Which Robust Versions of Sample Variance and Sample Covariance Are Most Appropriate for Econometrics: Symmetry-Based Analysis

Songsak Sriboonchitta[1], Ildar Batyrshin[2], and Vladik Kreinovich[3]

[1]Faculty of Economics, Chiang Mai University
Chiang Mai, Thailand
songsakecon@gmail.com
[2]Centro de Investigación en Computación (CIC)
Instituto Politécnico Nacional (IPN)
México, D.F., batyr1@gmail.com
[3]Department of Computer Science
University of Texas at El Paso,
500 W. University, El Paso, Texas 79968, USA,
vladik@utep.edu

Abstract

In many practical situations, we do not know the shape of the corresponding probability distributions and therefore, we need to use robust statistical techniques, i.e., techniques that are applicable to all possible distributions. Empirically, it turns out the the most efficient robust version of sample variance is the average value of the p-th powers of the deviations $|x_i - \hat{a}|$ from the (estimated) mean $\hat{a}$. In this paper, we use natural symmetries to provide a theoretical explanation for this empirical success, and to show how this optimal robust version of sample variance can be naturally extended to a robust version of sample covariance.

1 Formulation of the Problem

Need to determine a parameter: traditional case. Often, we observe a sample of several instances $x_1, \ldots, x_n$ of a random variable X.

In many practical situations, we know that the random variable X has a distribution with the probability density function (pdf) $\rho(x) = \rho_0(x - a)$, where $\rho_0(x)$ is a known function, and a is an unknown parameter. For example, X may be the measurement result, which can be represented as $X = a + X_0$, where

a is the actual (unknown) value of the corresponding quantity, and X_0 is the measurement error with a known pdf $\rho_0(x)$.

In such situations, to estimate the value a based on the observations $x_1, \ldots, x_n$, we can use, e.g., the maximum likelihood method, i.e., find the value a for which the product

$$L \stackrel{\text{def}}{=} \prod_{i=1}^{n} \rho(x_i) = \prod_{i=1}^{n} \rho_0(x_i - a)$$

is the largest possible; see, e.g., [21].

The corresponding optimization problem is equivalent to minimizing the sum

$$-\ln(L) = \sum \psi_0(x_i - a),$$

where $\psi_0(x) \stackrel{\text{def}}{=} -\ln(\rho_0(x))$. This is the equivalent form most frequently used for optimization, since most optimization techniques involve differentiation of the objective function, and differentiating the sum is much easier than differentiating the product – we get fewer terms in the expression for the derivative.

In particular, in the frequent case when the distribution $\rho_0(x)$ is Gaussian, i.e., when

$$\rho_0(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \exp\left(-\frac{x_0^2}{2\sigma^2}\right)$$

with a known standard deviation σ, the maximum likelihood method $L \to \max$ is equivalent to $-\ln(L) \to \min$ and is, thus, equivalent to the Least Square method

$$\sum_{i=1}^{n} (x_i - a)^2 \to \min_a.$$

For this problem, the Least Squares method results in the known estimate

$$\hat{a} = \frac{x_1 + \ldots + x_n}{n}.$$

Once we have found the estimate $\hat{a}$ for the parameter a, we can then estimate the variance as

$$\hat{V} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \hat{a})^2.$$

This value is proportional to X^2, so, as a measure of deviation of the random variable X from $\hat{a}$, we can take $\hat{\sigma} \stackrel{\text{def}}{=} \sqrt{\hat{V}}$.

If we have two random variables X and Y, with parameters a and b, then their covariance C can be estimated, similarly, as

$$\hat{C} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \hat{a}) \cdot (y_i - \hat{b}).$$

Gaussian distributions are ubiquitous. Their ubiquity comes from the fact that, according to the Central Limit Theorem, the distribution of the sum of a large number of independent small random components is close to Gaussian [21]. Not surprisingly, the empirical analysis of measuring instruments shows that for about 60% of them, the corresponding probability distribution is close to Gaussian [17, 18]. Because of this ubiquity, the Gaussian-motivated formulas are the ones (and often the only ones) that engineering and science students learn in their studies, and these formulas are the ones most frequently used in practice.

Need for robust estimations. In many practical situations, we do not know the shape $\rho_0(x)$ of the corresponding probability distribution. In such situations, several different probability distributions are consistent with our knowledge.

Sometimes, in such situations, practitioners use Gaussian-motivated estimators – since, as we have mentioned, these are the only estimators that these practitioners know. The results are often misleading, even for distributions which are rather close to Gaussian.

As an example, let us consider the case when 99% percent of the values X are normally distributed with mean 0 and standard deviation 0.1, but 1% represent outliers with standard deviation 1000. Then, with one such outlier of size 1000 in a sample of size 100, the arithmetic average $\hat{a}$ of this sample will be close to 10 – very far away from the actual 0 mean. The resulting estimate $\hat{V}$ for the variance will also be very misleading.

The results are also misleading for the case of heavy-tailed distributions, a very typical situation in econometrics (see, e.g., [3, 4, 5, 11, 13, 15, 22, 23]) and in many other application areas [2, 7, 12, 14, 20]. For heavy-tailed distributions, variance is infinite. So, due to the Large Numbers Theorem, the corresponding variance $\hat{V}$ tends to infinity as the sample size n grows – and thus, does not provide us with any meaningful measure of how far the random variable deviates from $\hat{a}$.

To cover such cases, we need to use techniques which are applicable not only for one of the possible distributions, but rather for all possible distributions. Such techniques are known as *robust*; see, e.g., [8].

The classical example of a robust estimate is the median, that corresponds to minimizing the sum $\sum\limits_{i=1}^{n} |x_i - a|$. The resulting smallest value of this sum can serve – after dividing by n – as a robust estimate of how far the random variable deviates from $\hat{a}$:

$$\hat{V} = \frac{1}{n} \cdot \sum_{i=1}^{n} |x_i - \hat{a}|.$$

Many other robust techniques have been proposed.

Bayesian approach to robust estimations: main ideas. The main idea behind the Bayesian approach is that in situations in which several different alternatives are consistent with our knowledge, we select a *prior* probability distribution on the set of all possible alternatives.

In our problem, alternatives are different probability distributions.

- In some practical situations – e.g., when we estimate variance-type characteristics – we are interested in the properties of a single random variable. In such situations, the unknown probability distribution $\rho(z)$ corresponds to a single variable z.
- In other situations – e.g., when we estimate covariance-type characteristics – we are interested in the relation between two random variables. In such situation, the unknown probability distribution corresponds $\rho(z)$ describes the joint distribution of a pair $z = (x, y)$ of random variables.
- In general, we are interested in the probability distribution $\rho(z)$ over some tuples z.

For our problem, the Bayesian approach means that we select a prior distribution on the set of all possible probability distributions $\rho(z)$.

We want to estimate some characteristic $c(\rho)$ of the distribution ρ. To compute such an estimate, we can use a sample $z_1, \ldots, z_n$ taken from the (unknown) probability distribution $\rho(z)$. Let us denote the corresponding estimate by $s(z_1, \ldots, z_n)$.

Which function $s(z_1, \ldots, z_n)$ should we choose? According to decision making theory (see, e.g., [6, 10, 16, 19]), a rational decision maker always maximizes the expected value of the objective appropriate function called *utility*. In our case, the utility function $u(s, c)$ depends on how close the estimate $s(z_1, \ldots, z_n)$ is to the actual value $c(\rho)$ of the desired characteristic. For each probability distribution ρ, the expected value of the utility is equal to

$$E(\rho_0) = \int u(s(z_1, \ldots, z_n), c(\rho)) \cdot \rho(z_1) \cdot \ldots \cdot \rho(z_n)\, dz_1 \ldots dz_n,$$

and the overall expected utility $E(s)$ can be obtained if we average the above utility $E(\rho)$ over all possible distributions ρ – using the prior distribution on the set of all possible distributions $\rho(z)$:

$$E(s) = \int E(\rho)\, d\rho =$$

$$\int d\rho \int u(s(z_1, \ldots, z_n), c(\rho)) \cdot \rho(z_1) \cdot \ldots \cdot \rho(z_n)\, dz_1 \ldots dz_n.$$

We should select the function $s(z_1, \ldots, z_n)$ for which the expected utility $E(s)$ is the largest possible.

Bayesian approach to robust estimations: main challenges. How do we select a prior distribution? In some cases, we can extract the prior probabilities from the expert – for example, from the bets that the expert is willing to pay when betting on one alternative against another one. However, in practice, such situations are rare, and so, we face a challenge of selecting the appropriate prior distribution.

For different priors, we get different estimates, some are empirically better, some empirically worse. It is known that under certain reasonable assumptions, as the sample size increases, the influence of the prior decreases – in the sense that the posterior distribution tends to the actual one. However, for finite samples, the difference can be drastic.

Selecting an appropriate prior is one of the main challenges of the Bayesian approach. There are many known ways to make this selection, both empirical and theory-based. In the empirical approach, we select a prior that leads to best results. In the theory-based approaches:

- we can select a prior that has the largest possible value of the entropy (see, e.g., [9]), or
- we can select the prior which is invariant with respect to corresponding symmetries, etc.

Of course, whatever theoretical approach we use, we want to make sure that empirically, the resulting method works well. Thus, to select an estimate in our case, let us recall the available empirical evidence.

Robust estimations: empirical data. Most known robust methods come from selecting among the methods optimal for some distribution $\psi_0(x)$, and thus, have the form $\sum\limits_{i=1}^{n} \psi_0(x_i - a) \to \min\limits_{a}$ for an appropriate function $\psi_0(x)$. These are the methods that we will analyze in this paper.

Empirically, the most efficient techniques are the so-called ℓ^p-techniques in which we minimize the sum $\sum\limits_{i=1}^{n} |x_i - a|^p$; see, e.g., [8]. A natural analogue of the sample variance is then the value

$$\hat{V} = \sum_{i=1}^{n} |x_i - \hat{a}|^p.$$

This value is proportional to X^p, so we can estimate the deviation of the random variable x from $\hat{a}$ by the value $\hat{\sigma} = (\hat{V})^{1/p}$.

Remaining problems. While the above estimate $\hat{a}$ (and the related estimates $\hat{V}$ and $\hat{\sigma}$) work well in many practical situations, there is no convincing theoretical explanation for this success. As a result, it is not clear whether the corresponding function $\psi_0(x) = |x|^p$ is indeed the best – or it is simply empirically the best among the few functions that were tried, and a different function $\psi_0(x)$ may be even better.

Another problem is that while we have a good robust version of the sample variance, it is not clear how to transform it into a robust version of sample covariance – and covariance is an important statistical characteristics describing the relation between two random variables.

What we do in this paper. In this paper, we show that both problems can be resolved if we apply an idea which frequently used in statistics – namely, if we use natural symmetries.

Comment. Based on the above analysis, a natural idea would be:

- to use symmetries for selecting a prior, and then
- to use the resulting prior to come up with an appropriate robust versions of sample variance and sample covariance.

What we show in this paper is that for our specific problem, we can simplify this idea and use symmetries to directly derive the corresponding robust functions $s(z_1, \ldots, z_n)$.

2 Natural Symmetries: Main Idea

What are natural symmetries. Numerical values x_i of physical quantities depend on the choice of a measuring unit. If instead of the original measuring unit, we use a new unit which is λ times smaller, then all the numerical values will be multiplied by λ.

For example, if we use centimeters instead of meters, with $\lambda = 100$, then a height of $x = 1.7$ m becomes $x_{\text{new}} = \lambda \cdot x = 170$ cm.

In econometric problems, where money-valued quantities like price, income, profit, etc., are important, similarly, the numerical values of the corresponding money quantities depend on the choice of the monetary unit: the salary in dollars has a different numerical value when translated into Euros.

The numerical value of a quantity also depends on the starting point: e.g., for C and F temperature scales, the starting points are different. However, when we consider the different $x_i - a$ between two values of the same quantity, this difference disappears, and the only natural symmetry is *scaling* $x \to \lambda \cdot x$.

Natural symmetries for utilities. Econometrics is about human economic behavior. As we have mentioned, according to decision theory, rational human behavior can be described in terms of an appropriate utility function.

The utility function can be determined by observing the person's decisions. It is known that this determination is not unique: indeed, if we apply a linear transformation $u \to u_{\text{new}} = \lambda \cdot u + c_a$, with $\lambda > 0$, to the original utility function, then the expected value of the re-scaled utility $E_{\text{new}}(s)$ can be obtained from the expected value $E(s)$ of the original utility by the same linear transformation: $E_{\text{new}}(s) = \lambda \cdot E(s) + c_a$. Linear transformation with $\lambda > 0$ preserves the order: if we had $E(s) > E(s')$, then we have $E_{\text{new}}(s) > E_{\text{new}}(s')$ and vice versa. Thus the order between different alternatives s and s' remains the same – and therefore, both the original utility function u and the re-scaled function u_{new} lead to the same decisions. Thus, the utility is determined uniquely modulo a linear transformation $u \to u_{\text{new}} = \lambda \cdot u + c_a$. So, when we describe the differences $x_i - a$, it also makes sense to consider scalings.

In general, there is no easy way to compare individual utilities; the utility of each person can be independently re-scaled. Thus, if we have several differences $x_i - a$ and $y_i - b$, then it makes sense to consider re-scalings $x_i - a \to \lambda \cdot (x_i - a)$ and $y_i - b \to \mu \cdot (y_i - b)$ for different values λ and μ.

3 Symmetry-Invariance: From Idea to Precise Definitions

What is symmetry-invariance. We select the value a for which the sum $\sum\limits_{i=1}^{n} \psi_0(\Delta_i)$ is the smallest possible, where we denoted $\Delta_i \stackrel{\text{def}}{=} x_i - a$. Thus, we compare the two tuples $\Delta = (\Delta_1, \ldots, \Delta_n)$ and $\Delta' = (\Delta'_1, \ldots, \Delta'_n)$ by the value of the corresponding sum.

If we apply scaling to both tuples Δ and Δ', we get different numerical values Δ_i and Δ'_i. However, these new numerical values describe the same two situations as the original values Δ_i and Δ'_i. It is therefore reasonable to require that the result of this comparison be the same whether we apply the scaling or not, i.e., whether we use the old or the new units to describe the corresponding quantities.

How to describe symmetry-invariance in precise terms. For $n = 2$, symmetry-invariance means, in particular, that if the two tuples Δ and Δ' are equivalent, i.e., if

$$\psi_0(\Delta_1) + \psi_0(\Delta_2) = \psi_0(\Delta'_1) + \psi_0(\Delta'_2),$$

then they should remain equivalent after re-scaling, i.e., we should have

$$\psi_0(\lambda \cdot \Delta_1) + \psi_0(\lambda \cdot \Delta_2) = \psi_0(\lambda \cdot \Delta'_1) + \psi_0(\lambda \cdot \Delta'_2).$$

Monotonicity. If the random variable is located at the value a with probability 1, i.e., if all the values x_i are equal to a, then the minimization of the sum $\sum\limits_{i=1}^{n} |x_i - a|$ should result in $\hat{a} = a$.

In mathematical terms, the tuple $(0, \ldots, 0)$ (which corresponds to $\hat{a} = a$) should have a smaller value of the sum than a tuple $(c, \ldots, c)$ corresponding to any other constant $c = a - \hat{a}$.

This requirement implies that $n \cdot \psi_0(0) < n \cdot \psi_0(c)$, i.e., that $\psi_0(0) < \psi_0(c)$ for all $c \neq 0$.

Differentiability. For simplicity of analysis, we will also assume that the function $\psi_0(x)$ is twice differentiable for $x > 0$.

We can make this assumption without losing generality, since any continuous function can be approximated, with any given accuracy on any given interval, by a twice-differentiable function – for example, by a polynomial.

Sign-invariance. In many physical situations, the sign of the quantity is also chosen arbitrarily: e.g., traditionally we consider the flow of electrons as a negative current, but we could have as well treat it as a positive one.

Because of this, it is reasonable to require that the value of the minimized function not change if we simply change the sign, i.e., that we should have $\psi_0(-x) = \psi_0(x)$ – i.e., that the function $\psi_0(x)$ is even.

Now, we are ready for formulate our first result.

4 First Result: Explaining Empirically Success of Robust ℓ_p Techniques

Definition 1. *We say that an even function $\rho_0(x)$ which is twice differentiable for $x > 0$ is* scale-invariant *if for every Δ_1, Δ_2, and $\lambda > 0$, if*

$$\psi_0(\Delta_1) + \psi_0(\Delta_2) = \psi_0(\Delta_1') + \psi_0(\Delta_2'),$$

then

$$\psi_0(\lambda \cdot \Delta_1) + \psi_0(\lambda \cdot \Delta_2) = \psi_0(\lambda \cdot \Delta_1') + \psi_0(\lambda \cdot \Delta_2').$$

Definition 2. *We say that a scale-invariant function $\rho_0(x)$ is* monotonic *if $\psi_0(0) < \psi_0(c)$ for all $c \neq 0$.*

Definition 3. *We say that two functions $\psi_0(x)$ and $\psi(x)$ are* equivalent *if for all possible tuples Δ and Δ', the condition $\sum\limits_{i=1}^{n} \psi_0(\Delta_i) \geq \sum\limits_{i=1}^{n} \psi_0(\Delta_i')$ is equivalent to $\sum\limits_{i=1}^{n} \psi(\Delta_i) \geq \sum\limits_{i=1}^{n} \psi(\Delta_i')$.*

Proposition 1. *Every monotonic scale-invariant function $\psi_0(x)$ is equivalent to either $|x|^p$ for some $p > 0$ or to $\ln(|x|)$.*

Comment. The function $\psi_0(x) = \ln(x)$ can be viewed as a limit of x^p when $p \to 0$. Indeed, in this case,

$$x^p = \exp(p \cdot \ln(x)) = 1 + p \cdot \ln(x) + o(p).$$

Thus, for small p, minimizing the sum $\sum\limits_{i=1}^{n} |x_i|^p$ is equivalent to minimizing the sum of the logarithms.

If we impose an additional condition that the function $\psi_0(x)$ is continuous for all x, then we only get $|x|^p$.

Proof. Let us consider the case when $\Delta_1' = \Delta_1 + \delta$ and $\Delta_2' = \Delta_2 + k \cdot \delta$ for some small δ and for an appropriate value k. In this case,

$$\psi_0(\Delta_1') = \psi_0(\Delta_1 + \delta) = \psi_0(\Delta_1) + \delta \cdot \psi_0'(\Delta_1) + o(\delta),$$

where $\psi_0'(x)$ denotes the derivative of the function $\psi_0(x)$. Similarly,

$$\psi_0(\Delta_2') = \psi_0(\Delta_2 + k \cdot \delta) = \psi_0(\Delta_2) + \delta \cdot k \cdot \psi_0'(\Delta_2) + o(\delta).$$

Thus, the original equality $\psi_0(\Delta_1) + \psi_0(\Delta_2) = \psi_0(\Delta_1') + \psi_0(\Delta_2')$ takes the form $\psi_0'(\Delta_1) \cdot \delta + \psi_0'(\Delta_2) \cdot k \cdot \delta + o(\delta) = 0$. Dividing both sides by δ, we get

$$\psi_0(\Delta_1) + k \cdot \psi'(\Delta_2) + o(1) = 0.$$

Thus, when $\delta \to 0$, this equality holds for

$$k = -\frac{\psi_0'(\Delta_1)}{\psi_0'(\Delta_2)}.$$

Similarly, the equality

$$\psi_0(\lambda \cdot \Delta_1) + \psi_0(\lambda \cdot \Delta_2) = \psi_0(\lambda \cdot \Delta_1') + \psi_0(\lambda \cdot \Delta_2'),$$

which is obtained after λ-rescaling, implies that

$$k = -\frac{\psi_0'(\lambda \cdot \Delta_1)}{\psi_0'(\lambda \cdot \Delta_2)},$$

for the same value k. Therefore, for every Δ_1, Δ_2, and λ, we have

$$\frac{\psi_0'(\lambda \cdot \Delta_1)}{\psi_0'(\lambda \cdot \Delta_2)} = \frac{\psi_0'(\Delta_1)}{\psi_0'(\Delta_2)}.$$

This equality can be represented in the following equivalent form:

$$\frac{\psi_0'(\lambda \cdot \Delta_1)}{\psi_0'(\Delta_1)} = \frac{\psi_0'(\lambda \cdot \Delta_2)}{\psi_0'(\Delta_2)}.$$

This means that the ratio

$$\frac{\psi_0'(\lambda \cdot \Delta)}{\psi_0'(\Delta)}$$

does not depend on Δ, it only depends on λ. Let us denote this ration by $r(\lambda)$. From

$$\frac{\psi_0'(\lambda \cdot \Delta)}{\psi_0'(\Delta)} = r(\lambda),$$

we conclude that

$$\psi_0'(\lambda \cdot \Delta) = r(\lambda) \cdot \psi_0'(\Delta). \tag{1}$$

Let us consider the case when we first re-scale by λ_2 and then by λ_1. In this case, we have

$$\psi_0'(\lambda_2 \cdot \Delta) = r(\lambda_2) \cdot \psi_0'(\Delta),$$

and thus,

$$\psi_0'(\lambda_1 \cdot \lambda_2 \cdot \Delta) = r(\lambda_1) \cdot \psi_0'(\lambda_2 \cdot \Delta) = r(\lambda_1) \cdot r(\lambda_2) \cdot \psi_0'(\Delta). \tag{2}$$

On the other hand, the same result can be obtained if we re-scale by $\lambda_1 \cdot \lambda_2$:

$$\psi_0'(\lambda_1 \cdot \lambda_2 \cdot \Delta) = r(\lambda_1 \cdot \lambda_2) \cdot \psi_0'(\Delta). \tag{3}$$

Since the left-hand sides of the last two formulas (2) and (3) coincide, the right-hand sides must be equal as well, so we have

$$r(\lambda_1 \cdot \lambda_2) = r(\lambda_1) \cdot r(\lambda_2). \tag{4}$$

The function $r(\lambda)$ is a ratio of two differentiable functions and is, thus, differentiable itself. It is known (see, e.g., [1]) that all differentiable functions that satisfy the above equality (4) have the form $r(\lambda) = \lambda^q$ for some q. Substituting this expression and $\Delta = 1$ into the formula (1), we conclude that $\psi_0'(x) = C_1 \cdot x^p$ for $x > 0$, where $C_1 \stackrel{\text{def}}{=} \psi_0'(1)$.

Integrating, for $q \neq -1$, we get $\psi_0(x) = C \cdot x^p + c_1$ for $p = q+1$ and some C, and for $q = -1$, we get $\psi_0(x) = C_1 \cdot \ln(x) + c_1$. The fact that $\psi_0(x)$ is an even function enables us to get the values for $x < 0$ as $\psi_0(x) = \psi_0(|x|)$. Monotonicity implies that $C > 0$ and $C_1 > 0$, and thus, these functions are indeed equivalent to $|x|^p$ and $\ln(|x|)$.

The proposition is proven.

5 Second Result: Invariant Generalizations of Sample Covariance

Discussion. Let us consider expressions of the type $\sum\limits_{i=1}^{n} f(a_i, b_i)$, where $a_i \stackrel{\text{def}}{=} x_i - a$ and $b_i \stackrel{\text{def}}{=} y_i - b$.

Similar to the standard covariance, we want the expression $f(x, y)$ to change sign when we change the sign of either x or y: $f(-x, y) = f(x, -y) = -f(x, y)$. We also want this expression to be symmetric: $f(x, y) = f(y, x)$.

And, of course, we want the resulting comparison to be scale-invariant, i.e., if $\sum\limits_{i=1}^{n} f(a_i, b_i) = \sum\limits_{i=1}^{n} f(a_i', b_i')$, then $\sum\limits_{i=1}^{n} f(\lambda \cdot a_i, \mu \cdot b_i) = \sum\limits_{i=1}^{n} f(\lambda \cdot a_i', \mu \cdot b_i')$. Similarly to the previous section, in the following definition, we will use the $n = 2$ case of this requirement.

When $x_i = y_i$, we want the sample covariance to be positive. Thus, we arrive at the following definitions.

Definition 4. *We say that a function $f(x, y)$ is a* covariance function *if it is continuous, twice differentiable for $x \neq 0$ and $y \neq 0$, and satisfies the following conditions for all x and y:*

- $f(x, y) = f(y, x)$,
- $f(-x, y) = f(x, -y) = -f(x, y)$, *and*
- $f(x, x) > 0$ *for* $x \neq 0$.

Definition 5. *We say that a covariance function $f(x, y)$ is* scale-invariant *if for every combination of a_i, b_i, $\lambda > 0$, and $\mu > 0$, the equality*

$$f(a_1, b_1) + f(a_2, b_2) = f(a_1', b_1') + f(a_2', b_2')$$

implies that

$$f(\lambda \cdot a_1, \mu \cdot b_1) + f(\lambda \cdot a_2, \mu \cdot b_2) = f(\lambda \cdot a_1', \mu \cdot b_1') + f(\lambda \cdot a_2', \mu \cdot b_2').$$

Proposition 2. *Every scale-invariant covariance function has the form* $f(x,y) = \text{sign(x)} \cdot \text{sign}(y) \cdot (|x| \cdot |y|)^q$ *for some real number* $q > 0$.

Discussion. Thus, as a robust version of estimating covariance, we can take the expression

$$\hat{C} = \frac{1}{n} \cdot \sum_{i=1}^{n} \text{sign(x}_\text{i}) \cdot \text{sign}(y_i) \cdot (|x_i| \cdot |y_i|)^q.$$

If we additionally require that for $y_i = x_i$, the resulting version of sample covariance coincides with the above invariant version of sample variance $\frac{1}{n} \cdot \sum_{i=1}^{n} |x_i|^p$, then we conclude that $q = p/2$. In particular, for the median case $p = 1$, we should thus consider

$$\hat{C} = \frac{1}{n} \cdot \sum_{i=1}^{n} \text{sign(x}_\text{i}) \cdot \text{sign}(y_i) \cdot \sqrt{|x_i| \cdot |y_i|}.$$

Comment. Please note that while the original covariance function $f(x,y) = x \cdot y$ is associative, the new function is, in general, *not* associative. Indeed, for $x, y, z > 0$, we have

$$f(f(x,y),z) = f((x \cdot y)^q, z) = ((x \cdot y)^q \cdot z)^q = x^{q^2} \cdot y^{q^2} \cdot z^q,$$

while

$$f(x, f(y,z)) = f(x, (y \cdot z)^q) = (x \cdot (y \cdot z)^q)^q = x^q \cdot y^{y^2} \cdot z^{q^2},$$

i.e., a different expression.

However, we still have some weaker form of associativity: namely, for every four values x, y, z, and t, the value $f(f(x,y), f(z,t))$ does not change if we permute these four values.

Proof of Proposition 2. Let us first consider transformations with $\mu = 1$, i.e., transformations that do not change y. Then, for a fixed $y > 0$, arguments provided in the proof of Proposition 1 imply that for $x > 0$, we have $f(x,y) = C(y) \cdot x^{p(y)} + c_1(y)$, where the parameters C, p, and c_1 are, in general, dependent on y. By continuity, we get a similar expression for all $x \geq 0$.

Please note that since we explicitly required that the function $f(x,y)$ be continuous for all x and y, we no longer have the logarithm option.

The requirement that $f(-x,y) = -f(x,y)$ implies that $f(0,y) = 0$, so $c_1(y) = 0$ and $f(x,y) = C(y) \cdot x^{p(y)}$. By taking logarithm of both sides, we conclude that

$$\ln(f(x,y)) = p(y) \cdot \ln(x) + \ln(C(y)),$$

i.e., that $\ln(f(x,y))$ is a linear function of $\ln(x)$.

Similarly, we can conclude that $\ln(f(x,y))$ is a linear function of $\ln(y)$. Thus, $\ln(f(x,y))$ is a bilinear function of the two variables $\ln(x)$ and $\ln(y)$. Due to symmetry, we thus have

$$\ln(f(x,y)) = k_0 + k_1 \cdot \ln(x) + k_1 \cdot \ln(y) + k_2 \cdot \ln(x) \cdot \ln(y) \tag{5}$$

for some parameters k_i.

In particular, for $x = y$, we get

$$\ln(f(x,y)) = k_0 + 2k_1 \cdot \ln(x) + k_2 \cdot (\ln(x))^2.$$

We can also consider the case when we take $y = x$ and apply the same rescaling to both variables. In this case, we get $f(x,x) = C \cdot x^p$ for some p, i.e., in logarithm terms,

$$\ln(f(x,x)) = \ln(C) + p \cdot \ln(x). \tag{6}$$

By comparing the expressions (5) and (6), we conclude that $k_2 = 0$. Thus,

$$\ln(f(x,y)) = k_0 + k_1 \cdot \ln(x) + k_1 \cdot \ln(y)$$

and $f(x,y) = \text{const} \cdot (x \cdot y)^{k_1}$. Positivity implies that the constant is positive, and thus, the expression is equivalent to $(x \cdot y)^{k_1}$ for $x \geq 0$ and $y \geq 0$.

By using the equalities $f(-x,y) = f(x,-y) = f(x,y)$, we can extend this expression to all possible values of x and y.

The proposition is proven.

Acknowledgments

We acknowledge the partial support of the Center of Excellence in Econometrics, Faculty of Economics, Chiang Mai University, Thailand. This work was also supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, by an award "UTEP and Prudential Actuarial Science Academy and Pipeline Initiative" from Prudential Foundation, and by the Institoto Politecnico Nacional, Mexico.

References

[1] J. Aczél, *Lectures on Functional Equations and Their Applications*, Dover, New York, 2006.

[2] J. Beirlant, Y. Goegevuer, J. Teugels, and J. Segers, *Statistics of Extremes: Theory and Applications*, Wiley, Chichester, 2004.

[3] B. K. Chakrabarti, A. Chakraborti, and A. Chatterjee, *Econophysics and Sociophysics: Trends and Perspectives*, Wiley-VCH, Berlin, 2006.

[4] A. Chatterjee, S. Yarlagadda, and B. K. Chakrabarti, *Econophysics of Wealth Distributions*, Springer-Verlag Italia, Milan, 2005.

[5] J. D. Farmer and T. Lux (eds.), *Applications of statistical physics in economics and finance*, a special issue of the *Journal of Economic Dynamics and Control*, 2008, Vol. 32, No. 1, pp. 1–320.

[6] P. C. Fishburn, *Utility Theory for Decision Making*, John Wiley & Sons Inc., New York, 1969.

[7] C. P. Gomez and D. B. Shmoys, "Approximations and Randomization to Boost CSP Techniques", *Annals of Operations Research*, 2004, Vol. 130, pp. 117–141.

[8] P. J. Huber and E. M. Ronchetti, *Robust Statistics*, Wiley, Hoboken, New Jersey, 2009.

[9] E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

[10] R. D. Luce and R. Raiffa, *Games and Decisions: Introduction and Critical Survey*, Dover, New York, 1989.

[11] B. Mandelbrot, "The variation of certain speculative prices", *J. Business*, 1963, Vol. 36, pp. 394–419.

[12] B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, San Francisco, California, 1983.

[13] B. Mandelbrot and R. L. Hudson, *The (Mis)behavior of Markets: A Fractal View of Financial Turbulence*, Basic Books, 2006.

[14] N. Markovich (ed.), *Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice*, Wiley, Chichester, 2007.

[15] J. McCauley, *Dynamics of Markets, Econophysics and Finance*, Cambridge University Press, Cambridge, Massachusetts, 2004.

[16] H. T. Nguyen, O. Kosheleva, and V. Kreinovich, Decision making beyond Arrow's 'impossibility theorem', with the analysis of effects of collusion and mutual attraction, *International Journal of Intelligent Systems*, 2009, Vol. 24, No. 1, pp. 27–47.

[17] P. V. Novitskii and I. A. Zograph, *Estimating the Measurement Errors*, Energoatomizdat, Leningrad, 1991 (in Russian).

[18] A. I. Orlov, "How often are the observations normal?", *Industrial Laboratory*, 1991, Vol. 57. No. 7, pp. 770–772.

[19] H. Raiffa, *Decision Analysis*, Addison-Wesley, Reading, Massachusetts, 1970.

[20] S. I. Resnick, *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*, Springer-Varlag, New York, 2007.

[21] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.

[22] S. V. Stoyanov, B. Racheva-Iotova, S. T. Rachev, and F. J. Fabozzi, "Stochastic models for risk estimation in volatile markets: a survey", *Annals of Operations Research*, 2010, Vol. 176, pp. 293–309.

[23] P. Vasiliki and H. E. Stanley, "Stock return distributions: tests of scaling and universality from three distinct stock markets", *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, 2008, Vol. 77, No. 3, Pt. 2, Publ. 037101.