# The Gerber Statistic: A Robust Measure of Correlation

Sander Gerber Babak Javid Harry Markowitz Paul Sargen
David Starer

February 21, 2019

**Abstract**

We introduce the Gerber statistic, a robust measure of correlation. The statistic extends Kendall's Tau by counting the proportion of simultaneous co-movements in series when their amplitudes exceed data-dependent thresholds. This is unlike the standard Pearson correlation that is sensitive to outliers or the Spearman correlation that relies on ranking observations. Since the statistic is neither affected by extremely large or extremely small movements, it is especially suited to financial time series since these can exhibit extreme movements as well as a great amount of noise. Therefore, the statistic can advantageously be converted into a robust estimate of a covariance matrix that is suitable for portfolio optimization.

## 1 Introduction

In this paper, we introduce the Gerber Statistic (GS); a robust measure of time-series correlation. The GS is designed to recognize co-movement between series when the movements are substantial (in a sense to be defined later) and to be insensitive to small co-movements that may be due to noise alone. The GS is a generalization of Kendall's Tau [Kendall, 1938]. The latter measures correlation between two sets as the ratio of the difference between the number of concordant and discordant pairs in the sets divided by the sum of the number of concordant and discordant pairs in the sets. The GS generalizes Kendall's Tau by including thresholds such that only co-movements that simultaneously exceed these thresholds will be recognized as concordant or discordant.

## Precursors

Portfolio construction [Markowitz, 1952, 1959] relies heavily on the availability of the matrix of covariances between securities' returns. Often the sample covariance matrix is used as an estimate for the actual covariance matrix. But as early as Sharpe [1963], various models have been used to ease the computational burden and to improve statistical properties of covariance matrix estimates. Nevertheless, a central problem still exists with many estimates

of covariance matrices: They use product-moment-based estimates that inherently are not robust. This is particularly troublesome if the underlying distribution is prone to containing extreme measurements or outliers.

Robust estimators, based on the pioneering work of Tukey [1960], Huber and Ronchetti [2009] and Hampel [1968, 1974], have largely overcome this problem. Shevlyakov and Smirnov [2011] provide a thorough examination of modern robust methods for computing correlations.

However, financial time series have characteristics that make even standard robust techniques unsuitable. Financial time series are particularly noisy, and this noise can be easily misinterpreted as information. One consequence, for example, is that correlation matrix estimates (even ones built using robust techniques) often have non-zero entries corresponding to series that in fact have no meaningful correlation. On the other hand, the correlation estimates can also be distorted if the series contains extremely large (positive or negative) observations. The GS provides a robust method for computing correlations that ignore fluctuations below a certain threshold, while simultaneously limiting the effects of extreme movements.

# 2 Theory

Let $r_{tk}$ be the return of security $k$ at time $t$, for $k = 1, \ldots, K$ securities and $t = 1, \ldots, M$ time periods. For every pair $(i, j)$ of assets for each time $t$, we convert each return observation pair $(r_{ti}, r_{tj})$ to a joint observation $m_{ij}(t)$ defined as:

$$m_{ij}(t) = \begin{cases} +1 \text{ if } r_{ti} \geq +T_i \text{ and } r_{tj} \geq +T_j, \\ +1 \text{ if } r_{ti} \leq -T_i \text{ and } r_{tj} \leq -T_j, \\ -1 \text{ if } r_{ti} \geq +T_i \text{ and } r_{tj} \leq -T_j, \\ -1 \text{ if } r_{ti} \leq -T_i \text{ and } r_{tj} \geq +T_j, \\ \quad 0 \text{ otherwise,} \end{cases}$$

where $T_k$ is a threshold for security $k$. The joint observation $m_{ij}(t)$ is therefore set to $+1$ if the series $i$ and $j$ simultaneously pierce their thresholds in the same direction at time $t$, to $-1$ if they pierce their thresholds in opposite directions at time $t$, or to zero if at least one of the series does not pierce its threshold at time $t$

We refer to a pair for which both components pierce their thresholds while moving in the same direction as a *concordant* pair, and to one whose components pierce their thresholds while moving in opposite directions as a *discordant* pair.

We set the threshold $T_k$ for security $k$ to be

$$T_k = c\sigma_k$$

where $c$ is some fraction (say $1/2$) and $\sigma_k$ is the sample standard deviation of the return of security $k$. More robust measures than standard deviation can, of course, be used for the threshold computation.

The Gerber statistic for a pair of assets is then defined as

$$g_{ij} = \frac{\sum_{t=1}^{M} m_{ij}(t)}{\sum_{t=1}^{M} \left| m_{ij}(t) \right|}. \tag{1}$$

Letting $n_{ij}^c$ be the number of concordant pairs for series $i$ and $j$, and letting $n_{ij}^d$ be the number of discordant pairs, it can be shown that Equation (1) is equivalent to

$$g_{ij} = \frac{n_{ij}^c - n_{ij}^d}{n_{ij}^c + n_{ij}^d}.$$

This statistic is identical to Kendall's Tau if the threshold $T_k$ is set to zero for all $k$.

Since this statistic relies on counts of the number of simultaneous piercings of thresholds, and not on the extent to which the thresholds are pierced, it is insensitive to extreme movements that distort product-moment-based measures. At the same time, since a series must exceed its threshold before it becomes a candidate to be counted, the measure is also insensitive to small movements that may simply be noise.

Define $\boldsymbol{R} \in \mathbb{R}^{M \times K}$ as the return matrix having $r_{tk}$ in its $t$-th row and $k$-th column. Also define $\boldsymbol{U}$ as a matrix with the same size as $\boldsymbol{R}$ having entries $u_{tj}$ such that

$$u_{tj} = \begin{cases} 1 & \text{if } r_{tj} \geq +T_j, \\ 0 & \text{otherwise.} \end{cases}$$

With this definition, the matrix of the number of samples that exceed the upper threshold is

$$\boldsymbol{N}^{\text{UU}} = \boldsymbol{U}^\top \boldsymbol{U}.$$

Specifically, we have the useful property that the $ij$ element $n_{ij}^{\text{UU}}$ of $\boldsymbol{N}^{\text{UU}}$ is the number of samples for which both time series $i$ exceeds the upper threshold and for which time series $j$ simultaneously exceeds the upper threshold.

Similarly, define $\boldsymbol{D}$ as the matrix with the same size as $\boldsymbol{R}$ having entries $d_{tj}$ such that

$$d_{tj} = \begin{cases} 1 & \text{if } r_{tj} \leq -T_j, \\ 0 & \text{otherwise.} \end{cases}$$

With this definition, the matrix of the number of samples that go below the lower threshold is

$$\boldsymbol{N}^{\text{DD}} = \boldsymbol{D}^\top \boldsymbol{D}.$$

Again, we have the useful property that the $ij$ element $n_{ij}^{\text{DD}}$ of $\boldsymbol{N}^{\text{DD}}$ is the number of samples for which both time series $i$ goes below the lower threshold and for which time series $j$ simultaneously goes below the lower threshold.

The matrix containing the numbers of concordant pairs is therefore,

$$\boldsymbol{N}_{\text{CONC}} = \boldsymbol{N}^{\text{UU}} + \boldsymbol{N}^{\text{DD}} = \boldsymbol{U}^\top \boldsymbol{U} + \boldsymbol{D}^\top \boldsymbol{D}.$$

It can be shown that the matrix containing the numbers of discordant pairs is

$$\boldsymbol{N}_{\text{DISC}} = \boldsymbol{U}^\top \boldsymbol{D} + \boldsymbol{D}^\top \boldsymbol{U}.$$

We can now write the Gerber matrix $\boldsymbol{G}$ (i.e., the matrix that contains $g_{ij}$ in its $i$-th row and $j$-th column) in the equivalent matrix form

$$\boldsymbol{G} = (\boldsymbol{N}_{\text{CONC}} - \boldsymbol{N}_{\text{DISC}}) \oslash (\boldsymbol{N}_{\text{CONC}} + \boldsymbol{N}_{\text{DISC}})$$

where the symbol $\oslash$ represents the Hadamard (elementwise) division.

To simplify the description of various versions of the Gerber statistic, it is useful to consider the following graphical representation for the relationship between two securities:

$$\begin{array}{ccc} UD & UN & UU \\ ND & NN & NU \\ DD & DN & DU \end{array}$$

Let the rows here represent categorizations of security $i$, and let the columns represent categorizations of security $j$. The boundaries between the rows and the columns are the chosen thresholds. The letter $U$ represents the case in which a security's return lies above the upper threshold (i.e., is up). The letter $N$ represents the case in which a security's return lies between the upper and lower thresholds (i.e., is neutral). The letter $D$ represents the case in which a security's return lies below the lower threshold (i.e., is down).

For example, if at time $t$, the return of security $i$ is above the upper threshold, this observation lies in the top row. If, at the same time $t$, the return of security $j$ lies between the two thresholds, this observation lies in the middle column. Therefore, this observation lies in the UN region.

Over the history, $t = 1, \ldots, M$, there will be observations scattered over the nine regions. Let $n_{ij}^{pq}$ be the number of observations for which the returns of securities $i$ and $j$ lie in regions $p$ and $q$, respectively, for $p, q \in \{U, N, D\}$. With this notation, we can write another equivalent expression for the Gerber statistic as

$$g_{ij} = \frac{n_{ij}^{\text{UU}} + n_{ij}^{\text{DD}} - n_{ij}^{\text{UD}} - n_{ij}^{\text{DU}}}{n_{ij}^{\text{UU}} + n_{ij}^{\text{DD}} + n_{ij}^{\text{UD}} + n_{ij}^{\text{DU}}}.$$

The correlation matrix constructed from the original Gerber statistic as defined in Equation (1) was often not positive semidefinite (PSD). We therefore tested some alternative versions that were PSD. A first alternative version is

$$g_{ij}^{(1)} = \frac{\sum_{t=1}^{M} m_{ij}(t)}{M - n_{ij}^{NN}}.$$

This can be written in terms of the alternative notation as

$$g_{ij}^{(1)} = \frac{n_{ij}^{\text{UU}} + n_{ij}^{\text{DD}} - n_{ij}^{\text{UD}} - n_{ij}^{\text{DU}}}{M - n_{ij}^{\text{NN}}}.$$

A second alternative version is

$$g_{ij}^{(2)} = \frac{n_{ij}^{\text{UU}} + n_{ij}^{\text{DD}} - n_{ij}^{\text{UD}} - n_{ij}^{\text{DU}}}{\sqrt{n_{ij}^{(\text{A})} n_{ij}^{(\text{B})}}}$$

where

$$n_{ij}^{(A)} = n_{ij}^{\text{UU}} + n_{ij}^{\text{UN}} + n_{ij}^{\text{UD}} + n_{ij}^{\text{DU}} + n_{ij}^{\text{DN}} + n_{ij}^{\text{DD}}$$
$$n_{ij}^{(B)} = n_{ij}^{\text{UU}} + n_{ij}^{\text{NU}} + n_{ij}^{\text{UD}} + n_{ij}^{\text{UD}} + n_{ij}^{\text{ND}} + n_{ij}^{\text{DD}}.$$

Let $\boldsymbol{H} = \boldsymbol{N}_{\text{CONC}} - \boldsymbol{N}_{\text{DISC}}$, and let $\boldsymbol{h} = \sqrt{\text{diag}(\boldsymbol{H})}$ be the vector of square roots of the diagonal elements of $\boldsymbol{H}$ (which are all positive). It can be shown that the second alternative version of the Gerber statistic can be written in the matrix form

$$\boldsymbol{G}^{(2)} = (\boldsymbol{N}_{\text{CONC}} - \boldsymbol{N}_{\text{DISC}}) \oslash (\boldsymbol{h}\boldsymbol{h}^\top).$$

Written differently, letting $\boldsymbol{J} = \boldsymbol{J}^\top$ be the diagonal matrix with the inverse of the $i$-th element of $\boldsymbol{h}$ in its $i$-th diagonal position, we have

$$\boldsymbol{G}^{(2)} = \boldsymbol{J}^\top(\boldsymbol{N}_{\text{CONC}} - \boldsymbol{N}_{\text{DISC}})\boldsymbol{J}.$$

Portfolio optimizers require the covariance matrix of securities' returns to be positive semidefinite. Our intention is to use the Gerber matrix as a robust version of the correlation matrix from which a corresponding robust version of the covariance matrix can be constructed. We will use this version of the covariance matrix in a portfolio optimizer. We therefore require the Gerber matrix to be positive semidefinite.

In all forms, the Gerber matrix can be viewed as a matrix ratio whose numerator matrix is $\boldsymbol{H} = \boldsymbol{N}_{\text{CONC}} - \boldsymbol{N}_{\text{DISC}}$ and whose denominator matrix depends on the particular alternative chosen. If the numerator matrix is not positive semidefinite, the Gerber matrix cannot be positive semidefinite even if the denominator matrix is positive definite. Therefore, a common requirement for all alternatives is for the numerator matrix to be positive semidefinite. The requirement is satisfied, as shown below.

From the definitions of $\boldsymbol{N}_{\text{CONC}}$ and $\boldsymbol{N}_{\text{DISC}}$, the numerator matrix can be written in the following squared form:

$$\begin{aligned}\boldsymbol{H} &= \boldsymbol{N}_{\text{CONC}} - \boldsymbol{N}_{\text{DISC}} \\ &= \boldsymbol{U}^\top\boldsymbol{U} + \boldsymbol{D}^\top\boldsymbol{D} - \boldsymbol{U}^\top\boldsymbol{D} - \boldsymbol{D}^\top\boldsymbol{U} \\ &= (\boldsymbol{U} - \boldsymbol{D})^\top(\boldsymbol{U} - \boldsymbol{D}).\end{aligned}$$

Therefore, for arbitrary but non-zero $\boldsymbol{x}$,

$$\boldsymbol{x}^\top\boldsymbol{H}\boldsymbol{x} = \boldsymbol{x}^\top(\boldsymbol{U} - \boldsymbol{D})^\top(\boldsymbol{U} - \boldsymbol{D})\boldsymbol{x} = \boldsymbol{u}^\top\boldsymbol{u} \geq 0.$$

Thus, the numerator matrix is positive semidefinite.

For certain cases, it is possible to extend this analysis to show that the Gerber matrix itself is positive semidefinite. For example, in the second alternative form,

$$\begin{aligned}\boldsymbol{x}^\top\boldsymbol{G}^{(2)}\boldsymbol{x} &= \boldsymbol{x}^\top\boldsymbol{J}^\top\boldsymbol{H}\boldsymbol{J}\boldsymbol{x} \\ &= \boldsymbol{x}^\top\boldsymbol{J}^\top(\boldsymbol{U} - \boldsymbol{D})^\top(\boldsymbol{U} - \boldsymbol{D})\boldsymbol{J}\boldsymbol{x} = \boldsymbol{u}^\top\boldsymbol{u} \geq 0.\end{aligned}$$

This shows that the second alternative form of the Gerber matrix is positive semidefinite.

The first alternative form also produces positive semidefinite correlation matrices. This can be proved by noting that the numerator matrix $\boldsymbol{H}$ is postive demi-definite as shown above, and the Hadamard denominator matrix is a totally positive matrix (see, for example, Fallat and Johnson [2011]). Its element-wise inverse exists and is positive definite. Therefore, by the Schur Product Theorem [Schur, 1911], the product of the elementwise inverse and the numerator matrix is positive semidefinite.

# 3 Conclusion

In this paper, we introduced the Gerber statistic; a robust measure of correlation. The statistic is useful for especially assessing correlation between financial time series because it is insensitive to extremely large co-movements that distort product-moment-based measures, while also being insensitive to small movements that are likely to be noise.

We propose three variations on the measure and prove that one produces a positive semidefinite correlation matrix. A second variation has been shown in numerous experiments (not reported here) also to produce positive semidefinite correlation matrices.

The properties of the measure make it ideal to act as the kernel for constructing robust covariance matrices for use in portfolio construction. In future research, we intend to test the efficacy of the measure for this purpose.

# References


Shaun M. Fallat and Charles R. Johnson. *Totally Nonnegative Matrices*. Princeton University Press, 2011.

Frank R. Hampel. *Contributions to the Theory of Robust Estimation*. PhD thesis, University of California, Berkeley, 1968.

Frank R. Hampel. The Influence Curve and Its Role in Robust Estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. John Wiley, second edition, 2009.

Maurice G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1/2):81–93, June 1938.

Harry M. Markowitz. Portfolio Selection. *Journal of Finance*, 7(1):77–91, 1952.

Harry M. Markowitz. *Portfolio Selection: Efficient Diversification of Investments*. Basil Blackwell, Cambridge, MA, 1959.

Issai Schur. Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen. *Journal für die reine und Angewandte Mathematik*, 140:1–28, 1911.

William F. Sharpe. A Simplified Model for Portfolio Analysis. *Management Science*, 9(2): 277–293, January 1963.

Georgy Shevlyakov and Pavel Smirnov. Robust Estimation of the Correlation Coefficient: An Attempt of Survey. *Austrian Journal of Statistics*, 40(1&2):147–156, 2011.

John W. Tukey. *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, chapter A Survey of Sampling from Contaminated Distributions. Stanford University Press, 1960.