A note on modelling cross correlations: hyperbolic secant regression

Gordon K. Smyth
Department of Mathematics, University of Queensland,
St Lucia, Q 4072, Australia.

30 March 1994

Abstract

The problem of determining if a bivariate normal correlation changes with respect to time or some other covariate is considered. It is assumed that the means and standard deviations of the normal random variables can be consistently estimated from the entire data run, and do not need to be re-estimated for each covariate value. A new estimator of a bivariate normal correlation is given that has useful performance down to samples of size one. This allows regression type modelling of the correlation without unnecessary loss of resolution. The arc-tanh transformation of this estimator has a symmetric Fisher's z-distribution about the arc-tanh correlation. A method of smoothing the correlation estimates is given using moving average smoothers of the sufficient statistics from which the correlation estimator is calculated.

Keywords: correlations; moving averages; z-distribution; z-transformation; hyperbolic secant distribution.

1 Introduction

Since the discovery of the bivariate normal distribution, correlation coefficients have been the most popular method of measuring the strength of relationships between approximately normal variables. In studies which focus on the stability of relationships over time or with respect to uncontrolled variables, it is of interest to determine if correlation coefficients change with respect to these variables. When a sample of bivariate observations of reasonable size is available for each value of the covariate, the Pearson correlation coefficient r can be calculated for each sample. Differences can then be estimated or tested for using Fisher's result that $z = \tanh^{-1} r$ is approximately normally distributed with approximate mean $\zeta = \tanh^{-1} \rho$, where ρ is the true correlation, and variance approximately constant with respect to ρ (Fisher, 1925; Hotelling, 1953; Johnson and Kotz, 1970, p. 229;

Mudholkar, 1983). See Rao (1973, p. 432) for a biological application and Haney and Lloyd (1978), Watson (1980), Maldonado and Saunders (1981) and Lerman and Schechtman (1989) for applications to financial statistics. Campbell (1981) gives a graphical procedure for comparing correlations. Lerman and Schechtman (1989) and Hawkins (1989) use Fisher's z-transform to test for a correlation change at an unknown time. Muirhead (1982) develops a different method of testing for a correlation change, using cusums of statistics based on the log-likelihood ratio.

In this note it is assumed that the bivariate normal means and variances can be consistently estimated from the entire data run, and do not need to be re-estimated for each covariate value. It is of interest therefore to consider correlation estimators assuming the means and standard deviations to be given. An apparently new correlation estimator $\tilde{\rho}$ is given in Section 2 that has useful performance down to samples of size one. This allows regression type modelling of the correlation without unnecessary loss of resolution. The new estimator is more accurate than r for any sample size and exactly unbiased and constant variance on the arc-tanh scale. It reduces to r when sample means and variances are re-estimated from the same data sample. It is equivalent or superior in performance to the maximum likelihood estimator assuming known means and variances except when the sample size and $|\rho|$ are both reasonably large.

Time series smoothing and regression modelling of the correlations is considered in Section 3. It is shown that to obtain efficient smoothed estimators it is necessary to smooth the sufficient statistics from which the correlations are calculated rather than smooth the correlations themselves.

2 A correlation estimator

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a bivariate normal sample with $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$ and $\mathrm{corr}(X, Y) = \rho$. If μ_X, μ_Y, σ_X and σ_Y are considered known, then $P = \sum X_i Y_i$ and $S = \sum(X_i^2 + Y_i^2)$ are together sufficient for ρ, and the maximum likelihood estimator $\hat{\rho}$ is a root of the cubic polynomial

$$\rho^3 - \rho^2 \frac{1}{n} P + \rho(\frac{1}{n} S - 1) - 1$$

(Kendall and Stuart, 1961, p. 39). The transformed maximum likelihood estimate $m = \tanh^{-1} \hat{\rho}$ has variance $1/[n(1+\rho^2)] + O(1/n^2)$. It is unbiased and symmetric for $\rho = 0$; otherwise the bias and third cumulant are $O(1/n)$ and $O(1/n^3)$ respectively. For large n, the polynomial is likely to have only one real root, but in general there may be three real roots in the admissible range and the likelihood itself must be evaluated to distinguish them.

A closed form estimator of ρ which is superior to $\hat{\rho}$ in small samples can be constructed by observing that positively correlated observations (X, Y) will tend to lie further from the origin in the $(1, 1)$ direction than in the $(1, -1)$ direction, and vice versa for negatively correlated observations. Let $U_i = (X_i + Y_i)/\sqrt{2}$ be the projection of (X_i, Y_i) onto the $(1,1)$ line and let $V_i = (X_i - Y_i)/\sqrt{2}$ be the projection onto the $(1,-1)$ line. Then U_i and V_i are independent with variances $1+\rho$ and $1-\rho$ respectively. The sum of the squared projected lengths in the $(1,1)$ direction relative to that in the $(1,-1)$ direction is $\sum U_i^2 / \sum V_i^2$, which has $(1+\rho)/(1-\rho)$ times an $F_{n,n}$ distribution. Taking the logarithm,

$$h = \frac{1}{2} \log \frac{\sum U_i^2}{\sum V_i^2}$$

is distributed as $\frac{1}{2}\log F$ plus $\frac{1}{2}\log[(1+\rho)/(1-\rho)] = \tanh^{-1}\rho$, where F has an $F_{n,n}$ distribution.

The distribution of $\frac{1}{2}\log F$ is often called Fisher's z-distribution in the literature, following Fisher (1924). A recent reference in Barndorff-Nielsen, Kent and Sorensen (1982). The probability density of h is

$$p_H(h) = \frac{1}{2^{n-1} B(\frac{n}{2}, \frac{n}{2})} \operatorname{sech}^n(h - \zeta)$$

where $\zeta = \tanh^{-1}\rho$. This distribution is symmetric about ζ with variance $\frac{1}{2}\psi'(n/2)$ and fourth cumulant $\frac{1}{8}\psi^{(3)}(n/2)$ where $\psi(\cdot)$ is the digamma function. See Johnson and Kotz (1970, p. 78). For $n = 1$, the distribution is hyperbolic secant with density

$$p_H(h) = \frac{1}{\pi} \operatorname{sech}(h - \zeta)$$

and variance $\pi^2/4$. The hyperbolic secant distribution was introduced by Perks (1932) and Talacko (1956), and is discussed by Johnson and Kotz (1970, p. 15) and Manoukian and Nadeau (1988). For $n = 2$, the distribution is logistic with density

$$p_H(h) = \frac{1}{2} \operatorname{sech}^2(h - \zeta)$$

and variance $\pi^2/12$, and this distribution is discussed by Johnson and Kotz (1970, Chapt. 22). The distribution of h approaches normality rapidly as n increases. The approximation to normality is already good for the logistic distribution, as discussed by Johnson and Kotz (1970, pp. 5–6).

The distribution of h can be compared with that of the sample correlation coefficient

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{[\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2]^{1/2}}$$

which is the maximum likelihood estimator for ρ with μ_X, μ_Y, σ_X and σ_Y considered unknown. The sample correlation returns useful estimates for $n \geq 3$. The probability density function of r is given for example by Hotelling (1953). The density of $z = \tanh^{-1} r$ can be written as

$$p_Z(z) = c(\tanh\zeta \tanh z) \operatorname{sech}^{n-1}\zeta \operatorname{sech}^{n-2} z$$

Table 1: Bias and standard deviation of two estimators of $\tanh^{-1}\rho$ for several samples sizes and values of ρ. m is the maximum likelihood estimator and h is the hyperbolic secant unbiased estimator.

		$n=1$		$n=2$		$n=3$	
Estimator	ρ	bias	std	bias	std	bias	std
h	Any	0.00	1.57	0.00	0.91	0.00	0.68
m	0.0	0.00	1.93	0.00	1.13	0.00	0.84
m	0.5	−0.03	1.96	−0.01	1.12	0.01	0.83
m	0.9	0.09	2.00	0.14	1.02	0.13	0.66

where $c(\cdot)$ is an infinite order polynomial or power series. When $\rho = 0$, the distribution of z is the same as that for h, but with $n-2$ substituted for n, an identity that was observed by Irwin (1953). Otherwise, z is slightly biased and skew. Asymptotic expressions for the moments of z are given by Hotelling (1953) and Johnson and Kotz (1970, p. 229).

We may also express h as a function of the sufficient statistics, since $\sum U_i^2 = \sum(X_i^2 + Y_i^2) + 2\sum X_iY_i = S + 2P$ and $\sum V_i^2 = S - 2P$. Therefore $\sum U_i^2/\sum V_i^2 = (1+\tilde\rho)/(1-\tilde\rho)$, with $\tilde\rho = 2P/S = \tanh h$. The expression $\tilde\rho = 2P/S$ makes it clear that $\tilde\rho$ reduces to r if the observations are standardized using sample means and standard deviations, i.e., if X_i and Y_i are replaced by $(X_i - \bar X)/\hat\sigma_X$ and $(Y_i - \bar Y)/\hat\sigma_Y$ respectively, where $\hat\sigma_X^2$ is any multiple of $\sum(X_i - \bar X)^2$ and $\hat\sigma_Y^2$ is the same multiple of $\sum(Y_i - \bar Y)^2$. If the variances is standardized but the variables are not mean corrected, i.e., if X_i and Y_i are replaced by $X_i/\hat\sigma_X$ and $Y_i/\hat\sigma_Y$, then $\tilde\rho$ has the same distribution as r but with $n-1$ in place of $n-2$. In a precise sense then, one degree of freedom is lost if we need to estimate the variances from the same data, and a second is lost if we need to estimate the means as well.

It is easily seen that h is invariant under rescaling of the bivariate data. The X_i and Y_i may therefore have any common and constant variance without affecting the distribution of h. For $n = 1$, the estimator $\tilde\rho$ may be written as $\sin 2\theta$ in terms of the spherical coordinate representation $X = a\cos\theta$, $Y = a\sin\theta$.

For large n, the efficiency of h relative to m is $1/(1+\rho^2)$. For small n though the picture is different. The bias and standard deviation of h and m are given in Table 1 for sample sizes $n = 1$, 2 and 3 and for $\rho = 0$, 0.5 and 0.9. The hyperbolic estimator h has the smaller mean square error for these very small sample sizes. Values for m were obtained from simulation using the matrix programming language Matlab (Mathworks, 1991) and may differ by at most 0.01 from true values due to round-off and sampling errors.

Table 2: The efficiency of $\bar{h}$ for estimating a constant correlation when the h_i are calculated from windows of size n, relative to h calculated from the entire data set.

n	1	2	3	4	5	6	7	8	9	10
$(n/2)\psi'(n/2)$	2.47	1.64	1.40	1.29	1.23	1.18	1.16	1.14	1.12	1.11

3 Applications

Suppose that $x_1, \ldots, x_N$ and $y_1, \ldots, y_N$ are prewhitened sequences standardized to have zero mean and unit variance, and that $\mathrm{corr}(X_i, Y_i) = \rho_i$. In principle we may simply calculate the correlation response $h_i = \tanh^{-1}[2x_i y_i/(x_i^2 + y_i^2)]$ for each bivariate observation and apply regression methods to model the correlations. If it can be assumed that $\zeta_i = \tanh^{-1}\rho_i = \beta^T w_i$, where w_i is a vector of covariates, then the least squares estimator

$$\hat{\beta} = (W^T W)^{-1} W^T h$$

where W is the matrix with ith row w_i and $h = (h_1, \ldots, h_N)^T$, is unbiased and consistent for β with covariance matrix $(W^T W)^{-1}\pi^2/4$. Also $\hat{\beta}$ is likely to be very nearly normally distributed. Manoukian and Nadeau (1988) show that the mean of a hyperbolic secant sample is closely normal even for very small n. In a similar way, standard nonlinear least squares methods can be used to estimate β given a more general correlation function $\zeta_i = g(w_i, \beta)$ where $g(\cdot)$ is some known function.

The approach based on individual hyperbolic secant correlation responses however is inefficient. Let $\zeta_i = \beta_0 + \beta^T w_i$ and suppose that the covariates w_i have been the mean corrected, i.e., that the matrix W with ith row w_i has all column sums zero. Then the maximum likelihood estimator of β has asymptotic covariance matrix $(W^T W)^{-1}(1 + \tanh^2\beta_0)^{-1}$ and the asymptotic relative efficiency of the hyperbolic secant least squares estimator is $4\pi^{-2}(1 + \tanh^2\beta_0)^{-1}$, which is about 41% for β_0 near zero and decreases to half that for $|\beta_0|$ large. Most of the lost information can be recovered by calculating h from larger windows of observations. To quantify this, suppose that the correlations ρ_i are constant and that we estimate the arc-tanh correlation by averaging N/n values h_i calculated from distinct sets of n bivariate observations. The variance of the resulting estimator $\bar{h}$ is $n/(2N)\psi'(n/2)$, which has a minimum of about $1/N$ at $n = N$. Relative to this minimum the variance is given in Table 2. This shows that in aggregating the correlations it is important to average or smooth the U_i^2 and V_i^2 from which the sufficient statistics are calculated rather than to average or smooth the correlation responses themselves.

For example, consider the following synthetic data sequence. Standard normal observations X_i and Y_i, $i = 1, \ldots, 200$ were simulated so that $\mathrm{corr}(X_i, Y_i) = 0.6$ for $76 \leq i \leq 125$ and $\mathrm{corr}(X_i, Y_i) = 0$ otherwise. The sequences $U_i^2 = (X_i + Y_i)^2$ and $V_i^2 = (X_i - Y_i)^2$ were then smoothed using unweighted moving average filters of various window widths n, producing smoothed sequences U_i^{*2} and V_i^{*2},

$i = 1, \ldots, 200 - n + 1$. For each window width, the smoothed correlation responses $h_i = \frac{1}{2}\log(U_i^{*2}/V_i^{*2})$ were calculated and plotted. Under the assumption of constant correlation, the h_i should have a z-distribution on n, n degrees of freedom about the arc-tanh correlation. The window width was gradually increased until a clear picture emerged. Not surprisingly, since there are 50 unusual observations in the middle of the sequence, the most interesting pictures emerged for window widths around 50. The smoothed correlation responses in Figure 1 are for $n = 60$. Also given in Figure 1 are approximate 95% and 99% confidence bands under the assumption of constant correlation. The height of the bands is $\bar{h} \pm \frac{1}{2}\log f$ where f is the $(1-p)$th quantile of the $F_{n,n}$ distribution. For a $(1-\alpha)100\%$ confidence band, p was set to $1 - (1-\alpha/2)^{n/N}$ where $N = 200$ is the sample size. The confidence bands are strictly appropriate for correlation responses calculated from non-overlapping windows of observations, and are slightly optimistic in the current case. Experimentation with other simulated data sets suggests that the level of optimism is small.

The above smoothing technique was applied to data from two variables measured simultaneously on a continuously operating ICI (Imperial Chemical Industries) production process. The two series of 100 observations each are given in graphical form in Muirhead (1982). For the current analysis the series were prewhitened using univariate AR(1) models, after removing linear trends, as described by Muirhead. Smoothed correlation response plots were then formed for various window sizes. For any window size between about 10 and 22 the plot shows that the correlation is decreasing at the end of the sequence; see Figure 2 which is for $n = 15$. Muirhead was concerned with testing whether the cross-correlation between the innovation sequences in this series of data was less than the long run value of 0.49. From Figure 2 it appears that not only is this so but a further decrease is discernible during the run. Introduction of measuring errors towards the end of the period is one possible explanation.

A circumstance in which it may be practical and beneficial to calculate unsmoothed correlation responses is the availability of very long data series. Consider the wind speed data analysed by Haslett and Raftery (1989) consisting of daily mean wind speeds at 12 meteorological stations in Ireland during the period 1961–1978. Haslett and Raftery omitted the Rosslare site on the south east coast from the main analysis, concluding that it may be subject to meteorological influences different from those at the other sites. Here we consider cross-correlations between Rosslare and its nearest inland neighbour Kilkenny. Haslett and Raftery model the series using fractionally differenced AR models. Here we transform to normality and remove seasonal trends as described by Haslett and Raftery, but do not pre-whiten the series. Unsmoothed correlation responses h_i, $i = 1, \ldots, 6574$ were calculated after mean correcting using the sample means and standardizing the variances using the sample standard deviations. The correlation responses represent a coloured process, since the original bivariate series was coloured, but are unbiased for the arc-tanh correlations of the wind speeds about their seasonal trends. Ordinary least squares methods suggest that the cross-correlation does not drift linearly over time but there is a significant annual cycle. Annual harmonics (with a sin and a cos term for each harmonic) were fitted to the h_i by ordinary least squares. The first three harmonics were highly significant on the basis of the usual least squares calculations. The annual trend line is shown in Figure 3 together with the mean correlation response for each day of the year. The residual mean square error from the regression is 2.434, close to the value of 2.467 that would be expected on the basis of hyperbolic secant errors. This analysis is crude and could be refined by modelling serial dependence in the h_i using univariate time series methods or by using a hyperbolic secant likelihood instead of least squares, but the conclusion seems clear enough. The cross-correlation decreases in summer to around tanh(0.66) or 0.58 and increases in winter to around tanh(1.25) or 0.85. Since the mean wind speeds also fall in summer and rise in winter, it appears that wind speeds are less correlated during the season when they are lower, and this may have implications for wind power generation.

Acknowledgements

This work was completed while the author was visiting the School of Statistics at the University of Minnesota. The Irish wind data was obtained from the Statlab network database at Carnegie-Mellon University and Matlab computer programs for the calculation of F distribution quantiles were obtained from the Netlib database at the Oak Ridge National Laboratory. The author wishes to thank Chris Fraley of the University of Washington for contributing the wind data and Peter Shaw of the Woods Hole Oceanographic Institution for contributing the Matlab routines.

References

Barndorff-Nielsen, O., Kent, J. ad Sorensen, M. (1982). Normal variance-mean mixtures and z distributions. *Int. Statist. Rev.*, **50**, 145–59.

Campbell, N. A. (1981). Graphical comparison of covariance matrices. *Austral. J. Statist.*, **23**, 21-37.

Fisher, R. A. (1924). On a distribution yielding the error functions of several well known statistics. *Proc. Int. Math. Congr. Toronto*, **2**, 805–813.

Fisher, R. A. (1925). Statistical methods for research workers. London: Oliver & Boyd.

Haney, R. L. and Lloyd, W. P. (1978). An examination of the stability of intertemporal relationships among national stock market indices. *Nebraska J. Econ. Bus.*, **17**, 55–65.

Haslett, J. and Raftery, A. E. (1989). Space-time modelling with long-memory dependence: assessing Ireland's wind power resource. *Appl. Statist.*, **38**, 1–50.

Hawkins, D. L. (1989). A note on the asymptotic distribution of a statistic for testing stability of a correlation coefficient. *Statist. Prob. Letters*, **9**, 149–54.

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. With discussion. *J. Roy. Statist. Soc. B*, **15**, 193–232.

Irwin, J. O. (1953). Discussion on Professor Hotelling's paper. *J. Roy. Statist. Soc. B*, **15**, 228.

Johnson, N. L. and Kotz, S. (1970). *Distributions in statistics: continuous distributions, Vol. 2.* New York: Wiley.

Kendall, M. G. and Stuart, A. (1961). The advanced theory of statistics, Vol. 2. New York: Hafner Publishing.

Lerman, Z. and Schechtman, E. (1989). Detecting a change in the correlation coefficient in a sequence of bivariate normal variables. *Commun. Statist. — Simula. Comput.*, **18**, 589–99.

Maldonado, R. and Saunders, A. (1981). International portfolio diversification and the inter-temporal stability of international stock market relationships, 1957–78. *Financial Analysts J.* **37**, 54–63.

Manoukian, E. B. and Nadeau, P. (1988). A note on the hyperbolic-secant distribution. *American Statist.*, **42**, 77–79.

Mathworks (1991). *Matlab user's guide.* South Natick, MA: Mathworks.

Mudholkar, G. S. (1983). Fisher's z-transformation. In: Kotz, S., Johnson, N. L. and C. B. Read (eds.). *Encyclopedia of Statistical Science, Vol. 3.* New York: Wiley, pp. 130–5.

Muirhead, C. R. (1982). Sequential detection of changes in the cross-correlation coefficient. In: O. D. Anderson (ed.), *Time series analysis: theory and practice 2*, New York: North-Holland, pp. 161–170.

Perks, W. F. (1932). On some experiments in the graduation of mortality statistics. *Institute Actuaries J.*, **58**, 12–57.

Rao, C. R. (1973). *Linear statistical inference and its applications.* New York: Wiley.

Talacko, J. (1956). Perks' distributions and their role in the theory of Wiener's stochastic variables. *Trabajos de Estadistica*, **17**, 159–74.

Watson, J. (1980). The stationarity of inter-country correlation coefficients: a note. *J. Bus. Finance Account.*, **7**, 297–303.