Multivariate Gaussians

We're going to use Gaussian distributions as parts of models of data, and to represent beliefs about models. Most models and algorithms in machine learning involve more than one scalar variable however. (A scalar meaning a single number, rather a vector of values.) Multivariate Gaussians generalize the univariate Gaussian distribution to multiple variables, which can be dependent.

[This note assumes that you know the background material on expectations of random variables.]

Independent Standard Normals

We could sample a vector $\mathbf{x}$ by independently sampling each element from a standard normal distribution, $x_d \sim \mathcal{N}(0,1)$. Because the variables are independent, the joint probability is the product of the individual or marginal probabilities:

$$p(\mathbf{x}) = \prod_{d=1}^{D} p(x_d) = \prod_{d=1}^{D} \mathcal{N}(x; 0, 1).$$

Usually I recommend that you write any Gaussian PDFs in your maths using the $\mathcal{N}(x; \mu, \sigma^2)$ notation unless you have to expand them. It will be less writing, and clearer. Here, I want to combine the PDFs, so will substitute in the standard equation:

$$p(\mathbf{x}) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi}} \, e^{-x_d^2/2} = \frac{1}{(2\pi)^{D/2}} \, e^{-\frac{1}{2} \sum_d x_d^2}$$

$$= \frac{1}{(2\pi)^{D/2}} \, e^{-\frac{1}{2} \mathbf{x}^\top \mathbf{x}}.$$

The PDF is proportional to the Radial Basis Functions (RBFs) we've used previously. Here the normalizer $1/(2\pi)^{D/2}$ means that the PDF integrates to one.

Like an RBF centred at the origin, this density function only depends on the (square) distance or radius of $\mathbf{x}$ from the origin. Any point in a spherical shell (or a circular shell in 2-dimensions) is equally probable. Therefore if we simulate points in 2-dimensions and draw a scatter plot:

```
% Matlab/Octave
N = 1e4; D = 2;
X = randn(N, D);
plot(X(:,1), X(:,2), '.');
axis('square');
```

Or

```
# Python
N = int(1e4); D = 2
X = np.random.randn(N, D)
plt.plot(X[:,0], X[:,1], '.')
plt.axis('square')
```

we will see a diffuse circular spray of points. The spherical symmetry is a special property of Gaussians. If you were to draw independent samples from, say, a Laplace distribution you would see a non-circular distribution, which has more density close to the axes.

Covariance

The multivariate generalization of variance, is *covariance*, which is represented with a matrix. While a variance is often denoted σ^2, a covariance matrix is often denoted Σ — not to be confused with a summation $\sum_{d=1}^{D} \dots$

The elements of the covariance matrix for a random vector $\mathbf{x}$ are:

$$\mathrm{cov}[\mathbf{x}]_{ij} = \mathbb{E}[x_i x_j] - \mathbb{E}[x_i]\mathbb{E}[x_j].$$

On the diagonal, where $i=j$, you will see that this definition gives the (marginal) variance $\mathrm{var}[x_i]$ of the elements of the vector. We can write the whole matrix with a linear algebra expression:

$$\mathrm{cov}[\mathbf{x}] = \mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^\top.$$

Question: What is the covariance S of the spherical distribution of the previous section? (I will reserve Σ for the covariance of the general Gaussian in the next section.)

Answer: The first term is $\mathbb{E}[x_i x_j] = \mathbb{E}[x_i]\mathbb{E}[x_j]$ if x_i and x_j are independent, which they are if $i \neq j$. Thus $S_{i \neq j} = 0$. The diagonal elements $S_i i$ are equal to the variances of the individual variables, which are all equal to one. Therefore, $S_{ij} = \delta_{ij}$, where δ_{ij} is a Kroneker delta. Or as a matrix, $S = \mathbb{I}$, the identity matrix.

Transforming and Rotating: general Gaussians

As with one-dimensional Gaussians we can generalize the standard zero-mean, unit-variance Gaussian by a linear transformation and a shift.

If we generated $\mathbf{x}$ from independent $\mathcal{N}(0,1)$ draws as above, we could form a linear combination of these outcomes:

$$\mathbf{y} = A\mathbf{x}.$$

To keep the discussion simpler, I will assume that A is square and invertible, so $\mathbf{y}$ has the same dimensionality as $\mathbf{x}$.

Question: What is the covariance Σ of the new variable $\mathbf{y}$?

Answer: Simply substitute $\mathbf{y}$ into the definition:

$$\begin{aligned}
\mathrm{cov}[\mathbf{y}] &= \mathbb{E}[\mathbf{y}\mathbf{y}^\top] - \mathbb{E}[\mathbf{y}]\mathbb{E}[\mathbf{y}]^\top \\
&= \mathbb{E}[A\mathbf{x}\mathbf{x}^\top A^\top] - \mathbb{E}[A\mathbf{x}]\mathbb{E}[A\mathbf{x}]^\top \\
&= A\mathbb{E}[\mathbf{x}\mathbf{x}^\top]A^\top - A\mathbb{E}[\mathbf{x}](A\mathbb{E}[\mathbf{x}])^\top.
\end{aligned}$$

Because $\mathbb{E}[\mathbf{x}]$ is zero, the second term is zero, and the expectation in the first term is equal to $\mathrm{cov}[\mathbf{x}] = \mathbb{I}$. Therefore,

$$\mathrm{cov}[\mathbf{y}] = \Sigma = AA^\top.$$

As we're assuming A is invertible, we can compute the original vector from the transformed one: $\mathbf{x} = A^{-1}\mathbf{y}$. Substituting that expression into the PDF for $\mathbf{x}$ we can see the shape of the new PDF:

$$\begin{aligned}
p(\mathbf{y}) &\propto e^{-\frac{1}{2}(A^{-1}\mathbf{y})^\top (A^{-1}\mathbf{y})} \\
&\propto e^{-\frac{1}{2}\mathbf{y}^\top A^{-\top} A^{-1}\mathbf{y}}.
\end{aligned}$$

However, if we have stretched out the PDF we must scale it down to maintain probability mass. If we apply a linear transformation A to a volume of points, then the volume is multiplied by $|A|$, the determinant of the matrix.[1] Therefore,

$$p(\mathbf{y}) = \frac{1}{|A|(2\pi)^{D/2}} e^{-\frac{1}{2}\mathbf{y}^\top A^{-\top} A^{-1}\mathbf{y}}.$$

---

1. Here $|A|$ is the "Jacobian of the transformation", although confusingly "Jacobian" can refer to both a matrix and its determinant.

Usually this expression is re-written in terms of the covariance of the vector. Noticing that

$$\Sigma^{-1} = A^{-\top} A^{-1}, \text{ and}$$

$$|\Sigma| = |AA^\top| = |A||A^\top| = |A|^2,$$

we can write:

$$p(\mathbf{y}) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{D/2}} \, e^{-\frac{1}{2}\mathbf{y}^\top \Sigma^{-1}\mathbf{y}} = |2\pi\Sigma|^{-1/2} \, e^{-\frac{1}{2}\mathbf{y}^\top \Sigma^{-1}\mathbf{y}}.$$

As demonstrated above, there are different equivalent ways to write the normalizing constant, and different books will choose different forms.

Finally, we can shift the distribution to have non-zero mean:

$$\mathbf{z} = \mathbf{y} + \boldsymbol{\mu}.$$

Shifting the PDF does not change its normalization, so we can simply substitute $\mathbf{y} = \mathbf{z} - \boldsymbol{\mu}$ into the PDF for $\mathbf{y}$:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \Sigma) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{D/2}} \, e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{z}-\boldsymbol{\mu})}.$$

Here we've generalized the $\mathcal{N}$ notation for Gaussian distributions to take a mean vector and a matrix of covariances. In one-dimension, these still correspond to the scalar mean, and the variance.

It's a common mistake to forget the matrix inverse inside the exponential. The inverse covariance matrix Σ^{-1}, is also known as the precision matrix.

Check your understanding

A special case of a general transformation A, is a diagonal matrix, that simply stretches each variable independently: $\Lambda_{ij} = \delta_{ij}\sigma_i$. For three dimensions the transformation would be:

$$\Lambda = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix}.$$

What is the covariance matrix Σ? If $\mathbf{x}$ is generated with standard normals, and $\mathbf{z} = \Lambda\mathbf{x}$, can you write $p(\mathbf{z})$ as a product of univariate Gaussian distributions? Are the z_i variables all independent?

You could use Matlab/Octave or Python to sample some points from different multivariate Gaussians, and see how the covariance affects the cloud of points.

For example, you could use a family of transformations parameterized by a:

$$A = \begin{bmatrix} 1 & 0 \\ a & 1-a \end{bmatrix},$$

What does this transformation do? Is it clear why the variables are dependent for $a \neq 0$? When are the variables maximally dependent? What happens to the PDF as $a \rightarrow 1$ and why? Does the covariance matrix have an inverse when $a = 1$?

The shape of a two-dimensional Gaussian is often sketched using a contour of its PDF. The contours of the radially symmetric Gaussian are circular. So if you compute the (x_1, x_2) coordinates on a circle, these can be joined up to plot a contour of the Gaussian with identity covariance. You can then transform these points, just like sample positions, with a matrix A, to plot a contour of $N(\mathbf{x}; \mathbf{0}, AA^\top)$.

A puzzle: Given a covariance matrix Σ, the transformation A is not uniquely defined. For example

$$A = \left[\begin{array}{cc} 2 & 0 \\ 0 & 2 \end{array} \right], \quad \text{and} \quad A = \left[\begin{array}{cc} \sqrt{3} & 1 \\ -1 & \sqrt{3} \end{array} \right],$$

both give the same product $AA^\top$. Can you explain how to generate an arbitrary number of transformations A that share the same $\Sigma = AA^\top$?

Further reading

https://en.wikipedia.org/wiki/Multivariate_normal_distribution

Barber Section 8.4 starts with the definition that this note builds up to, and then works in the reverse direction from there to build up an interpretation. The section then goes further than this note, and there are some Gaussian-based exercises at the end of the Chapter. The treatment in Murphy, Section 2.5.2, is rather more terse!

Transforming the PDF of the spherical distribution required getting the normalization correct due to the change of variables. You would have to do more work if the transformation was non-linear. The maths for transforming a PDF due to a change of variables is quickly reviewed in Barber Section 8.2, Result 8.1. The treatment is longer in Murphy this time, in Section 2.6.