Jouko Vilmunen – Peter Palmroos

Closed form solution of correlation in doubly truncated or censored sample of bivariate log-normal distribution

EUROJÄRJESTELMÄ
EUROSYSTEMET

Bank of Finland Research
Discussion Papers
17 • 2013

Closed Form Solution of Correlation in Doubly Truncated or Censored Sample of Bivariate Log-Normal Distribution

Jouko Vilmunen

Bank of Finland, PO BOX 160, 00101 Helsinki, Finland

Tel.: +358-10-831-2594

Peter Palmroos*

Financial Supervisory Authority, PO BOX 103 00101 Helsinki, Finland

Tel.: +358-10-831-5356

Fax: +358-10-831-5238

Abstract

In this study we present a closed form solution to the moments and, in particular, correlation of two log-normally distributed random variables, where the underlying log-normal distribution is potentially truncated and censored at both tails. Throughout the analysis we further assume that the parameters of the unconstrained bivariate log-normal distribution are known. The closed form solution also covers the cases where one tail is truncated and the other is censored.

Keywords: Bivariate log-normal distribution, Pearson's product-moment correlation, Truncated, Censored, Tail correlation, Solvency II

JEL Classification C18, C46, G28

*Corresponding author

Email addresses: jouko.vilmunen@bof.fi (Jouko Vilmunen), peter.palmroos@bof.fi (Peter Palmroos)

1. Introduction

Log-normality is a common assumption that researchers, both in banking and actuarial applications, impose in models of risk measurement. Furthermore, these models almost invariably use the Pearson's product-moment correlation to measure linear dependence between the variables under scrutiny, even in cases where normality of the variables cannot be assumed.

The prudential requirements and modelling options embedded in the Solvency II and Basel II frameworks, which are applied in regulating insurance and banking sectors, have greatly increased the interest in models pertaining to dependencies between extreme or tail events as well as in measures that capture these dependencies in different samples. Furthermore, truncated and/or censored samples are common in follow-up and survey studies in e.g. economics and biomedicine, as well as in actuarial applications.

The validity or otherwise of log-linearity is an issue that can be debated on several grounds, but even if one can justify its use, the possibility of non-linear dependencies between log-normally distributed variables complicates the interpretation of linear correlation measures like (estimates of) the Pearson's correlation coefficient. Moreover if the distribution of one of the variables is truncated or censored, the correlation coefficient may not be the most natural choice for a measure of dependence and, indeed, it may be very difficult to interpret.

Moments and correlation of variables from truncated or censored bivariate normal and log-normal distributions have been the subject of previous studies. Johnson and Kotz (1972) and Lien (1985) for example have derived closed form solutions to the asymptotic distributions of the parameters from single truncated samples. Kotz et al. (2000) has derived the solution to the moments and correlation for random variables from a truncated bivariate normal distribution. Among others, Zhao et al. (2011) lists some papers that apply left truncation and right censoring (LTRC) in the models these papers study.

However the general closed form solution for moments, covariance and correlation of random variables from the constrained bivariate log-normal distribution has not been presented.

In this paper we derive a closed form solution to the moments and correlation of two log-normally distributed random variables, when the bivariate log-normal distribution is potentially once or twice truncated or censored. We analyze the case where the parameters of the bivariate log-normal dis-

tribution are known. Our method of computing the correlation covers the particular case, when one tail of the log-normal distribution is truncated and the other is censored. Our method also makes it possible to calculate bounds on correlations, which has the benefit that it clarifies the interpretation of estimated correlations in observed bounded[1] samples.

The next section derives the closed form solution and section three gives examples of the kind of effects truncation and censoring may have on the correlation between two random variables.

2. The closed form solution

Assume that the random vector $\underline{u} = (x, y)^T = (\ln X, \ln Y)^T$ is jointly normally distributed with a known mean vector and variance-covariance matrix:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \ln X \\ \ln Y \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \sigma_x \sigma_y \rho \\ \sigma_x \sigma_y \rho & \sigma_y^2 \end{pmatrix} \right) \tag{1}$$

Using the Cholesky decomposition of variance-covariance matrix, the components of the random vector $\underline{u}$ can be represented as

$$\begin{aligned} x &= \mu_x + \sigma_x z_x, \ z_x \sim N(0,1) \\ y &= \mu_y + \rho \sigma_y z_x + \sigma_y \sqrt{1-\rho^2} z_y, \ z_y \sim N(0,1), \ z_x \perp z_y \end{aligned} \tag{2}$$

Pearson's (conditional) correlation coefficient between x and y, conditional on bounds imposed on the distribution of X is defined in the usual way as

$$\rho_{Q,D} = \frac{cov(X, Y|Q, D)}{\sqrt{Var(X|Q, D)}\sqrt{Var(Y|Q, D)}} \tag{3}$$

Here Q is defined as $Q = (L, U)$, where L and U denote, respectively, the value of the lower and the upper boundary of X, and the indicator variable $D = (D_L, D_U)$, where $D = (0, 0)$ for the truncated lower and upper boundary L and U, $D = (0, 1)$ for truncated lower boundary and censored upper boudary L and U and so on.

[1] To make the text more readable, bounded sample will be used as a synonym to all the doubly of single truncated or censored samples. Analogously, 'bounded variable' is used to refer to a random variable from a truncated or censored distribution.

The first moments of the distribution of the truncated or censored variable X are equal in the bivariate and in the one dimensional cases. The solution to the doubly truncated single variable sample can be found, for example, in the paper by Bebu and Mathew (2009).

In the case of a doubly censored or a censored and truncated distribution we need the portions of observations belonging in to the censored tails.These portions are marked in the following formulas as T_L and T_U.

$$\begin{aligned} T_L &= \frac{D_L\Phi(L_1)}{D_L\Phi(L_1) + [\Phi(U_1) - \Phi(L_1)] + D_U\left[1 - \Phi(U_1)\right]} \\ T_U &= \frac{D_U\left[1 - \Phi(U_1)\right]}{D_L\Phi(L_1) + [\Phi(U_1) - \Phi(L_1)] + D_U\left[1 - \Phi(U_1)\right]} \end{aligned} \tag{4}$$

where Φ denotes, as usual, the cumulative distribution function of the standardised normal distribution.

Now the variance of X, conditional on the bounds, can be calculated via the usual decomposition

$$Var(X|Q,D) = E(X^2|Q,D) - \left[E(X|Q,D)\right]^2 \tag{5}$$

The expected values of the first moments of the bounded X, $E(X^k|Q,D)$ with $(k = 1, 2)$, can be expressed as

$$E(X^k|Q,D) = T_L L^k + (1 - T_L - T_U)\left\{e^{k\mu_x + k^2\frac{\sigma_x^2}{2}}\left[\frac{\Phi\left(U_{2,k}\right) - \Phi\left(L_{2,k}\right)}{\Phi\left(U_1\right) - \Phi\left(L_1\right)}\right]\right\} + T_U U^k \tag{6}$$

where

$$\begin{aligned} U_1 &= \frac{\ln U - \mu_x}{\sigma_x};\ L_1 = \frac{\ln L - \mu_x}{\sigma_x} \\ U_{2,k} &= \frac{\ln U - (\mu_x + k\sigma_x^2)}{\sigma_x};\ L_{2,k} = \frac{\ln L - (\mu_x + k\sigma_x^2)}{\sigma_x} \end{aligned} \tag{7}$$

The solution to the first k moments of conditional Y, on the other hand, takes the following form

$$
\begin{aligned}
E(Y^k|Q,D) &= e^{k\mu_y+k^2\frac{\sigma_y^2}{2}} \\
&\quad \cdot\left\{\frac{[D_U\Phi(\infty)+(1-D_U)\Phi(U_{3,k})]-[D_L\Phi(-\infty)+(1-D_L)\Phi(L_{3,k})]}{[D_U\Phi(\infty)+(1-D_U)\Phi(U_1)]-[D_L\Phi(-\infty)+(1-D_L)\Phi(L_1)]}\right\} \\
&= e^{k\mu_y+k^2\frac{\sigma_y^2}{2}}\left\{\frac{[D_U+(1-D_U)\Phi(U_{3,k})]-[(1-D_L)\Phi(L_{3,k})]}{[D_U+(1-D_U)\Phi(U_1)]-[(1-D_L)\Phi(L_1)]}\right\} \qquad (8)
\end{aligned}
$$

where

$$
U_{3,k} = \frac{\ln U-(\mu_x+k\rho\sigma_x\sigma_y)}{\sigma_x};\ L_{3,k}=\frac{\ln L-(\mu_x+k\rho\sigma_x\sigma_y)}{\sigma_x} \qquad (9)
$$

Finally the variance of Y can be calculated as in formula (5).

If the correlation between X and Y differs from the zero, then the correlation between $X|Q,D$ and $Y|Q,D$ also differs from zero as long as $L\neq U$. Since the covariance between the two random variables can be calculated using the formula

$$
Cov(X,Y|Q,D)=E(XY|Q,D)-E(X|Q,D)E(Y|Q,D) \qquad (10)
$$

the only missing term consequently is $E(XY|Q,D)$. This too can be obtained from the moment generating function of the bivariate normal as follows

$$
\begin{aligned}
E(XY|Q,D) &= T_LLe^{\mu_y+\frac{1}{2}\sigma_y^2}\left[\frac{\Phi(L_{3,1})-\Phi(-\infty)}{\Phi(L_1)-\Phi(-\infty)}\right]+T_UUe^{\mu_y+\frac{1}{2}\sigma_y^2}\left[\frac{\Phi(\infty)-\Phi(U_{3,1})}{\Phi(\infty)-\Phi(U_1)}\right] \\
&\quad +\left\{(1-T_L-T_U)e^{\mu_x+\mu_y+\frac{1}{2}\left[(\sigma_x+\rho\sigma_y)^2+\sigma_y^2\left(1-\rho^2\right)\right]}\left[\frac{\Phi\left(U_4\right)-\Phi\left(L_4\right)}{\Phi\left(U_1\right)-\Phi\left(L_1\right)}\right]\right\} \\
&= T_LLe^{\mu_y+\frac{1}{2}\sigma_y^2}\left[\frac{\Phi(L_{3,1})}{\Phi(L_1)}\right]+T_UUe^{\mu_y+\frac{1}{2}\sigma_y^2}\left[\frac{1-\Phi(U_{3,1})}{1-\Phi(U_1)}\right] \\
&\quad +\left\{(1-T_L-T_U)e^{\mu_x+\mu_y+\frac{1}{2}\left[(\sigma_x+\rho\sigma_y)^2+\sigma_y^2\left(1-\rho^2\right)\right]}\left[\frac{\Phi\left(U_4\right)-\Phi\left(L_4\right)}{\Phi\left(U_1\right)-\Phi\left(L_1\right)}\right]\right\} \\
&= e^{\mu_y+\frac{1}{2}\sigma_y^2}\left\{T_LL\left[\frac{\Phi(L_{3,1})}{\Phi(L_1)}\right]+T_UU\left[\frac{1-\Phi(U_{3,1})}{1-\Phi(U_1)}\right]\right\} \qquad (11) \\
&\quad +\left\{(1-T_L-T_U)e^{\mu_x+\mu_y+\frac{1}{2}\left[(\sigma_x+\rho\sigma_y)^2+\sigma_y^2\left(1-\rho^2\right)\right]}\left[\frac{\Phi\left(U_4\right)-\Phi\left(L_4\right)}{\Phi\left(U_1\right)-\Phi\left(L_1\right)}\right]\right\}
\end{aligned}
$$

Table 1: Population and asymptotic sample correlations with minimum and maximum boundaries

Population / Sample	Asymptotic Correlation	Asymptotic Minimum	Asymptotic Maximum
Bivariate normal population (x, y)	-0.791	-1.000	1.000
Bivariate log-normal population (X, Y)	-0.700	-0.882	0.909
Doubly truncated sample	-0.448	-0.989	0.991
Doubly censored sample	-0.500	-0.595	0.907
Left truncated and right censored sample	-0.529	-0.990	0.847
Left censored and right truncated sample	-0.467	-0.569	0.954

When $ln(X) \sim N(0.094, 0.045), ln(Y) \sim N(-0.117, 0.652), T_L = 50\%$ and $T_U = 5\%$

where

$$U_4 = \frac{\ln U - (\mu_x + \sigma_x^2 + \rho\sigma_x\sigma_y)}{\sigma_x};\ L_4 = \frac{\ln L - (\mu_x + \sigma_x^2 + \rho\sigma_x\sigma_y)}{\sigma_x} \quad (12)$$

At this stage we have all the terms needed in the correlation equation (3), and the asymptotic sample correlation can be calculated.

It should be noticed that when $L \to 0$ and $U \to \infty$, then

$$\rho_{Q,D} \underset{\substack{U\to\infty \\ L\to 0}}{\longrightarrow} \rho_{XY}$$

This is shown in Appendix A.

3. The numerical examples

In addition to population parameters, the truncation and censoring points also affect to the observed sample correlation. Depending on the portions that have been truncated or censored, the difference between the sample and the population correlations can be remarkably large. Lien (1985), for example, reports the difference between the truncated tail correlations and correlation of the population using the bivariate log-normal distribution.

Table 1 presents the population and asymptotic sample correlations and asymptotic minimum and maximum correlations in all four combinations of double truncation and censoring. As can be seen from the figures, the difference between sample and population correlations can be considerable.

In the case of a zero population correlation, the sample correlation is also zero in all bounded samples. This can be proved using formula 10.

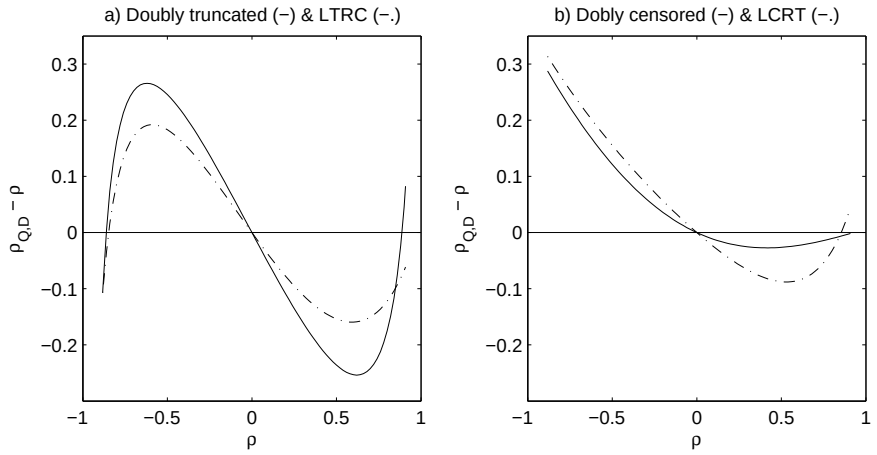

Figure 1: Differense between a) the population correlation and the correlation of the doubly truncated and the LTRC samples and b) the population correlation and the correlation of the doubly censored and the LCRT samples as a function of population correlation

When the population correlation differs from zero, the difference between unconstrained population correlation and bounded correlation can be of either sign and can be verified via calculation only after the parameters of the bounded distribution are fixed. Consequently, the asymptotic bounded sample correlation can be smaller or larger than the correlation of the unconstrained population. This is true for all combinations of truncation or censoring.

Figures 1a and 1b present the difference between the correlation coefficient of the unconstrained and bounded distribution ($\rho_{Q,D} - \rho$) as a function of ρ, when the other parameters are as in Table 1. If the absolute value of the bounded correlation is smaller than the population correlation, the line will stay above the horizontal axis when $\rho < 0$ and under when $\rho > 0$.

Thus researchers should pay attention to the results derived in applications, where samples from the underlying bounded distribution are analyzed. For example caution should be exercised when performing and interpreting simulations with truncated or censored multivariate distributions or when estimating correlation matrices from truncated or censored data. Otherwise one could well run the risk of being misled and drawing wrong conclusions. The effects on correlation of bounds in the underlying distribution should also be duly controlled when, for example due to the small sample, the correlation structure cannot be estimated and must be evaluated using expert

opinion.

References

Bebu, I., Mathew, T. (2009) Confidence intervals for limited moments and truncated moments in normal and lognormal models. Statistics and Probability Letters, 79, 375-380.

Johnson, N.L., Kotz, S. (1972) Distributions in Statistics: Continuous Multivariate Distributions. New York, John Wiley & Sons, Inc.

Kotz, S., Balakrishnan, N., Johnson, N.L. (2000) Continuous Multivariate Distributions, Volume 1: Models and Applications. New York, John Wiley & Sons, Inc.

Lien, D-H.D. (1985) Moments of truncated bivariate log-normal distributions. Economics Letters,19, 243-247.

Zhao, M., Bai, F., Zhou, Y. (2011) Relative deficiency of quantile estimators for left truncated and right censored data. Statistics and Probability Letters, 81, 1725-1732.

Appendix A. Correlation when $L \to 0$ and $U \to \infty$

Now

$$\begin{aligned}
Var\left(X|Q\right) &= E\left(X^2|Q\right) - \left[E\left(X|Q\right)\right]^2 \\
&= e^{2\mu_x+2\sigma_x^2}\left[\frac{\Phi\left(U_3\right)-\Phi\left(L_3\right)}{\Phi\left(U_1\right)-\Phi\left(L_1\right)}\right] - e^{2\mu_x+\sigma_x^2}\left[\frac{\Phi\left(U_2\right)-\Phi\left(L_2\right)}{\Phi\left(U_1\right)-\Phi\left(L_1\right)}\right]^2 \\
&= e^{2\mu_x+\sigma_x^2}\left\{e^{\sigma_x^2}\left[\frac{\Phi\left(U_3\right)-\Phi\left(L_3\right)}{\Phi\left(U_1\right)-\Phi\left(L_1\right)}\right] - \left[\frac{\Phi\left(U_2\right)-\Phi\left(L_2\right)}{\Phi\left(U_1\right)-\Phi\left(L_1\right)}\right]^2\right\} \\
&\doteqdot e^{2\mu_x+\sigma_x^2}\Phi_2 \underset{\substack{U\to\infty \\ L\to 0}}{\longrightarrow} e^{2\mu_x+\sigma_x^2}\left(e^{\sigma_x^2}-1\right) \qquad \text{(A.1)} \\
Var\left(Y|Q\right) &= e^{2\mu_y+\sigma_y^2}\left\{e^{\sigma_y^2}\left[\frac{\Phi\left(U_6\right)-\Phi\left(L_6\right)}{\Phi\left(U_1\right)-\Phi\left(L_1\right)}\right] - \left[\frac{\Phi\left(U_5\right)-\Phi\left(L_5\right)}{\Phi\left(U_1\right)-\Phi\left(L_1\right)}\right]^2\right\} \\
&\doteqdot e^{2\mu_y+\sigma_y^2}\Phi_3 \underset{\substack{U\to\infty \\ L\to 0}}{\longrightarrow} e^{2\mu_y+\sigma_y^2}\left(e^{\sigma_y^2}-1\right) \qquad \text{(A.2)}
\end{aligned}$$

and

$$\begin{aligned}
Cov\left(X,Y|Q\right) &= E\left(XY|Q\right) - E\left(X|Q\right)E\left(Y|Q\right) \\
&= e^{\mu_x+\mu_y+\frac{1}{2}\left(\sigma_x^2+\sigma_y^2\right)}\left\{e^{\rho\sigma_x\sigma_y}\left[\frac{\Phi\left(U_4\right)-\Phi\left(L_4\right)}{\Phi\left(U_1\right)-\Phi\left(L_1\right)}\right]\right. \\
&\quad \left. - \left[\frac{\Phi\left(U_2\right)-\Phi\left(L_2\right)}{\Phi\left(U_1\right)-\Phi\left(L_1\right)}\right]\left[\frac{\Phi\left(U_5\right)-\Phi\left(L_5\right)}{\Phi\left(U_1\right)-\Phi\left(L_1\right)}\right]\right\} \\
&\doteqdot e^{\mu_x+\mu_y+\frac{1}{2}\left(\sigma_x^2+\sigma_y^2\right)}\Phi_1 \\
&\quad \underset{\substack{U\to\infty \\ L\to 0}}{\longrightarrow} e^{\mu_x+\mu_y+\frac{1}{2}\left(\sigma_x^2+\sigma_y^2\right)}\left(e^{\rho\sigma_x\sigma_y}-1\right) \qquad \text{(A.3)}
\end{aligned}$$

Finally we see that

$$\rho_Q \underset{\substack{U\to\infty \\ L\to 0}}{\longrightarrow} \frac{\left(e^{\rho\sigma_x\sigma_y}-1\right)}{\sqrt{\left(e^{\sigma_x^2}-1\right)}\sqrt{\left(e^{\sigma_y^2}-1\right)}} \qquad \text{(A.4)}$$

BANK OF FINLAND RESEARCH DISCUSSION PAPERS

ISSN 1456-6184, online

1/2013 Tuomas Takalo **Rationales and instruments for public innovation policies.** 2013. 29 p. ISBN 978-952-462-000-0, online.

2/2013 Tuomas Takalo – Tanja Tanayama – Otto Toivanen **Market failures and the additionality effects of public support to private R&D: Theory and empirical implications.** 2013. 40 p. ISBN 978-952-462-001-7, online.

3/2013 Martin Ellison – Andreas Tischbirek **Unconventional government debt purchases as a supplement to conventional monetary policy.** 2013. 26 p. ISBN 978-952-462-002-4, online.

4/2013 Fabio Verona – Manuel M. F. Martins – Inês Drumond **(Un)anticipated monetary policy in a DSGE model with a shadow banking system.** 2013. 40 p. ISBN 978-952-6699-04-2, online.

5/2013 Fabio Verona – Maik H. Wolters **Sticky information models in Dynare.** 2013. 17 p. ISBN 978-952-6699-08-0, online.

6/2013 Sami Oinonen – Maritta Paloviita – Lauri Vilmi **How have inflation dynamics changed over time? Evidence from the euro area and USA.** 2013. 36 p. ISBN 978-952-6699-09-7, online.

7/2013 Iftekhar Hasan – Matej Marinč **Should competition policy in banking be amended during crises? Lessons from the EU.** 2013. 35 p. ISBN 978-952-6699-10-3, online.

8/2013 Bill Francis – Iftekhar Hasan – Qiang Wu **The benefits of conservative accounting to shareholders: Evidence from the financial crisis.** 2013. 44 p. ISBN 978-952-6699-09-7, online.

9/2013 Juha Kilponen – Jouko Vilmunen – Oskari Vähämaa **Estimating intertemporal elasticity of substitution in a sticky price model.** 2013. 27 p. ISBN 978-952-6699-12-7, online.

10/2013 Owain ap Gwilym – Qingwei Wang – Iftekhar Hasan – Ru Xie **In search of concepts: The effects of speculative demand on returns and volume.** 2013. 35 p. ISBN 978-952-6699-13-4, online.

11/2013 Matti Viren **Sensitivity of fiscal-policy effects to policy coordination and business cycle conditions.** 2013. 21 p. ISBN 978-952-6699-20-2, online.

12/2013 Seppo Honkapohja **The euro crisis: a view from the North.** 2013. 28 p. ISBN 978-952-6699-23-3, online.

13/2013 Stergios Leventis – Iftekhar Hasan – Emmanouil Dedoulis **The cost of sin: The effect of social norms on audit pricing.** 2013. 58 p. ISBN 978-952-6699-27-1, online.

14/2013 Chrysovalantis Gaganis – Iftekhar Hasan – Fotios Pasiouras **Efficiency and stock returns: Evidence from the insurance industry.** 2013. 33 p. ISBN 978-952-6699-28-8, online.

15/2013 Chung-Hua Shen – Iftekhar Hasan – Chih-Yung Lin **The government's role in government-owned banks.** 2013. 51 p. ISBN 978-952-6699-29-5, online.

16/2013 Fabio Verona **Lumpy investment in sticky information general equilibrium.** 2013. 37 p. ISBN 978-952-6699-30-1, online.

17/2013 Jouko Vilmunen – Peter Palmroos **Closed form solution of correlation in doubly truncated or censored sample of bivariate log-normal distribution.** 2013. 9 p. ISBN 978-952-6699-31-8, online.

http://www.suomenpankki.fi/en • email: Research@bof.fi
ISBN 978-952-6699-31-8, ISSN 1456-6184, online