Correlated, Uniform, Random Values

Andrew Cooke*

November 2009

Abstract

I describe two ways to generate pairs of psuedo–random values, each distributed uniformly, but which are also mutually correlated.

1 Introduction

It is easy to generate two correlated, normally distributed variates (section 2), and the algorithm is well known. Unfortunately the same approach does not work for *uniformly* distributed variates. This led me to assume that the problem was more difficult. In fact, it is not — there is a very simple approach that works well (section 3).

That solution, however, "feels wrong". So I also provide a second method that addresses some of the problems with the first (section 4). But even that solution is not, in my opinion, perfect. My final hope is that by providing two imperfect solutions I can help uncover a better approach (section 5).

2 Normal (Gaussian) Variates

Figure 1 shows (on the left) a sample of points drawn from a population whose x and y coordinates have a normal distribution (truncated by the plot window) and whose values are correlated — in other words they are sampled from a bivariate gaussian distribution.

The histogram (on the right) shows the distribution of y values. The values are the number of points for all x in the given y bin.

The same data could be generated by sampling from two independent distributions and then rotating the coordinate system by 45 degrees (from u, v to x, y). This is the well–known algorithm I refer to in the introduction (the amount of rotation and the relative widths of the two populations changes the degree of correlation).

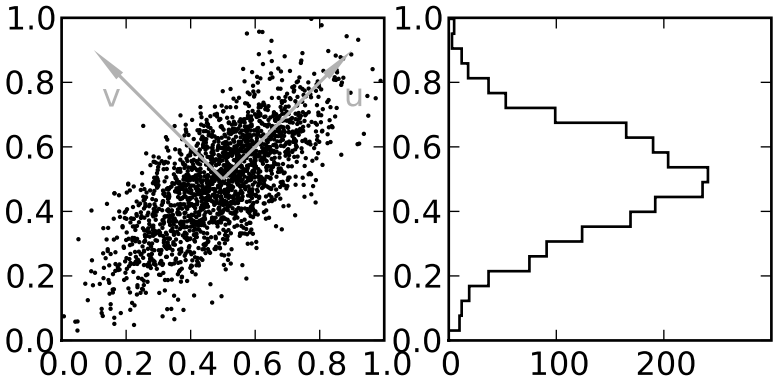

Figure 1: Points sampled from a bivariate gaussian whose prinicipal axes (u, v) are not parallel to x, y give correlated, normally distributed variates. The histogram shows the y distribution.

3 Uniform Variates — Blending Extremes

Unfortunately the "trick" used for normal variates — a coordinate transformation from a "diagonal" distribution — does not work for uniform distributions. For uniformly distributed points in two dimension only the two extreme cases, no– and complete–correlation, give uniform one–dimensional variates (figure 2).

But the two extremes are all we need. Combining samples from the two gives one-dimensional variates that remain uniform, but which have an intermediate degree of correlation.

For example, to generate data with an r (correlation coefficient) of 0.7, take 70% of points from a line ($r = 1$) and 30% from a uniform rectangle ($r == 0$). See figure 3.

4 Uniform Variates — Banding

4.1 The Problem

In section 3 I combined samples from two "completely different" populations and then took the x and y coordinates to generate two uniform, correlated variates.

There is, perhaps, something "ugly" or "unnatural" about this approach — the scatterplot in figure 3 is a "Frankenstein" combination of two distributions that are "quite distinct".

Whether or not the objections above are valid there is also a practical concern — the correlated variables may be used to simulate a process for which the joint distribution is unsuitable. Perhaps the data are displayed

*andrew@acooke.org

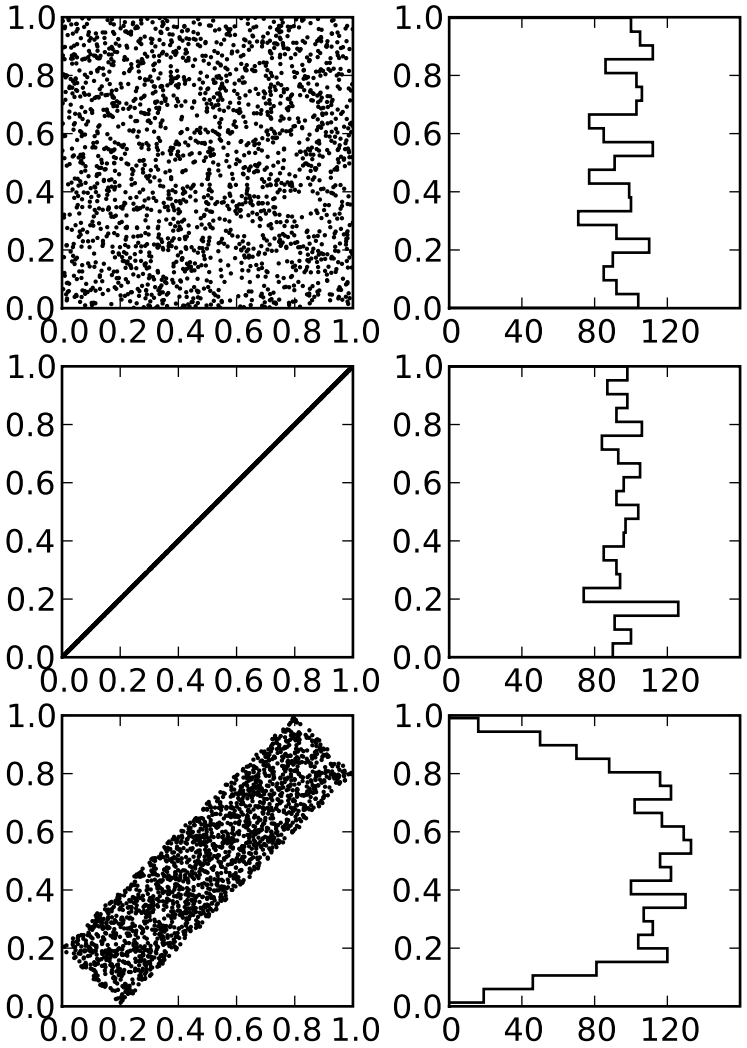

Figure 2: The y value (and, by symmetry, x) from a uniformly filled rectangle, or a uniformly sampled diagonal line, is uniformly distributed. But in the case of a diagonal rectangle, it is not (compare with the normal distribution in figure 1).

in a way that clearly shows the two "underlying" distributions, giving a

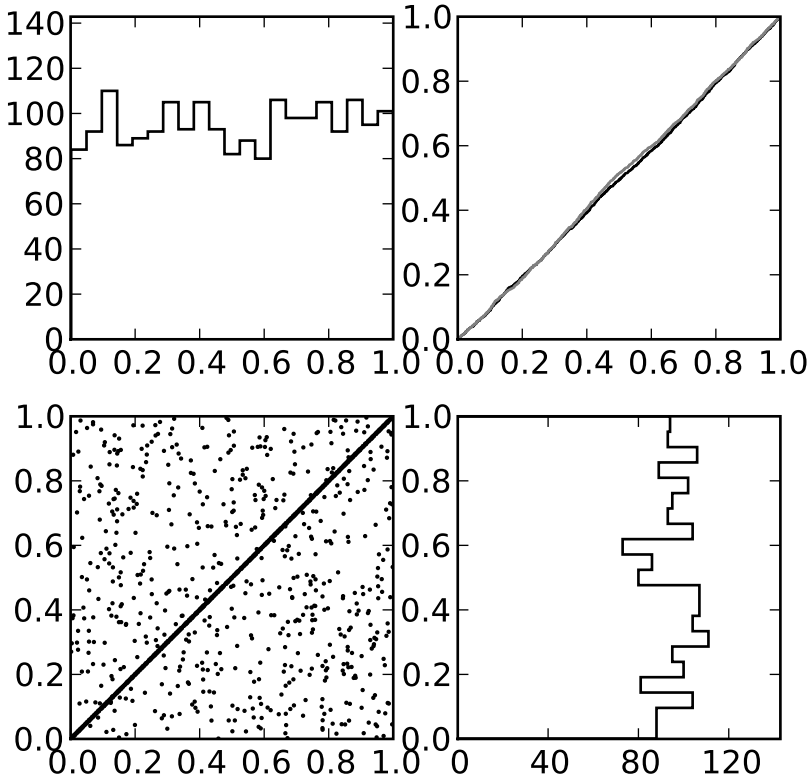

Figure 3: 2000 points sampled from two different populations (bottom left). Each sample is taken from a completely correlated population with a probability of 0.7 and from a completely uncorrelated population with a probability of 0.3. The result, projected onto the x (top left) and y (bottom right) axes, are two variates with $r = 0.7$. The top right plot shows the cumulative distribution for the two variates.

misleading "feature".

If the second problem is relevant then there must be additional constraints, in addition to "uniform and correlated". Without knowing what these constraints are I cannot address them here in detail, but I can still illustrate how they might be managed by a second algorithm.

4.2 A Second Approach

A more "natural" approach to generating "uniform, correlated" variates might be based on the banded distribution shown in figure 2. That does not work as it stands, but perhaps it can be modified in some way?

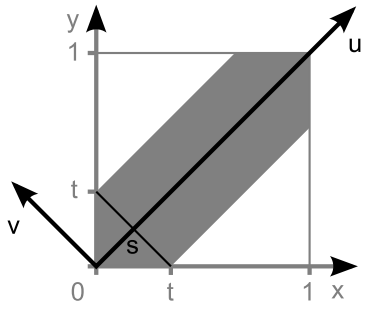

Figure 4: Axes and points used in the calculation below.

The problem with the rectangle approach was too few points from the "end regions". If the rectangle is extended as shown in figure 4 the result is closer, but even that is insufficient.

Rather than continuing to guess I will parameterize the density distribution and then solve for that, given the constraint that the x and y coordinates must be uniformly distributed. To obtain both variates from the same calculation (which reduces the work involved considerably) the distribution must be a function of u only. Furthermore, I will focus only on the bottom left corner ($u < s$) and assume that for $u > s$ the desity (ρ) is a constant, ρ_0 (the top right corner can be inferred from symmetry later).

Given the above, for some x, where $x < t$, the constraint is:

$$\int_{y=0}^{t+x} \rho(y)\, dy = 2t\rho_0$$

And for $y > t - x$, $\rho(y) = \rho_0$, so:

$$\int_{y=0}^{t-x} \rho(y)\, dy + 2x\rho_0 = 2t\rho_0$$

Transforming variables:

$$\int_{u=x/\sqrt{2}}^{t/\sqrt{2}} \rho(u)\sqrt{2}\, du + 2x\rho_0 = 2t\rho_0$$

If the solution is a polynomial, $\rho(u) = \sum_{i=0}^{n} a_i u^i$, then:

$$\left[\sum_{i=0}^{n} a_i \frac{u^{i+1}}{i+1}\right]_{u=x/\sqrt{2}}^{t/\sqrt{2}} = \sqrt{2}\rho_0(t-x)$$

and comparing coefficients of x^i it is clear that $a_i = 0$ when $i > 0$, leaving:

$$
\begin{aligned}
a_0 \frac{t-x}{\sqrt{2}} &= \sqrt{2}\rho_0(t-x) \\
a_0 &= 2
\end{aligned}
$$

which is so simple that you can see it "geometricaally" by considering similar triangles in figure 5.

4.3 Summary

The x and y coordinates of points sampled at random from the density distribution shown in figure 5 are uniformly distributed and correlated (the degree of correlation depends on the value of t; see the next section).

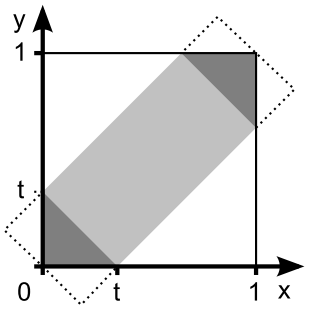

Figure 5: A density distribution for generating uniform, correlated, random variates. The darker area has twice the density of the lighter area.

This can be achieved by selecting points at random within the extended rectangle indicated by the dotted lines (using a uniform density) and then "folding" those that lie outside the unit square into the more darkly shaded areas.

Example data for 2000 points when $t = 0.3$ are shown in figure 6.

4.4 Correlation Coefficient

Figure 7 plots r against t for simulated data using 10,000 samples per measurement. The grey line is the curve $r = 1-2t^2+t^3$. This is *not derived from the measurements in a rigorous manner*, but appears to adequately model the relationship.

5 Discussion

In sections 3 and 4 I have shown two different methods to generate two samples of uniformly distributed values that are correlated.

The second approach also illustrated a more general approach to the problem: normalisation of a suitably symmetric distribution function. There is nothing "special" about the banded form used there, except that

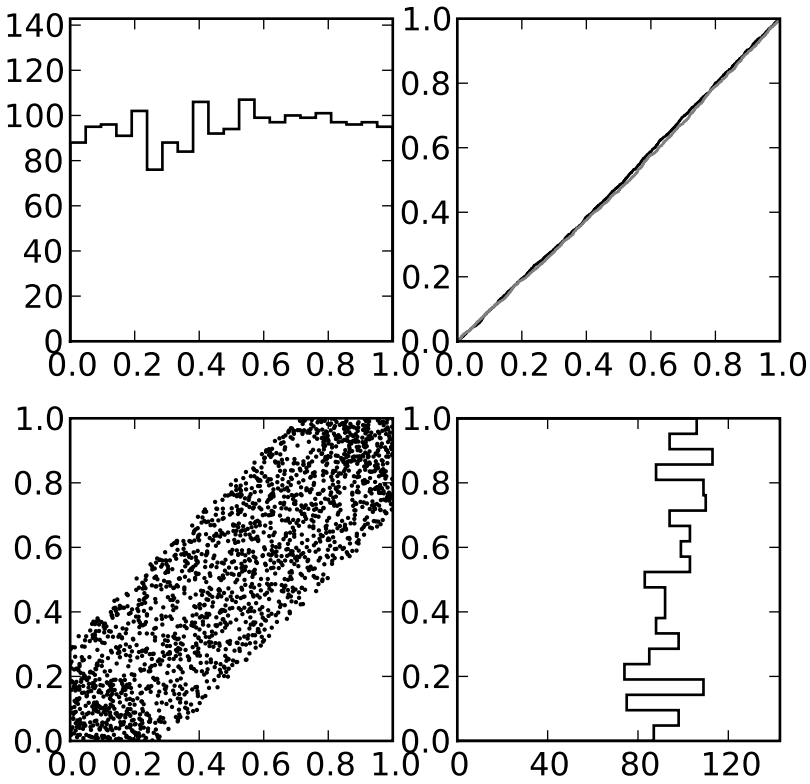

Figure 6: 2000 points (bottom left) taken from the "banded" distribution described in section 4.3 for the case where $t = 0.3$. The histograms show the distribution of x (top left) and y (bottom right) coordinates; their cumulative distributions are shown top right.

it simplified the calculations. A variety of other distributions could be used instead.

Neither example, however, is as "natural" as the approach for normally distributed variates. In that case (section 2) the initial and final distributions shared the same basic form. Perhaps there exists a shape within the unit square which, when sampled with uniform density, generates uniform variates? I cannot see any argument against this, but nor can I see a simple way to derive that form.

(It would be nice if someone could derive the relationship between r and t that is suggested by figure 7.)

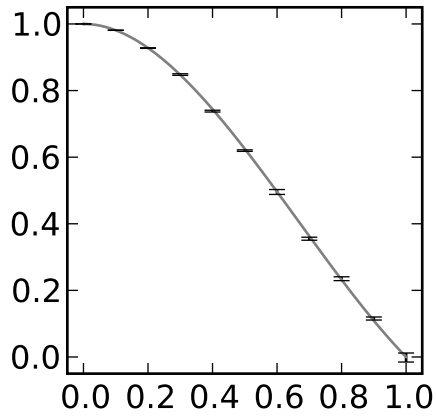

Figure 7: r (correlation coefficient, y axis) against t (x axis) for simulated data sets of 10,000 points each. The grey line is $r = 1 - 2t^2 + t^3$

6 Acknowledgements

This work was inspired by a question posted by "Gideon" on Stack Overflow[1].

References

[1] Stack Overflow — Generating Correlated Numbers.
http://stackoverflow.com/questions/1717907