NISS

Sampling from a Bivariate Distribution with Known Marginals

Vincent Granville

Technical Report Number 47
August, 1996

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Sampling from a bivariate distribution with known marginals

VINCENT GRANVILLE*

August 27, 1996

Abstract

We propose a general algorithm to sample from a bivariate distribution with given marginals and arbitrary dependence structure. We mainly focus on positive random variables with a highly negative coefficient of correlation.

Key words: order statistics, exponential distribution, minimum correlation, resampling.

1 Introduction

We consider the problem of sampling from a joint bivariate distribution, assuming the marginals are known. Bivariate models with known marginals have already been thoroughly investigated e.g. in the framework of extreme value theory [15],[16], but essentialy for positive random variables with a positive correlation. Our approach allows in particular to deal easily with positive random variables negatively correlated.

For this purpose, we have designed an algorithm which incorporates three parameters: an integer $n \geq 2$, a permutation σ on $\{1, \cdots, n\}$, and a parameter p. Both n, p and σ determine the correlation between the two components of the sampled bivariate random vector. The p.d.f. we are sampling from will thus be parametrized by n, p and σ. When $p = 1$, independence between the two components is guaranteed. When $n = 2$, our distribution corresponds to the Farlie-Gumbel-Morgenstern model [10].

We shall be interested in the behaviour of the underlying joint distribution when $n \to \infty$. In particular, we show how to sample from a bivariate distribution either with a minimum or maximum coefficient of correlation ρ. We also prove that if the marginals are positive variables, then the lower bound for ρ may be greater than -1. We compute explicitly this lower bound.

2 Algorithm

Although there is an extensive literature on simulation [5], only a few authors consider the problem of sampling from a bivariate distribution with given marginals [4],[11].

Let f_X, f_Y be univariate p.d.f. Here, we propose an algorithm to sample from a joint p.d.f. with given marginals f_X and f_Y. We proceed as follows. Let $n \geq 2$, let σ be a permutation on $\{1, \cdots, n\}$, and let $0 \leq p \leq 1$.

1. **Sample independent deviates $X_1, \cdots, X_n$ from f_X.**

2. **Sample independent deviates $Y_1, \cdots, Y_n$ from f_Y.**

*National Institute of Statistical Sciences, Research Triangle Park, NC, USA.

3. **Let $X_{(i)}, Y_{(j)}$, $1 \leq i,j \leq n$, denote the order statistics. Then the final bivariate sampled vector is (U,V) with**

$$(U,V) = \begin{cases} (X_1, Y_1) & \textbf{with probability } p \\ (X_{(1)}, Y_{(\sigma(1))}) & \textbf{with probability } (1-p)/n \\ (X_{(2)}, Y_{(\sigma(2))}) & \textbf{with probability } (1-p)/n \\ \vdots & \\ (X_{(n)}, Y_{(\sigma(n))}) & \textbf{with probability } (1-p)/n \end{cases}$$

This algorithm can easily be generalized to the fully multivariate case (see section 4), but here we only focus on the bivariate case.

As $n \to \infty$, using appropriate permutations σ, we can sample from any arbitrary joint p.d.f. with fixed marginals f_X and f_Y. It is also clear that if $p = 1$, then U and V are independent. Furthermore, minimizing or maximizing $\rho(U,V)$ will be achieved with $p = 0$. To proceed any further, we need two lemmas.

Lemma 2.1 *Let n be fixed , and let $x_1, \cdots, x_n$ and $y_1, \cdots, y_n$ be two sequences of arbitrary real numbers. Then, the optimum of the expression $x_{(1)}y_{(\sigma(1))} + \cdots + x_{(n)}y_{(\sigma(n))}$ is attained by the following permutations: $\sigma(i) = n - i + 1$ for the minimum, and $\sigma(i) = i$ for the maximum.*

Lemma 2.2 *The joint theoretical distribution we are sampling from, using the previous algorithm, is given by the following mixture:*

$$P(U < u, V < v) = pP(X_1 < u)P(Y_1 < v) + \frac{1-p}{n}\sum_{i=1}^{n} P(X_{(i)} < u)P(Y_{(\sigma(i))} < v) \qquad (1)$$

Hence,

$$\begin{aligned} E(UV) &= pE(X_1)E(X_2) + \frac{1-p}{n}E\Big[\sum_{i=1}^{n} X_{(i)}Y_{(\sigma(i))}\Big] && (2) \\ &= pE(X_1)E(X_2) + \frac{1-p}{n}\sum_{i=1}^{n} E(X_{(i)})E(Y_{(\sigma(i))}) && (3) \end{aligned}$$

and therefore the correlation between U and V is equal to

$$\rho = -\frac{1-p}{\sqrt{\mathrm{Var}(X_1)\mathrm{Var}(Y_1)}}\Big\{E(X_1)E(Y_1) - \frac{1}{n}\sum_{i=1}^{n} E(X_{(i)})E(Y_{(\sigma(i))})\Big\} \qquad (4)$$

From lemma 2.1 combined with (2) and (3), we find that if n is fixed, the correlation is minimum and denoted as ρ_{min} (resp. maximum and denoted as ρ_{max}) if $\sigma(i) = n - i + 1$ (resp. $\sigma(i) = i$), $i = 1, \cdots, n$. From now on, we shall only consider these two permutations.

3 Asymptotics

The asymptotic distributions for $X_{(i)}$ and $Y_{(\sigma(i))}$ are normal, see [3], page 469. Here, we are interested in looking at ρ_{min} and ρ_{max} as $n \to \infty$. We have ([3], page 469):

$$E(X_{(i)}) \approx F_X^{-1}\Big(\frac{i}{n+1}\Big), \qquad E(Y_{(\sigma(i))}) \approx F_Y^{-1}\Big(\frac{\sigma(i)}{n+1}\Big),$$

where F_X, F_Y are the c.d.f. associated with the marginals f_X, f_Y. As $n \to \infty$, assuming F_X and F_Y have finite first and second order moments, we finally find:

$$\rho_{\text{min}} \quad \to \quad -\frac{1-p}{\sqrt{\text{Var}(X_1)\text{Var}(Y_1)}}\Big\{E(X_1)E(Y_1) - \int_0^1 F_X^{-1}(x)F_Y^{-1}(1-x)dx\Big\}, \tag{5}$$

$$\rho_{\text{max}} \quad \to \quad -\frac{1-p}{\sqrt{\text{Var}(X_1)\text{Var}(Y_1)}}\Big\{E(X_1)E(Y_1) - \int_0^1 F_X^{-1}(x)F_Y^{-1}(x)dx\Big\}. \tag{6}$$

The condition for (5) or (6) to hold is that the Riemann integral in the right hand side exists.

Now, assume that $p = 0$. In particular, if $F_X = F_Y$, with the change of variable $x = F_X(y)$ in (6), we find that $\rho_{\text{max}} \to 1$. Also, if F_X and F_Y are nonidentical exponential distributions, then $\rho_{\text{max}} \to 1$. If $F_X = F_Y$ and F_X is a symmetric distribution, then $\rho_{\text{min}} \to -1$, and this can be proved using the change of variable $x = F_X(y)$ in (5) together with the fact that for a symmetric distribution, $F_Y^{-1}(1-x) = 2E(Y) - F_Y^{-1}(x)$. For positive random variables, the lower bound ρ_{min} may be greater than -1. In particular, we get this surprising result:

Theorem 3.1 *If U and V are correlated random variables with exponential marginal distributions and possibly different marginal means, then*

$$\rho(U,V) \geq -1 + \int_0^1 \log(x)\log(1-x)dx = 1 - \frac{\pi^2}{6} \approx -0.644.$$

The lower bound in theorem 3.1 can be attained, see section 5.1. Note that positive correlated random variables have received some attention in the literature during the last 30 years at least. The reader is referred to [6],[7],[10],[11],[15],[16]. See also [2],[9],[14] for interesting material on multivariate exponential distributions.

4 Trivariate distributions

The algorithm easily generalizes to the d-dimensional case. If $d = 3$, we need two permutations σ, τ and the algorithm is as follows, with a straightfoward extension of notations:

1. **Sample independent deviates $X_1, \cdots, X_n$ from f_X.**

2. **Sample independent deviates $Y_1, \cdots, Y_n$ from f_Y.**

3. **Sample independent deviates $Z_1, \cdots, Z_n$ from f_Z.**

4. **Let $X_{(i)}, Y_{(j)}, Y_{(k)}$, $1 \leq i,j,k \leq n$, denote the order statistics. Then the final trivariate sampled vector is (U,V,W) with**

$$(U,V,W) = \left\{ \begin{array}{ll} (X_1, Y_1, Z_1) & \textbf{with probability } p \\ (X_{(1)}, Y_{(\sigma(1))}, Z_{(\tau(1))}) & \textbf{with probability } (1-p)/n \\ (X_{(2)}, Y_{(\sigma(2))}, Z_{(\tau(2))}) & \textbf{with probability } (1-p)/n \\ \vdots & \\ (X_{(n)}, Y_{(\sigma(n))}, Z_{(\tau(n))}) & \textbf{with probability } (1-p)/n \end{array} \right.$$

Lemma 4.1 *The joint theoretical distribution we are sampling from, using the previous algorithm, is given by the following mixture:*

$$\begin{aligned} P(U<u, V<v, W<w) &= pP(X_1<u)P(Y_1<v)P(Z_1<w) + \\ &\qquad \frac{1-p}{n}\sum_{i=1}^{n} P(X_{(i)}<u)P(Y_{(\sigma(i))}<v))P(Z_{(\tau(i))}<w) \end{aligned}$$

5 Applications

We first assess our algorithm on simulated data, then we provide an example based on true data. An important case corresponds to $n = 2$, with a minimum correlation. In that case, the joint distribution we are sampling from has a simple c.d.f.:

$$F(u, v) = F_X(u)F_Y(v)\{p + (1 - p)(F_X(u) - F_X(u)F_Y(v) + F_Y(v))\}.$$

This is just a Farlie-Gumbel-Morgenstern system of distributions [10]. Using the notation $\bar{F}_X = 1 - F_X, \bar{F}_Y = 1 - F_Y$, we can still write:

$$F(u, v) = F_X(u)F_Y(v)\{1 - (1 - p)\bar{F}_X(u)\bar{F}_Y(v)\}, \tag{7}$$

and now the analogy with formula (1) in [10] is straightfoward. Similarly, if $n = 2$ and if the correlation is maximum, we find

$$F(u, v) = F_X(u)F_Y(v)\{1 + (1 - p)\bar{F}_X(u)\bar{F}_Y(v)\}, \tag{8}$$

corresponding to the other case of the Farlie-Gumbel-Morgenstern system of distributions.

For $n > 2$, our family of bivariate distributions does not fall within the class of the iterated Farlie-Gumbel-Morgenstern distributions. Let us consider the case $n = 3$ with minimum correlation and exponential marginals. Table 1 shows that this correlation is less than the minimum correlation obtained with $n = 2$. If our bivariate distribution is of iterated Farlie-Gumbel-Morgenstern type, then it is clear that $n = 3$ should correspond to the single iteration of that model [10]. But Huang and Kotz have proved that the single iteration can not decrease the correlation of the noniterated (or ordinary Farlie-Gumbel-Morgenstern) case. Thus our distribution (with $n = 3$ and miminum correlation) does not fall into the class of distributions studied by Huang and Kotz. And remarquably, we decrease the minimum correlation by increasing n from 2 to 3. This contrasts with the results obtained for the single iteration model in [10].

5.1 Simulations

Only a few of the simulations we have performed are reported here. For instance, we have partly omitted to incorporate the tests which show good fit with the target marginal distributions. Also, the case when the random variables are positively correlated has been successfully handled, but the results are not reported here.

We only present the results connected with the most challenging problem: assessing the lower bound of -0.644 on practical simulations for the exponential case. Let us mention that in this case, the limiting distribution is degenerated, with $UV = 0$. Also note that with gamma marginals, we were able to obtain a coefficient of correlation as low as -0.75. All the simulations were performed on a basis of 20,000 generated deviates per test.

When $n = 2$, it is easy to prove that $\rho_{\text{min}} = -(1 - p)/4$ and $\rho_{\text{max}} = (1 - p)/4$ for exponential marginals. In table 1, the simulations are performed for exponential marginals of mean $E(X_1) = 1$ and $E(Y_1) = 0.5$. The parameter p is set to zero.

5.2 Real Data

Theorem 3.1 has a significant importance in some contexts. For instance, let us consider the stochastic process of storms and cells investigated by Rodriguez *et al* [12],[13] to model rainfall precipitations. Each cell has an exponential duration and an exponential depth. In

n	$E(U)$	$E(V)$	$\mathrm{Var}(U)$	$\mathrm{Var}(V)$	ρ
2	1.013	0.496	1.018	0.249	-0.253
3	0.996	0.501	0.971	0.248	-0.368
4	0.998	0.496	0.993	0.245	-0.427
10	1.004	0.504	1.017	0.259	-0.548
400	1.005	0.500	1.006	0.257	-0.640

Table 1: Simulations: computation of the moments and correlation of the sampled marginals.

	$E(U)$	$E(V)$	$E(UV)$	ρ	$E(Z)$	$\mathrm{Std}(Z)$
Data	0.69	2.44	1.34	-0.22	0.072	0.33
$n = 2, p = 1$	0.69	2.28	1.70	$+0.07$	0.082	0.59
$n = 2, p = 0.06$	0.67	2.49	1.29	-0.21	0.071	0.45

Table 2: Model fitting by comparing estimates of different moments for rainfall data. Note the dramatic improvement obtained with $n = 2$ and $p = 0.06$ over $n = 2$ and $p = 1$.

the original model, cell depths and cell durations are assumed to be independent. But in fact, this assumption is irrealistic, since heavy rains are likely to have a short duration and conversely. As a result, this model exhibits a lack of fit with true data. In order to overcome this deficiency, we are currently investigating a joint model for cell depths and cell durations, with exponential marginals and a negative coefficient of correlation. The theory developed here shows that this coefficient must necessarily be greater than -0.644, otherwise we would have to investigate e.g. gamma marginals instead of exponential marginals. Fortunately, in this example, $\rho \approx -0.22$ and thus we can use our algorithm even with $n = 2$. The dataset investigated here is studied in [8].

We denote cell depths by U and cell durations by V. Three sets of statistics are compared: the statistics computed on real data via Markov chain Monte-Carlo methods (this methodology allows us to recover hidden features of the process such as cell durations and cell depths), the statistics obtained under the assumption of independence (i.e. with $p = 1$), and then the statistics computed with $n = 2$ and $p = 0.06$. The statistics computed here are essentially moments of (U, V) (mean, correlation), but we have also considered the hourly rainfall aggregate Z as defined in [8]. The hourly aggregates are a complex function of U, V and other hidden features of the model developed in [8].

Table 2 shows a dramatic improvement with $n = 2$ and $p = 0.06$, over $p = 1$. Still, the standard deviation for Z is quite poor. There is about 460 cells, and each cell has a duration which depends on its parent storm. Since there is on average one cell per storm, we can not expect to get a high degree of precision in the statistics computed here. We should allow for an error equal to about 5%.

In a similar context, Bacchi *et al.* have investigated another bivariate exponential model with a negative correlation for modeling storm depths (also called *intensities* in their paper) and durations [1]. This model, namely

$$F(u,v) = 1 - \exp(-\alpha u) - \exp(-\beta v) + \exp(-\alpha u - \beta v - \alpha\beta\delta uv), \qquad 0 \leq \delta \leq 1 \qquad (9)$$

(see [9]), has a minimum correlation (when $\delta = 1$) equal to

$$\rho = -1 + \int_0^\infty \frac{\exp(-x)}{1+x} dx \approx -0.404. \qquad (10)$$

The authors were not aware of the fact that other bivariate exponential models with a negative correlation, such as the model derived from (7), were already available in the literature. In an application on a real dataset, they find that the correlation is close to the lower bound, $\rho = -0.404$, and some of the correlations they have computed and reported in table 2 in their paper are as low as -0.591. In such a case, it might be better to use a model which allows for highly negative correlations, rather than using (9). For instance, one might investigate our model with gamma marginals. Or our model with exponential marginals and $n = 4$, which has a minimum correlation equal to $\rho = -0.427$, as can be seen from table 1. It would be interesting to check whether our model could improve the results obtained in [1], if we consider a sufficiently large value for n.

6 Acknowledgment

This work has been financially supported by NSF-DMS grant # 9208758 offered to NISS and by an EPSRC grant while the author was at the Statistical Laboratory, Cambridge. I am also very grateful to R.L. Smith.

References

[1] **Bacchi B., Becciu G., Kottegoda N.T.:** Bivariate exponential model applied to intensities and durations of extreme rainfall. *J. of Hydrology*, 155(1994), 225-236.

[2] **Balakrishnan N.** *The Exponential Distribution: Theory, Methods and Applications.* Gordon and Breach, 1995.

[3] **Cox D.R., Hinkley D.V.:** *Theoretical Statistics.* Chapman and Hall, London, 1974.

[4] **Damien P., Müller P.:** A Bayesian bivariate failure time regression model. Technical Report 94-36, ISDS, Duke University, 1994.

[5] **Devroye L.** *Non uniform random variate generation.* Springer-Verlag, New York, 1986.

[6] **Ferguson T.S.:** A class of symmetric bivariate uniform distributions. Technical Report 07-08-94, Dept. of Stat., UCLA, 1994.

[7] **Galambos J.:** *The Asymptotic Theory of Extreme Order Statistics.* R.E. Krieger Publishing Company, 1987.

[8] **Granville V., Smith R.L.:** Disaggregation of rainfall time series via Gibbs sampling. Working paper in progress, 1996.

[9] **Gumbel E.J.:** Bivariate exponential distributions. *J. American Stat. Assoc.*, 55(1960), 698–707.

[10] **Huang J.S., Kotz S.:** Correlation structure in iterated Farlie-Gumbel-Morgenstern distributions. *Biometrika*, 71(1984), 633–636.

[11] **Marshall A.W., Olkin I.:** Families of multivariate distributions. *J. American Stat. Assoc.*, 83(1988), 834–841.

[12] **Rodriguez-Iturbe I., Cox D.R., Isham V.** Some models for rainfall based on stochastic point processes. *Proc. Royal Soc. London A*, 410(1987), 269–288.

[13] **Rodriguez-Iturbe I., Cox D.R., Isham V.** A point process model for rainfall: further developments. *Proc. Royal Soc. London A*, 417(1988), 283–298.

[14] **Sarkar S.K.:** A continuous bivariate exponential distribution. *J. American Stat. Assoc.*, 82(1987), 667–675.

[15] **Smith R.L.:** Multivariate threshold methods. *In*: Extreme Value Theory and Applications, J.Galambos (Ed.), Kluwer Academic Publishers, 1994.

[16] **Tiago de Oliveira J.:** Bivariate models for extremes; statistical decision. *In*: Statistical Extremes and Applications, J. Tiago de Oliveira (Ed.), D. Reidel Publishers, Dordrecht, 1984.