Effective Dimensionality: A Tutorial

Effective Dimensionality: A Tutorial

Marco Del Giudice

University of New Mexico

In press: Multivariate Behavioral Research

Marco Del Giudice, Department of Psychology, University of New Mexico. Address: Logan Hall, 2001 Redondo Dr. NE, Albuquerque, NM 87131, USA; email: marcodg@unm.edu

Abstract

The topic of this tutorial is the *effective dimensionality* (ED) of a dataset, i.e., the equivalent number of orthogonal dimensions that would produce the same overall pattern of covariation. The ED quantifies the total dimensionality of a set of variables, with no assumptions about their underlying structure. The ED of a dataset has important implications for the "curse of dimensionality;" it can be used to inform decisions about data analysis and answer meaningful empirical questions. The tutorial offers an accessible introduction to ED, distinguishes it from the related but distinct concept of *intrinsic dimensionality* (ID), critically reviews various ED estimators, and gives indications for practical use with examples from personality research. An R function is provided to implement the techniques described in the tutorial.

Keywords: Correlation; curse of dimensionality; effective dimensionality; entropy; intrinsic dimensionality.

1. Introduction

Questions about the dimensionality of data are pervasive in multivariate applications, and become especially critical in fields such as molecular biology, ecology, and neuroscience—where the number of measured variables is often orders of magnitude higher than the degrees of freedom of the system under study. However, the problem of dimensionality is central even in less data-intensive areas, as for example research on personality and individual differences. In this tutorial I introduce the concept of the *effective dimensionality* (ED) of a dataset, and demonstrate how it can be applied in empirical research. In a nutshell, the ED of a set of correlated variables is the equivalent number of orthogonal dimensions that would produce the same overall pattern of covariation. The ED is a basic index of the *total* dimensionality of the data: as such, it makes no assumptions about the underlying structure of the variables and does not attempt to distinguish between "signal" and "noise." This can be contrasted with other approaches to dimensionality that seek to recover a smaller number of latent variables, and/or separate the major features of the data from those deemed trivial or negligible. Of particular interest, the ED of a dataset plays a major role in determining the severity of the "curse of dimensionality"—a shorthand for the statistical challenges that arise as the space of the data becomes increasingly high-dimensional (Aggarwal et al., 2001; Altman & Krzywinski, 2018; Giraud, 2015).

The ED is a simple metric that can be estimated with minimal assumptions and without complex computations; it deserves to be included in the basic toolkit of multivariate statistics. Unfortunately, the literature on this topic is scattered across disciplines, frustratingly disconnected, and marred by confusing terminology. This tutorial provides a one-stop resource on this little-known topic and a function for ED estimation in the R environment (R Core Team, 2019). I begin by defining effective dimensionality (Section 2) and demarcating it from the related but distinct concept of *intrinsic dimensionality* (Section 3). Next, I review the available indices of ED, discuss their rationale and limitations, and compare their behavior in different scenarios (Section 4); I then address some important practical issues such as sample size and measurement error (Section 5). Finally, I illustrate the use of ED indices with two empirical examples from personality psychology (Section 6).

2. What is effective dimensionality?

The definition of ED rests on the notion that the structure of a set of K variables—as described by the correlation or covariance matrix—can be summarized by an equivalent number n of orthogonal dimensions, with equal variance along each dimension (isotropy). The number n can vary continuously from 1 to K, and quantifies the ED of the original variables (Bretherton et al., 1999; Gnedenko & Yelnik, 2016; Pirkl et al., 2012; Roy & Vetterli, 2007). The stronger the correlational structure of the variables, the smaller the equivalent number of dimensions; in the

limit, a set of perfectly correlated variables can be represented by just one dimension of variation ($n = 1$). At the other extreme are cases in which the ED equals the number of original variables. The exact conditions under which $n = K$ depend on whether the ED is based on the correlation matrix (the variables must be all orthogonal) or the covariance matrix (the variables must be orthogonal *and* have the same variance).

Measuring dimensionality as a continuous quantity is a powerful idea, but also one that can be puzzling when encountered for the first time. Figure 1 offers an intuitive geometric illustration. The four ellipsoids have the same volume, and represent the distribution of variation along three orthogonal axes (x, y, and z). In Figure 1a, variation is the same in all directions, and the ellipsoid is a sphere with an ED of 3. In Figure 1b, variation along the z axis is restricted and the ellipsoid becomes flattened—that is, *effectively* more two-dimensional than a sphere. Stated differently, it takes less than three full dimensions to describe this pattern of variation; according to one of the estimators I introduce in Section 4 (n_1), the ED of the flattened ellipsoid in the figure is exactly 2.5. In Figure 1c, variation is restricted along both the y and z axes, and the ellipsoid has exactly two effective dimensions according to the n_1 index. Note that an ED of 2 does *not* imply that the structure of variation is well described by a flat, two-dimensional surface. This is a useful warning that the number of effective dimensions is not a straightforward description of the geometry of the data—an important point that I discuss again in Section 3. In Figure 1c, most of the variance lies along the x axis and there is relatively little variation in the other directions. The ellipsoid begins to look approximately one-dimensional, and the ED gets closer to 1 (specifically, $n_1 = 1.5$).

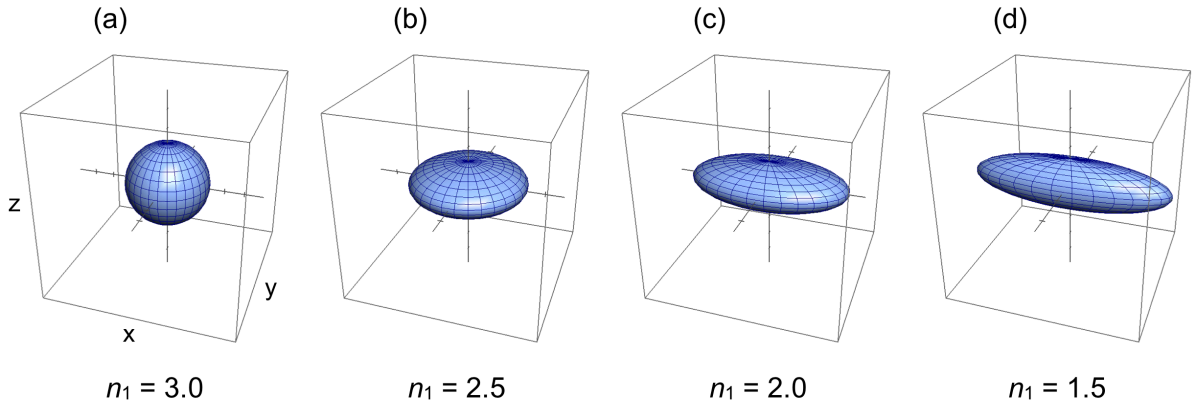

Figure 1. Geometric illustration of effective dimensionality (ED). The ellipsoids have the same volume, but different patterns of variation along the three axes. Index n_1 is an entropy-based estimator of ED, as described in Section 4.

2.1. Implications for the curse of dimensionality

The so-called "curse of dimensionality" is a set of statistical phenomena that occur in high-dimensional spaces, violate geometric intuitions that work well in low dimensions, and can make the analysis of large-scale data particularly challenging (for a detailed overview see Giraud, 2015). For example, a surprising property of high-dimensional distributions is that combinations of rare values can become extremely common: as dimensionality increases, a larger and larger proportion of the mass of the distribution becomes concentrated in the tails, where the probability density is low (Giraud, 2015). This property of multivariate distributions led van Tilburg (2019) to note that, as the number of traits used to describe personality increases, the frequency of "average" personality profiles (i.e., those close to the distribution centroid according to their Euclidean distance) is going to decrease rapidly. With enough traits, one may end up in a paradoxical situation in which almost every individual in the population is highly "unusual" when compared with the average.

More troubling, the vastness and sparsity of high-dimensional spaces make the very notions of distance and similarity problematic and ill-defined (Aggarwal et al., 2001; Altman & Krzywinski, 2018). As the number of dimensions becomes larger, the minimal distance between two points increases, and all the points become approximately equally distant to one another (this is known as the "distance concentration effect"). Since many algorithms for search, classification, and outlier detection rely on distance metrics to quantify the similarity between data points, their performance in high-dimensional spaces tends to drop sharply unless sample size becomes exponentially larger (Aggarwal et al., 2001; Beyer et al., 1999; Houle et al., 2010; Zimek et al., 2012). (Other problems that are sometimes discussed in this context are overfitting in regression models—estimation errors can become large as small fluctuations cumulate across predictors—and the fact that computational complexity may increase nonlinearly as the number of dimensions grows; see Giraud, 2015).

In practice, however, the curse of dimensionality is often less severe than one might expect—even in datasets with hundreds or thousands of variables that might seem hopelessly high-dimensional (e.g., gene expression data; Durrant & Kabán, 2009; Zollanvari et al., 2011). As it turns out, the impact of distance-related phenomena does not just depend on the number of variables but also on their statistical overlap. If the variables share a strong correlational structure, the concentration of Euclidean distances takes place at a much slower pace than expected; conversely, the effect becomes more severe if the dataset includes many irrelevant or noisy variables that weaken the correlational structure and increase the total dimensionality (Durrant & Kabán, 2009; Zimek et al., 2012). Similarly, the average Euclidean distance from the centroid is reduced if the variables are not orthogonal but correlated (van Tilburg, 2019). In other words, the key governing factor is the *effective* number of independent dimensions in the dataset—precisely the quantity measured by ED.

3. Effective vs. intrinsic dimensionality

The ED of a set of variables is a continuous measure of its total dimensionality, without distinction between signal and noise. In contrast, intrinsic dimensionality (ID) is defined as the minimum number of variables needed to accurately describe the important features of the system (Campadelli et al., 2015; Carreira-Perpiñán, 1996). From a geometric point of view, this informal concept can be made more rigorous by defining the ID as the dimensionality of the manifold that approximately embeds the data, and is itself embedded in the higher-dimensional space of the original variables (Campadelli et al., 2015; Carreira-Perpiñán, 1996; Facco et al., 2017; Zwiggelaar, 2014). To illustrate with a particularly clear-cut example, the points in Figure 2a are identified by three coordinates; however, they lie entirely on a plane within the three-dimensional space. Since a plane is a two-dimensional manifold, their intrinsic dimensionality is 2 instead of 3. In Figure 2c, all the points lie on a line, and their ID equals 1 even if they are described by three coordinates. Note that the geometry of ID does not have to be linear as in Figure 2; in principle, the embedding manifold can be curved and twisted into complex shapes (see Carreira-Perpiñán, 1997; Facco et al., 2017). The ID of a set of variables contributes to determine the severity of the curse of dimensionality. For example, simulations show that algorithms based on similarity can perform well in high-dimensional datasets, provided that the ID of the latter (estimated with the fractal dimension methods discussed in Section 3.1) is sufficiently low (Korn et al., 2001).

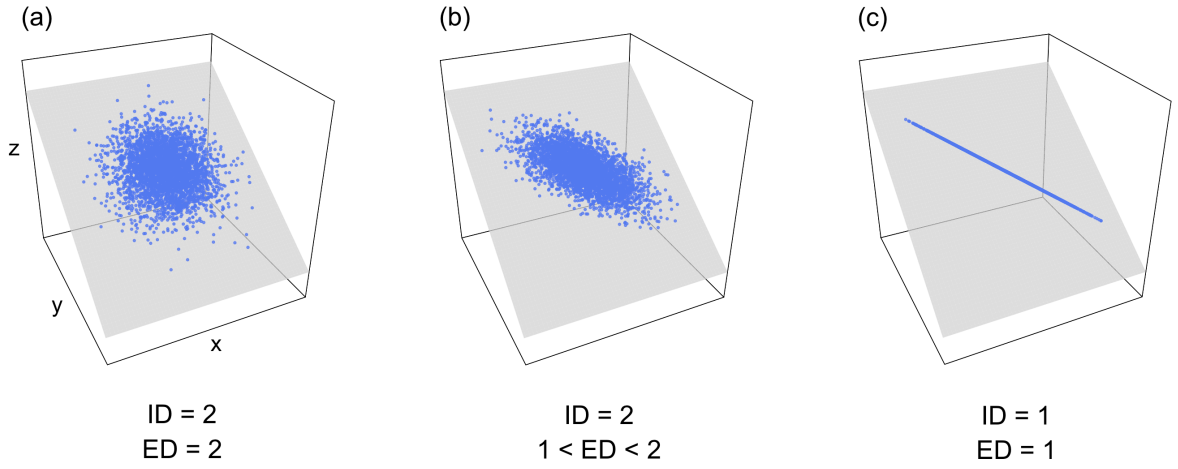

Figure 2. Illustration of the difference between intrinsic dimensionality (ID) and effective dimensionality (ED). In both (a) and (b), the points are embedded in a plane (a two-dimensional manifold) and their ID is 2. The ED does not just reflect the embedding of the points but also their correlational structure, and is lower in (b) (nonzero correlation) than in (a) (zero correlation, ED = 2). The points in (c) are embedded in a line (a one-dimensional manifold), the coordinates on the three axes are perfectly correlated, and both ID and ED equal 1.

Under the formal definition, the ID does not depend on the correlational structure of the variables, but only on the dimensionality of their embedding manifold. Consider the ellipsoids in Figure 1: they all have the same ID of 3, but the number of effective dimensions changes depending on how variation is distributed along the three axes. Figure 2 is also instructive in this regard. ED and ID agree in the case of Figure 2c, in which the manifold is one-dimensional and all the variables are perfectly correlated with one another. In Figure 2a, x is uncorrelated with y and z, while y and z are perfectly correlated; accordingly, both the ID and the ED equal 2. The points of Figure 2b still lie on a plane (ID = 2), but the variables are correlated and thus partially redundant, and their ED is lower than 2. Crucially, the summary provided by the ED is not intended as a direct representation of the geometry of the data. For example, the ellipsoid in Figure 1c has two effective dimensions, but it would be a mistake to infer that its shape is that of a two-dimensional disc. If anything, this description is better approximated by the flattened ellipsoid of Figure 1b, which however has an ED of 2.5.

Informally, ID quantifies the number of variables needed to accurately describe the important features of a system. This broader definition implies that redundancies among the original variables may mask a simpler underlying structure; however, ID and ED remain critically different. In particular, the purpose of ID is to distinguish between the "important" or "relevant" features of the data and the "trivial" or "irrelevant" ones. In contrast, ED is purely descriptive: the number n summarizes the overall structure of the variable set, including possible sources of noise such as measurement error and the presence of irrelevant features—unless these have been statistically controlled for (Section 5.5). As the amount of noise increases, correlations among variables become weaker and the estimated ED increases accordingly (see Cangelosi & Goriely, 2007).

In sum, ED and ID answer different questions about the data and must not be confused with one another. Consider a dataset that, according to a given criterion, can be adequately described by m variables (so that ID = m). To the extent that the dataset includes additional "minor" dimensions of variation and/or measurement error, the ED will tend to be larger than m. But to the extent that the m variables are correlated and hence partially redundant, the ED will tend to be smaller than m (see Figure 2). As a result, the ED of a dataset can be larger, smaller, or equal to the ID. The point is that ED is not a simpler or approximate version of ID, but a conceptually distinct quantity with its own interpretation.

3.1. Methods for estimating intrinsic dimensionality

Even though ID is not the main subject of this tutorial, it can be useful to briefly review the methods used to estimate it in practice. Both *exploratory factor analysis* (EFA) and *principal component analysis* (PCA) can be used to reduce the dimensionality of a dataset by retaining a smaller number of meaningful dimensions (factors or components). EFA assumes an underlying

causal model in which correlations among observed variables are determined by unobserved latent variables; PCA is purely a data reduction technique, and the components are linear combinations of the original variables rather than latent constructs (see e.g., Fabrigar et al., 1999). Both methods are implemented in several R packages, including the user-friendly *psych* (Revelle, 2019). While standard EFA and PCA are linear techniques, there are nonlinear extensions that can model a curved manifold (e.g., Carreira-Perpiñán, 1997; Yalcin & Amemiya, 2001). A number of those extensions can be found in the R package *Rdimtools* (Suh & You, 2018).

In both PCA and EFA, the ID is estimated by deciding how many factors or components should be retained. Dozens of decision algorithms have been proposed and tested over the decades (see e.g., Cangelosi & Goriely, 2007; Peres-Neto et al., 2005). Some consist of simple and often arbitrary rules of thumb: for example, retaining enough components to account for 80% or 90% of the total variance, or discarding the components that explain less variance than a preset threshold. More sophisticated methods employ significance testing, randomization, or model selection to identify the appropriate number of dimensions to retain (see Peres-Neto et al., 2005; Ruscio & Roche, 2012). *Parallel analysis* and its variants rank among the best-performing algorithms of this kind (Lim & Jahng, 2019; Ruscio & Roche, 2012). Of note, parallel analysis and other algorithms that rely on eigenvalues (see Section 4) may underestimate the ID if the underlying factors have a strong correlational structure, and hence a low ED (see Zopluoglu & Davenport, 2017). From a Bayesian perspective, various methods have been developed to select the number of dimensions with the highest posterior probability (e.g., Conti et al., 2014; Minka, 2001; Nakajima et al., 2011; Seghouane & Cichocki, 2007).

Besides PCA, other projection methods that can be applied to ID estimation include *independent component analysis* (ICA) and *multidimensional scaling* (MDS; see Carreira-Perpiñán, 1997). The R package *Rdimtools* (Suh & You, 2018) can be used to estimate ID with these techniques. More recently, *exploratory graph analysis* (EGA) has been proposed as a method for dimensionality estimation based on network theory (Golino & Epskamp, 2017); EGA is implemented in the R package *EGAnet* (Golino et al., 2019).

An alternative approach to ID estimation is based on *fractal dimensions*. Intuitively, the dimensionality of the embedding manifold can be estimated by how completely the data fill the space as one moves from larger to increasingly smaller scales of analysis (Campadelli et al., 2015; Einbeck & Kalantan, 2013; Zwiggelaar, 2014). Fractal dimension estimators capture the self-similarity of data at different scales (Korn, 2001) and are theoretically attractive, but demand massive amounts of data to perform well (Zwiggelaar, 2014). Yet another family of ID methods is that of *nearest neighbors-based* estimators. These methods exploit the distribution of the data at a local scale (e.g., the distribution of distances between neighboring data points) to estimate the dimensionality of the entire manifold (Campadelli et al., 2015; Facco et al., 2017;

Zwiggelaar, 2014). The *Rdimtools* package (Suh & You, 2018) offers an extensive collection of ID estimation tools, including several linear and nonlinear projection methods, fractal dimension estimators, and nearest neighbors-based estimators. In general, there is no "gold standard" for estimating ID; which method performs best depends on the specific features of the dataset in ways that are often hard to anticipate (e.g., Zwiggelaar, 2014). For empirical comparisons of alternative methods, see Campadelli et al. (2015), Einbeck and Kalantan (2013), and van der Maaten et al. (2009).

4. Estimating effective dimensionality

The most common indices of ED (see Table 1 for an overview) are based on the eigenvalues of the correlation or covariance matrix, which collectively are known as the *spectrum* of the matrix. There are as many eigenvalues as variables ($\lambda_1 \ldots \lambda_N$), and their magnitude quantifies the variance in the direction of the corresponding eigenvectors. The sum of the eigenvalues equals the sum of the variances of the original variables; if the correlation matrix is used, the variables are standardized to unit variance, and the sum of the eigenvalues is simply their number (K). These notions should be familiar to readers acquainted with the basics of PCA (e.g., the spectrum of the correlation or covariance matrix is displayed in the "scree plot;" eigenvalues are used to compute the proportion of variance explained by each component). For an accessible explanation of eigenvalues and eigenvectors, see Strang (2016).

When all the variables are perfectly correlated, the number of effective dimensions is 1; there is only one nonzero eigenvalue, whose magnitude is the sum of the variances (Figure 3a). When the variables are all orthogonal (and have equal variance if the covariance matrix is used), the number of effective dimensions is K and all the eigenvalues have the same magnitude (Figure 3c). When there are n clusters of variables that are perfectly correlated within each cluster, but orthogonal across clusters (and the clusters have the same total variance if the covariance matrix is used), the number of effective dimensions is n. In this case, the spectrum contains n nonzero eigenvalues of equal magnitude followed by $(K - n)$ zeroes (Figure 3b). Note that the determinant of the matrix equals the product of the eigenvalues, and is zero whenever one or more eigenvalues are zero (see Strang, 2016).

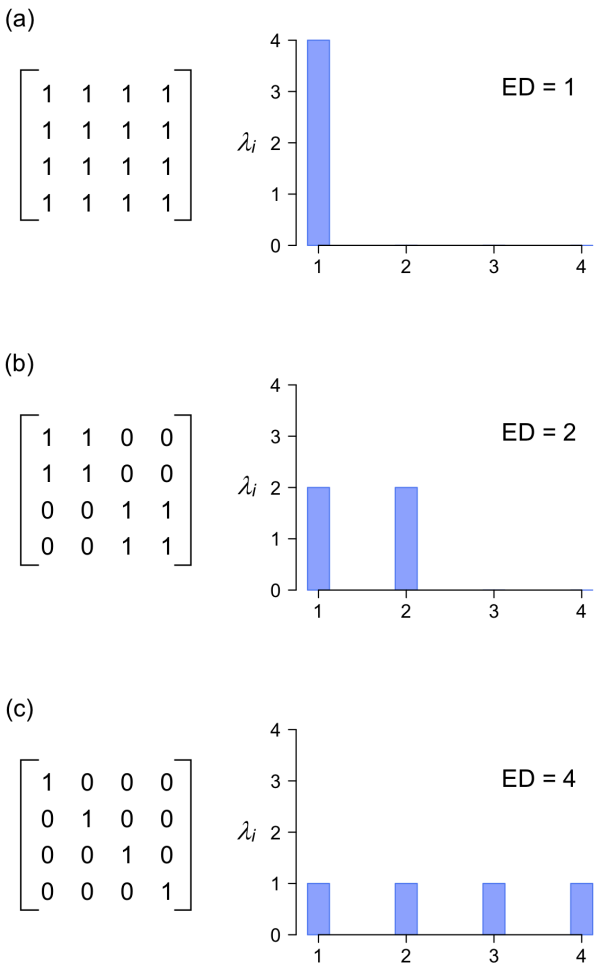

Figure 3. Examples of 4×4 correlation matrices (left) and their spectra (right). (a) The variables are perfectly correlated; there is only one nonzero eigenvalue (λ) and the effective dimensionality (ED) is 1. (b) The variables form two clusters, with perfect correlations within cluster and zero correlations between clusters. There are two equal nonzero eigenvalues and the ED is 2. (c) The variables are all orthogonal; the four eigenvalues are all equal and the ED is 4. Note that the sum of the eigenvalues equals the sum of variances (4 in this case).

4.1. Entropy-based estimators

The standard approach to ED estimation is based on the information-theoretic concept of *entropy*, which in this context can be defined intuitively as the information content of a probability distribution. For discrete distributions, the information content is maximized when all the values are equally probable, and hence equally "surprising" (uniform distribution). Conversely, if one particular value occurs with a probability of 1 while all the others have zero probability, the distribution carries no information and its entropy is zero. See Stone (2015) for a tutorial introduction to information theory, and Stone (2019) for a condensed version.

The first step in the derivation of ED estimators is to recast the spectrum λ of a correlation or covariance matrix as a discrete probability distribution. To achieve this, the eigenvalues are normalized by their sum:

$$p_i = \frac{\lambda_i}{\sum_{j=1}^{K} \lambda_j} \quad (1)$$

The resulting (pseudo-)probabilities can then be used to calculate the information entropy associated with the spectrum (*spectral entropy*). The spectral entropy is maximized when the distribution is uniform (i.e., the eigenvalues are all equal, and the variables are all orthogonal). As the correlational structure gets stronger and variables become more redundant, each variable provides less unique information and entropy diminishes accordingly. If there is only one nonzero eigenvalue, the spectral entropy becomes zero since the variables are completely redundant with one another.

Once the spectral entropy of a correlation or covariance matrix has been calculated, it is easy to find the equivalent number of orthogonal dimensions that would result in the same amount of entropy, and use that number as an estimate of ED. This approach has been used for decades in ecology to estimate the "effective number of species," a basic index of ecological diversity within a community (Hill, 1973; Jost, 2006; Tuomisto, 2010). From a social sciences perspective, Budescu and Budescu (2012) proposed the equivalent entropy as a summary measure of ethnic diversity. The key decision is which measure of entropy to use among the many possible alternatives. The Rényi entropy (see Bromiley et al., 2004; Rényi, 1961) is a generalized entropy that includes the familiar Shannon entropy as a special case:

$$H_q = \frac{1}{1-q} \log\left(\sum_{i=1}^{K} p_i^q\right) \quad (2)$$

The limit of Eq. 2 when the order parameter q tends to 1 is the Shannon entropy (H_1). In the Shannon entropy, the normalized eigenvalues are weighted in proportion to their size. The value $q = 2$ yields the Rényi entropy of order 2 or *quadratic entropy* (H_2). The entropies H_1 and H_2 are both maximized in uniform distributions and have the same maximum value, but H_2 assigns disproportionally more weight to the larger eigenvalues while discounting the smaller ones. As a result, H_2 drops more steeply than H_1 as the spectrum deviates from a uniform distribution (Figure 4). This means that the same non-uniform spectrum of eigenvalues will have a higher entropy (and more equivalent dimensions) if H_1 is used, and a lower entropy (and fewer equivalent dimensions) if H_2 is used instead. Higher values of q assign progressively more weight to the larger eigenvalues. The limit of Eq. 2 when α tends to infinity is the *min-entropy* (H_∞), which is entirely determined by the largest eigenvalue (see Figure 4). H_∞ discounts all the information in the spectrum beyond the largest eigenvalue, and yields the minimum estimate of entropy (and hence the minimum number of equivalent dimensions).

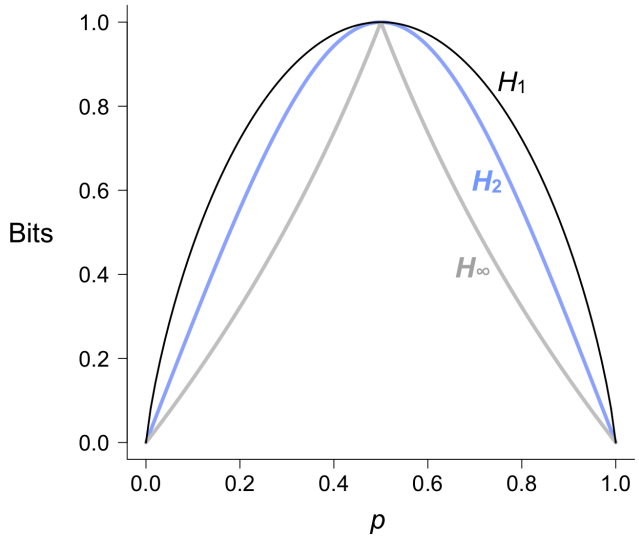

Figure 4. Illustration of the Shannon entropy (H_1), quadratic entropy (H_2), and min-entropy (H_∞) in the case of a random variable with two outcomes and probabilities p and $(1 - p)$. The entropy (i.e., the average amount of information provided by the outcome) is always maximized when $p = 0.5$ (1 bit) and minimized when $p = 0$ or $p = 1$ (0 bits). As outcome probabilities deviate from 0.5, H_2 decreases more steeply than H_1, and H_∞ decreases more steeply than H_2.

The principle of equivalent entropy for estimating ED was employed independently by Cangelosi and Goriely (2007), who called the estimator "information dimension;" Roy and Vetterli (2007), who interpreted it as "effective rank" (a continuous extension of the rank of a matrix, which can only take integer values); and more recently Gnedenko and Yelnik (2016), who labeled it simply as "effective dimensionality." All these authors used the Shannon entropy H_1 in their derivations, and the resulting index can be labeled n_1, consistent with the usage in ecology (Hill, 1973):

$$n_1 = \prod_{j=1}^{K} \left(\frac{\lambda_j}{\sum_{i=1}^{K} \lambda_i} \right)^{-\frac{\lambda_j}{\sum_{i=1}^{K} \lambda_i}} \tag{3}$$

Because H_1 is a "balanced" entropy that does not assign disproportionately more weight to the larger eigenvalues, n_1 is suitable as a general-purpose estimator of ED.

Drawing on previous work by Fraedrich et al. (1995) and Bretherton et al. (1999), Pirkl et al. (2012) derived another entropy-based index of ED, which they called the "effective number of uncorrelated measurements." This estimator is based on H_2 instead of H_1, and can be labeled n_2:

$$n_2 = \frac{\left(\sum_{i=1}^{K} \lambda_i\right)^2}{\sum_{i=1}^{K} \lambda_i^2} \quad (4)$$

The choice of H_2 means that n_2 is generally more conservative than n_1. While Eqs. 3 and 4 give identical results in the special cases illustrated in Figure 3, n_2 yields lower estimates of ED in most realistic scenarios.

Finally, Kirkpatrick (2009) proposed a simple ED estimator as the sum of the eigenvalues divided by the largest eigenvalue. Although this was not the original rationale, Kirkpatrick's index turns out to be the equivalent entropy estimator for the min-entropy H_∞; accordingly, it can be labeled n_∞:

$$n_\infty = \frac{\sum_{i=1}^{K} \lambda_i}{\max_i \lambda_i} \quad (5)$$

The n_∞ index discards all the information in the spectrum beyond the first eigenvalue, effectively assuming a scenario like the one in Figure 3b. As a result, it lacks sensitivity and yields extremely conservative estimates of ED.

To summarize, the behavior of entropy-based indices depends on the order of the corresponding Rényi entropy. The n_1 index can be used in most situations as a general-purpose estimator of ED. The n_2 index is appropriate when one seeks a more conservative estimate, or a reasonable lower bound on the dimensionality of a dataset. The estimates provided by n_∞ are extremely conservative, and too insensitive to be of use in most practical contexts.

4.2. Other estimators

Cheverud (2001) proposed a non-entropy-based ED index as a method to correct for the effective degrees of freedom in multiple significance testing. The same estimator was then used by Wagner et al. (2008) to measure the "effective number of traits" in a correlation matrix. This estimator can be labeled n_C to distinguish it from its entropy-based counterparts:

$$n_C = K - \mathrm{Var}(\lambda) \quad (6)$$

where K is the number of variables. The rationale for n_C is that, in a correlation matrix, the variance of the eigenvalues is $K - 1$ when there is only one nonzero eigenvalue as in Figure 3a (hence $n_C = 1$), and zero when the eigenvalues are all equal as in Figure 3c (hence $n_C = K$). Interpolating between these two extremes yields a continuous estimate of ED.

The original formula shown in Eq. 6 only works with correlation matrices, but a simple adjustment makes it equally applicable to covariance matrices:

$$n_{\mathrm{C}} = K - \frac{K^2}{\left(\sum_{i=1}^{K} \lambda_i\right)^2} \mathrm{Var}(\lambda) \quad (7)$$

The n_{C} index has two main limitations. First, it is not well justified for intermediate values between 1 and K; and second, it systematically overestimates the ED, often by a large margin (Li & Ji, 2005). For these reasons, n_{C} is not recommended for practical use, and is only reviewed here for completeness.

Table 1. Overview of four estimators of effective dimensionality (ED).

ED index	Formula	Notes
n_1	$\prod_{j=1}^{K} \left(\frac{\lambda_j}{\sum_{i=1}^{K} \lambda_i} \right)^{-\frac{\lambda_j}{\sum_{i=1}^{K} \lambda_i}}$	- Rationale: equivalent spectral entropy (Shannon entropy, H_1) - Balanced, general-purpose estimator
n_2	$\frac{\left(\sum_{i=1}^{K} \lambda_i\right)^2}{\sum_{i=1}^{K} \lambda_i^2}$	- Rationale: equivalent spectral entropy (quadratic entropy, H_2) - More conservative than n_1
n_∞	$\frac{\sum_{i=1}^{K} \lambda_i}{\max_i \lambda_i}$	- Rationale: equivalent spectral entropy (min-entropy, H_∞) - Extremely conservative; not recommended
n_{C}	$K - \frac{K^2}{\left(\sum_{i=1}^{K} \lambda_i\right)^2} \mathrm{Var}(\lambda)$	- Rationale: interpolation between 1 and K - Typically overestimates ED; not recommended

Legend: λ = eigenvalues of the correlation or covariance matrix. K = number of variables in the set.

4.3. An illustrative comparison

Figure 5 compares the estimators in three illustrative scenarios, based on the correlation matrix of four variables. The largest eigenvalue is 2 in all the scenarios. The spectrum in Figure 5a reproduces that of Figure 3b, with two nonzero eigenvalues of equal magnitude. This is a trivial special case, and all the estimators agree on a dimensionality of 2.00. In Figure 5b, there are three nonzero eigenvalues, but the first explains twice as much variance as the other two. The lack of sensitivity of n_∞ is apparent, as it yields the same value as in the first scenario (2.00). On the other hand, n_{C} grossly overestimates the ED, and returns a number of dimensions larger than the number of nonzero eigenvalues (3.50). Both n_1 and n_2 estimate an ED of more than 2.5 and less than 3, with n_2 predictably smaller than n_1. Figure 5c shows an even more realistic spectrum

with four nonzero eigenvalues of decreasing magnitude. Since n_∞ only considers the largest eigenvalue, it continues to indicate an ED of 2, whereas n_C yields the highest estimate of the set (3.63). The estimates provided by n_1 and n_2 are somewhat above and below 3, respectively. The relative difference between the two entropy-based estimators is predictably larger in this scenario, since H_1 and H_2 diverge more strongly for distributions with many intermediate values.

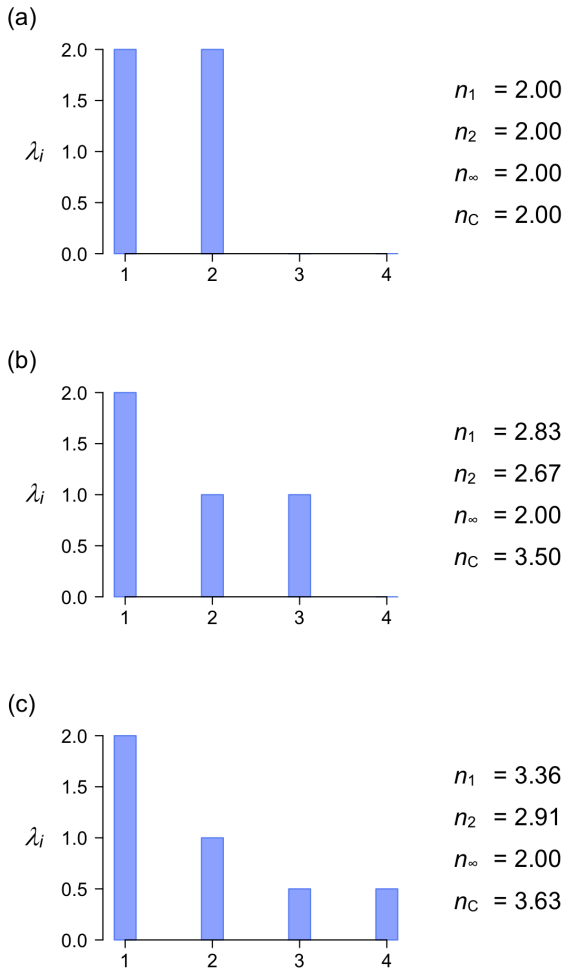

Figure 5. Comparison of four estimators of effective dimensionality (ED) in three simple scenarios. The n_2 index (based on the quadratic entropy) is more conservative than n_1 (based on the Shannon entropy). The n_∞ index (based on the min-entropy) depends only on the largest eigenvalue, yields highly conservative estimates of the ED, and is insensitive to differences between scenarios. The n_C index typically overestimates the ED and often returns more dimensions than nonzero eigenvalues, as in panel (b).

4.4. R code

The R function *estimate.ED* is available at https://doi.org/10.6084/m9.figshare.11954661. This function computes the four estimators reviewed in this section—either from raw data or correlation/covariance matrices—and implements the error-correction techniques discussed in Section 5.

5. Practical issues

5.1. Potential uses of ED

ED indices can be employed in a variety of research contexts. To begin, ED provides an initial summary of the correlational structure of the data, and an indication of the likely severity of the curse of dimensionality. This can be useful to decide whether approaches such as dimension reduction (e.g., via PCA) or variable selection (e.g., via regularization; see James et al., 2013; Lever et al., 2016) should be employed to alleviate the analytic problems described in Section 2.1. Even *after* dimension reduction, the ED of the reduced data (e.g., a set of correlated factor scores) can be informative, especially if the number of retained dimensions is large. Alternatively, researchers may want to analyze the data "as is" without recourse to dimension reduction, for example to preserve the meaning of the original variables. In such cases, ID becomes less relevant but ED remains a viable measure of dimensionality. When ED and ID are used in combination, discrepancies between the two can be explored and may suggest new insights into the data.

The fact that ED is a continuous measure is an advantage when one wants to compare the correlational structure of the same set of variables across multiple groups, contexts, experimental conditions, or time points (e.g., in longitudinal studies). Most methods for ID estimation only yield discrete values, and are insensitive to gradual change; moreover, the exact number of dimensions selected by algorithms such as parallel analysis can depend on minor fluctuations in the data. In contrast, ED provides a fine-grained assessment of dimensionality and is naturally suited to comparative research, as demonstrated in Section 6.2.

5.2. Correlation or covariance?

Effective dimensionality can be calculated from either correlations or covariances, raising the question of which option is more appropriate in a given case. This is a familiar problem in PCA, where principal components can be extracted from the correlation or the covariance matrix. In the covariance matrix, the variables with the largest variance dominate the overall structure, and tend to overshadow the contribution of the other variables. In the correlation

matrix, all variances are standardized to unity, meaning that each variable carries the same weight as the others regardless of its original scale.

When the scales of different variables in the set are arbitrary (e.g., rating scales that use different numerical ranges) or incommensurable (e.g., variables measuring age, height, and income), this may be the only reasonable options. When differences in scale are meaningful and non-arbitrary (e.g., a set of variables measuring the length of different anatomical traits), researchers need to decide whether it makes sense to equalize the contributions of different variables or let the largest variances determine the dimensionality of the dataset. In any event, it is important to clearly specify the source of the eigenvalues whenever ED is estimated.

5.3. Linearity and normality

The ED estimators reviewed in this tutorial are based on the spectrum of the correlation or covariance matrix, which is a complete description of the data only if the latter follow a multivariate normal distribution. If the data are characterized by nonlinear dependencies, ED estimators will only capture those aspects of the structure that are reflected in the correlation or covariance matrix. Note that the same limitation applies to linear techniques used to estimate ID, such as standard PCA and EFA (Section 3.1).

If the distribution of the variables deviates from normality, sample correlations may be systematically inflated or deflated compared with their population value. Simulations show that, in a range of plausible scenarios, biases due to non-normality tend to become negligible when sample size is larger than about 100-200 (Bishara & Hittner, 2015). However, there are cases (for example involving pairs of lognormal distributions) in which bias remains substantial even with sample sizes in the hundreds of thousands (e.g., Lai et al., 1998). This is not a problem when researchers are only interested in the particular sample under consideration. However, in most cases the quantity of interest is the ED of the population (see Sections 5.4 and 5.5 for more discussion). One should keep in mind that, especially in small samples, the correlational structure of the sample may not reflect that of the population if there are marked deviations from normality.

A related issue arises when some variables in the dataset are not continuous but categorical, either dichotomous (binary) or polytomous (three or more ordered levels). While categorical variables do not follow a normal distribution, it is possible to compute *tetrachoric* and *polychoric* correlations, which estimate the correlation coefficient under the assumption that the observed categories reflect a continuous and normally distributed latent variable (Drasgow, 1988). Tetrachoric and polychoric correlations can be calculated with packages *psych* (Revelle, 2019) and *polycor* (Fox, 2019). The main problem with this method is that the resulting correlation matrices may be indefinite—that is, some of the eigenvalues may be negative. A

practical solution is to approximate the original matrix with the nearest positive definite matrix. Approximation methods include those by Higham (2002) and Knol and ten Berge (1989). The function *estimate.ED* automatically detects indefinite matrices and applies Higham's (2002) method, as implemented in the *Matrix* package (Bates & Maechler, 2019).

5.4. Correcting for small-sample bias

The eigenvalues of a sample correlation or covariance matrix are not unbiased estimators of the corresponding population values. Specifically, the sample eigenvalues tend to be more spread out than those in the population, so that estimates of large eigenvalues are biased up whereas those of small eigenvalues are biased down (see Lim & Jahng, 2019; Mestre, 2008). As a result, ED estimators computed from sample data generally underestimate the effective number of dimensions in the population, particularly when correlations among variables are small and the population spectrum is close to uniform (see Figure 6 for an illustration). This bias becomes more pronounced as sample size gets small relative to the number of variables in the set, and can be severe when the number of variables is comparable to (or even larger than) the number of observations.

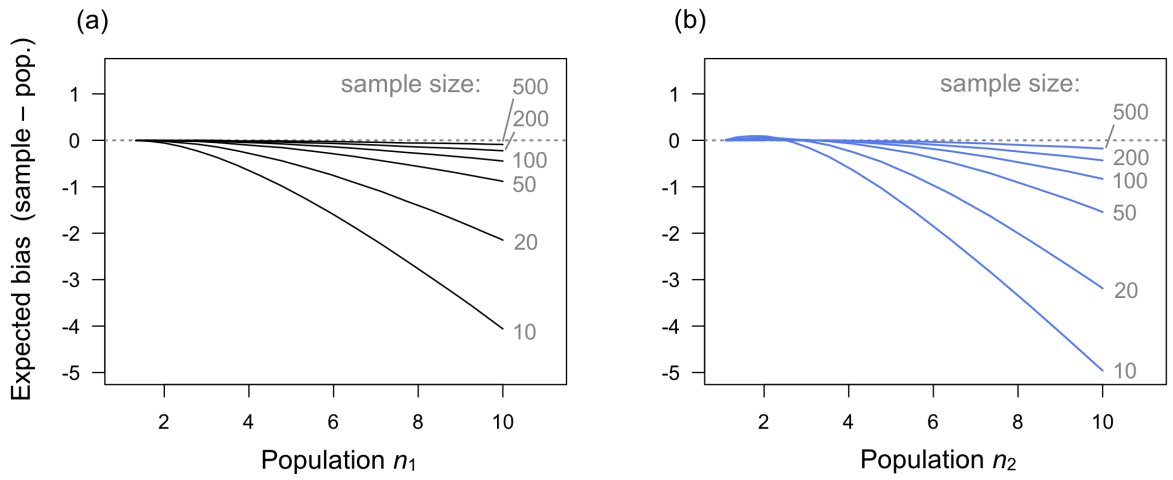

Figure 6. Illustration of small-sample bias in ED estimation with indices n_1 (a) and n_2 (b). Lines show the amount and direction of bias (i.e., the average sample estimate minus the population value) as a function of population ED and sample size. The simulation is based on uniform correlation matrices for 10 variables, with correlations ranging from 0 (ED = 10) to 1 (ED = 1). Bias becomes stronger as sample size decreases and correlations among variables get smaller (higher ED values).

As noted in Section 5.3, the relevance of this issue depends on whether the research question concerns the dimensionality of the specific sample at hand, or the dimensionality of the same variables in the population. Consider a scenario in which researchers wish to compare the correlational structure of the same set of variables across different groups. For example,

Lukaszewski et al. (2017) investigated how the degree of covariation among personality traits varies across countries (see Section 6.2 for a reanalysis). If there are marked differences in sample size between groups, ED estimates are going to be confounded, since—all else being equal—smaller samples tend to show smaller values of ED.

A solution to this problem is to use corrected population estimates of ED in place of unccorrected sample values. Fortunately, it is easy to correct the small-sample bias of ED estimators with shrinkage methods that appropriately reduce the larger eigenvalues and increase the smaller ones, thus bringing them closer to their population values. Two examples are the adjustment method by Mestre (2008) and the nonlinear shrinkage estimator by Ledoit and Wolf (2012, 2015). The latter performs particularly well if the number of variables is comparable to (or even larger than) the sample size, and is implemented in the R package *nlshrink* (Ramprasad, 2016). The function *estimate.ED* allows the user to correct for small-sample bias, using Ledoit and Wolf's method if the raw data are available and Mestre's adjustment otherwise.

5.5. Correcting for measurement error

As noted throughout this tutorial, ED describes the total correlational structure of the data without distinction between signal and noise. However, it is always possible to apply corrections for measurement error *before* computing ED to reduce the amount of noise included in the estimates. Measurement error adds unsystematic variance and attenuates the correlational structure of the data; hence, estimates of ED can be expected to decrease after correction. Such corrected estimates approximate the dimensionality that the data *would* have, if the variables had been measured without error. Naturally, the resulting ED estimates refer to a hypothetical scenario rather than to the actual data at hand. Corrections for measurement error can be useful when one's research question concerns the dimensionality of the variables as idealized constructs. For example, researchers may want to compare the correlational structure of the same variables in different samples or at different time points, while adjust the estimated ED values for systematic changes in measurement quality (see Section 6.2 for an example).

If the reliability of the variables in the dataset is known, correlations can be disattenuated by simply dividing them by the square root of the product of the reliabilities. For example, a raw correlation of $r = .30$ between two variables with reliabilities .70 and .80 would become $r = .40$ after disattenuation. Indices of reliability include Cronbach's *alpha* (α), as well as McDonald's *omega total* (ω_t) and *omega hierarchical* (ω_h). (For in-depth discussion of these and other indices, see Dunn et al., 2014; McNeish, 2018; Revelle & Condon, 2018; Zinbarg et al., 2005.) Generally speaking, reliability indices seek to quantify the proportion of variance attributable to the construct being measured ("true score variance," as contrasted with "error variance"; see Revelle & Condon, 2018). Whereas α and ω_t regard all the variance shared among the items as true score variance, ω_h only considers the variance that can be attributed to a single general

factor underlying the items (estimated through hierarchical EFA). In the context of dimensionality estimation, this makes ω_h an especially attractive option; the reason is that ω_h can be used to adjust correlations for irrelevant specific factors that are confounded with the construct of interest, in addition to the unsystematic error associated with individual items. The function *estimate.ED* can disattenuate the correlation matrix with a vector of reliabilities supplied by the user.

An alternative, more sophisticated approach is to use latent variable methods (most commonly *structural equation modeling* or SEM) to explicitly model the factor structure of the measures, estimate correlations between latent variables instead of observed scores, and calculate the ED using the latent correlation matrix. If the factor structure is correctly specified, latent variable modeling overcomes the limitations of simple reliability indices, and can achieve a virtually error-free estimate of the correlation matrix (see Brown, 2015; Kline, 2016).

6. Empirical examples

6.1. Dimensionality of a large-scale personality dataset

The following example demonstrates ED estimation and correction for measurement error with a large personality dataset (Kaiser, 2019; original data by Johnson, 2015). Specifically, the present analysis focuses on the United States subsample of the dataset, which comprises $N = 617{,}180$ online respondents (379,323 females; for details see Kaiser, 2019). Personality was assessed with the 120-item version of the IPIP-NEO (Johnson, 2014; see http://personal.psu.edu/~j5j/IPIP/). The items (on a 1-5 scale from "very inaccurate" to "very accurate") measure 30 narrow facets of the Big Five domains (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism; Costa & McCrae, 1992), with six facets per domain (e.g., Extraversion comprises Friendliness, Gregariousness, Assertiveness, Activity level, Excitement seeking, and Cheerfulness). The 30 facet scores were calculated as averages of the four corresponding items. The dataset was retrieved from https://osf.io/9kpc5.

To estimate the ED of this dataset, n_1 was calculated from the correlation matrix (pooled from the male and female subsamples) with the function *estimate.ED*. The R script of the analysis is available at https://doi.org/10.6084/m9.figshare.11954667. Observed facet scores yielded $n_1 = 17.47$, indicating that the structure of the data is markedly lower-dimensional than suggested by the number of observed variables. This is unsurprising, since different facets of the same domain are expected to correlate with one another. For comparison, the other ED indices were $n_2 = 10.78$, $n_\infty = 4.44$, and $n_C = 28.22$. Note that this sample is very large relative to the number of variables, so that correcting for small-sample bias would have a negligible effect on the eigenvalues.

What are the implications for the curse of dimensionality? From the standpoint of data analysis, the practical impact of the phenomena described in Section 2.1 depends on the number of dimensions in the data, but also on the size of the sample and the details of the statistical model one employs (see Giraud, 2015). While there are no simple rules of thumb, the scale of this dataset (about 35,000 observations per effective dimension) should minimize the severity of the curse for many standard analyses. That said, high-dimensional phenomena also have theoretical and empirical implications that do not depend on sample size. For example, the concentration of probability in the outer edge of the distribution means that, as the number of measured traits increases, the proportion of individuals with "average" personality profiles will quickly become vanishingly small (van Tilburg, 2019). With 30 orthogonal dimensions of variation, one would expect this effect to be rather extreme, as shown in Figure 7. However, observed scores have an ED of about 17; as a result, the average distance from the distribution centroid becomes noticeably smaller, and the number of profiles in the vicinity of the centroid increases accordingly (Figure 7).

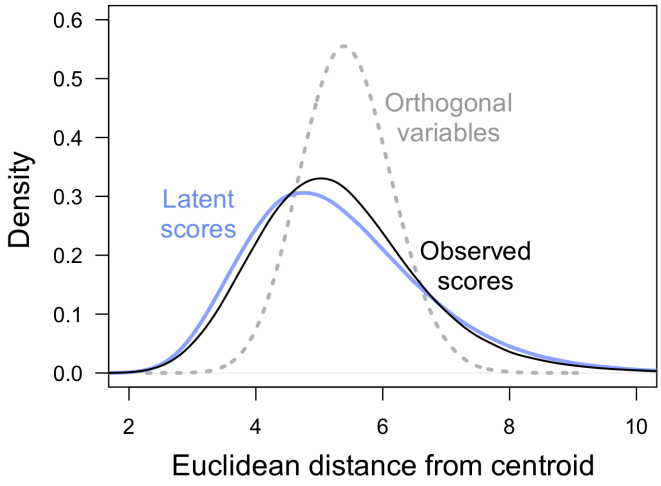

Figure 7. Density plots of Euclidean distances from the distribution centroid, based on 30 personality facets. The dotted line shows the expected distribution for 30 orthogonal variables. The thin line shows empirical distances calculated from observed scores; the thick line shows distances from the simulated distribution of latent scores.

Of course, observed scores in this dataset include a certain amount of measurement error, which contributes to increase the dimensionality of the dataset. From a theoretical standpoint, it can be interesting to estimate the dimensionality of the 30 personality facets as idealized, error-free constructs. To illustrate the difference between alternative correction methods, the observed correlation matrix was first disattenuated with Cronbach's α (obtained with package *psych* v. 1.8.12; Revelle, 2019). Then, latent correlation matrices for males and females were estimated from a multigroup confirmatory factor analysis model fit to item-level data. The details of the analysis are reported in Kaiser (2019); the matrices were obtained from the author of the original

study (Tim Kaiser, personal communication, June 20, 2019). The results of this analysis are summarized in Figure 8.

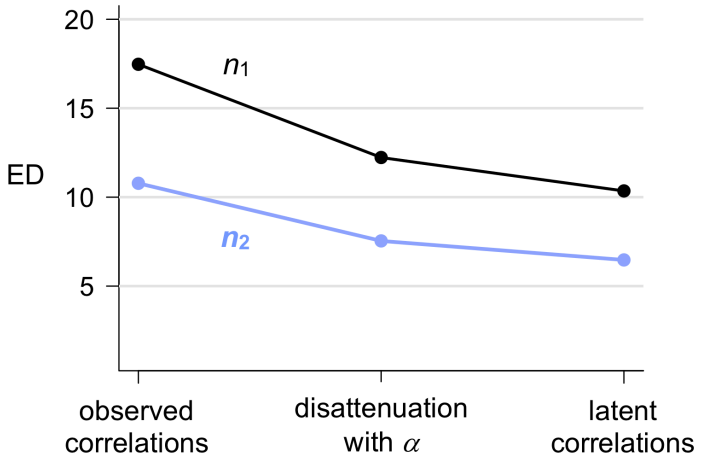

Figure 8. Effective dimensionality (ED) estimated with indices n_1 and n_2 from the correlations among 30 personality facets. Estimates based on observed correlations are compared with those obtained after disattenuation with Cronbach's α and latent variable modeling (confirmatory factor analysis).

After disattenuation with α, the estimated ED decreased to $n_1 = 12.23$. As expected, the change was even larger using latent correlations, which yielded $n_1 = 10.35$. Computing n_2 as a lower-bound estimate yielded 7.54 with α disattenuation and 6.47 with latent correlations. These results indicate that, once measurement error is accounted for, the dimensionality of the 30 facets is effectively equivalent to about 10 orthogonal dimensions, with about 6 dimensions as a conservative lower bound. In light of these findings, one may further reconsider the implications of van Tilburg's (2019) argument about the unusualness of personality profiles. If the effective number of independent dimensions at the level of latent traits is about 10, the "true" personality profiles of most people are going to be even closer to the centroid than their observed profiles based on questionnaire scores, which have an ED of about 17 (see Figure 7).

To illustrate the difference between ED and ID in this dataset, parallel analysis was used to estimate the number of reliable components in PCA (*psych* package). The results suggested 6 components; note that, in large samples such as the present one, parallel analysis converges with the classic Kaiser-Guttman rule of retaining the components with eigenvalues > 1 (Guttman, 1954; see Revelle, 2019). Parallel analysis for EFA is less straightforward, as it requires *a priori* assumptions about the underlying factor structure (see Revelle, 2019). The one-factor approach that is the default in the *psych* package suggested 8 factors. The comparison between the ID estimated with parallel analysis (6-8) and the ED estimated after error correction (~10) is potentially informative. The n_1 index was larger than both the PCA- and EFA-based estimates, and the PCA-based estimate was close to the lower bound indicated by n_2. The extra

dimensionality detected by ED is not readily explained by measurement error in the observed variables, since ED indices were calculated from error-corrected matrices. Thus, the discrepancy between ED and ID might simply reflect the idiosyncratic content of individual facets, but might also point to the presence of meaningful constructs that are not adequately captured by the first 6-8 dimensions of variation. This possibility is plausible in light of other studies of the Big Five model, which have identified 10 intermediate "aspects" of personality between the level of narrow facets and that of broad domains (DeYoung et al., 2007).

6.2. Cross-cultural differences in personality covariation

The next example shows how ED can be employed to study patterns of variation in the dimensionality of a set of variables—in this case, across multiple samples. Using the cross-cultural data by Schmitt et al. (2007), Lukaszewski et al. (2017) calculated the degree of covariation among the Big Five domains (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) in 55 countries from across the world. The index of covariation chosen by the authors was the average r^2 between pairs of traits, which ranged from .01 to .21 (mean $r^2 = .05$). The main hypothesis tested in the study was that personality traits would be more differentiated (i.e., less strongly correlated) in countries with higher levels of socioecological complexity. Consistent with this hypothesis, the correlation between a composite index of socioecological complexity and the average r^2 was –.53 (Spearman's rank correlation was $\rho = -.49$). Complexity remained a significant predictor in more complex statistical models that will not be discussed here (for details see Lukaszewski et al., 2017).

While the average r^2 is a sensible measure of covariation, the ED provides an attractive alternative for this kind of study. ED has an intuitive interpretation as the effective number of independent personality dimensions in a country; arguably, this provides a more meaningful summary of covariation patterns than the average pairwise r^2. Conveniently, ED values can be easily corrected for small-sample bias in addition to measurement error. In this study, the sample size for different countries showed a dramatic range of variation, with $N = 62$ to 2,793 (median $N = 216$); bias correction can ensure that ED estimates remain fully comparable between small and large samples.

The data for key study variables were obtained from the supplementary material in Lukaszewski et al. (2017). Correlations among personality traits were used to compute index n_1 with the function *estimate.ED.* Correlation matrices were disattenuated using the values of α in different world regions reported in the original study by Schmitt et al. (p. 185); Mestre's (2008) method was used to correct for small-sample bias. The dataset and R script used for the analysis are available at https://doi.org/10.6084/m9.figshare.11954667.

With 5 personality traits, the maximum ED is 5 when all the traits are perfectly orthogonal; smaller ED values indicate a stronger degree of personality covariation. Across the 55 countries, n_1 ranged from 2.11 (Tanzania) to 4.95 (France), with a mean of 4.15. In other words, the average dimensionality of the Big Five domains across countries was equivalent to slightly more than 4 orthogonal dimensions. Note that this is probably an overestimate, because disattenuation with α is typically less effective than other methods (e.g., disattenuation with ω_h or latent variable modeling).

Predictably, n_1 showed a strong negative correlation of –.95 with the average r^2 calculated by Lukaszewski et al. (Figure 9). Socioecological complexity was more strongly associated with the corrected n_1 (r = .66; Spearman's ρ = .59) than with the average r^2 employed in the original study (r = –.53; Spearman's ρ = –.49; all correlations p < .001). The same pattern remained after controlling for sample size (log-transformed): the partial correlations of socioecological complexity were .60 with n_1 and –.49 with the average r^2. As it turns out, this improvement was due to the correction for measurement error: when the average r^2 was computed from disattenuated correlations, it performed similarly to the corrected n_1 (r = –.64; partial r = –.58; Spearman's ρ = –.61). In conclusion, this example shows that ED can be usefully employed as a comparative measure of trait covariation. ED indices have an intuitive interpretation, and can be easily adjusted for unequal sample sizes and/or differences in measurement quality.

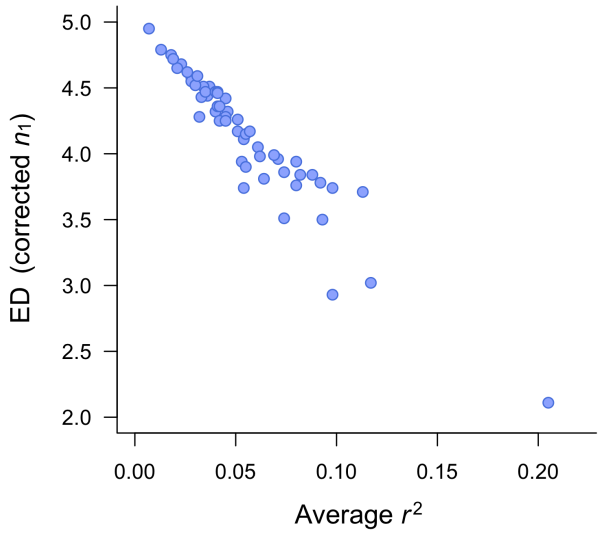

Figure 9. The relation between two indices of personality covariation across 55 countries (data from Lukaszewski et al., 2017). The indices are the average r^2 between pairs of traits (used in the original study), and the effective dimensionality (ED) estimated with n_1. The n_1 index was corrected for small-sample bias and measurement error (disattenuation with Cronbach's α).

7. Conclusion

The effective dimensionality of a set of variables is a useful but underutilized measure of correlational structure. Alone or in combination with estimates of intrinsic dimensionality, ED indices can be used to inform decisions about data analysis and answer meaningful empirical questions. Hopefully, this tutorial will encourage more researchers to incorporate this versatile tool in their own statistical practice.

References

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In J. Van den Bussche & V. Vianu (Eds.), *International conference on database theory 2001* (pp. 420-434). Berlin, Germany: Springer. https://doi.org/10.1007/3-540-44503-X_27

Altman, N., & Krzywinski, M. (2018). The curse(s) of dimensionality. *Nature Methods, 15,* 399-400. https://doi.org/10.1038/s41592-018-0019-x

Bates, D., & Maechler, M. (2019). Matrix v. 1.2-17. URL: https://CRAN.R-project.org/package=Matrix

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is "nearest neighbor" meaningful? In C. Beeri & P. Buneman (Eds.), *7th International conference on database theory* (pp. 217-235). Berlin, Germany: Springer. https://doi.org/10.1007/3-540-49257-7_15

Bishara, A. J., & Hittner, J. B. (2015). Reducing bias and error in the correlation coefficient due to nonnormality. *Educational and Psychological Measurement, 75,* 785-804. https://dx.doi.org/10.1177/0013164414557639

Bretherton, C. S., Widmann, M., Dymnikov, V. P., Wallace, J. M., & Bladé, I. (1999). The effective number of spatial degrees of freedom of a time-varying field. *Journal of Climate, 12,* 1990-2009. https://doi.org/10.1175/1520-0442(1999)012<1990:TENOSD>2.0.CO;2

Bromiley, P. A., Thacker, N. A., & Bouhova-Thacker, E. (2004). Tina Memo 2004-04: Shannon entropy, Rényi entropy, and information. *Technical report, Imaging Science and Biomedical Engineering Division, Medical School, University of Manchester*. URL: http://www.tina-vision.net/docs/memos/2004-004.pdf

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: Guilford.

Budescu, D. V., & Budescu, M. (2012). How to measure diversity when you must. *Psychological Methods, 17,* 215-227. https://dx.doi.org/10.1037/a0027129

Campadelli, P., Casiraghi, E., Ceruti, C., & Rozza, A. (2015). Intrinsic dimension estimation: Relevant techniques and a benchmark framework. *Mathematical Problems in Engineering, 2015,* 759567. https://dx.doi.org/10.1155/2015/759567

Cangelosi, R., & Goriely, A. (2007). Component retention in principal component analysis with application to cDNA microarray data. *Biology Direct, 2,* 2. https://doi.org/10.1186/1745-6150-2-2

Carreira-Perpiñán, M. A. (1996). A review of dimension reduction techniques. *Technical Report of the Department of Computer Science, University of Sheffield, CS-96-09,* 1-69.

Cheverud, J. M. (2001). A simple correction for multiple comparisons in interval mapping genome scans. *Heredity, 87*, 52-58. https://doi.org/10.1046/j.1365-2540.2001.00901.x

Conti, G., Frühwirth-Schnatter, S., Heckman, J. J., & Piatek, R. (2014). Bayesian exploratory factor analysis. *Journal of Econometrics, 183,* 31-57. https://doi.org/10.1016/j.jeconom.2014.06.008

Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment, 4,* 5–13.

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology, 93,* 880-896. https://doi.org/10.1037/0022-3514.93.5.880

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105,* 399-412. https://doi.org/10.1111/bjop.12046

Durrant, R. J., & Kabán, A. (2009). When is "nearest neighbor" meaningful: A converse theorem and implications. *Journal of Complexity, 25,* 385-397. https://doi.org/10.1016/j.jco.2009.02.011

Einbeck, J., & Kalantan, Z. (2013). Intrinsic Dimensionality estimation for high-dimensional data sets: New approaches for the computation of correlation dimension. *Journal of Emerging Technologies in Web Intelligence, 5,* 91-97. https://doi.org/10.4304/jetwi.5.2.91-97

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4,* 272-299. https://doi.org/10.1037/1082-989X.4.3.272

Facco, E., d'Errico, M., Rodriguez, A., & Laio, A. (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports, 7,* 12140. https://doi.org/10.1038/s41598-017-11873-y

Fox, J. (2019). polycor v. 0.7-10. URL: https://CRAN.R-project.org/package=polycor

Fraedrich, K., Ziehmann, C., & Sielmann, F. (1995). Estimates of spatial degrees of freedom. *Journal of Climate, 8,* 361-369. https://doi.org/10.1175/1520-0442(1995)008<0361:EOSDOF>2.0.CO;2

Giraud, C. (2015). *Introduction to high-dimensional statistics*. Boca Raton, FL: CRC Press.

Gnedenko, B., & Yelnik, I. (2016). Minimum entropy as a measure of effective dimensionality. *SSR.* https://dx.doi.org/10.2139/ssrn.2767549

Golino, H. F. (2019). *EGAnet v. 0.8*. URL: https://CRAN.R-project.org/package=EGAnet

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PLoS One, 12,* e0174035. https://doi.org/10.1371/journal.pone.0174035

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika, 19,* 149 –161. https://dx.doi.org/10.1007/BF02289162

Higham, N. J. (2002). Computing the nearest correlation matrix—a problem from finance. IMA Journal of Numerical Analysis, 22, 329-343. https://doi.org/10.1093/imanum/22.3.329

Hill, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology, 54,* 427-432. https://doi.org/10.2307/1934352

Houle, M. E., Kriegel, H. P., Kröger, P., Schubert, E., & Zimek, A. (2010). Can shared-neighbor distances defeat the curse of dimensionality? In M. Gertz & B. Ludäscher (Eds.), *International Conference on Scientific and Statistical Database Management 2010* (pp. 482-500). Berlin, Germany: Springer. https://doi.org/10.1007/978-3-642-13818-8_34

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality, 51,* 78–89. https://doi.org/10.1016/j.jrp.2014.05.003

Johnson, J. A. (2015). *Johnson's IPIP-NEO data repository*. Retrieved February 26, 2018, from the Open Science Framework website. URL: https://osf.io/tbmh5/

Jost, L. (2006). Entropy and diversity. *Oikos, 113,* 363-375. https://doi.org/10.1111/j.2006.0030-1299.14714.x

Kaiser, T. (2019). Nature and evoked culture: Sex differences in personality are uniquely correlated with ecological stress. *Personality and Individual Differences, 148,* 67-72. https://doi.org/10.1016/j.paid.2019.05.011

Kirkpatrick, M. (2009). Patterns of quantitative genetic variation in multiple dimensions. *Genetica, 136,* 271-284. https://doi.org/10.1007/s10709-008-9302-6

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York: Guilford.

Korn, F., Pagel, B. U., & Faloutsos, C. (2001). On the "dimensionality curse" and the "self-similarity blessing". *IEEE Transactions on Knowledge and Data Engineering, 13,* 96-111. https://doi.org/10.1109/69.908983

Lai, C. D., Rayner, J. C., & Hutchinson, T. P. (1998). Properties of the sample correlation of the bivariate lognormal distribution. *Proceedings of the ICOTS, 5,* 312-318.

Ledoit, O., & Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Annals of Statistics, 40,* 1024-1060. https://dx.doi.org/10.1214/12-AOS989

Ledoit, O., & Wolf, M. (2015). Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. *Journal of Multivariate Analysis, 139,* 360-384. https://doi.org/10.1016/j.jmva.2015.04.006

Lever, J., Krzywinski, M., & Altman, N. (2016). Points of significance: Regularization. *Nature Methods, 13,* 803-804. https://doi.org/10.1038/nmeth.4014

Li, J., & Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity, 95,* 221-227. https://doi.org/10.1038/sj.hdy.6800717

Lukaszewski, A. W., Gurven, M., von Rueden, C. R., & Schmitt, D. P. (2017). What explains personality covariation? A test of the socioecological complexity hypothesis. *Social Psychological and Personality Science, 8,* 943-952. https://doi.org/10.1177/1948550617697175

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23,* 412-433. https://doi.org/10.1037/met0000144

Minka, T. P. (2001). Automatic choice of dimensionality for PCA. *Advances in Neural Information Processing Systems* (pp. 598-604). Boston, MA: MIT Press.

Nakajima, S., Sugiyama, M., & Babacan, S. D. (2011). On Bayesian PCA: Automatic dimensionality selection and analytic solution. In L. Getoor & T. Scheffer (Eds.), *28th International Conference on Machine Learning* (pp. 497-504). Madison, WI: Omnipress.

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis, 49,* 974-997. https://doi.org/10.1016/j.csda.2004.06.015

Pirkl, R. J., Remley, K. A., & Patané, C. S. L. (2012). Reverberation chamber measurement correlation. *IEEE Transactions on Electromagnetic Compatibility, 54*, 533-545. https://doi.org/10.1109/TEMC.2011.2166964

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

Ramprasad, P. (2016). *nlshrink* v. 1.0.1. URL: https://CRAN.R-project.org/package=nlshrink

Rényi, A. (1961). On measures of entropy and information. In J. Neyman (Ed.), *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, Vol. 1: Contributions to the theory of statistics* (pp. 547-561). University of California Press, Berkeley, CA. URL: https://projecteuclid.org/euclid.bsmsp/1200512181

Revelle, W. (2019). *psych v. 1.8.12*. URL: https://CRAN.R-project.org/package=psych

Revelle, W., & Condon, D. M. (2018). Reliability. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing* (pp. 709-749). Hoboken, NJ: Wiley.

Roy, O., & Vetterli, M. (2007). The effective rank: A measure of effective dimensionality. In M. Domański, R. Stasiński, & M. Bartkowiak (Eds.), *15th European Signal Processing Conference* (pp. 606-610). Poznań, Poland: PTETiS.

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment, 24,* 282-292. https://doi.org/10.1037/a0025697

Schmitt, D. P., Allik, J., McCrae, R. R., & Benet-Martínez, V. (2007). The geographic distribution of Big Five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology, 38,* 173-212. https://doi.org/10.1177/0022022106297299

Seghouane, A. K., & Cichocki, A. (2007). Bayesian estimation of the number of principal components. *Signal Processing, 87,* 562-568.

Stone, J. V. (2015). Information theory: A tutorial introduction. *Sebtel Press*.

Stone, J. V. (2019). Information theory: A tutorial introduction. *arXiv*, 1802.05968v3. URL: https://arxiv.org/abs/1802.05968

Strang, G. (2016). *Introduction to linear algebra* (5th ed.). Wellesley, MA: Wellesley-Cambridge Press. URL: https://math.mit.edu/linearalgebra

Suh, C., & You, K. (2018). *Rdimtools v. 0.4.2*. URL: https://CRAN.R-project.org/package=Rdimtools

Tuomisto, H. (2010). A diversity of beta diversities: Straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. *Ecography, 33,* 2-22. https://doi.org/10.1111/j.1600-0587.2009.05880.x

van der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: A comparative review. *Tilburg University Technical Report,* 2009-005. URL: https://lvdmaaten.github.io/publications/papers/TR_Dimensionality_Reduction_Review_2009.pdf

van Tilburg, W. A. (2019). It's not unusual to be unusual (or: A different take on multivariate distributions of personality). *Personality and Individual Differences, 139,* 175-180. https://doi.org/10.1016/j.paid.2018.11.021

Wagner, G. P., Kenney-Hunt, J. P., Pavlicev, M., Peck, J. R., Waxman, D., & Cheverud, J. M. (2008). Pleiotropic scaling of gene effects and the "cost of complexity". *Nature, 452,* 470-472. https://doi.org/10.1038/nature06756

Yalcin, I., & Amemiya, Y. (2001). Nonlinear factor analysis as a statistical method. *Statistical Science, 16,* 275-294. URL: https://projecteuclid.org/euclid.ss/1009213729

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70,* 123-133. https://doi.org/10.1007/s11336-003-0974-7

Zimek, A., Schubert, E., & Kriegel, H. P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining, 5,* 363-387. https://doi.org/10.1002/sam.11161

Zollanvari, A., Saccone, N. L., Bierut, L. J., Ramoni, M. F., & Alterovitz, G. (2011). Is the reduction of dimensionality to a small number of features always necessary in constructing predictive models for analysis of complex diseases or behaviours? *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 3573-3576). https://dx.doi.org/10.1109/IEMBS.2011.6090596

Zopluoglu, C., & Davenport Jr., E. C. (2017). A Note on Using Eigenvalues in Dimensionality Assessment. *Practical Assessment, Research & Evaluation, 22,* 7. URL: https://pareonline.net/getvn.asp?v=22&n=7

Zwiggelaar, R. (2014). Intrinsic dimensionality. In H. Strange and R. Zwiggelaar (Eds.), *Open problems in spectral dimensionality reduction* (pp. 41-52). New York: Springer. https://doi.org/10.1007/978-3-319-03943-5_4