IMPROVING THE SMALL-SAMPLE EFFICIENCY OF A ROBUST CORRELATION MATRIX: A NOTE

Eric Blankmeyer
Department of Finance and Economics
McCoy College of Business Administration
Texas State University – San Marcos
San Marcos, TX 78666
Email eb01@txstate.edu
April 2007

The correlation matrix plays a central role in multivariate analysis, and for Gaussian distributions the maximum-likelihood estimator (mle) is easily computed. However, it is quite vulnerable to outlying observations. The challenge is to specify a highly robust estimator that has acceptable statistical efficiency when the data in fact resemble a multinormal sample. From Kendall and Spearman onward, many researchers have addressed this problem. In recent years, interest has focused on high-breakdown, affine-equivariant estimators of the correlation matrix, of which the best known is perhaps Rousseeuw's Minimum Covariance Determinant (mcd) estimator (Rousseeuw and Leroy 1987, Rousseeuw and Van Driessen 1999). Using resampling algorithms, the mcd searches for the correlation matrix with the smallest volume that contains a fraction h of the observations ($1/2 < h < 1$). The tuning parameter h reflects the researcher's belief that the proportion of contaminated data does not exceed 1-h.

Louphaa and Rousseeuw (1991), Butler et al (1993), and Croux and Haesbroeck (1997, 1999) are among the authors who explore the mcd's large-sample properties and make proposals to enhance its precision at Gaussian distributions. However, asymptotic efficiency may be irrelevant to researchers who must work with small samples. In fact, Rousseeuw and van Zomeren (1990) remark that, for sparse observations, the robust correlation matrix suffers from a

"curse of dimensionality"; they recommend at least five observations for each variable in the matrix.

It is known that a pairwise difference transformation (pdt) can improve the efficiency of some robust estimators. Given n observations on a random variable x, the pdt produces n(n-1)/2 values $z_{ij} = x_i - x_j$ where $i < j$. Using the pdt, Rousseeuw and Croux (1993) develop robust, highly efficient estimates of scale (dispersion), while Croux et al (1994) and Hossjer et al (1994) show that the pdt can greatly enhance the asymptotic efficiency of least median of squares regression. Stromberg et al (2000) obtain a similar result for robust regression by least trimmed squares. The effect of the pdt is to smooth a robust estimator's influence function and also to reduce skewness in the data. (On the other hand, it is easily verified that the pdt leaves the Gaussian mle unchanged, so no efficiency gain is possible.) It seems plausible that the pdt can improve the mcd's precision in small samples. This note reports some exploratory simulations based on the hypothetical correlation matrix shown below:

Table 1. A correlation matrix in three variables

1.000	0.700	0.200
0.700	1.000	0.300
0.200	0.300	1.000

We use the version of mcd in the S-Plus 6.2 Robust Library (Insightful Corporation 2002) with default settings. One thousand samples, each containing 15 observations, are drawn from a multinormal distribution; and the simulation is repeated for samples of 30 observations. Table 2 displays the average correlations before transformation ("mcd") and after the pdt has been applied ("mcd-pdt"). It appears that both estimators are biased upward, but the mcd-pdt has a smaller bias. Pison et al (2002) investigate the small-sample bias in mcd and propose correction factors.

Table 2. Multivariate normal simulation of a correlation matrix in three variables

Sample size n:	15 mcd	15 mcd-pdt	30 mcd	30 mcd-pdt
Average r_{12}	0.761	0.734	0.749	0.720
Average r_{13}	0.242	0.209	0.224	0.213
Average r_{23}	0.344	0.315	0.313	0.309
Variance of z	0.545	0.292	0.306	0.101
Efficiency of z relative to 1/(n-3)	15.3%	28.6%	12.0%	36.5%

As for efficiency, the sampling distribution of a correlation coefficient is truncated and skewed, so we use the well-known Fisher z transformation (the arctangent of the correlation coefficient), whose variance in Gaussian samples of size n is 1/(n - 3). Table 2 suggests that the pdt almost doubles the efficiency of the mcd when n = 15 and triples the efficiency when n = 30. Of course, these simulations are very limited in scope; it will be necessary to explore other correlation matrices, more variables, and different sample sizes. In addition, the pdt should be applied to contaminated samples; and alternatives to the mcd should be examined (e. g., Maronna and Yohai 1995).

References

Butler, R. W., P. L. Davies, and M. Jhun (1993). Asymptotics for the Minimum Covariance Determinant Estimator. *Annals of Statistics* 21, 1385-1400.

Croux, Christophe and Gentiane Haesbroeck (1997). An Easy Way to Increase the Finte-Sample Efficiency of the Resampled Minimum Volume Ellipsoid Estimator. *Computational Statistics and Data Analysis* 25, 125-141.

Croux, Christophe and Gentiane Haesbroeck (1999). Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator. *Journal of Multivariate Analysis* 71, 161-190.

Croux, Christophe, Peter J. Rousseeuw, and Ola Hossjer (1994). Generalized S-estimators. *Journal of the American Statistical Association* 89, 1271-1281.

Hossjer, Ola, Christophe Croux, and Peter J. Rousseeuw (1994). Asymptotics of generalized S-estimators. *Journal of Multivariate Analysis* 51, 148-177.

Insightful Corporation (2002). *S-Plus 6 Robust Library User's Guide*. Insightful Corporation. Seattle, WA.

Lopuhaa, H. P., and P. J. Rousseeuw (1991). Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices. *Annals of Statistics* 19, 229-248.

Maronna, Ricardo A., and Victor J. Yohai (1995). The behavior of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association* 90, 330-341.

Pison, G., S. Van Aelst, and G. Willems (2002). Small sample corrections for LTS and MCD. *Metrika* 55, 111-123.

Rousseeuw, Peter J. and Annick M. Leroy (1987). *Robust regression and outlier detection*. New York: John Wiley.

Rousseeuw, Peter J., and Bert C. van Zomeren (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association* 85, 633-651.

Rousseeuw, Peter J., and Christophe Croux (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association* 88, 1273-1283.

Rousseeuw, Peter J. and Katrien Van Driessen (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41, 212-223.

Stromberg, Arnold, Ola Hossjer, and Douglas M. Hawkins (2000). The least trimmed differences regression estimator and alternatives. *Journal of the American Statistical Association* 95, 853-864.