A Lot of Data

Kent Johnson[†]

q1 This article encourages the use of explicit methods in linguistics by attempting to estimate the size of a linguistic data set. Such estimations are difficult because redundant data can easily pad the data set. To address this, I offer some explicit operationalizations of the data and their features. For linguistic data, negative associations do not indicate true redundancy, and yet for many measures they can be mathematically impossible to ignore. It is proven that this troublesome phenomenon has positive Lebesgue measure and is monotonically increasing and that these two features hold robustly in four different ways.

1. Introduction. Studying how evidence is produced and related to theory is an important part of developing and maintaining a discipline. Fields such as psychology, economics, biology, and chemistry have tenured appointments, conferences, societies, and journals (e.g., *Psychometrica*, *Econometrica*, *Biometrika*, *Journal of Chemometrics*) dedicated to the study and improvement of their methods.

Attention to methods gets increasingly important as matters get more complex and are less well developed. For example, the methods of some parts of economics cry out for attention more so than those of some parts of physics. This is not only because physics is an older discipline and has had time for its methods to mature but also because economics often deals with highly complicated phenomena laden with enormous uncertainties. In such situations, little headway is made without using (and studying, improving, etc.) explicit, typically quantitative, methods for relating evidence to theory.

Whatever else theorizing in linguistics is, it is complex (even if human language is built from just a few relatively simple structural items). Nearly all currently active linguistic projects involve complicated, untamed, and

[†] To contact the author, please write to: Department of Logic and Philosophy of Science, 3151 SSPA, University of California, Irvine, Irvine, CA 92697-5100; e-mail: johnsonk@uci.edu.

Philosophy of Science, 78 (December 2011) pp. 000–000. 0031-8248/2011/7805-0042$10.00

uncharted regions of human language. Moreover, linguistics is a relatively young discipline. Thus, linguistics would seem to be yet another field that would benefit from the incorporation of explicit methods that themselves can be studied, criticized, improved, and so forth. In this article I explore this issue. In section 2, I consider a very basic issue: assessing the size of a given linguistic data set. In section 3, I take an initial step toward explicitly addressing this matter, by suggesting an operationalized characterization of an expression type. In section 4, I consider how this operationalization might help with our assessment. Structurally, the situation is similar to familiar cases involving multivariate data sets, with one exception: the irrelevance of negative associations. This exception, I argue, changes matters considerably. A theorem, proven in the appendix, shows that these negative correlations cannot simply be ignored and that this undesirable phenomenon is quite robust. I conclude in section 5.

The size of one's data set is important; however, my broader goal is to promote greater methodological research in linguistics. Currently, linguists routinely assess large and diverse bodies of evidence almost entirely by informal, holistic, "expert judgments." Famously, however, in situations vastly less complex than linguistics, expert judgments are much less reliable than they are typically assumed to be (e.g., Dawes 1979; Johnson 2009). q2

2. Target Issue: Individuating Linguistic Data. I turn now to the basic matter of estimating the size of a linguistic data set. Such a notion is employed, implicitly or explicitly, whenever one judges that a journal article, research project, presentation, and so forth, used "a lot" of data, "not enough" data, a "wide range" of "diverse" data, "more" data than a rival hypothesis is founded on, and so on. Such comments are clearly meant to summarize certain aspects of the evidence and to be part of the overall normative judgment regarding the theory. If "a lot" of data were used, how much was that? More than 15 data points? What were those data points, anyway, and what makes any two of them (if there is more than one) distinct from one another?

The amount of data used in the construction or confirmation of a theory is an utterly fundamental matter across the empirical sciences, especially in those areas where there is great uncertainty and complexity. I will largely take it for granted that it is of similar importance to linguistics to have some (explicit) means for estimating the size of a data set. In general, it is hard to see how progress could be made toward an explicit linguistic methodology if one cannot even say how large one's sample is. (For example, not knowing the size of one's sample severely limits any further analysis or assessment in most experimental designs.)

To begin, let us consider the general type of evidence that mainstream linguists typically employ in actual practice. Simplifying, we can assume the data are linguistic expressions. They are not the psychological data structures realized (under appropriate idealizations, etc.) inside speakers' heads (e.g., Chomsky 1986, 25–26). Instead, they are expression types, such as the type of which the physical inscription "the cat is on the mat" is an instance. I will assume that such types can be unproblematically individuated.

In asking how much data were employed by a given linguistic project, a natural first thought would be to just count how many expression types were offered in the project. For example, consider a linguist developing a theory of control, illustrated by sentence 1a.

1a. Sue wants to win.
1b. Sue^i wants [PRO^i to win].

Control structures are noteworthy in that they contain a clause (here, *to win*) that does not overtly contain a subject. However, in sentence 1a, the subject of this lower clause can only be *Sue*. This and other such phenomena have led linguists to posit a phonologically null but syntactically and semantically active element, PRO, as the subject of the lower clause (cf. 1b). PRO is "controlled" by *Sue*, thus mechanically and automatically determining the correct interpretation of the sentence. Suppose that a linguist develops a theory concerning a fragment of human language that includes control structures, which focuses on the phenomenon of "partial control," illustrated in sentence 2a:

2a. The chair wanted to meet on Tuesday.
2b. Sue wanted to meet on Tuesday.
2c. The chair hoped to meet on Tuesday.

Sentence 2a is noteworthy in that its most natural interpretation is that the chair wanted a group of people, only one of whom is the chair herself, to meet on Tuesday. Thus, *the chair* only partially determines the subject of the lower clause (i.e., the value of PRO; Landau 2000).

The central difficulty with determining how much evidence is used in a theory is that new, redundant data are all too easy to generate. For example, one gathers no new evidence for a theory by adding sentence 2b to a data set containing sentence 2a. Sentences 2a–2b are simply too relevantly similar to count as distinct data points. Of course, sentences 2a–2b have different grammatical properties; for example:

3a. *Crazy old the chair wanted to meet on Tuesday.
3b. Crazy old Sue wanted to meet on Tuesday.

However, it is unlikely that such a difference would be relevant to a theory of control.

What about sentence 2c? The extent to which this expression provides the theory with something new may depend on the nature of the particular theory. As sentences 4 and 5 suggest, sentences 2a and 2c are somewhat different as regards the licensing of possible complement structures:

4a. The chair wanted the committee to meet on Tuesday.
4b. *The chair hoped the committee to meet on Tuesday.

5a. *The chair wanted the committee would meet on Tuesday.
5b. The chair hoped the committee would meet on Tuesday.

However, sentence 6 shows that they behave similarly in other respects, so—depending on the details of the theory at hand—there may be some amount of (relevant) redundancy present.

6a. It was wanted for the committee to meet on Tuesday.
6b. It was hoped for the committee to meet on Tuesday.

These examples show how a pair of expressions may exhibit some degree q3 of redundancy. Importantly, however, redundancy is a holistic affair, potentially involving most or all of the data set. For example, if expressions 7a and 7b are already in the data set, then 7c adds no new information. Similarly, depending on the nature of one's project, expressions 7d and q4 7e may also be highly redundant with (7a–7b) collectively, although not so much with either one individually.

7a. Sue crashed while PRO biking.
7b. Kim wants PRO to be recognized *t*.
7c. Sue crashed while PRO biking, and Kim wants PRO to be recognized *t*.
7d. Kim wants PRO to be recognized *t* while PRO biking.
7e. Kim went unnoticed while PRO wanting PRO to be recognized *t*.

In sum, an expression may exhibit some degree of redundancy with respect to other elements of a data set. Any such redundancy will always be relative to both a given data set and the particular theory at hand. I will call this phenomenon the "problem of redundancy." Since linguistic data are used in the construction of a theory as well as its confirmation, what properties of an expression are relevant to an assessment of redundancy—for example, does it license direct objects with controlled complements 4 or tensed clausal complements 5—may not be known a priori. Instead, determining the relevant properties of expressions may be a matter of a "bootstrap" procedure (Glymour 1980) as the theory is developed over time.

The problem of redundancy shows that determining whether one has used "a lot of data" in the construction/confirmation of a linguistic theory is a highly nontrivial matter. If redundancy is not addressed, then there is no difference between any finite data set and a countably infinite one (e.g., augment expression 2a–2b with *DP wanted to meet on Tuesday*, for every DP). If there is no difference between finite and infinite data sets, then all distinctions regarding the amount of data used collapse.

I suspect that some linguists would see the problem of redundancy as not particularly serious. I also suspect that some linguists feel that as they are constructing or evaluating a theory, they notice such correlations in the behavior of the data and take this into account in an implicit and intuitive manner. In the next two sections, I will attempt to render explicit this purported practice and analyze it.

3. Operationalizing Theoretical Types of Expressions and Their Properties. In linguistic theorizing, we want to relate the evidence of concrete expressions to the theoretical models that produce psychological expressions (e.g., Chomsky 1986, 25–26). However, we have seen that redundancy q5 threatens to undermine, partially or completely, one of the most basic features of a body of evidence, namely, its size. In this section, I outline a strategy for operationally characterizing the relevant theoretical types of expressions and the relevant properties of these types. Then, in section 4, I consider the problem of redundancy explicitly.

Two key factors motivate our operationalization. First, as sentences 2a–2b and 6 showed, the expressions used as evidence have a great deal of structure, only some of which is relevant to a given project (since we currently have only very partial theories of human language). More generally, a project on control is not likely to be concerned with the highly detailed structures sketched in expressions 8a–8b. Instead, it is more likely to focus on certain schematic aspects of structure that are hypothesized to be those aspects relevant to control, similarly sketched in 8c:

8a. [*TP* [*DP* [*D′* the chair]] [*T′* wanted [*VP* [*CP*[*TP* PRO [*T′* to [*VP* PRO [*V′* meet [*PP* on Tuesday]]]]]]]]]

8b. [*TP* [*DP* Sue] [*T′* wanted [*VP* [*CP*[*TP* PRO [*T′* to [*VP* PRO [*V′* meet [*PP* on Tuesday]]]]]]]]]

8c. [*TP* DP [*T′* Verb {F_1,. . .,F_n} [*VP* [*CP*[*TP* PRO [*T′* to [*VP* PRO [*V′*]]]]]]]]

The structures in expression 8 are merely illustrative—different theories q6 would posit different structures. However, they show that part of analyzing linguistics is isolating those structural elements that are/are not relevant to a given project. Because both 2a and 2b have the structure given in expression 8, they count as the same type, and so adding 2b to a data set that contains 2a should not increase the amount of evidence considered.

TABLE 1. Hypothetical Organization of Expression Types.

Expression Type	P_1	P_2	P_3	. . .	P_k
The chair wanted. . .	1	0	1		0
Susan wanted. . .	1	0	1		0
The chair hoped. . .	0	1	1		1
⋮	⋮	⋮	⋮		⋮
John tried. . .	0	0	0		1

Note.—Hypothetical explicit organization of nominally distinct expression types, with relevant properties. The first three properties correspond to expressions 4–6 above, namely, the ability to accept direct object control, finite 0-complement, and *for*-IP, respectively. The final *k*th property concerns the ability to have a complement of the form *for*-DP, as in **The chair wanted for a short meeting* and *The chair hoped for a short meeting*.

Of course, we do not have immediate access to the structures in expression 8; determining them is a big part of what linguistics is all about. Here too, linguists often bootstrap into increasingly better theories: the linguist uses current theory to hypothesize some relevant structure thought to be shared by some expressions and thus individuates the data and explores the results so as to arrive at a new (hopefully improved) theory. q7 This new theory is then part of the input that allows the hypothesis of what the relevant types of expressions are that then allows for a new individuation of the data, and so on.

I call theoretical items represented by expression 8c "(theoretical) types." The difficult task of determining what structure a theoretical type contains belongs mainly to linguists. From the present perspective, however, we can assume that such concrete decisions for particular theories have been made.

The second key fact is that we use expressions like those in sentences 4–6 to explore the nature of theoretical types. In that sense, sentences 4–6 represent what I call "properties" of the types. For example, suppose that the relevant expression types differentiate 2a and 2c—that is, the different kinds of relevant structures that the theory posits are fine-grained enough that one of them applies to 2a and a different one applies to 2c. In such a case, sentence 4 tells us that the type represented with 2a allows direct objects, but the one that 2c represents does not. Similarly, sentence 5 tells us that of the two, only the 2a type allows tensed clauses.

Eventually, in the simplest case, the linguist would posit *n* theoretical types each examined from the perspective of *k* properties. The data can then be represented by an $n \times k$ matrix, where the *ij*th element contains a 1 if the *i*th expression type possesses the *j*th property, and a 0 otherwise (or by some other coding system if greater discrimination is used). The q8 outcome of this stage of the analysis is represented by table 1: The *i*th expression type can be operationally defined as the *i*th row of this matrix, and the *j*th property can similarly be operationally defined as the *j*th column. Thus, a theoretical type can also be (operationally) thought of as an equivalence class of expressions all of whose members behave identically across the properties that the linguist has identified as relevant. How reasonable these operationalizations are is largely an empirical matter, to be addressed mainly by linguists.

4. The Problem of Negative Correlations. In section 2, we saw that distinguishing various data points in a linguistic data set is neither trivial nor an all-or-nothing affair. We need to recognize that there can be redundancy in our data sets and that a nominally new addition to a data set may not contribute all that much novel information. Moreover, we saw that the redundancy of an expression may be spread across multiple elements of the data set. We would like to somehow account for this redundancy so as to estimate the true size of our data set.

Fortunately, the issue to which we have reduced our linguistic problem is a familiar one. Intercorrelations in multivariate data sets are a bread-and-butter issue for many fields, and there are numerous statistical techniques for dealing with them. Because we now have explicit representations (from sec. 3) of the relevant aspects of the expressions under study, we can make an assessment of redundancy by considering the amount of "overlap" between the various pairs of theoretical types of expressions. A natural first step would be to consider the correlations between the various types, represented by the *n* row vectors described above. (In a complete analysis, the *k* column vectors would be analyzed as well, as there could easily be unwanted redundancy in the relevant properties of the theoretical types. For simplicity's sake, I ignore this matter here.) Such correlations are often the inputs to various techniques for treating redundancy (e.g., Jolliffe 2010).

Before turning to these techniques, however, one final empirical consideration must be addressed. The relevant type of redundancy here is only that which corresponds to "positive" correlation. That is, we only care about the extent to which two (operationalized) expression types exhibit the same behavior. In particular, the extent to which they are negatively associated is the same, for the present estimation purposes, as having no association at all. For example, across the right range of properties, pronouns and anaphors are highly negatively correlated. But this high correlation does not suggest that they are redundant; rather, it shows how importantly different they are.

Because negatively correlated theoretical types share no structural redundancy, we might try to disregard them by treating the relevant theoretical types as independent and hence uncorrelated (i.e., having a correlation of zero). At first, this idea seems simple and natural, and it does work in two or three dimensions. Unfortunately, with more than three expressions, there is no guarantee that this strategy will work: the resulting set of correlations often cannot be simultaneously realized. A few remarks may give a sense of why this is so.

When $n > 3$, there are $\binom{n}{2} = n[(n-1)/2] > n$ possible correlations. Thus, in fixing these correlations, we have $\binom{n}{2}$ equations of the form $r(\mathbf{x}_i, \mathbf{x}_j) = r_{ij}$ (where r is the correlation function), but only n (vectors of) variables to work with. Thus, there is no guarantee that a solution to these equations will always exist.

More can be said. Let $p = \binom{n}{2}$, and for any set of n vectors, let $\mathbf{c} = \langle c_1, \ldots, c_p \rangle$ be the correlations between each pair of expression types, that is, each pair of kk-dimensional vectors. (I assume some canonical ordering of all sets of vectors and of the elements of $\mathbf{c}$.) The "nonnegative variant" of $\mathbf{c}$ is the sequence that is just like $\mathbf{c}$ except that the negative correlations have been replaced with zeroes. Let us say that $\mathbf{c}$ is *bad* if no set of n vectors could have the correlations of its nonnegative variant. Finally, let $B = \{\mathbf{c}{:}\mathbf{c} \text{ is bad}\}$ We then have the following robustness theorem, proven in the appendix:

> **Theorem 1**. (i) $B \neq \emptyset$, and more generally, (ii) within $\mathbb{R}^p$, B has positive Lebesgue measure; (iii) B is monotonically increasing in n. Furthermore, facts ii and iii: iv do not depend on any probability distributions, v do not depend on the operational characterization of expression types given in section 3, vi hold for any choice of inner product used as the measure of association (of possibly scaled data), and vii have corresponding versions for other means of reducing redundancy, such as singular value decomposition, that do not obviously depend on pairwise associations.

The results just listed apply immediately to the position described at the end of section 3. There I imagined a linguist saying that the problem of redundancy can be dealt with by just noticing and keeping track of the correlations between the data, and accounting for this accordingly, as part of a holistic expert judgment. However, such a strategy simply will not work for a very broad range of data sets and a similarly broad range of measures of association. The reason why has nothing to do with a linguist's expertise; rather, it is a fact about what is mathematically possible.

5. Conclusion. Where does this leave us? We started off by isolating a fundamental issue regarding the evaluation of linguistic theories, counting data. Along the way we helped ourselves to many simplifications, finally ending up with a task vastly simpler than what is routinely performed in linguistic inference. Unfortunately, that simple task is often impossible.

The persistent adherent of informal, holistic expert judgment faces the challenge of demonstrating the accuracy of this method. This may be difficult, as tasks generally do not get easier as they get more complex, and the simple task is not possible. Given the remarkable weakness of such holistic expert judgments in simpler inferential situations, I am not optimistic about these prospects. A more promising strategy, I submit, is to view the present problem as encouragement to follow the other sciences, and begin taking the difficult steps necessary to develop explicit linguistic methods. A first step would be to find a way to estimate the size of a data set, which requires some metric of association to factor out redundancy. I believe common metrics like the correlation are inappropriate for linguistics, which is good since we have just seen that they do not work. But what appropriate metric(s) will work is only one of many questions yet to be addressed.

Appendix: Proof of Theorem 1

Proof of i. Recall that our operationalized initial data were a set x_1, x_n of vectors in R^k. We can rescale our data, setting $z_{ij} = [(x_{ij} - \bar{x}_i)/\sqrt{k}s_i]$, where x_{ij} is the jth element of x_i, and $\bar{x}_i$ and s_i are the mean and standard deviation of $\mathbf{x}_i$. Then the (Pearson) correlation coefficient is given by

$$r_{ij} = \frac{\mathrm{Cov}\,(\mathbf{x}_i,\ \mathbf{x}_j)}{s_i s_j} = \frac{1}{k}\frac{\sum_{h=1}^{k}(x_{ih} - \bar{x}_i)(x_{jh} - \bar{x}_j)}{s_i s_j} = \sum\nolimits_{h=1}^{k} z_{ih} z_{jh} \qquad \text{(A1)}$$

$$= (\mathbf{z}_i,\ \mathbf{z}_j).$$

Thus, the correlation between x_i and x_j is also the usual inner product between the standardizations z_i and z_j. Suppose, for example, $n = 4$, and that the four vectors are related as in (A2a). The nonnegative variant of (A2a) is then given in (A2b):

$$\begin{aligned} a.\quad & r_{12} = -.4, \quad r_{13} = .8, \quad r_{14} = .1, \\ & r_{23} = .1, \quad r_{24} = .8, \quad r_{34} = .6. \\ b.\quad & r_{12} = 0, \quad r_{13} = .8, \quad r_{14} = .1, \\ & r_{23} = .1, \quad r_{24} = .8, \quad r_{34} = .6. \end{aligned} \qquad \text{(A2)}$$

Regardless of the nature of the original x_i, we now show that no set of four vectors $z_1, \ldots, z_4$ can have the correlations in (A2b). The proofs below follow easily from some well-known results in matrix analysis, which are covered in many standard textbooks (e.g., Horn and Johnson 1985).

To begin, consider the correlation matrices for (A2a) and (A2b), where the *ij*th entry of (A3a) corresponds to the correlation between x_i and x_j:

$$a.\begin{pmatrix} 1 & -.4 & .8 & .1 \\ -.4 & 1 & .1 & .8 \\ .8 & .1 & 1 & .6 \\ .1 & .8 & .6 & 1 \end{pmatrix} \quad b.\begin{pmatrix} 1 & 0 & .8 & .1 \\ 0 & 1 & .1 & .8 \\ .8 & .1 & 1 & .6 \\ .1 & .8 & .6 & 1 \end{pmatrix}. \qquad \text{(A3)}$$

For $n = 4$, $p = 6$; fix some unproblematic bijection f between $\mathbb{R}^p$ and the symmetric $n \times n$ matrices whose diagonal elements are uniformly 1 and whose off-diagonals are the p elements of the argument. Thus, if **x** and **y** are the vectors corresponding to the (ordered) lists in (A2a) and (A2b), then $f(\mathbf{x}) = (a)$ and $f(\mathbf{y}) = (b)$. Let d be a function from $n \times n$ matrices to $n \times n$ matrices such that $d(M)$ is exactly like M, except that any negative entries in M have been replaced with zeros. Thus, $d(M)$ is then the "nonnegative variant" of M. Thus, $d(f(\mathbf{x})) = (b)$ is a function from a set of correlations to the correlation matrix of its nonnegative variant.

By equation (A1), the *ij*th entry of equation (A3a) is also the inner product of $\mathbf{z}_i$ and $\mathbf{z}_j$. Suppose for a moment that equation (A2b) is a possible set of correlations. Then equation (A3b) is a Gram matrix, that is, a matrix whose *ij*th entry is the inner product between two vectors, for some fixed set of n vectors. A Gram matrix G is positive semidefinite (PSD), i.e., a symmetric $n \times n$ matrix G such that for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}^T G \mathbf{x} \geq 0$. Importantly, all n eigenvalues of a PSD matrix are real and nonnegative. The converses of these implications hold as well, meaning that G is a Gram matrix if and only if (iff) it has no negative eigenvalues. The smallest eigenvalues of (A3a) and (A3b) are .039 and -0.062, respectively. Thus, there exist four vectors $\mathbf{z}_1, \ldots, \mathbf{z}_4$ such that (A2a) is their correlation matrix; no such vectors have (A3b) as their correlation matrix. Thus, $B \neq \emptyset$. QED.

Proof of ii. The eigenvalues of a square matrix are a continuous function of the matrix's components. Therefore, for any $\mathbf{c} \in B$, there is an open ball $D \in \mathbb{R}^p$ of radius $\varepsilon > 0$, centered at **c**, such that for any $\mathbf{y} \in D$: $\mathbf{y} \in [-1, 1]^p$, and $f(\mathbf{y})$ is PSD, but $d(f(\mathbf{y}))$ is not. To see this notice that there are open balls E, F, G (of $\mathbb{R}^p$), centered at **x**, such that $\mathbf{y} \in E$ iff the smallest eigenvalue of $f(\mathbf{y})$ is nonnegative,

$\mathbf{y} \in F$ iff the smallest eigenvalue of $d(f(\mathbf{y}))$ is negative, and $G \subseteq (-1, 1)^p$. Let $D = E \cap F \cap G$. Clearly, $D \subseteq B$. Thus, B has positive Lebesgue measure.

Proof of iii. Pick any $\mathbf{c} \in B$, and let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be any n vectors with correlation matrix $f(\mathbf{c})$. Pick any $\mathbf{x}_{n+1} \in \mathbb{R}^n$, and let N be the correlation matrix of $\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}_{n+1}$. Clearly, N is PSD. Let λ and μ be the smallest eigenvalues of $d(f(\mathbf{c}))$ and $d(N)$, respectively. Since $\mathbf{c} \in B$, $\lambda < 0$. But by the interlacing theorem for bordered matrices, $\mu \leq \lambda$, and so $\mu < 0$. Thus, Lebesgue measure is preserved when we move to the set B' of bad matrices for $\mathbb{R}^{p+n}$, and any "new" regions of bad correlations will only increase the size of B'. QED.

Proof of iv and v. The discussion so far, while employing some techniques widely used in statistics, has been purely algebraic. No use of probability distributions, implicit or otherwise, has been made; this proves iv. Moreover, no essential use was made of the operationalized notion of an expression type from section 3. Any means of comparing the basic evidential units of linguistic theorizing can establish the same results, provided only that they ultimately determine a Gram matrix. This yields v. QED.

Proof of vi. Pick any inner product $\langle \cdot , \cdot \rangle$, and any $\mathbf{c} \in B$; let $f(\mathbf{c}) = C$. There exist vectors $\mathbf{z}_1, \ldots, \mathbf{z}_4$ such that $C = Z^T Z$, where $Z = [\mathbf{z}_1, \ldots, \mathbf{z}_4]$ is the $k \times 4$ matrix composed of the $\mathbf{z}_i$s. Since $\langle \cdot , \cdot \rangle$ is an inner product, $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T P \mathbf{y}$, for some $k \times k$ positive definite P. Thus, $G = X^T P X$ is the Gram matrix for $X = \mathbf{x}_1 \ldots \mathbf{x}_n$. Since P is positive definite, there exists a nonsingular $k \times k$ Q such that $P = Q^T Q$. Since Q is nonsingular, there exists a $k \times 4$ R such that $QR = Z$. Thus, $C = Z^T Z = (QR)^T QR = R^T Q^T QR = R^T PR$. Since $d(C)$ has a negative eigenvalue, it is not the Gram matrix with respect to any inner product, including $\langle \cdot , \cdot \rangle$. Since $\langle \cdot , \cdot \rangle$ is a continuous function of the components of its two input vectors, versions of ii and iii follow for $\langle \cdot , \cdot \rangle$.

It is worth noting that many measures of association can be represented as an inner product. For example, covariance is just a rescaling of the correlation, and proportions of agreement can be represented by setting $x_{ij} = 1/\sqrt{k}$ if the ith element possesses the jth property, and $x_{ij} = -(1/\sqrt{k})$ otherwise. In this case, $\langle \mathbf{x}, \mathbf{y} \rangle \in [-1, 1]$ is positive when **x** and **y** agree on most properties, 0 when they are evenly split, and negative when they disagree on most properties. (In general, we may assume that the vectors **x** and **y** are scaled so that they are unassociated iff $\langle \mathbf{x}, \mathbf{y} \rangle = 0$.) QED.

Proof of vii. Finally, it might be thought that we can avoid these problems by moving from the restrictive case of square, symmetric ($n \times n$) matrices to the more general case of rectangular ($n \times k$) matrices, which is the form of our original data set. Techniques such as the singular value decomposition are commonly used to eliminate redundancy by operating directly on the data matrix, not on its correlation matrix. Perhaps we could decompose some minor $n \times k$ variant of our original data in a way that ignores the negative correlations. The results above show that this is impossible. If a set of correlations is bad, then no vectors will collectively realize the correlations of its denegativized variant. In particular, no vectors in R^k will do this. Thus, there simply does not exist an appropriate $n \times k$ matrix, made up of n such vectors, to decompose. QED.

REFERENCES

Chomsky, N. 1986. *Knowledge of Language*. Westport, CT: Praeger.

Dawes, R. 1979. "The Robust Beauty of Improper Linear Models in Decision Making." *American Psychologist* 34: 571–82.

Glymour, C. 1980. *Theory and Evidence*. Princeton, NJ: Princeton University Press.

Horn, R. A., and C. R. Johnson. 1985. *Matrix Analysis*. Cambridge: Cambridge University Press.

Johnson, K. 2009. "The Need for Explicit Inferential Methods in Linguistics." In *Language and Linguistics Emerging Trends*, ed. C. R. Dreyer 193–208. New York: Nova.

Jolliffe, I. 2010. *Principal Component Analysis*. 2nd ed. New York: Springer.

Landau, I. 2000. *Elements of Control: Structure and Meaning in Infinitival Constructions*. Dordrecht: Kluwer.

QUERIES TO THE AUTHOR

q1. Au: Abstract: Changed “This article motivates using explicit methods” to “This article encourages the use of explicit methods...”; intended meaning kept?

q2. Au: In sentence “Famously, however, in situations...” changed “typically thought to be” to “typically assumed to be”; change okay?

q3. Au: The journal prefers to not use italics when the meaning is clear from the context; “holisitic” changed accordingly.

q4. Au: In expressions 7b, 7c, and 7e, what does “t” indicate? I have italized “t”; is it a variable?

q5. Au: In sentence “In linguistic theorizing...,” changed “partly or totally” to “partially or completely”; change okay?

q6. Au: Changed “they show that part of doing linguistics” to “they show that part of analyzing linguistics...”; intended meaning kept?

q7. Au: Edits made to sentence “This new theory is then...” for greater clarity; please check that your meaning has been retained.

q8. Au: It is journal style to cite tables at least once in the run of text. A reference to table 1 has been placed here; please revise as needed.