Dave Blei:
Scribe: Greg Detre

Lecture # 20

---

1 Multivariate Gaussian distributions

The multivariate Gaussian can be defined in terms of its mean, μ, a p x 1 vector, and its covariance, Σ, p x p positive definite, symmetrical, invertible matrix.

The covariance for a pair of components i and j:

$$\sigma_{ij} = E[x_i x_j] - E[x_i]E[x_j] \tag{1}$$

The variance for a single i^{th} component:

$$\sigma_{ii} = E[x_i^2] - E[x_i]^2 \tag{2}$$

The density for a given x is given by:

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \right) \tag{3}$$

The covariance matrix Σ describes the shape of the multivariate Gaussian distribution. We can visualize it by drawing contours of constant probability in p dimensions:

$$F(x) = 1/2(x-\mu)^T \Sigma^{-1}(x-\mu) \tag{4}$$

The simplest covariance matrix to think about is an identity matrix. This yields a circular Gaussian distribution in 2 dimensions, or a hypersphere in higher dimensions, where each component has a variance of 1, e.g.

$$\begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{matrix}$$

If you scale the individual components, this will cause the distribution to be ellipsoid, but still oriented along the axes, e.g.

$$\begin{matrix} 1.5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 5 \end{matrix}$$

If some of the off-diagonal values are non-zero, this covariance between components skews the orientation of the Gaussian so that it is no longer necessarily oriented along the axes.

In other words, the off-diagonal covariance values tell you that some of the components are non-independent, i.e. they vary with respect to one another, e.g.

$$\begin{matrix} 1.5 & 0 & 0.3 \\ 0 & 1 & 0 \\ 0 & 0 & 5 \end{matrix}$$

The mean μ defines the offset of the whole distribution, shifting the whole thing in space.

If we want to find the maximum likelihood estimate of the parameters of a multivariate Gaussian distribution, given some $X_1 ... X_n$ p-dimensional data points:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{5}$$

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})(x_i - \hat{\mu})^T \tag{6}$$

These reduce down to the standard MLE estimates for a Gaussian in one dimension.

Now let's take a multivariate Gaussian and divide it into two pieces, X_1 and X_2, that are themselves multivariate Gaussians.

$$[\vec{X}] = [\vec{X_1} \quad \vec{X_2}]$$

$p(\vec{x}_1, \vec{x}_2)$ is a Gaussian too.

We can now decompose μ into two parts:

$$\vec{\mu} = < \vec{\mu}_1, \vec{\mu_2} > \tag{7}$$

and likewise decompose the covariance matrix into 4 parts:

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

By the chain rule, we know that:

$$p(x_1, x_2) = p(x_2) p(x_1 | x_2) \tag{8}$$

The marginal of x_2 is Gaussian:

$$\mu_m = \mu_2 \tag{9}$$

$$\Sigma_m = \Sigma_{22} \tag{10}$$

The conditional $x1|x2$ is Gaussian:

$$\mu_c = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \tag{11}$$

$$\Sigma_c = \Sigma_{11} - \Sigma_{12}\Sigma^{-1}\Sigma_{21} \tag{12}$$

Knowing x_2 tells us something about the distribution of x_1.

Sidenote: the Gaussian distribution is *conjugate to itself* - that is, when you have a Gaussian, and you condition on a Gaussian, then you get another Gaussian.

2 Factor analysis

Factor analysis is another dimensionality reduction algorithm, that uses latent variables. It's a lot like PCA, except that it's probabilistic. It was used a lot in the 1950s, but fell out of favour because it doesn't find a global minimum solution, is not identifiable, and therefore is difficult to interpret in a causal way (as was the style). However, it's a perfectly good uninterpreted dimensionality reduction technique.

$$\vec{Z}_n \sim N_q(\vec{O}, I) \tag{13}$$

where:

Z is the low-dimensional space we're trying to project into

$\vec{0}$ is a vector of zeros

I = the identity matrix

$$\vec{X}_n \sim N_p(\mu + \Lambda z, \Psi) \tag{14}$$

where:

μ is the mean, a (q x 1) vector

Λ is a (p x q) matrix

Ψ is diagonal, i.e. all the components are independent

Let's assume a zero mean for simplicity from now on.

$$\vec{X}_n \sim N_p(\Lambda z, \Psi) \tag{15}$$

In PCA:

$$X = Z_1\vec{\lambda}_1 + Z_2\vec{\lambda}_2 + ... + Z_q\vec{\lambda}_q \tag{16}$$

where:

$\vec{Z}$ is the low-dimensional representation of $\vec{X}$

On the other hand, in factor analysis, each component can have its own variability and these λ's don't have to be orthogonal:

$$X \sim N(Z_1\vec{\lambda}_1 + Z_2\vec{\lambda}_2 + ... + Z_q\vec{\lambda}_q, \Psi) \tag{17}$$

Now, imagine we're trying to project a 3D space onto a 2D plane.

$$p = 3, q = 2 \tag{18}$$

In other words, we're assuming that there are 2 'factors', but your data are in a 3rd dimension as a result of noise. The Z's reflect common sources of variation amongst the data and account for its correlation structure. Uncorrelated $\vec{\epsilon}$s are unique to each diension and pick up remaining variation that's not accounted for.

The covariance matrix is going to define some kind of sphere or ellipsoid. We can contrast this with PCA which requires the covariance to be a sphere, since the components all have the same variability in the lower dimensional space.

The joint distribution of $(\vec{Z}, \vec{X})$ is a $(p+q)$-dimensional Gaussian.

$$\mu_J =< \vec{0}_q, \vec{0}_p > \tag{19}$$

$$\Sigma_J = \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix}$$

where:

I is $var(Z)$

Λ is $cov(Z, X)$

Λ^T is $cov(X, Z)$

$\Lambda\Lambda^T + \Psi$ is $var(X)$

From this, we know that:

$$\vec{X} \sim N(\vec{0},\ \Lambda\Lambda^T + \Psi) \tag{20}$$

$$\vec{Z}|\vec{X} \sim N(\Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}\vec{x},\ (I + \Lambda^T\Psi^{-1}\Lambda)^{-1}) \tag{21}$$

Reminder: we're assuming that the low-dimensional components are independent, although of course $\vec{z}_n|\vec{x}_n$ are not independent.

2.1 Estimating Λ with EM

Notice that:

$$x_n = \Lambda z_n + \epsilon \tag{22}$$

This looks just like linear regression. If we knew Z, we'd simply be doing multivariate regression - but we don't, so we're going to use EM to find Λ and Ψ.

2.1.1 E-step

Compute:

$$p(\vec{Z}_n|\vec{X}_n, \Lambda, \Psi) \tag{23}$$

2.1.2 M-step

Calculate the MLE with

$$\vec{Z}_n, \vec{Z}_n\vec{Z}_n^T \tag{24}$$

replaced by their posterior expectations:

$$\hat{\Lambda} = (\sum_{n=1}^{N}(\vec{x}_n E[z_n|x_n]^T)) + (\sum_{n=1}^{N} E[z_n z_n^T|x_n])^{-1} \tag{25}$$

In linear regression, you've observed the covariates, but in the M-step, you replace those with the expectations you computed in the E-step.