BioMath - Probability - additional Gaussian question - August 29, 2014

1. **Multidimensional Gaussian distributions**

In this problem, we'll build up to teaching you about multidimensional Gaussian distributions, which are widely used.

First off, let's start with a uniform distribution in two dimensions. Each trial of an experiment produces two outcomes, x_1 and x_2, which are independent, and each of which is uniformly distributed between 0 and 10.

• Use Matlab to sample 100 times from this joint distribution. Make a figure in which for each one of your samples, place a blue dot at its (x_1, x_2) position. In addition, place red dots at $(0, 0)$, $(0, 10)$, $(10, 0)$, and $(10, 10)$. Use 'MarkerSize', 10 for your blue dots and 'MarkerSize' 26 for your red dots. After plotting, set the axes limits and aspect ratio with `axis equal; xlim([-1 11]); ylim([-1 11]);`

• Now, in a new figure, transform your points tot he new coordinates given by $\mathbf{x_{new}} = A\mathbf{x}$ where

$$A = \begin{pmatrix} 0.5 & 0.25 \\ 0.25 & 0.5 \end{pmatrix} \tag{1}$$

And replot them in a new figure. Once again, use `axis equal; xlim([-1 11]); ylim([-1 11]);` so that the two figures are directly comparable.

• What happened to the density of points in the new figure? Did it go up or down? By what factor did it go up or down? (Hint: it's related to the determinant of A.)

• What was the probability density $P(x, y)$ in the original coordinates? To answer this, remember that it was a uniform distribution, and that by definition of probability distributions, since they have to add up to 1, we know that

$$\int_{x=0}^{\infty} \int_{y=0}^{\infty} P(x, y) \mathrm{d}x \mathrm{d}y = 1 \tag{2}$$

• In the new corrdinates, within the parallelepiped given by the red dots, the probability density is still uniform, meaning, that it is the same at every point within the parallelepiped. But what is the value of the probability density in the new coordinates?

Now let's define a Gaussian in one and two dimensions. We'll define a one-dimensional Gaussian with mean μ and variance σ^2 as the probability density

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \tag{3}$$

We'll assert (please believe us) that the integral of this over x is 1. We'll mostly be working $\mu = 0$.

Let's take two such Gaussians, independent from each other, each with mean zero and variance equal to 1. This defines a two-dimensional density,

$$P(x_1, x_2) = \frac{1}{2\pi} e^{-(x_1^2+x_2^2)/2} \tag{4}$$

Let's write $\mathbf{x} = [x_1 x_2]^T$, to get

$$P_x(\mathbf{x}) = \frac{1}{2\pi} e^{\frac{1}{2}\mathbf{x}^T\mathbf{x}} \tag{5}$$

• Show that the covariance matrix for this distribution $C_x = \langle \mathbf{x}\mathbf{x}^T \rangle = I$.

• Use Matlab's `randn` to create one hundred samples from this distribution, and in a new figure, plot them all using a blue dot for each sample, similarly to what you did for the uniform distribution.

• Now transform each point with $\mathbf{y} = A\mathbf{x}$, and plot the the $\mathbf{y}$ points. Can you see the Gaussian ellipse shape?

• Show, using linear algebra, that the covariance matrix of $\mathbf{y}$ is $C_y = \langle \mathbf{y}\mathbf{y}^T \rangle = AA^T$.

• Now show, again using linear algebra, that $C_y^{-1} = (A^{-1})^T A^{-1}$. We will further assert that $|C_y| = |A|^2$.

Let's all the probability density in the $\mathbf{x}$ space $P_x(\mathbf{x})$, and let's call the probability density in the $\mathbf{y}$ space $P_y(\mathbf{y})$. We'll furthermore assert (and hope that you can follow the logic in the line below that lets us do this) that $P_y(\mathbf{y})$ is given by

$$P_y(\mathbf{y}) = \frac{1}{|A|} P_x(\mathbf{x}(\mathbf{y})) = \frac{1}{|A|} P_x(A^{-1}\mathbf{y}) = \frac{1}{2\pi|A|} e^{\frac{1}{2}\mathbf{y}^T(A^{-1})^T A^{-1}\mathbf{y}} \tag{6}$$

• Use your results to show that in the case $N = 2$, and N-dimensional Gaussian in the vector space $\mathbf{y}$, with mean zero and covariance C_y has probability density

$$P(\mathbf{y}) = \frac{1}{\sqrt{2\pi|C_y|}} e^{\frac{1}{2}\mathbf{y}^T C_y^{-1}\mathbf{y}} \tag{7}$$

The nice thing about this formula is that although you developed it for 2 dimensions, it holds for N dimensions! Not only that, but you now know how to generate samples from such a Gaussian: you first sample $\mathbf{x}$ from the N-dimensional independent Gaussian with variances equal to 1 (which is easy because the N dimensions are independent); and then you transform your samples using $\mathbf{y} = A\mathbf{x}$, where $A = C^{0.5}$. And you're done.

(You might ask, "suppose I know C, how do I get $C^{0.5}$?". As usual, when in doubt, diagonalize).