An L1-norm analog of covariance

A.J. Allinger
Design Service Corporation
2013-2025

Abstract: The statistic s = (1/N) SUM (max(min(x, y), min(-x, -y))) is proposed as measure of linear dependence. Three distinct lines of reasoning converge on this result: from the orthogonality of vectors in a non-Euclidean space, by analogy with the standard covariance, and as an equivalence function in fuzzy logic. It is empirically validated to give reasonable results on real-world data.

Background

A fundamental task in statistics is to seek relations between variables as a means of explaining relationships and variations within datasets. For this purpose, the classical covariance statistic, based on the ℓ^2 norm (Euclidean distance), has long been a cornerstone of statistical analysis, providing insights into the linear dependence between variables. This article proposes a statistic rooted in the ℓ^1 norm (Manhattan distance).

The ℓ^2 standard deviation has its ℓ^1 counterpart in the mean absolute deviation. The proposed statistic extends this correspondence by introducing an ℓ^1 counterpart to the covariance, to be called the codeviation.

The following sections will give a derivation of codeviation, point out its significance in non-Euclidean geometry, and study its feasibility for use as an alternative to covariance in a conventional linear discriminant analysis classifier.

Definition

The codeviation may be defined as a scalar-valued function of two vectors.

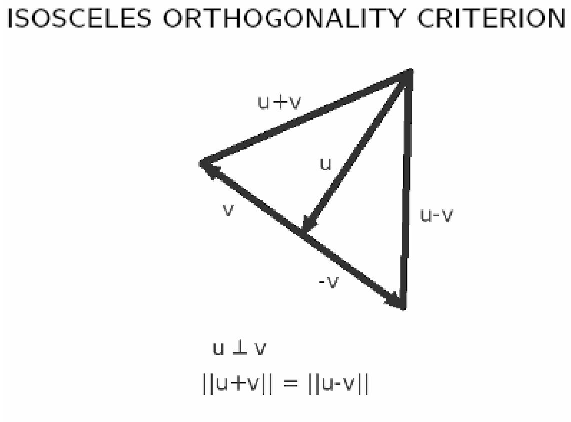
$$\sigma_1(\mathbf{x}, \mathbf{y}) = \frac{1}{2N}\sum_{i=1}^{N}\left(|x_i + y_i| - |x_i - y_i|\right)$$

This function is not an inner product. It is symmetric such that $\sigma_1(\mathbf{x},\mathbf{y})=\sigma_1(\mathbf{y},\mathbf{x})$, but not homogeneous, so that for scalar *a*

$$\sigma_1(a\mathbf{x},\mathbf{y}) \neq a\sigma_1(\mathbf{x},\mathbf{y})$$

The isosceles orthogonality criterion

When two vectors have zero covariance, they are orthogonal. A criterion for orthogonality of vectors **u** and **v** is that the distance from **u** to **v** must equal the distance from **u** to -**v**. (Roberts, 1934)



Applying the ℓ^1 metric, this means that

$$\sum|u_i+v_i| = \sum|u_i-v_i|$$
$$\sum(|u_i+v_i| - |u_i-v_i|) = 0$$

A constant factor of $\frac{1}{2N}$ does not affect the equality, and so the critical expression may be written $\frac{1}{2N}\sum(|u_i+v_i| - |u_i-v_i|)$.

The Gnanadesikan-Kettenring construction

In a study of robust statistics, Gnanadesikan and Kettenring (1972) introduced a generally useful technique for deriving a measure of relation between variables from a measure of dispersion. (page 90)

> A simple idea for estimating the covariance between two variables Y_1 and Y_2 is based on the identity
>
> $$\operatorname{cov}(Y_1, Y_2) = ¼[\operatorname{var}(Y_1+Y_2) - \operatorname{var}(Y_1-Y_2)]$$

> One robust estimator, s_{12}^{*}, of the covariance between Y_1 and Y_2 may, therefore, be obtained from
>
> $$s_{12}^{*} = ¼\left(\hat{\sigma}_1^{*^2} - \hat{\sigma}_2^{*^2}\right)$$
>
> where $\hat{\sigma}_1^{*^2}$ and $\hat{\sigma}_2^{*^2}$ are robust estimators of the variances of $Y_1 + Y_2$ and $Y_1 - Y_2$, respectively....

This is not equal to the proposed statistic. However, the aim of the present article is not an estimate of covariance as such, but rather a low-order analog of covariance. Taking their result as motivational, consider the formula for covariance of the centered variables where $\tilde{x} = x-\mu_x$ and $\tilde{y} = y-\mu_y$

$$\sigma_{xy} = \frac{1}{N}\sum \tilde{x}\,\tilde{y}$$

Compare this to the identity:

$$xy = (½(x+y))^2 - (½(x-y))^2$$

Notice that the covariance is the difference of an agreement and a disagreement. Extending the analogy of mean absolute deviation to standard deviation, change from squaring the quantities to taking the absolute value, and define:

$$\sigma_{xy}^{1} = \frac{1}{2N}\sum \left(|x+y| - |x-y|\right)$$

as a new measure of the relationship between x and y. The factor ½ ensures that for x = y and centered data, the expression reduces to mean absolute deviation:

$$\sigma_1 = \frac{1}{N}\sum |\tilde{x}|$$

Thus, the mean absolute deviation is a special case of the ℓ^1 codeviation, just as the ℓ^2 variance is a special case of the ℓ^2 covariance.

Argument from fuzzy logic

In their work on fuzzy logic, Klir and Yuan (1995) observed that when the variables are the Boolean {0,1}, then logical AND is equivalent to the minimum, and logical OR is equivalent to the maximum. To obtain the Manhattan codeviation, begin with an expression for fuzzy logical equivalence. (Theodoridis & Koutroumbas, 2006)

$$x \Leftrightarrow y = (x \wedge y) \vee (\neg x \wedge \neg y) = \max(\min(x, y), \min(1-x, 1-y))$$

Switching from logical variables to scalar variables, the effects of the min and max operations are unchanged. For negation, values must be reflected across 0 instead of ½, thus ¬x becomes -x rather than 1-x, resulting in:

$$x \Leftrightarrow y \;=\; \max(\min(x,\ y),\ \min(-x, -y))$$

Equivalence of these approaches

The expression derived from covariance and the expression derived from fuzzy logic are identical.

$$\tfrac{1}{2}(|x+y| - |x-y|) \;\equiv\; \max(\min(x,\ y),\ \min(-x, -y))$$

which can be seen by expanding

$$\tfrac{1}{2}(|x+y| - |x-y|) \;=\; \tfrac{1}{2}\left[\left\{\begin{matrix} x+y, & x>-y \\ -x-y, & x<-y \end{matrix}\right\} - \left\{\begin{matrix} x-y, & x>y \\ -x+y, & x<y \end{matrix}\right\}\right]$$

$$= \left\{\begin{matrix} x, & x>-y, & x<y \\ y, & x>-y, & x>y \\ -x, & x<-y, & x>y \\ -y, & x<-y, & x<y \end{matrix}\right\}$$

and again

$$\max(\min(x,\ y), \min(-x, -y)) \;=\; \max\left(\left\{\begin{matrix} x, & x<y \\ y, & x>y \end{matrix}\right\}, \left\{\begin{matrix} -x, & x>y \\ -y, & x<y \end{matrix}\right\}\right)$$

$$= \left\{\begin{matrix} x, & x>-y, & x<y \\ y, & x>-y, & x>y \\ -x, & x<-y, & x>y \\ -y, & x<-y, & x<y \end{matrix}\right\}$$

In the case $x=y$,

$$\tfrac{1}{2}(|x+x| - |x-x|) \;=\; \max(\min(x, x),\ \min(-x, -x)) \;=\; |x|$$

This demonstrates a remarkable relationship between fuzzy logic and ℓ^1 geometry. For the purposes of computation, the min/max formula is to be preferred, since it is not susceptible to subtractive cancellation.

Meaning in Planar Geometry

Angles in the ℓ^1 plane are measured according to arc length along the unit parallelogram. Functions $\sin_t$ and $\cos_t$ may be defined analogously to their ℓ^2 counterparts. (Akça & Kaya,

1997) The unit parallelogram has a circumference of 8 rather than 2π. It can be shown for unit vectors **u** and **v** separated by an angle θ_t that

$$\mathbf{u} \lozenge \mathbf{v} \;=\; \cos_t(\theta_t)$$

where $\mathbf{u} \lozenge \mathbf{v}$ is the codeviation, and $\cos_t$ is the taxicab cosine of θ_t, given by

$\cos_t(\theta_t) = 1 - \left|\frac{\theta_t}{2}\right|, \quad -4 \le x < 4$. The subscript *t* denotes the "taxicab" or ℓ^1 metric. The relationship is not as general as might be hoped for: It applies only to unit vectors, and it has not been determined if the result applies to dimensions higher than N=2.

Correlation and Generalized Distance

For a measure of dependence that is unaffected by scale, define ℓ^1 correlation

$$r_1 \;=\; \frac{\sum |x+y| - |x-y|}{\sum |x| + |y|}$$

For *p* variables, there is a *p*×*p* codeviation matrix S. An expression for ℓ^1 generalized statistical distance is

$$d_1 \;=\; \sum_i \left| \sum_j (S^{-1})_{ij} \; x_j \right|$$

Classification Experiment

To test the effectiveness of ℓ^1 generalized distance, traditional linear discriminant analysis (LDA) was compared to the ℓ^1 variant in classifying 19 data sets from the UCI repository (Lichman, 1987) and one other. The results are disappointing, showing that the ℓ^1 variant is usually quite similar to traditional LDA and generally inferior. The table below result lists the results for classification accuracy, and the difference. The 95% confidence interval was computed by a paired Student's t test, according to the procedure described by Mitchell (1997). 10-fold cross-validation was used, except as noted.

Data set	**variables**	**objects**	**classes**	**L2**	**L1**	**dif**	**interval**	**note**
iris	4	150	3	**0.980**	0.967	-0.013	0.000	*in 5-fold cross-validation
balance-scale	4	625	3	**0.851**	0.838	-0.013	0.085	
Banknote authentication	4	1372	2	**0.976**	0.971	-0.005	0.000	
wilt	5	4839	2	0.943	**0.957**	0.014	0.000	
Wholesale customers	6	440	2	0.848	**0.861**	0.014	0.037	*ignoring the nominal variable "region"
seed	7	210	3	**0.962**	0.933	-0.029	0.025	*in 5-fold cross-validation

E. coli	7	336	8	**0.881**	0.834	-0.048	0.025	
pima-indians-diabetes	8	768	2	**0.778**	0.721	-0.057	0.028	
yeast	8	1484	10	**0.587**	0.580	-0.007	0.020	
fertility	9	100	2	0.830	**0.840**	0.010	0.113	*in 3-fold cross-validation
glass	9	214	6	**0.631**	0.594	-0.037	0.033	*in 5-fold cross-validation, omitting 1 empty class
breast-cancer-wisconsin	9	683	2	**0.960**	0.917	-0.043	0.021	*16 missing data deleted
hockey	10	796	2	0.685	**0.764**	0.079	0.137	*new data set
wine	13	178	3	**0.989**	0.977	-0.011	0.040	*in 5-fold cross-validation
leaf	14	340	30	**0.812**	0.768	-0.044	0.042	*30 non-empty classes. Not to be confused with 'One-hundred species plant leaves' or 'folio' data sets
parkinsons	22	195	2	**0.846**	0.790	-0.056	0.047	*in 5-fold cross-validation
ionosphere	34	351	2	**0.877**	0.795	0.083	0.050	
Landsat satellite	36	6435	6	**0.838**	0.680	-0.158	0.013	
musk	166	6598	2	**0.945**	0.893	-0.051	0.000	
isolet	617	7797	26	**0.944**	0.858	-0.087	0.011	
average				**0.858**	0.828	-0.030		

Discussion

The ℓ^1 codeviation seems to be chiefly of theoretical interest. It is insensitive to extreme values. This makes it robust against outliers, but that is not always an advantage. These measures are equal:

$$\frac{1}{2}(|x+y|-|x-y|)$$
$$\max(\min(x,y),\min(-x,-y))$$
$$\operatorname{sgn}(xy)\cdot\min(|x|,|y|)$$

The last form makes the behavior explicit. If the signs of x and y are the same, the result is positive, and if the signs of x and y are different, the result is negative. The magnitude is the lesser of x and y.

For example, these data sets have the same codeviation:

{(1,1), (2,2), (3,3)} and

{(1,1), (2,2), (3,99)}.

The codeviation has been given a formal derivation as a property of vectors which are orthogonal under the ℓ^1 norm. This argument is buttressed by supporting arguments from fuzzy logic and by comparison to the covariance and the mean absolute deviation, which demonstrates convincingly that codeviation is the correct ℓ^1 measure of linear dependence. It is easy to compute, and can be applied with reasonable success on real-world data. A practical application remains to be discovered.

References

B. D. Roberts, "On the geometry of abstract vector spaces", *Tôhuku Math. Jour*. v.39 pp.49-59, 1934. cited in: Javier Alonso and Carlos Benítez, "Orthogonality in Normed Linear Spaces: A Survey: Part I: Main Properties", *Extracta Mathematicae* v.3 n.1 pp.1-15, 1988.

R. Gnanadesikan and J. R. Kettenring, “Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data”, *Biometrics*, March 1972, Vol. 28, No. 1, pp. 81-124 https://www.jstor.org/stable/2528963

M. Lichman (admin.), UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science, 1987-present

George Klir and Bo Yuan, *Fuzzy Sets and Fuzzy Logic*, Prentice Hall, 1995.

Ziya Akça and Rüstem Kaya, "On the Taxicab Trigonometry", Jour. Of Inst. of Math. & Comp. Sci. (Math. Ser.) v.10 n.3 pp.151-159, 1997

Tom M. Mitchell, *Machine Learning*, New Delhi: McGraw Hill Education, 2013. [1997]

Sergios Theodoridis and Konstantinos Koutroumbas, *Pattern Recognition*, Fourth Edition, Academic Press/Elsevier, 2009.

Appendix: Gram-Schmidt Process under the L1 norm

Let $\mathbf{u} \lozenge \mathbf{v}$ denote $\frac{1}{2}\sum\left(|u_i+v_i| - |u_i-v_i|\right)$. To orthogonalize **v** against **u**, subtract the parallel component from **v**. Seek a scalar c such that

$(\mathbf{v} - c\mathbf{u}) \lozenge \mathbf{u} = 0$

This reduces the problem to solving an equation of a single variable. The function

$$f(c) \;=\; (\mathbf{v} - c\mathbf{u}) \lozenge \mathbf{u} \;=\; 0$$

may be solved by bisection for *c*.

A set of vectors may be made orthogonal by a variant of Gram-Schmidt orthogonalization. The normalization step must be deferred, because

$$(c\,\mathbf{x}) \lozenge \mathbf{y} \;\neq\; c\,(\mathbf{x} \lozenge \mathbf{y})$$

so normalizing the vectors destroys orthogonality. However, an iterative method may be employed, alternately normalizing a vector, then subtracting parallel components.

Unlike in Euclidean geometry, the shortest path from a point to a hyperplane is not orthogonal to the hyperplane. This means that making a vector **b** orthogonal by the ℓ1 isosceles criterion to a set of vectors A does not lead to a solution to the linear system $\mathrm{A}\,\mathbf{x} = \mathbf{b}$ in the sense of minimizing the residual $\sum|\mathrm{A}\,\mathbf{x} - \mathbf{b}|$.

Existence and uniqueness of the orthogonal vector

From the orthogonality criterion, vector $\mathbf{v}' = \mathbf{v} - c\,\mathbf{u}$ will be orthogonal to **u** when

$$(\mathbf{v} - c\,\mathbf{u}) \lozenge \mathbf{u} \;=\; 0$$

$$f(c) \;=\; \frac{1}{2}\sum\; |(v_i - cu_i) + u_i| \;-\; |(v_i - cu_i) - u_i| \;=\; 0$$

To show that the function is monotonic, differentiate with respect to *c*. Expanding the absolute values gives

$$\frac{\partial f}{\partial c} \;=\; \frac{1}{2}\sum \begin{Bmatrix} -1, & v_i - cu_i + u_i < 0 \\ +1, & v_i - cu_i + u_i > 0 \end{Bmatrix} (-u_i) \;-\; \begin{Bmatrix} -1, & v_i - cu_i - u_i < 0 \\ +1, & v_i - cu_i - u_i > 0 \end{Bmatrix} (-u_i)$$

Since terms where u_i is zero vanish, it may be taken $u_i \neq 0$, and

$$\frac{\partial f}{\partial c} \;=\; \frac{1}{2}\sum \left\{\begin{matrix} -1, & c>\dfrac{v_i}{u_i}-1\,, & u_i>0 \\ -1, & c<\dfrac{v_i}{u_i}-1\,, & u_i<0 \\ +1, & c<\dfrac{v_i}{u_i}-1\,, & u_i>0 \\ +1, & c>\dfrac{v_i}{u_i}-1\,, & u_i<0 \end{matrix}\right\} u_i \;-\; \left\{\begin{matrix} -1\,, & c>\dfrac{v_i}{u_i}+1\,, & u_i>0 \\ -1\,, & c<\dfrac{v_i}{u_i}+1\,, & u_i<0 \\ +1\,, & c<\dfrac{v_i}{u_i}+1\,, & u_i>0 \\ +1\,, & c>\dfrac{v_i}{u_i}+1\,, & u_i<0 \end{matrix}\right\} u_i$$

By ordering these conditions according to *c*,

$$\frac{\partial f}{\partial c} \;=\; \frac{1}{2}\sum \left\{\begin{matrix} -|u_i|\,, & c>\dfrac{v_i}{u_i}-1 \\ |u_i| & c<\dfrac{v_i}{u_i}-1 \end{matrix}\right\} \;-\; \left\{\begin{matrix} -|u_i|, & c>\dfrac{v_i}{u_i}+1 \\ |u_i|, & c<\dfrac{v_i}{u_i}+1 \end{matrix}\right\}$$

$$\frac{\partial f}{\partial c} \;=\; \sum \left\{\begin{matrix} 0\,, & c<\dfrac{v_i}{u_i}-1 \\ -|u_i|\,, & \dfrac{v_i}{u_i}-1<c<\dfrac{v_i}{u_i}+1 \\ 0\,, & \dfrac{v_i}{u_i}+1<c \end{matrix}\right\}$$

The derivative with respect to *c* is negative or zero for all *c*. Therefore the function is monotonically decreasing, although not strictly so.

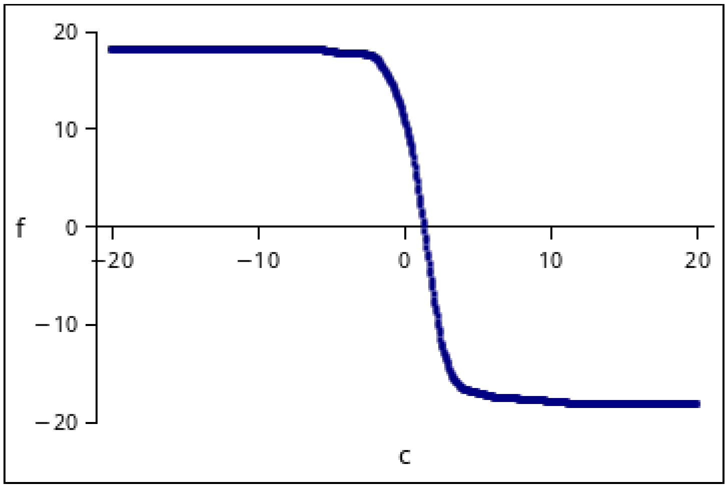
For sufficiently large negative values of *c*, the quantities inside the absolute value bars are dominated by the cu_i term and the expression simplifies

$$\lim_{c\to-\infty} f(c) \;=\; \frac{1}{2}\sum \left\{\begin{matrix} (v_i-cu_i+u_i)-(v_i-c\,u_i-u_i), & u_i>0 \\ -(v_i-cu_i+u_i)+(v_i-c\,u_i-u_i), & u_i<0 \end{matrix}\right\} \;=\; \sum |u_i|$$

For sufficiently large positive values of *c*, the absolute values again simplify

$$\lim_{c\to\infty} f(c) \;=\; \frac{1}{2}\sum \left\{\begin{matrix} -(v_i-cu_i+u_i)+(v_i-c\,u_i-u_i), & u_i>0 \\ (v_i-cu_i+u_i)-(v_i-c\,u_i-u_i), & u_i<0 \end{matrix}\right\} \;=\; -\sum |u_i|$$

For nonzero **u**, the function *f(c)* is continuous, monotonic, and has values less than and greater than zero. Its slope is zero only at its extreme values. Therefore the equation $f(c) = 0$ must have a unique solution.



A graph of a typical function *f(c)* illustrates the behavior.