Q_i}\left[z^{(i)}z^{(i)T}\right] &= \mu_{z^{(i)}|x^{(i)}}\mu_{z^{(i)}|x^{(i)}}^T + \Sigma_{z^{(i)}|x^{(i)}}.
\end{aligned}$$

The latter comes from the fact that, for a random variable Y, $\mathrm{Cov}(Y) = \mathrm{E}[YY^T] - \mathrm{E}[Y]\mathrm{E}[Y]^T$, and hence $\mathrm{E}[YY^T] = \mathrm{E}[Y]\mathrm{E}[Y]^T + \mathrm{Cov}(Y)$. Substituting this back into Equation (7), we get the M-step update for Λ:

$$\Lambda = \left(\sum_{i=1}^{n}(x^{(i)} - \mu)\mu_{z^{(i)}|x^{(i)}}^T\right)\left(\sum_{i=1}^{n}\mu_{z^{(i)}|x^{(i)}}\mu_{z^{(i)}|x^{(i)}}^T + \Sigma_{z^{(i)}|x^{(i)}}\right)^{-1}. \quad (8)$$

It is important to note the presence of the $\Sigma_{z^{(i)}|x^{(i)}}$ on the right hand side of this equation. This is the covariance in the posterior distribution $p(z^{(i)}|x^{(i)})$ of $z^{(i)}$ give $x^{(i)}$, and the M-step must take into account this uncertainty

about $z^{(i)}$ in the posterior. A common mistake in deriving EM is to assume that in the E-step, we need to calculate only expectation $E[z]$ of the latent random variable z, and then plug that into the optimization in the M-step everywhere z occurs. While this worked for simple problems such as the mixture of Gaussians, in our derivation for factor analysis, we needed $E[zz^T]$ as well $E[z]$; and as we saw, $E[zz^T]$ and $E[z]E[z]^T$ differ by the quantity $\Sigma_{z|x}$. Thus, the M-step update must take into account the covariance of z in the posterior distribution $p(z^{(i)}|x^{(i)})$.

Lastly, we can also find the M-step optimizations for the parameters μ and Ψ. It is not hard to show that the first is given by

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x^{(i)}.$$

Since this does not change as the parameters are varied (i.e., unlike the update for Λ, the right hand side does not depend on $Q_i(z^{(i)}) = p(z^{(i)}|x^{(i)}; \mu, \Lambda, \Psi)$, which in turn depends on the parameters), this can be calculated just once and needs not be further updated as the algorithm is run. Similarly, the diagonal Ψ can be found by calculating

$$\Phi = \frac{1}{n} \sum_{i=1}^{n} x^{(i)}x^{(i)^T} - x^{(i)} \mu_{z^{(i)}|x^{(i)}}^T \Lambda^T + \Lambda \mu_{z^{(i)}|x^{(i)}} x^{(i)^T} + \Lambda(\mu_{z^{(i)}|x^{(i)}} \mu_{z^{(i)}|x^{(i)}}^T + \Sigma_{z^{(i)}|x^{(i)}})\Lambda^T,$$

and setting $\Psi_{ii} = \Phi_{ii}$ (i.e., letting Ψ be the diagonal matrix containing only the diagonal entries of Φ).