In this month's blog, we explore some of the statistical results from multivariate Gaussian distributions. Typically, many of these results were developed in support of particular problems in science and engineering but they stand on their own in that the results flow from the logic of statistical thinking that can be cleanly separated from the application that birthed them.

The Monte Carlo method, which was the subject in a previous run of blogs (here to here), will be used here as an additional check on the techniques developed but there will be no emphasis on the mechanics of these techniques.

For a typical, motivating example, consider the expectation value presented in Section 1.2 of Brezin's text *Introduction to Statistical Field Theory*, which asks the reader to compute

\[ E[x^4 y^2] = \frac{1}{N} \int\_{-\infty}^{\infty} \int\_{-\infty}^{\infty} x^4 y^2 e^{-(x^2 + xy + 2 y^2)} \; , \]

where the normalization constant $N$ is given by

\[ N = \int\_{-\infty}^{\infty} \int\_{-\infty}^{\infty} e^{-(x^2 + xy + 2 y^2)} \; . \]

Examples like these arise in any discipline where the model describing a system depends several (perhaps a huge number) of parameters that manifest themselves as inter-correlated variables that we can measure. The cross-correlation between the variables $x$ and $y$ is display by the $xy$ term in the exponential.

The first task in determining the expectation value is to evaluate the normalization constant. The general method for solving this task will be focus of this installment.

The method for integrating Gaussian integrals in one dimension is well known (see, e.g., [here](http://underthehood.blogwyrm.com/?p=835)) but the approach for multi-variate distributions is not as common. The first point is to look at the quadratic form in the argument of the exponential as being built as the inner product between an array presenting the variables $x$ and $y$ and some matrix $A$. In the present case this is done by recognizing

\[ x^2 + xy + 2 y^2 = \left[ \begin{array}{cc} x & y \end{array} \right] \left[ \begin{array}{cc} 1 & 1/2 \\ 1/2 & 2 \end{array} \right] \left[ \begin{array}{c} x \\ y \end{array} \right] \; . \]

Looking ahead to higher dimensional cases, we can see that this expression can be written more generally and compactly as $r^T A r \equiv r\_i A\_{ij} r\_j$, where $r$ is a column array presenting the independent variable of the distribution (now imagined to be $n$ in number) and $A$ is a real $n \times n$ matrix; the superscript $T$ indicates that the transpose of the original array is being used.

With this new notation, any $n$-dimensional Gaussian distribution can now be written as

\[ e^{-(r^T A r)} \; . \]

Since this expression is symmetric in $r$, $A$ can be taken to be symmetric without loss of generality. This observation guarantees that $A$ can be diagonalized by an orthogonal matrix $M$ such that $M^T A M = D$, $M^T M = M M^T = 1$, and $D$ is a diagonal matrix. This in turn means that the quadratic form can be rewritten as

\[ r^T A r = r^T (M M^T) A (M M^T) r = (r^T M) (M^T A M) (M r) = (r^T M) D (M^T r) \; . \]

The next step is to define a new set of variables defined by $s = M^T r$, which then decouples the multi-variate distribution into a product of $D$ independent one-dimensional Gaussians:

\[ e^{ - (r^T A r) } = e^{ - (s^T D s) } = \prod\_{i=1}^{n} e^{-d\_i s\_i^2} \; , \]

where the $\{d\_i\}$ are the individual, on-diagonal elements of $D$ and the $\{s\_i\}$ For the integral to exist (i.e., converge) we need one additional stipulation that all of the $d\_i > 0$. Said differently, we require that $A$ have positive eigenvalues (i.e., be positive definite).

The final step is to recognize that the measure of the integral changes in a convenient way when we make the substitution $r = M^T s$ since

\[ d^n r = det(M^T) d^n s = d^n s \; , \]

where we've exploited the fact that since $M$ is orthogonal, its determinant is 1.

Putting all of these pieces together we now have

\[ \int\_{R^n} d^n e^{-(r^T A r} = \int\_{R^n} d^n s \prod\_{i=1}^n e^{-d\_i s\_i^2} = \prod\_{i=1}^{n} \int\_{-\infty}^\infty ds\_i e^{-d\_i s\_i^2} \; .\]

Since the right-hand side is a product of independent Gaussian integrals it simplifies to

\[\prod\_{i=1}^{n} \int\_{-\infty}^\infty ds\_i e^{-d\_i s\_i^2} = \prod\_{i=1}^{n} \sqrt{\frac{\pi}{d\_i}} = \frac{\pi^{n/2}}{\sqrt{d\_1 \cdots d\_n}} \; . \]

The denominator of the last expression is the determinant of the $D$, which is also the determinant of $A$. So, the final formula is

\[ \int\_{R^n} d^n r e^{-(r^T A r)} = \frac{\pi^{n/2}}{\sqrt{det(A)}} \; .\]

Returning to the original example, the formula tells us that since $det(A) = 7/4$ then the normalization constant $N = 2\pi/\sqrt(7) \approx 2.37482$. We can confirm this using brute force Monte Carlo by uniformly sampling across the ranges $x \in [-6,6]$ and $y \in [-12,12]$. The limits of these ranges where selected by going out roughly to 6 times the respective eigenvalue ($d\_x = 0.79$, $d\_y = 2.21$) but the estimated integral is insensitive to fine details as long as the computational domain is sufficiently large. That said, there is an art to picking the computational domain and sophisticated pieces of software are designed to adjust to the integrand. Implementing the brute force Monte Carlo in python gave an estimate of $N = 2.37661$ for 10 million trials, an error of less than $0.1\%$.

As a last sample, suppose we provide a bigger matrix given by:

\[ A = \left[ \begin{array}{cccc} 3 & 1 & 5 & 3\\ 1 & 5 & 2 & 0\\ 5 & 2 & 10 & 3 \\ 3 & 0 & 3 & 14 \end{array} \right] \; .\]

The Gaussian integral is

\[ N = \int\_{R^n} d^n r \, e^{-(3w^2 + 2wx + 10wy + 6w^2 + 5x^2 + 4xy + 10y^2 + 6yz + 14z^2)} \; .\]

The eigenvalues of $A$ are: $\{17.63,0.26,9.85,4.26\}$. Application of the formula give $N = 0.70497$ compared with a brute force Monte Carlo estimate of $0.71061$.

With this operation well in hand, next month's post will focus on calculating various moments of these distributions.