Quantifying the Informativeness of Similarity Measurements

Journal of Machine Learning Research 18 (2017) 1-61 Submitted 6/16; Revised 4/17; Published 7/17

Quantifying the Informativeness of Similarity Measurements

Austin J. Brockmeier AJBROCKMEIER@GMAIL.COM
Department of Computer Science
University of Liverpool
Liverpool L69 3BX, UK

Tingting Mu TINGTING.MU@MANCHESTER.AC.UK
Sophia Ananiadou SOPHIA.ANANIADOU@MANCHESTER.AC.UK
School of Computer Science
University of Manchester
Manchester M1 7DN, UK

John Y. Goulermas J.Y.GOULERMAS@LIVERPOOL.AC.UK
Department of Computer Science
University of Liverpool
Liverpool L69 3BX, UK

Editor: Christoph Lampert

Abstract

In this paper, we describe an unsupervised measure for quantifying the 'informativeness' of correlation matrices formed from the pairwise similarities or relationships among data instances. The measure quantifies the heterogeneity of the correlations and is defined as the distance between a correlation matrix and the nearest correlation matrix with constant off-diagonal entries. This non-parametric notion generalizes existing test statistics for equality of correlation coefficients by allowing for alternative distance metrics, such as the Bures and other distances from quantum information theory. For several distance and dissimilarity metrics, we derive closed-form expressions of informativeness, which can be applied as objective functions for machine learning applications. Empirically, we demonstrate that informativeness is a useful criterion for selecting kernel parameters, choosing the dimension for kernel-based nonlinear dimensionality reduction, and identifying structured graphs. We also consider the problem of finding a maximally informative correlation matrix around a target matrix, and explore parameterizing the optimization in terms of the coordinates of the sample or through a lower-dimensional embedding. In the latter case, we find that maximizing the Bures-based informativeness measure, which is maximal for centered rank-1 correlation matrices, is equivalent to minimizing a specific matrix norm, and present an algorithm to solve the minimization problem using the norm's proximal operator. The proposed correlation denoising algorithm consistently improves spectral clustering. Overall, we find informativeness to be a novel and useful criterion for identifying non-trivial correlation structure.

Keywords: correlation matrices, similarity information, kernel methods, information theory, quantum information theory, clustering

1. Introduction

Shannon's entropy measures the dispersion of a sample of objects among the possible elements of a discrete space. A sample consisting solely of repeated instances of the same element has minimal entropy, and a sample wherein each object is distinct from the rest has maximum entropy. If these samples correspond to the labels assigned by a clustering algorithm, then neither the minimal entropy sample corresponding to a single cluster, nor the maximal entropy sample, where each instance is in its own cluster, is very informative about any underlying organization within the sample. In this work, we investigate and introduce univariate measures of *informativeness* that are minimized for both the maximal and minimal entropy samples. Informativeness is based on the heterogeneity of the similarity measurements for a sample and is applicable to samples from any space that has a positive semidefinite correlation measure. The positive definiteness ensures that the similarity measurements can be represented as inner-products in a Hilbert-space, or more concretely, that a given sample can be embedded in Euclidean space regardless of the original space. The restriction to correlation measurements ensures each instance is represented by a vector with equal magnitude, which distinguishes the effect of differences in similarity from differences in scale or variance, and ensures that each object is equally represented. Any positive semidefinite matrix formed from pairwise similarity measurements can be normalized to obtain a correlation matrix.

The notion of informativeness coincides with hypothesis tests for the homogeneity of correlation coefficients, that is, how uniform the off-diagonal elements of a correlation matrix are (Bartlett, 1954; Anderson, 1963; Lawley, 1963; Gleser, 1968; Aitkin et al., 1968; Steiger, 1980; Brien et al., 1984). According to this notion, a *non-informative* correlation matrix has equal off-diagonal elements, which indicates that no pair of objects is more similar than any other pair. For example, a constant matrix of all ones implies the objects are indistinguishable, and an identity matrix implies the objects are all distinct with no similarity. While observing an identity correlation matrix may be informative itself, because the objects have no dependence, no hierarchical grouping or clustering of the objects is appropriate, and any reordering of the objects yields the same correlation matrix. This last property—invariance to permutations—provides another definition of a non-informative correlation matrix.

The null hypothesis of the aforementioned tests is that the expected value of the correlation matrix has equal off-diagonal elements, where the parameter of the off-diagonal elements is left unspecified. Tests of equality of correlation coefficients are a specific case of hypothesis tests for patterns within sets of correlation coefficients (Steiger, 1980). Instead of hypothesis testing, we are interested in using informativeness as an objective function for unsupervised learning and for making relative comparisons between correlation matrices of equal size—in particular, between matrices corresponding to the same sample, e.g., kernel matrices formed with different kernel functions, or correlation matrices obtained by applying different dimensionality reduction techniques. For instance, evaluating a Gaussian kernel on a sample of distinct points yields a non-informative correlation matrix when the kernel bandwidth is chosen either too large (resulting in a nearly constant matrix) or too small (resulting in an identity matrix). These extremes correspond to the minimal and maximal cases of matrix entropy as defined by von Neumann (1955) for quantum density matrices. While many heuristics can be used to select an appropriate kernel bandwidth or

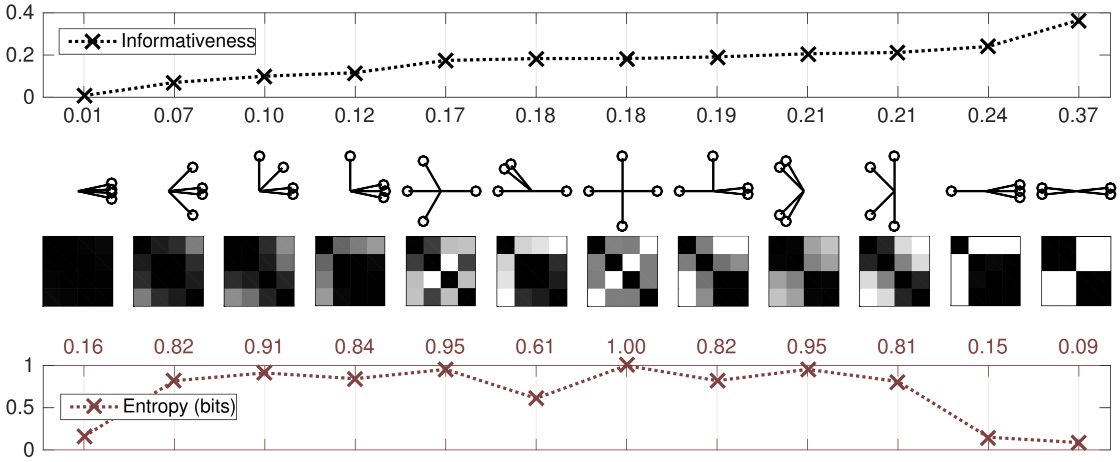

Figure 1: Informativeness versus von Neumann entropy for correlation matrices obtained from various configurations of four unit vectors. Both measures are minimal when the vectors are configured in a single cluster. Informativeness is higher for nontrivial clusterings, whereas entropy is maximized when the vectors are maximally separated.

the dimension of a lower-rank approximation of a kernel matrix, informativeness provides an unsupervised objective function for these selections.

Informativeness can be used to rank different samples and identify samples that have more structure. An example comparing informativeness to entropy for different samples of two-dimensional vectors is shown in Figure 1. Informativeness can also be applied to graphs, since the normalized graph Laplacian is a correlation matrix, and informativeness can be used to identify graphs that exhibit more structure, for example, regularity or clustering. As an objective function, informativeness can be used for enhancing the observed structure in a sample by searching for a matrix with more exaggerated differences in correlation coefficients around a target matrix.

While there are existing statistical tests to measure the heterogeneity of correlation coefficients, we define informativeness based on the distance between a given correlation matrix and the nearest non-informative matrix. In other words, informativeness is proportional to the distance between a given correlation matrix and the set of non-informative matrices. This definition is essential, since defining the distance to a particular non-informative matrix, either the identity matrix or a constant matrix, is insufficient as these distances will be proportional or inversely proportional to the von Neumann entropy. Even if two equal-sized correlation matrices are closest to distinct non-informative matrices, the distance to the set can be used to determine which matrix is more informative.

As defined, informativeness will be dependent upon the choice of the distance metric. Among other distances, we explore some from quantum information theory (Fuchs and van de Graaf, 1999; Nielsen and Chuang, 2000). These distances are applicable since any positive semidefinite matrix can be scaled to be a valid quantum density matrix. In particular, we find the Bures distance (Bures, 1969; Uhlmann, 1976), which generalizes the

Hellinger distance to matrices (Nielsen and Chuang, 2000; Koltchinskii and Xia, 2015), to have useful properties. Unlike many distance or similarity functions that are based on the Hilbert-Schmidt inner product, the similarity function for the Bures distance (known as fidelity (Jozsa, 1994)) relies on the trace norm. The Bures distance is both contractive and locally Riemannian, and can be considered the best-suited distance metric for the space of quantum density matrices (Bromley et al., 2014).

For several choices of distance metrics or dissimilarity functions, the minimum distance to the set of non-informative matrices has a closed-form expression, but this is not the case for the Bures distance. Nonetheless, using the sub-additive property of the trace norm, we derive a closed-form expression of the lower bound which is tight in most cases. Interestingly, the resulting measure leads to a convex cost function that is inversely proportional to the informativeness of a correlation matrix in terms of its Euclidean-space embedding. Besides this unique property, we find the measures of informativeness based on the Bures distance, and the other quantum distances, to perform better in machine learning applications than those based on the Hilbert-Schmidt inner product or the existing test statistics for homogeneous correlation coefficients (Bartlett, 1954; Lawley, 1963).

We organize the exposition of the paper as follows: In Section 2 we introduce some preliminaries regarding positive semidefinite, correlation, and quantum density matrices. In Section 3 we introduce the general definition of informativeness and specific measure instances based on various distance and dissimilarity functions. The derivations are included in Appendix B. We conclude the section with an illustration of the various measures on patterned correlation matrices corresponding to different partitions. In Section 4 we introduce the problem of finding a maximally informative correlation matrix nearby a target matrix, a problem we refer to as correlation matrix denoising. We also propose first-order optimizations of informativeness for a sample's correlation matrix parameterized in terms of a kernel function applied to the sample coordinates. For correlation matrix denoising, we also parameterize the optimization in terms of Euclidean-space embeddings and detail the cost functions that arise from maximizing the informativeness of the embeddings. Moreover, we show that the Bures-based informativeness is inversely proportional to a specific matrix norm. Finally, using the norm's proximal operator, we propose an algorithm to solve a relaxation of the correlation denoising problem. In Section 5 we explore different applications of informativeness for machine learning algorithms including correlation matrix denoising. A discussion of the results and general conclusions are given in Section 6.

2. Preliminaries

For a positive integer m, we use $[m]$ to denote the set $\{1, \dots, m\}$. We denote a scalar variable as x or X, a set as $\mathcal{X}$, a real-valued column vector as $\mathbf{x}$, a matrix as $\mathbf{X}$, and its jth column as $\mathbf{x}_j$. Vectorizing the diagonal of a matrix is denoted by $\operatorname{diag}(\mathbf{X})$, and $\operatorname{diag}(\mathbf{x})$ denotes a diagonal matrix with $[\operatorname{diag}(\mathbf{x})]_{i,i} = x_i$. Inner-product notation $\langle \cdot, \cdot \rangle$ is used for both vectors and matrices, according to $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_i x_i y_i$ and $\langle \mathbf{X}, \mathbf{Y} \rangle = \operatorname{tr}\big(\mathbf{X}^\top \mathbf{Y}\big) = \sum_{i,j} X_{i,j} Y_{i,j}$, respectively. The Euclidean norm of a vector is denoted by $\|\mathbf{x}\|_2$, and $\|\mathbf{X}\|_F^2 = \langle \mathbf{X}, \mathbf{X} \rangle$ denotes the square of the Frobenius norm, which is the sum of the squared entries of $\mathbf{X}$ and equals the sum of the squares of its singular values. We denote the trace norm (also called the nuclear norm or Schatten 1-norm) by $\|\mathbf{X}\|_*$, defined as the sum of all the singular values of $\mathbf{X}$, and by $\|\mathbf{X}\|_2$ the spectral (operator) norm which is the largest singular value of $\mathbf{X}$.

2.1 Similarity, Correlation, and Embedding Matrices

We assume that we are provided with measurements of the similarity between n objects. This similarity corresponds to either a non-negative symmetric bivariate measure, for which larger positive values indicate similarity between objects and values near zero indicate weak relationships, or if the concept of negative correlation is applicable, a symmetric bivariate measure where negative values indicate inversely correlated objects. The similarity values are represented by an $n \times n$ real symmetric matrix $\mathbf{A}$, where $A_{i,j}$ denotes the similarity between objects i and j. Assuming that the similarities of all objects to themselves are greater than zero, the similarity matrix can be rescaled such that all the self-similarities are one. That is, any symmetric matrix $\tilde{\mathbf{A}}$ with $\tilde{A}_{i,i} > 0$ can be symmetrically normalized via

$$A_{i,j} = \frac{\tilde{A}_{i,j}}{\sqrt{\tilde{A}_{i,i}\tilde{A}_{j,j}}}, \tag{1}$$

such that $A_{i,i} = 1, i \in [n]$, or equivalently as $\mathbf{A} = \mathbf{D}\tilde{\mathbf{A}}\mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix with entries $D_{i,i} = \frac{1}{\sqrt{\tilde{A}_{i,i}}}$. We assume that we have no further information about the objects or their original representations.

Many similarity measures, such as positive definite kernel functions, will yield a positive semidefinite similarity matrix. Positive semidefinite matrices can be represented by the embeddings of the objects in a Hilbert space for which the theory of reproducing kernel Hilbert spaces (Aronszajn, 1950) applies. That is, if a given similarity matrix is positive semidefinite, then there exists a set of n points in a Hilbert space $\mathcal{H}$, such that the inner product between any pair of points quantifies their similarity. Specifically, $A_{i,j} = \langle \mathbf{z}_i, \mathbf{z}_j \rangle$ where $\mathbf{z}_i \in \mathcal{H}$ with $i \in [n]$, denotes the point corresponding to the ith object. For finite dimensional embeddings in $\mathbb{R}^b$, we refer to the $b \times n$ matrix of concatenated vectors $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]$ as an embedding. This is not unique as any orthogonal transformation in $\mathbb{R}^b$ yields an equally valid representation.

We restrict our attention to *correlation matrices*,[1] which are matrices that are both positive semidefinite and normalized such that the diagonal elements are equal to one. Applying the symmetric normalization from Equation 1 to a positive semidefinite similarity matrix yields a correlation matrix, and in fact, the symmetric normalization yields the closest correlation matrix in terms of the Bures distance, which we show in Lemma 2 in Appendix A. A correlation matrix can be represented by a Hilbert space embedding where each object is represented by a point on a unit sphere. The convex set of all the $n \times n$ correlation matrices is known as the elliptope and it is defined as $\mathcal{E} = \left\{\mathbf{X} \in \mathbb{R}^{n\times n} : \mathbf{X} = \mathbf{X}^\top \succcurlyeq 0, X_{i,i} = 1, \forall i \in [n]\right\}$. It contains 2^{n-1} vertices that have full dimensional normal cones and correspond to rank-1 cut matrices of the form $\mathbf{v}\mathbf{v}^\top$ with $\mathbf{v} \in \{\pm 1\}^n$ (Laurent and Poljak, 1995, 1996). The convex hull of the cut matrices defines the cut polytope, and a strict subset of correlation matrices are members of this polytope. This relationship between the elliptope and the

[1] Their name is due to the fact they coincide with the set of Pearson correlation matrices for multivariate samples.

cut polytope is exploited in semidefinite programming relaxations (Saunderson et al., 2012) of combinatorial problems associated with dividing objects, such as finding the maximum cut of a graph (Goemans and Williamson, 1995). The set of embeddings of correlation matrices is the oblique manifold $\mathcal{OB} = \left\{\mathbf{Z} \in \mathbb{R}^{b\times n} : \|\mathbf{z}_i\|_2 = 1, \forall i \in [n]\right\}$ (Trendafilov and Lippert, 2002; Absil and Gallivan, 2006), where any correlation matrix can be written as $\mathbf{Z}^\top\mathbf{Z}$. When $b = 1$, the oblique manifold is simply the set of vectors with elements ± 1. The elements of a correlation matrix necessarily lie within $[-1, +1]$ since from the Cauchy-Bunyakovsky-Schwarz inequality we have $|\langle \mathbf{z}_i, \mathbf{z}_j\rangle| \leq \|\mathbf{z}_i\| \cdot \|\mathbf{z}_j\|$.

When the underlying objects are real-valued vectors, similarity measures that yield correlation matrices include the cosine similarity; various correlation coefficients, such as Pearson's, Spearman's, or Kendall's; and positive definite shift-invariant kernel functions, such as the Laplacian or Gaussian kernels. Efficient positive semidefinite kernels, which can generally be normalized, have also been developed for non-vectorial objects, such as strings (Lodhi et al., 2002), point sets (Kondor and Jebara, 2003), and the univariate point sets known in neuroscience as spike trains (Park et al., 2013). Nonetheless, some similarity measures yield normalized but indefinite matrices. To obtain a correlation matrix, one approach is to apply a non-negative threshold to the eigenvalues by setting any negative eigenvalues to zero (Higham, 1988) and then apply the symmetric normalization. An alternative is to find the nearest, in terms of Euclidean distance, correlation matrix (Higham, 2002; Qi and Sun, 2006). Both of these approaches, however, may distort the measures of similarity. Another approach is to treat each row of the similarity matrix as a vector-space representation of each object and compute a second-order similarity matrix using cosine similarity between these vectors. The result will be a correlation matrix. A final possibility motivated by spectral graph theory (Chung, 1997), is to treat the similarity matrix as the weighted affinity matrix of a graph, and then compute the normalized graph Laplacian, which is always a correlation matrix with non-positive off-diagonal entries.

2.2 Density Matrices

Quantum density matrices represent the probabilities associated with the outcomes of measurements on quantum states (Nielsen and Chuang, 2000; Hayashi, 2006). Ignoring their physical interpretations, density matrices are trace-normalized positive semidefinite Hermitian matrices. This means that statistical tools developed to analyze quantum density matrices can be applied to positive semidefinite matrices, including correlation matrices, after rescaling them by their trace.

Quantum information theory is based on the fact that a quantum density matrix is itself a non-commutative generalization of a probability mass function (Nielsen and Chuang, 2000). The positive semidefiniteness and unit-trace restrictions ensure that a matrix's eigenvalues are strictly positive and sum to 1; these are the same properties that a probability mass function has. This allows extensions of information theoretic quantities, such as entropy and mutual information to be defined for density matrices, and subsequently, trace-normalized positive semidefinite matrices (Sanchez Giraldo et al., 2015).

Our motivation for introducing density matrices is to leverage quantum information-based distance measures, which have unique properties compared to more familiar measures. Any correlation matrix can be converted to a density matrix by dividing by n. The set of rescaled correlation matrices forms a subset of symmetric density matrices with constant diagonals. In the next section, we introduce our distance-based framework for measures of informativeness that can use the quantum distance measures.

3. Defining and Measuring Informativeness

In this section, we propose a univariate measure of how informative a given correlation matrix is, based on its distance from the nearest non-informative correlation matrix. This distance can be measured using various distance/dissimilarity measures and for some of these measures we show that we can obtain closed-form expressions of informativeness.

3.1 Non-informative Correlation Matrices

A set of pairwise correlation coefficients is non-informative if it indicates that no pair of objects is any more similar than any other pair, i.e., there are no groups of more (or less) related objects. Two exemplary cases of non-informativeness correspond to

1. all objects being indistinguishable, and
2. all objects being distinct and equally dissimilar.

By utilizing correlation matrices, these cases can be represented by the constant matrix $\mathbf{J} = \mathbf{1}\mathbf{1}^\top$ and the identity matrix $\mathbf{I}$, respectively. In general, any correlation matrix with constant off-diagonal elements is non-informative. Reordering the objects, i.e., simultaneously permuting the rows and columns of the correlation matrix, will not change a non-informative matrix. The set of non-informative matrices forms the null hypothesis for tests of homogeneous correlation coefficients (Bartlett, 1954; Lawley, 1963). Like these test statistics, informativeness essentially measures how heterogeneous the off-diagonal entries of the correlation matrix are.

If we let ρ be the value of the off-diagonal elements, correlation matrices with constant off-diagonal elements can be expressed as $\mathbf{A}_\rho = \rho\mathbf{J} + (1-\rho)\mathbf{I}$, where the range of ρ must be restricted to ensure $\mathbf{A}_\rho$ is a valid $n \times n$ correlation matrix. Specifically, since $\mathbf{J}\mathbf{1} = n\mathbf{1}$, the spectrum of $\mathbf{A}_\rho$ is $\{\rho n + 1 - \rho, 1 - \rho\}$. For both eigenvalues to be non-negative, we must have $\frac{-1}{n-1} \leq \rho \leq 1$.

It it more convenient, however, to substitute ρ with a variable $a \in [0, 1]$, so that it parameterizes the expression of a non-informative matrix as a convex combination of two other non-informative matrices. Linearly mapping $\rho \in \left[\frac{-1}{n-1}, 1\right]$ to a, gives $a = \frac{\rho(n-1)+1}{n}$ or equivalently $\rho = \frac{an-1}{n-1}$. Substituting the latter within the expression for $\mathbf{A}_\rho$ gives $\frac{na-1}{n-1}\mathbf{J} + \frac{n(1-a)}{n-1}\mathbf{I}$, which is also equal to the expression $a\mathbf{J} + (1-a)\frac{n}{n-1}\mathbf{H}$, where $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{J}$ is the centering matrix. The parameter a is the mean of the non-informative matrix since $a = \frac{\mathbf{1}^\top\mathbf{A}_\rho\mathbf{1}}{n^2}$, whereas $\rho = \frac{\mathbf{1}^\top(\mathbf{A}_\rho - \mathbf{I})\mathbf{1}}{n(n-1)}$ is the mean of its off-diagonal elements.

In this work, we define the set of non-informative matrices to be

$$\mathcal{N} \equiv \left\{\mathbf{N}_a = a\mathbf{J} + (1-a)\frac{n}{n-1}\mathbf{H} : 0 \leq a \leq 1\right\}. \tag{2}$$

Under this parameterization, $\mathbf{N}_1 = \mathbf{J}$ is the rank-1 constant matrix, $\mathbf{N}_{\frac{1}{n}} = \mathbf{I}$ is the identity matrix, and $\mathbf{N}_0 = \frac{n}{n-1}\mathbf{H}$ is a scaled version of the centering matrix with rank $n-1$.

Geometrically, $\mathbf{J}$ is a vertex of the elliptope corresponding to the cut matrix $\mathbf{1}\mathbf{1}^\top$, $\mathbf{I}$ is the barycenter, $\frac{n}{n-1}\mathbf{H}$ is an extreme point with a one-dimensional normal cone, and all the non-informative matrices lie along the line connecting them.

From the characteristic polynomial of $\mathbf{N}_a$, it can be seen that it has one simple eigenvalue $\lambda_1 = an$ and one semisimple (due to the symmetry and, thus, diagonalizability of $\mathbf{N}_a$) eigenvalue $\lambda_2 = \frac{(1-a)n}{n-1}$ of multiplicity $n-1$. The corresponding eigenspaces $\mathrm{Null}(\mathbf{N}_a - \lambda_1\mathbf{I})$ and $\mathrm{Null}(\mathbf{N}_a - \lambda_2\mathbf{I})$ are orthogonal complementary subspaces and, since $(\lambda_1, \mathbf{1})$ is an eigenpair, they are common to the entire family of $\mathbf{N}_a$. Any vector within the eigenspace of λ_2 is an eigenvector. The existence of common eigenvectors is also supported by the fact that $\mathcal{N}$ is a commuting matrix family, which follows from the facts that $\mathbf{J}\mathbf{H} = \mathbf{0}$, and $\frac{1}{n}\mathbf{J}$ and $\mathbf{H}$ are symmetric and idempotent. Furthermore, for any given a, $\mathbf{N}_a$ can be expressed in terms of its eigenvalues as $\frac{\lambda_1}{n}\mathbf{J} + \lambda_2\mathbf{H}$, and since it is diagonalizable, for any function f defined at each λ_i, we have $f(\mathbf{N}_a) = \frac{f(\lambda_1)}{n}\mathbf{J} + f(\lambda_2)\mathbf{H}$. A particular instance of this, which we will use in the subsequent section, is the matrix square root given by $\sqrt{\mathbf{N}_a} = \sqrt{\frac{a}{n}}\mathbf{J} + \sqrt{\frac{(1-a)n}{n-1}}\mathbf{H}$.

3.2 Measuring the Distance Between Correlation Matrices

We propose to quantify the informativeness of a given correlation matrix $\mathbf{K}$ by its distance to the nearest non-informative correlation matrix. This measure of informativeness can be defined as

$$d_{\mathcal{N}}(\mathbf{K}) \equiv \min_{\mathbf{N}\in\mathcal{N}} d(\mathbf{K}, \mathbf{N}) = \min_{0\le a\le 1} d(\mathbf{K}, \mathbf{N}_a), \tag{3}$$

and can rely on various distance or dissimilarity measures applicable to matrices. To ensure informativeness is invariant to the ordering of the objects, the distance metric must be invariant to symmetric permutations of both arguments: $d(\mathbf{A}, \mathbf{B}) = d(\mathbf{\Pi}\mathbf{A}\mathbf{\Pi}^\top, \mathbf{\Pi}\mathbf{B}\mathbf{\Pi}^\top)$, where $\mathbf{\Pi}$ is a permutation matrix. For any non-informative matrix $\mathbf{N}$, we have $\mathbf{N} = \mathbf{\Pi}\mathbf{N}\mathbf{\Pi}^\top$. To ensure the informativeness measures can all be bounded between 0 and 1, we restrict our attention to bounded distance/dissimilarity measures.[2]

Table 1 contains distance metrics and dissimilarity functions—along with any underlying similarity measure—that are applicable to positive semidefinite matrices and can be used in the definition of $d_{\mathcal{N}}$. The table includes measures that have been widely applied in kernel-based machine learning, such as kernel alignment (Cristianini et al., 2002), which corresponds to the cosine similarity between the vectorized forms of matrices, the Hilbert-Schmidt independence criterion (Gretton et al., 2005), and the centered kernel alignment (Cortes et al., 2012), also known as distance correlation (Székely et al., 2007; Sejdinovic et al., 2013). The table also includes distances/dissimilarities used in quantum information theory, such as the trace distance, which is a matrix version of the Kolmogorov or total variation distance (Fuchs and van de Graaf, 1999); the quantum Jensen-Shannon divergence (Majtey et al., 2005); a distance based on the quantum Chernoff bound (Audenaert et al.,

[2] This restriction excludes the family of log-determinant divergences (Lee and Lim, 2008; Cherian et al., 2011; Chebbi and Moakher, 2012; Cichocki et al., 2015) defined on the set of symmetric positive definite matrices, e.g., the affine invariant Reimannian metric (AIRM) and other symmetric log-det divergences. In preliminary experiments, we found these metrics to have unique behavior; however, they do not perform consistently as measures of informativeness, primarily because they are unbounded and the input matrix needs to be regularized to ensure it is strictly positive definite.

Table 1: Possible distance metrics/dissimilarity functions for the realization of $d_{\mathcal{N}}(\cdot)$ in Equation 3. The first five are applicable to any conformable symmetric matrices $\mathbf{A}$ and $\mathbf{B}$, whereas the second five additionally assume positive semidefiniteness and unity trace.

Distance metric/dissimilarity function $d(\mathbf{A}, \mathbf{B})$:	Underlying similarity function:
Euclidean distance	Matrix dot-product
$\\|\mathbf{A} - \mathbf{B}\\|_F$	$\langle \mathbf{A}, \mathbf{B} \rangle$
Cosine distance	Cosine similarity, or kernel alignment
$\sqrt{2 - 2\frac{\langle \mathbf{A},\mathbf{B} \rangle}{\\|\mathbf{A}\\|_F \\|\mathbf{B}\\|_F}}$	$\frac{\langle \mathbf{A},\mathbf{B} \rangle}{\\|\mathbf{A}\\|_F \\|\mathbf{B}\\|_F}$
Hilbert-Schmidt independence criterion (HSIC)	
$\\|\mathbf{HAH} - \mathbf{HBH}\\|_F$	$\langle \mathbf{HAH}, \mathbf{HBH} \rangle$
Centered kernel alignment (CKA)	
$\sqrt{2 - 2\frac{\langle \mathbf{HAH},\mathbf{HBH} \rangle}{\\|\mathbf{HAH}\\|_F \\|\mathbf{HBH}\\|_F}}$	$\frac{\langle \mathbf{HAH},\mathbf{HBH} \rangle}{\\|\mathbf{HAH}\\|_F \\|\mathbf{HBH}\\|_F}$
Trace distance	
$\frac{1}{2}\mathrm{tr}\sqrt{(\mathbf{A} - \mathbf{B})^2} = \frac{1}{2}\\|\mathbf{A} - \mathbf{B}\\|_*$	
Quantum Jensen-Shannon (QJS) divergence	
$H(\frac{\mathbf{A}+\mathbf{B}}{2}) - \frac{1}{2}H(\mathbf{A}) - \frac{1}{2}H(\mathbf{B})$, where $H(\mathbf{A}) \triangleq -\langle \mathbf{A}, \ln \mathbf{A} \rangle$ is the von Neumann entropy	
Chernoff bound	
$\sqrt{2 - 2 \min\limits_{0 \le s \le 1} \langle \mathbf{A}^s, \mathbf{B}^{1-s} \rangle}$	$\min\limits_{0 \le s \le 1} \langle \mathbf{A}^s, \mathbf{B}^{1-s} \rangle$
Quantum Hellinger (QH) distance	Affinity
$\\|\sqrt{\mathbf{A}} - \sqrt{\mathbf{B}}\\|_F = \sqrt{2 - 2\langle \sqrt{\mathbf{A}}, \sqrt{\mathbf{B}} \rangle}$	$\langle \sqrt{\mathbf{A}}, \sqrt{\mathbf{B}} \rangle$
Bures distance (d_B)	Fidelity
$\sqrt{2 - 2\\|\sqrt{\mathbf{A}}\sqrt{\mathbf{B}}\\|_*}$	$\\|\sqrt{\mathbf{A}}\sqrt{\mathbf{B}}\\|_*^2$
Sub-Bures dissimilarity	Super-fidelity
$\sqrt{2 - 2\sqrt{\langle \mathbf{A}, \mathbf{B} \rangle + \sqrt{(1 - \\|\mathbf{A}\\|_F^2)(1 - \\|\mathbf{B}\\|_F^2)}}}$	$\langle \mathbf{A}, \mathbf{B} \rangle + \sqrt{(1 - \\|\mathbf{A}\\|_F^2)(1 - \\|\mathbf{B}\\|_F^2)}$

2007); the quantum Hellinger divergence (Luo and Zhang, 2004); the Bures distance (Bures, 1969; Uhlmann, 1976; Braunstein and Caves, 1994); and the sub-Bures dissimilarity[3] measure, which is a lower bound on the Bures distance (Mendonça et al., 2008; Miszczak et al., 2009).

Of the quantum information theoretic distances, the Bures distance can be considered the natural metric between trace-normalized positive semidefinite matrices (Bromley et al., 2014). It is contractive, locally Riemannian, invariant to the choice of embedding coordinates (Uhlmann, 1976), and in its infinitesimal form it coincides with the quantum Fisher information metric for quantum densities (Braunstein and Caves, 1994). Various elements

[3] The corresponding similarity measure, super-fidelity, can be used to define a metric distance, but it is not a lower bound on the Bures distance. The dissimilarity measure violates the triangle inequality for certain cases (Mendonça et al., 2008).

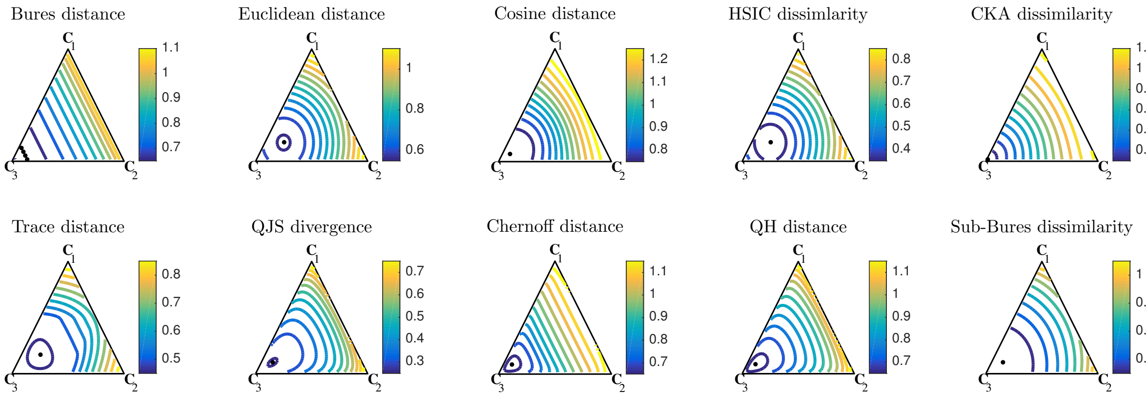

Figure 2: Equidistant contour lines for the measures in the marginal simplex of cut matrices. The Bures distance is the only distance invariant to the relative proportion of $\mathbf{C}_1$ and $\mathbf{C}_2$, which are not in the convex basis of the target matrix.

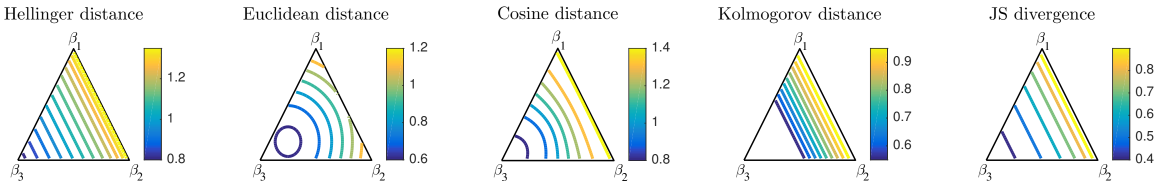

Figure 3: Equidistant contour lines for distances between points on the marginal simplex defined by $[\beta_1, \beta_2, \beta_3, 0]$ to the target point $[0, 0, 0.5, 0.5]$. The shape of the contours illustrates that most measures, besides the Euclidean and cosine distances, are invariant to the relative proportion of β_1 and β_2.

of the formulation and computation of the Bures distance and the associated fidelity similarity measure, which are relevant to the study of correlation matrices and informativeness, are presented in Appendix A.

Example 1 (Comparison of Distance/Dissimilarity Measures) *We compare the distance measures from Table 1 on correlation matrices with $n = 3$.*

In particular, we consider the set of correlation matrices $\mathbf{X}_{\boldsymbol{\beta}} = \beta_1\mathbf{C}_1 + \beta_2\mathbf{C}_2 + \beta_3\mathbf{C}_3 + \beta_4\mathbf{J}$ that can be represented as the convex combination of the elliptope vertices $\mathbf{C}_i = \mathbf{v}_i\mathbf{v}_i^\top$ and $\mathbf{J} = \mathbf{1}\mathbf{1}^\top$, with $\mathbf{v}_1 = [-1, 1, 1]^\top$, $\mathbf{v}_2 = [1, -1, 1]^\top$ and $\mathbf{v}_3 = [1, 1, -1]^\top$. Since $\sum_i \beta_i = 1$, each 3×3 matrix $\mathbf{X}_{\boldsymbol{\beta}}$ corresponds to a point of a 3-simplex. For visualization purposes, we calculate the distance between sets of correlation matrices $\mathbf{X}_{\boldsymbol{\beta}}$ and a target matrix $\mathbf{A}$. As an example, we set $\mathbf{A}$ to have as its convex basis (Brøndsted, 2012) the vertices $\mathbf{C}_3$ and $\mathbf{J}$, according to $\mathbf{A} = 0.5\ \mathbf{C}_3 + 0.5\ \mathbf{J}$, and restrict $\mathbf{X}_{\boldsymbol{\beta}}$ to lie on one face of the tetrahedron by fixing $\beta_4 = 0$. Figure 2 shows the equidistant contours for various measures.

The key distinction is how convex combinations of $\mathbf{C}_1$ and $\mathbf{C}_2$ affect the calculated distances $d(\mathbf{X}_{\boldsymbol{\beta}}, \mathbf{A})$. We can observe that the Bures distance is only a function of β_3, or equivalently of $\beta_1 + \beta_2$. That is, the Bures distance is invariant to the relative proportion of β_1 and β_2, since along lines where β_3 is fixed, every correlation matrix is equidistant from the target matrix $\mathbf{A}$. This invariance is desirable since neither $\mathbf{C}_1$ nor $\mathbf{C}_2$ is in the convex basis of the $\mathbf{A}$. An equivalent invariance occurs when the corresponding vector-based distances[4] are computed between the vector of coefficients $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3, 0]$, which defines a discrete probability distribution, and the target $[0, 0, 0.5, 0.5]$. Figure 3 shows the equidistant contours for various measures. Except for the Euclidean and cosine distances, the distance metrics are invariant to the relative proportion of β_1 and β_2, since these coefficients are not in the support of the target. In comparison, this behavior is not exhibited by any of the other matrix measures as evidenced by the nonlinear contours. For example, for fixed β_3 values, the Euclidean distance yields shorter distances to $\mathbf{A}$ when $\beta_1 = \beta_2$. The same applies to the other measures, although the effect is less obvious for the Chernoff-based measure, which behaves similar to the Bures distance for larger distances, but for shorter distances it is also susceptible to the same issue.

3.3 Derivation of Measures of Informativeness

We proceed to analytically derive closed-form expressions of informativeness for a subset of the distance/dissimilarity measures from Table 1. In the derivations we use the parametric form of the nearest non-informative matrix $\mathbf{N}_{a^\star}$, where $a^\star = \arg\min_{0 \leq a \leq 1} d(\mathbf{K}, \mathbf{N}_a)$ and $\mathbf{K} \in \mathcal{E}$ is the input correlation matrix. Given the non-informative matrix $\mathbf{N}_{a^\star}$, we define the informativeness measures as $d_{\mathcal{N}}(\mathbf{K}) = d(\mathbf{K}, \mathbf{N}_{a^\star})$ or $\frac{1}{2} d^2_{\mathcal{N}}(\mathbf{K}) = \frac{1}{2} d^2(\mathbf{K}, \mathbf{N}_{a^\star})$ for chordal distances defined on a unit sphere whose range is intrinsically $[0, \sqrt{2}]$; this ensures the informativeness measures all range between 0 and 1. The analytic expressions for $a^\star$ and $d_{\mathcal{N}}(\mathbf{K})$ are derived in Appendix B, and the expressions of informativeness are in Table 2.

These expressions assume the input matrix is a correlation matrix $\mathbf{K} \in \mathcal{E}$. If $\mathbf{K} \notin \mathcal{E}$ then the Euclidean, cosine, HSIC, and CKA measures compute the distance/dissimilarity between $\mathbf{K}$ and the set of matrices with a diagonal of all ones and constant off-diagonal elements, i.e., the set $\{\mathbf{N}_a\}$, where a is unconstrained. In this case, the nearest matrix $\mathbf{N}_{a^\star}$ is not necessarily a correlation matrix, e.g., $a^\star < 0$ or $a^\star > 1$.

We note that for the Chernoff bound the expression for informativeness does not have a closed form, but only requires the same computational complexity as a single evaluation of the Chernoff bound. Also, the sub-Bures informativeness is not strictly closed form, requiring the maximum of two algebraic expressions. We did not find a closed-form expression of informativeness for neither the trace distance nor the quantum Jensen-Shannon divergence. Consequently, computing these measures requires a search over the family of non-informative matrices, which is costly since each search iteration requires computing either the singular values or the eigenvalues of an $n \times n$ matrix.

Each measure in Table 2 is bounded between 0 and 1, and will necessarily reach the lower bound of 0 for non-informative matrices. However, the upper limit of 1 is a loose upper bound and may not be attainable. Because of this and the fact there are no units associated with informativeness;[5] the nominal value of informativeness is not meaningful for

[4] The Hellinger distance is the vector equivalent of the Bures and QH distances, and the Kolmogorov distance is the vector equivalent of the trace distance.

[5] In comparison, the units of entropy (bit, nat, or hartley) are gauged to the probability of an event.

Table 2: Informativeness measures based on the distance and dissimilarity measures from Table 1. All measures, apart from the Bures-based one, are derived in Appendix B. $\mathbf{K} \in \mathcal{E}$ is a correlation matrix and $\bar{k} = \frac{1}{n^2}\mathbf{1}^\top\mathbf{K}\mathbf{1}$ is the average of its entries. All measures are bounded within the range $[0, 1]$.

Measure:	Informativeness:	Simplified expression (=), or bound (≥):	Relevant equations:
Euclidean	$d_{\mathcal{N}}(\mathbf{K}) = \min\limits_{0\leq a\leq 1} \frac{1}{n}\|\mathbf{K} - \mathbf{N}_a\|_F$	$= \sqrt{\frac{1}{n^2}\|\mathbf{K}\|_F^2 - \frac{n\bar{k}^2-2\bar{k}+1}{n-1}}$	(B.1),(B.2)
Cosine	$\frac{1}{2}d^2_{\mathcal{N}}(\mathbf{K}) = 1 - \max\limits_{0\leq a\leq 1} \frac{\langle \mathbf{K},\mathbf{N}_a\rangle}{\|\mathbf{K}\|_F\|\mathbf{N}_a\|_F}$	$= 1 - \frac{n}{\|\mathbf{K}\|_F}\sqrt{\frac{n\bar{k}^2-2\bar{k}+1}{n-1}}$	(B.3),(B.4)
HSIC	$d_{\mathcal{N}}(\mathbf{K}) = \min\limits_{0\leq a\leq 1} \frac{1}{n}\|\mathbf{HKH} - \mathbf{HN}_a\mathbf{H}\|_F$	$= \sqrt{\frac{1}{n^2}\|\mathbf{HKH}\|_F^2 - \frac{(1-\bar{k})^2}{n-1}}$	(B.5),(B.6)
CKA	$\frac{1}{2}d^2_{\mathcal{N}}(\mathbf{K}) = 1 - \max\limits_{0\leq a\leq 1} \frac{\langle \mathbf{HKH},\mathbf{HN}_a\mathbf{H}\rangle}{\|\mathbf{HKH}\|_F\|\mathbf{HN}_a\mathbf{H}\|_F}$	$= 1 - \frac{n(1-\bar{k})}{\|\mathbf{HKH}\|_F\sqrt{n-1}}$	(B.7)
Chernoff	$d_{\mathcal{N}}(\mathbf{K}) = 1 - \max\limits_{0\leq a\leq 1} \min\limits_{0\leq s\leq 1} \frac{1}{n}\langle \mathbf{K}^s, \mathbf{N}_a^{1-s}\rangle$	$= 1 - q(\mathbf{K})$	(B.8–B.11)
QH	$\frac{1}{2}d^2_{\mathcal{N}}(\mathbf{K}) = 1 - \max\limits_{0\leq a\leq 1} \frac{1}{n}\langle \sqrt{\mathbf{K}}, \sqrt{\mathbf{N}_a}\rangle$	$= 1 - \sqrt{\frac{(\mathbf{1}^\top\sqrt{\mathbf{K}}\mathbf{1})^2}{n^3} + \frac{\mathrm{tr}^2(\mathbf{H}\sqrt{\mathbf{K}})}{n(n-1)}}$	(B.12),(B.13)
Bures	$\frac{1}{2}d^2_{\mathcal{N}}(\mathbf{K}) = 1 - \max\limits_{0\leq a\leq 1} \frac{1}{n}\|\sqrt{\mathbf{K}}\sqrt{\mathbf{N}_a}\|_*$	$\geq i(\mathbf{K}) = 1 - \sqrt{\bar{k} + \frac{\mathrm{tr}^2\sqrt{\mathbf{HKH}}}{n(n-1)}}$	(B.16),(B.17)
Sub-Bures	$\frac{1}{2}d^2_{\mathcal{N}}(\mathbf{K}) = 1 - \max\limits_{0\leq a\leq 1} \sqrt{G(\frac{1}{n}\mathbf{K}, \frac{1}{n}\mathbf{N}_a)}$	$= 1 - \sqrt{g(\mathbf{K})}$	(B.21–B.24)

comparing differently-sized correlation matrices; nonetheless, informativeness can be used for relative comparisons between correlation matrices of equal size, even if they are closest to different non-informative matrices.

For comparison, the test statistics for equality of correlation coefficients proposed by Bartlett (1954) and Lawley (1963) are included in Appendix C. The null hypothesis for these tests is that the multivariate sample, corresponding to the observed correlation matrix, is drawn from a distribution whose true correlation structure is a non-informative matrix. The main benefit of using Lawley's test statistic is that, asymptotically, its distribution under the null hypothesis has a χ^2 distribution regardless of which particular non-informative matrix corresponds to the true correlation matrix. The null distribution for the other measures (including Bartlett's) depends on the underlying non-informative matrix, which is in practice unknown.

Example 2 (Informativeness for 3×3 Correlation Matrices) *To gain a perspective of informativeness measures, we evaluate them for 3×3 correlations matrices.*

Since the Bures-based informativeness is a convex function (Theorem 4 in Appendix B.7), it is necessarily maximized at the elliptope's extreme points, which are correlation matrices with rank-r where $r(r+1) \leq 2n$ (Li and Tam, 1994). Figure 4 shows the Bures informativeness $i(\mathbf{K})$ across the parametric surface representing the surface of the elliptope. The measure is maximized at the three rank-1 correlation matrices besides $\mathbf{J}$. In general, the Bures-based informativeness is maximal for centered, rank-1 correlation matrices, as noted in Theorem 5 in Appendix B.7.

To compare the informativeness measures, we consider the subset of 3×3 correlation matrices $\mathbf{X}_{\boldsymbol{\beta}} = \beta_1\mathbf{C}_1 + \beta_2\mathbf{J} + \beta_3\frac{\mathbf{C}_2+\mathbf{C}_3}{2}$ with $\sum_i \beta_i = 1$, which are convex combinations

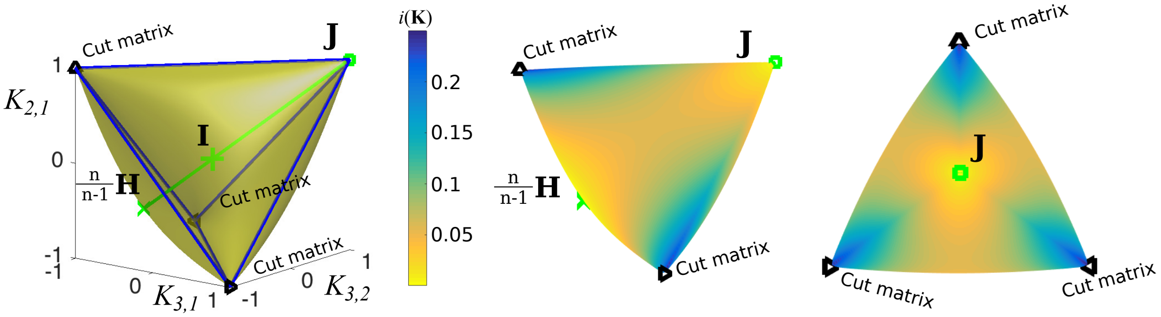

Figure 4: The Bures-based informativeness measure evaluated on the boundary of the elliptope. For $n = 3$, each correlation matrix $\mathbf{K}$ can be represented by a three-dimensional point that defines the three lower-triangular elements $K_{2,1}$, $K_{3,1}$, and $K_{3,2}$. (Left:) The family of non-informative matrices $\mathcal{N}$ is shown as a line across the elliptope passing through $\mathbf{N}_1 = \mathbf{J}$ and $\mathbf{N}_{1/n} = \mathbf{I}$ (discussed in Section 3.1). The outline of the cut polytope is shown as the wireframe connecting the four vertices. (Center, Right:) Informativeness peaks at the three extremal points corresponding to the three cut matrices.

of the elliptope vertices $\mathbf{C}_i = \mathbf{v}_i\mathbf{v}_i^\top$ and $\mathbf{J} = \mathbf{1}\mathbf{1}^\top$, with $\mathbf{v}_1 = [-1, 1, 1]^\top$, $\mathbf{v}_2 = [1, -1, 1]^\top$ and $\mathbf{v}_3 = [1, 1, -1]^\top$. Figure 5 illustrates the measures, along with Bartlett's and Lawley's test statistics, evaluated for each $\mathbf{X}_{\boldsymbol{\beta}}$ across the 2-simplex defined by $\boldsymbol{\beta}$. All measures are maximized when $\mathbf{X}_{\boldsymbol{\beta}} = \mathbf{C}_1$, but the shape of the level sets highlights the differences among the distance/dissimilarity measures underlying the informativeness measures.

Example 3 (Patterned Correlation Matrices) *We use* 16×16 *rank-1 and block diagonal correlation matrices to compare the measures of informativeness.*

The value of the informativeness measures for different partition sizes are shown in Figure 6. For binary partitions, the HSIC and Bures measures are uniquely maximized by balanced rank-1 matrices, and the measures—excepting the Euclidean, trace, and CKA measures—are maximized by balanced block-diagonal when there are two blocks. In general, balanced rank-1 matrices have higher informativeness than any of the block diagonal matrices (except for the CKA measure). For block-diagonal matrices, the Euclidean, HSIC, CKA, and sub-Bures measures are maximal for two equal-sized blocks, but the other measures (excepting the trace distance) are maximized when there are $\sqrt{n}$ equal-sized blocks. This latter disparity means that these two groups of measures will behave differently when evaluating correlation matrices with cluster structure.

On these patterned correlation matrices, we see that the Bures-based measure is a tight bound on the exact Bures informativeness measure. Additionally, the quantum Hellinger-based measure is a tight upper bound on the exact Bures measure except in the rank-1 case. The cosine-based measure also tracks the QH measure on these patterned matrices. The sub-Bures bound appears to be quite loose, except in the rank-1 case for which it is equal to the exact Bures measure.

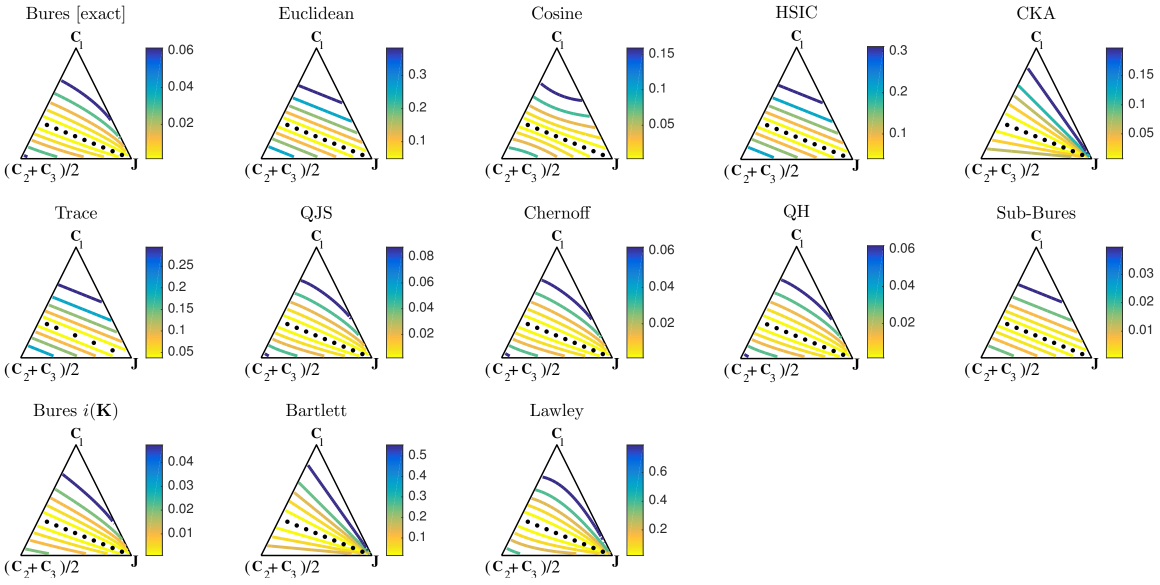

Figure 5: Equidistant contour lines for the measures on a 2-simplex of correlation matrices. The dotted line represents the non-informative matrix family $\mathcal{N}$. Informativeness is maximal at extremal points of the elliptope for convex distance metrics, but the choice of distance/dissimilarity measure affects the geometry of the level sets.

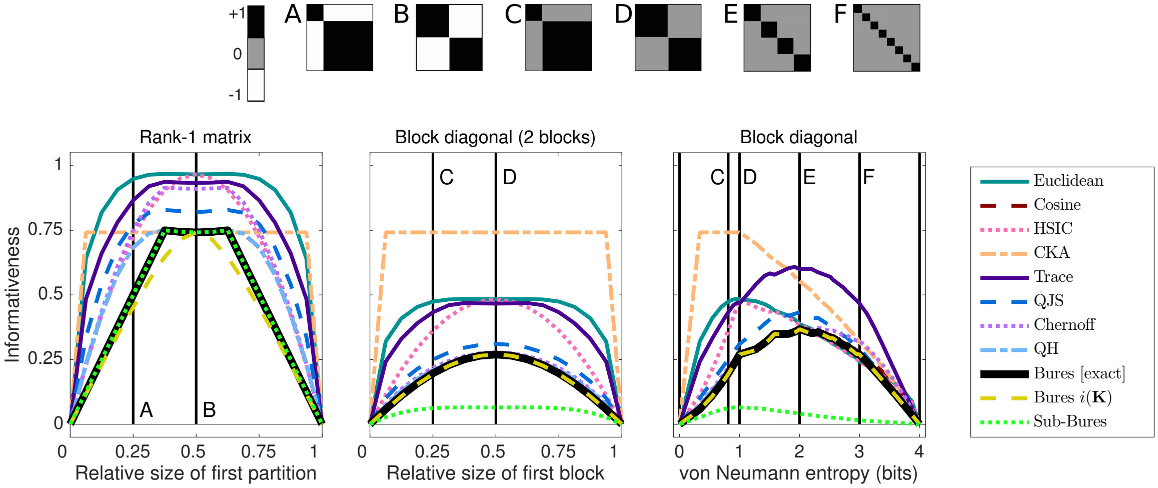

Figure 6: Comparison of informativeness measures for 16 × 16 correlation matrices with partition structure. Examples include rank-1 matrices (A, B); block diagonal matrices with two blocks (C, D); and block diagonal with increasing number of blocks (D, E, F) as indicated by the von Neumann entropy. (Note: The curves for the QH and cosine measures overlap on all plots, and the Bures curve covers the QH curve in the block diagonal plots.)

4. Optimization of Informativeness

In this section, we investigate constrained optimization problems based on finding a maximally informative correlation matrix close to a target matrix. We cover two types of cases; those optimizations parameterized in terms of a correlation matrix, and those in terms of a Euclidean embedding. In the latter case, we derive the analytic forms of cost functions for some measures of informativeness that can be applied directly to an embedding without requiring the explicit calculation of a correlation matrix. In particular, we show that when applied to an embedding, the Bures-based cost function is convex and corresponds to a novel matrix norm, which is a combination of trace seminorms on orthogonal subspaces. As this norm is non-differentiable, we describe its proximal operator, which requires the definition of the proximal operator for the squared trace norm. Using the Bures-based measure of informativeness as a cost function, we detail the optimization problem of finding a maximally informative correlation matrix nearby a target matrix, where nearness is assessed via the Euclidean distance between the embeddings. Although constraining the embedding to ensure it corresponds to a correlation matrix is non-convex, we relax this constraint to yield a completely convex optimization problem, the solutions of which can satisfy the original constraint. We propose an alternating direction method of multiplier (ADMM) algorithm that uses the proximal operator for the Bures-based cost function to solve this problem.

4.1 Maximally Informative Correlation Matrices

Firstly, we consider finding a maximally informative matrix that lies within an ϵ-radius ball of an arbitrary target matrix $\mathbf{A}$. This optimization can be written as

$$\begin{aligned} &\underset{\mathbf{K}\in\mathcal{E}}{\arg\max}\ d_{\mathcal{N}}(\mathbf{K}) \\ &\text{s.t. } \|\mathbf{A}-\mathbf{K}\|_F \le \epsilon, \end{aligned} \tag{4}$$

where $d_{\mathcal{N}}(\mathbf{K})$ is the distance to closest member of the family of $\mathbf{N}_a$ of non-informative matrices (or a lower bound on this distance in the case of the Bures distance). As noted in Theorem 4, the Bures-based informativeness is a convex function, as are the Euclidean and HSIC-based measures (their convexity can be verified by the properties of the Frobenius norm). Consequently, for these measures maximizing the informativeness is a convex maximization problem with multiple optima necessarily occurring at the extreme points of the elliptope, as evidenced by Figures 4. In particular, as stated in Theorem 5 in Appendix B.7, the Bures-based informativeness is maximized by the vertices of the elliptope corresponding to centered, rank-1 correlation matrices. Although there is multiple maxima, for appropriate choices of ϵ there may be a unique maximum within the local neighborhood.

For convenience, we instead define a minimization problem—replacing the ϵ-ball constraint with a penalty term scaled by ρ (where a large ρ corresponds to a small ϵ)—as

$$\underset{\mathbf{K}\in\mathcal{E}}{\arg\min}\ \rho\,\|\mathbf{A}-\mathbf{K}\|_F^2 + F\left(d_{\mathcal{N}}(\mathbf{K})\right), \tag{5}$$

where F is a monotonically decreasing function of the form $F : d \mapsto 1-\nu d^2$ with $\nu \in \{\frac{1}{2}, 1\}$ ($\nu = \frac{1}{2}$ for chordal distances in order to restrict the range of F to $[0, 1]$). The

Table 3: Expressions for finding the most informative correlation matrices via cost functions of the form $J_\rho(\mathbf{K}) = \rho \left\|\mathbf{A} - \mathbf{K}\right\|_F^2 + F\left(d_{\mathcal{N}}(\mathbf{K})\right)$. The forms are applicable to all positive semidefinite measures, even though informativeness is only defined for correlation matrices. Because of this generality, the Euclidean and cosine measures are written in terms of $\mathbf{N}_{a^\star} = a^\star\mathbf{J} + (1-a^\star)\frac{n}{n-1}\mathbf{H}$.

Measure:	$F(d)$:	$F\left(d_{\mathcal{N}}(\mathbf{K})\right)$:	$J_\rho(\mathbf{K})$ convex?
Euclidean	$1-d^2$	$1 - \frac{1}{n^2}\left\|\mathbf{K} - \mathbf{N}_{a^\star}\right\|_F^2, \quad a^\star = \frac{1}{n^2}\langle \mathbf{K}, \mathbf{J} - \mathbf{I}\rangle + \frac{1}{n}$	$\rho \geq \frac{1}{n^2}$
Cosine	$1-\frac{1}{2}d^2$	$\frac{\langle \mathbf{K}, \mathbf{N}_{a^\star}\rangle}{\|\mathbf{K}\|_F \|\mathbf{N}_{a^\star}\|_F}, \quad a^\star = \frac{1}{n}\frac{\langle \mathbf{K}, \mathbf{J}\rangle}{\langle \mathbf{K}, \mathbf{I}\rangle}$	—
HSIC	$1-d^2$	$1 + \frac{1}{n^2(n-1)}\langle \mathbf{K}, \mathbf{H}\rangle^2 - \frac{1}{n^2}\|\mathbf{HKH}\|_F^2$	$\rho \geq \frac{1}{n^2}$
CKA	$1-\frac{1}{2}d^2$	$\frac{\langle \mathbf{K}, \mathbf{H}\rangle}{\|\mathbf{HKH}\|_F\sqrt{n-1}}$	—
QH	$1-\frac{1}{2}d^2$	$\sqrt{\frac{1}{n^3}\langle\sqrt{\mathbf{K}}, \mathbf{J}\rangle^2 + \frac{1}{n(n-1)}\langle\sqrt{\mathbf{K}}, \mathbf{H}\rangle^2}$	no
Bures-based	$1-\frac{1}{2}d^2$	$\sqrt{\frac{1}{n^2}\langle \mathbf{K}, \mathbf{J}\rangle + \frac{1}{n(n-1)}\|\sqrt{\mathbf{K}}\mathbf{H}\|_*^2}$	no

function $F\left(d_{\mathcal{N}}(\mathbf{K})\right)$ simplifies for the measures of informativeness in Table 2 with closed-form expressions. The choice of F and the resulting expressions of $F\left(d_{\mathcal{N}}(\mathbf{K})\right)$ are listed in Table 3.

Since $1 - \nu d^2$ is concave and nonincreasing with respect to d^2, when $d_{\mathcal{N}}^2(\mathbf{K})$ is convex, $\rho\left\|\mathbf{A} - \mathbf{K}\right\|_F^2 + F\left(d_{\mathcal{N}}(\mathbf{K})\right)$ is a difference of convex functions. By inspection we see that the cost functions for the Euclidean and HSIC-based measures of informativeness are quadratic functions of $\mathbf{K}$ and convex[6] when $\rho \geq \frac{1}{n^2}$, i.e., this choice of ρ corresponds to a sufficiently small ϵ-radius ball, such that it contains a unique maximally informative correlation matrix. In contrast, the convexity of the cost function for the cosine and CKA-based measures depends on both the target matrix $\mathbf{A}$ and the regularization ρ; in certain cases the cost function may be convex. Due to the matrix square root, neither the QH nor the Bures-based measures are convex. In particular, $\frac{1}{n^2}\langle \mathbf{K}, \mathbf{J}\rangle + \frac{1}{n(n-1)}\|\sqrt{\mathbf{K}}\mathbf{H}\|_*^2$ is concave, as $\|\sqrt{\mathbf{K}}\mathbf{H}\|_*^2$ is proportional to the squared fidelity between $\frac{1}{n-1}\mathbf{H}$ and $\frac{1}{\mathrm{tr}\mathbf{K}}\mathbf{K}$ and squared fidelity is a concave function (Uhlmann, 1976) as noted in Theorem 3 in Appendix A. The function $\langle\sqrt{\mathbf{K}}, \mathbf{C}\rangle$ is also concave for any positive semidefinite $\mathbf{C}$, as the function $-\langle\sqrt{\mathbf{K}}, \mathbf{C}\rangle$ is convex (Borwein and Lewis, 2010, Section 3.1, Exercise 25c).

First-order gradient approaches may be applied to the expressions in Table 3 with some adjustments to ensure they are well-defined and smooth. Firstly, the cost functions corresponding to chordal distances (cosine, CKA, QH, and Bures) are squared to remove the scalar square roots from the expressions. Secondly, since the CKA measure is undefined at $\mathbf{K} = \mathbf{J}$, a log barrier of the form $-\log(1 - \bar{k})$ should be added to the cost function to prevent this case. Thirdly, for the QH measure to be differentiable, $\mathbf{K}$ must be strictly

[6] Letting $\mathbf{x} = \mathrm{vec}(\mathbf{K})$, these two cost functions can be written as $\mathbf{x}^\top\mathbf{M}\mathbf{x} + \mathbf{x}^\top\mathbf{u} + c$, which is a convex function if $\mathbf{M}$ is positive semidefinite. For the Euclidean measure, $\mathbf{M} = (\rho - \frac{1}{n^2})\mathbf{I} + \frac{1}{n^3(n-1)}\mathbf{v}\mathbf{v}^\top$ where $\mathbf{v} = \mathrm{vec}(\mathbf{J} - \mathbf{I})$. For the HSIC measure, $\mathbf{M} = \rho\mathbf{I} - \frac{1}{n^2}\mathbf{H} \otimes \mathbf{H}$, where $\otimes$ indicates the Kronecker product. In both cases, $\mathbf{M}$ is positive semidefinite if $\rho \geq \frac{1}{n^2}$.

Table 4: Gradients of cost functions derived from the closed-form informativeness measures with respect to $\mathbf{K}$, where $\mathbf{K} \in \mathcal{E}$ is a correlation matrix and $\bar{k} = \frac{1}{n^2}\mathbf{1}^\top\mathbf{K}\mathbf{1}$ is the average of its entries. The gradient for the QH measure is only valid if $\mathbf{K}$ is positive definite, which ensures $\sqrt{\mathbf{K}}$ and $-\sqrt{\mathbf{K}}$ have disjoint eigenvalues—guaranteeing a unique solution to the Sylvester equations (Bartels and Stewart, 1972).

Measure:	$\tilde{F}(\mathbf{K})$:	$\nabla_{\mathbf{K}}\tilde{F}(\mathbf{K})$:
Euclidean	$\frac{1}{2}\left(1 - d_{\mathcal{N}}(\mathbf{K})^2\right)$	$\frac{1}{n^2}\left(\frac{n\bar{k}-1}{n-1}\mathbf{J} - \mathbf{K}\right)$
Cosine	$\frac{1}{2}\left(1 - \frac{1}{2}d_{\mathcal{N}}(\mathbf{K})^2\right)^2$	$\frac{1}{\lVert\mathbf{K}\rVert_F^2}\left(\frac{n\bar{k}-1}{n^2}\mathbf{J} - \frac{n\bar{k}^2-2\bar{k}+1}{\lVert\mathbf{K}\rVert_F^2}\mathbf{K}\right)$
HSIC	$\frac{1}{2}\left(1 - d_{\mathcal{N}}(\mathbf{K})^2\right)$	$\frac{1}{n^2}\left(-\mathbf{HKH} - \frac{1-\bar{k}}{n-1}\mathbf{J}\right)$
CKA	$\frac{1}{2}\left(1 - \frac{1}{2}d_{\mathcal{N}}(\mathbf{K})^2\right)^2$	$\frac{\langle\mathbf{K},\mathbf{H}\rangle}{\lVert\mathbf{HKH}\rVert_F^2\sqrt{n-1}}\left(\mathbf{H} - \frac{\langle\mathbf{K},\mathbf{H}\rangle}{\lVert\mathbf{HKH}\rVert_F^2}\mathbf{HKH}\right)$
QH	$\frac{1}{2}\left(1 - \frac{1}{2}d_{\mathcal{N}}(\mathbf{K})^2\right)^2$	$\frac{\langle\sqrt{\mathbf{K}},\mathbf{J}\rangle}{n^3}\mathbf{S}_1 + \frac{\langle\sqrt{\mathbf{K}},\mathbf{H}\rangle}{n(n-1)}\mathbf{S}_2$, where
		$\sqrt{\mathbf{K}}\mathbf{S}_1 + \mathbf{S}_1\sqrt{\mathbf{K}} = \mathbf{J}, \quad \sqrt{\mathbf{K}}\mathbf{S}_2 + \mathbf{S}_2\sqrt{\mathbf{K}} = \mathbf{H}$
Bures-based	$\left(1 - \frac{1}{2}\tilde{d}_{\mathcal{N}}(\mathbf{K})^2\right)^2$	$\frac{1}{n^2}\mathbf{J} + \frac{S_\gamma(\mathbf{HKH})}{n(n-1)}\mathbf{S}'_\gamma(\mathbf{HKH})$
	$\frac{1}{2}\tilde{d}_{\mathcal{N}}(\mathbf{K})^2 = 1 - \sqrt{\frac{1}{n^2}\langle\mathbf{K},\mathbf{J}\rangle + \frac{1}{n(n-1)}S_\gamma^2(\mathbf{HKH})} \approx i(\mathbf{K})$	
	$S_\gamma(\mathbf{X}) = \sum_{i=1}^n s_\gamma\left(\lambda_i\right)$, where $\lambda_1, \ldots, \lambda_n$ are eigenvalues of $\mathbf{X}$	
	$s_\gamma(x) = \begin{cases} \sqrt{x} & x > \gamma \\ \frac{1}{4}\gamma^{-\frac{3}{2}}x^2 + 3\sqrt{\gamma} & x \leq \gamma \end{cases}$	
	$\mathbf{S}'_\gamma(\mathbf{X}) = \sum_{i=1}^n s'_\gamma(\lambda_i)\mathbf{u}_i\mathbf{u}_i^\top$, where $\mathbf{X} = \sum_{i=1}^n \lambda_i\mathbf{u}_i\mathbf{u}_i^\top$	

positive definite. Finally, although the Bures-based measure is not differentiable, a smooth function operating on the eigenvalues of $\mathbf{HKH}$, i.e., a spectral function (Lewis, 1996), can be used as an approximation to replace the trace of the matrix square root, which is not differentiable for any $\mathbf{K}$ since $\mathbf{HKH}$ has an eigenvalue of 0 corresponding to the constant eigenvector. For the smoothing function, we use a Huber-type approximation (Huber, 1964; Hintermüller and Wu, 2014) of the square root. Using standard identities for the differentiation of scalar-matrix functions (Lewis, 1996; Olsen et al., 2012), we derive the gradients of cost functions proportional to those in Table 3. The simplified expressions of the gradients are given in Table 4.

As informativeness is only meaningful for correlation matrices, the optimization variable must be constrained to lie within the elliptope. Restriction to correlation matrices can be enforced via a combination of a linear constraint (for the diagonal) and a semidefinite constraint. Finding the maximally informative correlation matrix is a type of semidefinite optimization (convex for the Euclidean and HSIC measures and appropriate choice of ρ). For the convex cases, the unique optimum can be sought using either a projected-gradient or a conditional gradient (Frank-Wolfe) algorithm. In the former case, the Euclidean projection to the elliptope can be used, but is non-trivial (Higham, 2002; Malick, 2004). The latter case requires solving a semidefinite program for the approximation step at each iteration. Furthermore, both cases require an order of n^2 variables. An alternative is to factorize the correlation matrix into a lower-dimensional Euclidean embedding (Journée et al., 2010). Factorizations have been used for solving the nearest correlation matrix problem (Grubišić and Pietersz, 2007) and other semidefinite programs (Burer and Monteiro, 2003). In Section 4.2, we pursue this factorization approach and express the cost functions corresponding to the measures of informativeness in terms of a Euclidean embedding.

When the correlation matrix is a function of parameters, informativeness can be used as an objective for optimizing these parameters. For instance, the correlation matrix can be defined parametrically in terms of the squared distances among a sample of points, passed through a differentiable function κ,

$$K_{i,j} = \kappa\left((\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{U}^\top \mathbf{U} (\mathbf{x}_i - \mathbf{x}_j)\right),$$

where $\mathbf{U}^\top\mathbf{U}$ defines a Mahalanobis distance metric as $d^2_{\mathbf{U}^\top\mathbf{U}}(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{U}\mathbf{x}_i - \mathbf{U}\mathbf{x}_j\|_2$ and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ is the sample of points. If $\kappa(|x-y|^2) = k(x,y) \quad \forall x, y \in \mathbb{R}$, where k is a positive semidefinite kernel function and $\kappa(0) = k(x,x) = 1 \quad \forall x$, then $\mathbf{K}$ is always a correlation matrix. Using this form, optimization of either $\mathbf{U}$ or $\mathbf{X}$ can be performed without additional constraints to ensure $\mathbf{K}$ is a correlation matrix. Optimizing informativeness with respect to $\mathbf{X}$ will shift the points in order to increase the informativeness, and optimization in terms of $\mathbf{U}$ corresponds to an unsupervised form of metric learning: a task that is typically supervised (Lowe, 1995; Xing et al., 2003; Fukumizu et al., 2004).

The gradient in terms of the coordinates of the sample $\mathbf{X}$ is

$$\nabla_{\mathbf{X}} \tilde{F}(\mathbf{K}) = -4\mathbf{U}^\top\mathbf{U}\mathbf{X}\left(\mathbf{B} - \operatorname{diag}(\mathbf{B}\mathbf{1})\right),$$

where $\mathbf{B} = \mathbf{K}' \circ \left[\nabla_{\mathbf{K}} \tilde{F}(\mathbf{K})\right]$, $\circ$ denotes the Hadamard product, $K'_{i,j} = \kappa'_{i,j}\left(d^2_{\mathbf{U}^\top\mathbf{U}}(\mathbf{x}_i, \mathbf{x}_j)\right)$, and κ' is the derivative of the scalar kernel function, e.g., for the Gaussian kernel, $k(x,y) = \kappa\left(|x-y|^2\right) = \exp\left(-|x-y|^2\right)$ and $\kappa'\left(|x-y|^2\right) = -\kappa\left(|x-y|^2\right)$. Similarly, the gradient with respect to $\mathbf{U}$ is $\nabla_{\mathbf{U}} \tilde{F}(\mathbf{K}) = -4\mathbf{U}\mathbf{X}\left(\mathbf{B} - \operatorname{diag}(\mathbf{B}\mathbf{1})\right)\mathbf{X}^\top$.

4.2 Maximally Informative Embeddings

When the corresponding embeddings are available and in cases when the dimensionality is lower than the number of objects, the informativeness can be evaluated with lower storage and computational requirements without unnecessarily forming the full correlation matrix. This can be realized as follows. Let $\mathbf{Z} \in \mathbb{R}^{b\times n}$ be a Euclidean space embedding, such that $\mathbf{Z}^\top\mathbf{Z} = \mathbf{K} \in \mathcal{E}$. Since we assume $\mathbf{K}$ is a correlation matrix, then $\mathbf{Z}$ belongs to the oblique manifold $\mathcal{OB}$. For each expression in Table 3, an equivalent expression of the cost function $F\left(d_{\mathcal{N}}(\mathbf{Z}^\top\mathbf{Z})\right)$ is listed in Table 5. It is noteworthy that the Bures-based measure yields a convex cost function, which we investigate in the next section. Empirically, the quantum Hellinger-based cost function appears convex, but we have not been able to prove its convexity. Besides these two, the other cost functions are not convex.

Table 5: Cost functions for finding a maximally informative Euclidean embedding $\mathbf{Z} \in \mathbb{R}^{b\times n}$ of the form $F\left(d_{\mathcal{N}}(\mathbf{Z}^\top\mathbf{Z})\right)$, where F is a monotonically decreasing function of the form $F : d \mapsto 1 - \nu d^2$ with $\nu \in \{\frac{1}{2}, 1\}$, and $d_{\mathcal{N}}(\mathbf{Z}^\top\mathbf{Z})$ is the distance or dissimilarity between the matrix $\mathbf{Z}^\top\mathbf{Z}$ and the family of non-informative matrices.

Measure:	$F\left(d_{\mathcal{N}}(\mathbf{Z}^\top\mathbf{Z})\right)$:	Convexity:
Euclidean	$1 - \frac{1}{n^2}\left\|\mathbf{Z}^\top\mathbf{Z} - \mathbf{N}_a\right\|_F^2, \quad a = \frac{1}{n^2}(\|\mathbf{Z1}\|_2^2 - \|\mathbf{Z}\|_2^2 + n)$	no
Cosine	$\frac{\langle\mathbf{Z}^\top\mathbf{Z}, \mathbf{N}_a\rangle}{\|\mathbf{Z}^\top\mathbf{Z}\|_F\|\mathbf{N}_a\|_F}, \quad a = \frac{1}{n}\|\mathbf{Z1}\|_2^2/\|\mathbf{Z}\|_F^2$	no
HSIC	$1 + \frac{1}{n^2(n-1)}\|\mathbf{ZH}\|_F^4 - \frac{1}{n^2}\|\mathbf{ZHZ}^\top\|_F^2$	no
CKA	$\frac{\|\mathbf{ZH}\|_F^2}{\|\mathbf{ZHZ}^\top\|_F\sqrt{n-1}}$	no
QH	$\sqrt{\frac{1}{n^3}\langle\mathbf{J}, \sqrt{\mathbf{Z}^\top\mathbf{Z}}\rangle^2 + \frac{1}{n(n-1)}\langle\mathbf{H}, \sqrt{\mathbf{Z}^\top\mathbf{Z}}\rangle^2}$	—
Bures-based	$\sqrt{\frac{1}{n^2}\|\mathbf{Z1}\|_2^2 + \frac{1}{n(n-1)}\|\mathbf{ZH}\|_*^2}$	yes

Given these cost functions, we can now pose the problem of finding a maximally informative embedding $\mathbf{Z}$ that lies within an ϵ-radius ball of a target embedding $\mathbf{T}$, as

$$\begin{aligned} \min_{\mathbf{Z}\in\mathcal{OB}} \; & F\left(d_{\mathcal{N}}(\mathbf{Z}^\top\mathbf{Z})\right) \\ \text{s.t.} \; & \|\mathbf{T} - \mathbf{Z}\|_F \leq \epsilon. \end{aligned} \tag{6}$$

The target embedding could arise from a low-rank approximation $\mathbf{T}^\top\mathbf{T} \approx \mathbf{A}$ of a target similarity matrix $\mathbf{A}$. In general, this optimization is non-convex because of the constraint that $\mathbf{Z} \in \mathcal{OB}$. In Section 4.2.2 we present a convex relaxation and optimization algorithm for the Bures-based measure of informativeness. The optimization uses properties of the Bures-based measure explored in Section 4.2.1.

4.2.1 Optimizing the Bures-based Measure of Informativeness for Embeddings

Due to its convexity, we concentrate on the Bures-based informativeness measure for embeddings and define

$$h(\mathbf{Z}) \equiv F^2\left(d_{\mathcal{N}}(\mathbf{Z}^\top\mathbf{Z})\right) = \left(1 - i(\mathbf{Z}^\top\mathbf{Z})\right)^2 = \tfrac{1}{n^2}\|\mathbf{Z1}\|_2^2 + \frac{1}{n(n-1)}\|\mathbf{ZH}\|_*^2, \tag{7}$$

where $d_{\mathcal{N}}(\mathbf{Z}^\top\mathbf{Z}) = \sqrt{2i(\mathbf{Z}^\top\mathbf{Z})}$ is the lower-bound on the Bures distance to the nearest non-informative matrix as defined in Equation B.17. The cost function $h(\cdot)$ has some interesting properties. Firstly, it is well-defined for an arbitrary matrix $\mathbf{X} \in \mathbb{R}^{b\times n}$, not necessarily on the oblique manifold. Secondly, it is convex, since it is the sum of two squared seminorms. Used as a penalty function, the two seminorms simultaneously penalize high-rank embeddings and embeddings which have a large mean, i.e., $h(\cdot)$ is minimal for rank-1, centered embeddings. Furthermore, since $\|\mathbf{X1}\|_2^2 = \frac{1}{n}\|\mathbf{X11}^\top\|_F^2 = n\|\mathbf{X}\frac{1}{n}\mathbf{J}\|_F^2 = n\|\mathbf{X}\frac{1}{n}\mathbf{J}\|_*^2$ both seminorms correspond to trace norms on the orthogonal subspaces defined by the projections $\mathbf{H}$ and $\frac{1}{n}\mathbf{J}$. We provide a proof that $\sqrt{h(\mathbf{X})} = \sqrt{\frac{1}{n}\|\mathbf{X}\frac{1}{n}\mathbf{J}\|_*^2 + \frac{1}{n(n-1)}\|\mathbf{XH}\|_*^2}$ is itself a matrix norm in Appendix D.

Due to the underlying trace norm, $h(\cdot)$ is not differentiable everywhere. Therefore, to enable the optimization we consider its proximal operator (Moreau, 1965)

$$\mathbf{prox}_{\lambda h}(\mathbf{Z}) \equiv \arg\min_{\mathbf{X}} \frac{1}{2}\|\mathbf{Z}-\mathbf{X}\|_F^2 + \lambda h(\mathbf{X}) \tag{8}$$

with $\lambda \geq 0$. Proximal operators enable first-order optimization involving non-smooth penalty functions (Nesterov, 2007; Wright et al., 2009; Beck and Teboulle, 2009; Combettes and Pesquet, 2011; Parikh and Boyd, 2014). To derive the proximal operator of $h(\cdot)$, we use the fact that both $h(\cdot)$ and the squared Frobenius norm (and Euclidean distance) can both be decomposed into the orthogonal subspaces defined by $\mathbf{H}$ and $\frac{1}{n}\mathbf{J}$, as

$$\begin{aligned} h(\mathbf{X}) &= \tfrac{1}{n}\|\mathbf{X}\tfrac{1}{n}\mathbf{J}\|_*^2 + \tfrac{1}{n(n-1)}\|\mathbf{XH}\|_*^2 = \tfrac{1}{n}\|\bar{\mathbf{X}}\|_*^2 + \tfrac{1}{n(n-1)}\|\tilde{\mathbf{X}}\|_*^2 \\ &= \|\bar{\mathbf{x}}\|_2^2 + \tfrac{1}{n(n-1)}\|\tilde{\mathbf{X}}\|_*^2, \end{aligned} \tag{9}$$

and

$$\|\mathbf{X}\|_F^2 = \left\|\mathbf{X}(\tfrac{1}{n}\mathbf{J}+\mathbf{H})\right\|_F^2 = \left\|\bar{\mathbf{X}}\right\|_F^2 + \|\tilde{\mathbf{X}}\|_F^2 = n\|\bar{\mathbf{x}}\|_2^2 + \|\tilde{\mathbf{X}}\|_F^2,$$

where $\tilde{\mathbf{X}} = \mathbf{XH}$ is a centered version of $\mathbf{X}$, $\bar{\mathbf{X}} = \bar{\mathbf{x}}\mathbf{1}^\top$, $\bar{\mathbf{x}}$ is the average of the columns of $\mathbf{X}$, and $\|\bar{\mathbf{X}}\|_* = \|\bar{\mathbf{X}}\|_F = \sqrt{n}\|\bar{\mathbf{x}}\|_2$ since $\bar{\mathbf{X}}$ is rank-1. Thus, the proximal solution can be written as the sum of the terms $\bar{\mathbf{Y}}$ and $\tilde{\mathbf{Y}}$, each computed independently as

$$\mathbf{prox}_{\lambda h}(\mathbf{Z}) = \bar{\mathbf{Y}} + \tilde{\mathbf{Y}}, \tag{10}$$

with the two terms calculated as

$$\begin{aligned} \bar{\mathbf{Y}} &= \arg\min_{\mathbf{X}} \frac{1}{2}\left\|\bar{\mathbf{Z}}-\mathbf{X}\right\|_F^2 + \frac{\lambda}{n}\|\mathbf{X}\|_F^2 = \frac{n}{n+2\lambda}\bar{\mathbf{Z}}, \\ \tilde{\mathbf{Y}} &= \arg\min_{\mathbf{X}} \frac{1}{2}\left\|\tilde{\mathbf{Z}}-\mathbf{X}\right\|_F^2 + \frac{\lambda}{n(n-1)}\|\mathbf{X}\|_*^2 = \mathcal{Z}_\beta(\tilde{\mathbf{Z}}), \end{aligned}$$

where $\beta = \frac{2\lambda}{n(n-1)}$ and $\mathcal{Z}_\beta(\cdot)$ is the proximal operator of the squared trace norm $\frac{\beta}{2}\|\cdot\|_*^2$, which we define in Equation E.1 in Appendix E along with a proof of its optimality.

4.2.2 Correlation Matrix Denoising

Given the proximal operator for the Bures-based informativeness, we proceed to solve the maximally informative embedding problem of Equation 6. We employ the optimization problem

$$\min_{\mathbf{Z}\in\mathcal{OB}} \tfrac{1}{2}\|\mathbf{T}-\mathbf{Z}\|_F^2 + \lambda h(\mathbf{Z}), \tag{11}$$

where the ϵ-radius constraint is replaced with the penalty term $\frac{1}{2}\|\mathbf{T}-\mathbf{Z}\|_F^2$ and the informativeness is scaled by λ. The minimization's cost function is convex, but the oblique manifold is not a convex set. We propose to relax the oblique manifold constraint, by replacing it with the constraint that each point in the embedding $\mathbf{Z}$ lies within the unit-sphere $\|\mathbf{z}_i\|_2 \leq 1, \forall i \in [n]$, rather than on it. The relaxed constraint corresponds to the convex hull of the oblique manifold, denoted by $\operatorname{conv}(\mathcal{OB})$. We then add the penalty term $-\frac{\mu}{2}\|\mathbf{Z}\|_F^2$ to ensure the norm of each column is maximized, which pushes the solution closer to the oblique manifold. The relaxed optimization is written as

$$\min_{\mathbf{Z}\in\operatorname{conv}(\mathcal{OB})} \tfrac{1}{2}\|\mathbf{T}-\mathbf{Z}\|_F^2 - \tfrac{\mu}{2}\|\mathbf{Z}\|_F^2 + \lambda h(\mathbf{Z}).$$

This is a convex optimization problem for $\mu \leq 1$. We choose the largest possible penalty $\mu = 1$ to satisfy convexity and approach feasibility, leading to the cost function $\langle \mathbf{Z}, -\mathbf{T}\rangle + \lambda h(\mathbf{Z})$.

To isolate the non-smooth term, we write an equivalent composite form of the optimization problem, using an auxiliary variable $\mathbf{X}$ as

$$\begin{aligned}&\min_{\substack{\mathbf{X}\in\mathbb{R}^{b\times n}\\ \mathbf{Z}\in\operatorname{conv}(\mathcal{OB})}} \quad \langle \mathbf{Z}, -\mathbf{T}\rangle + \lambda h(\mathbf{X}) \\ &\text{s.t. } \mathbf{X} = \mathbf{Z}.\end{aligned} \tag{12}$$

An approximate solution to the composite optimization can be found by removing the equality constraint and adding the penalty term $\frac{\rho}{2}\|\mathbf{Z}-\mathbf{X}\|_F^2$ and using alternating optimization to find a solution to the relaxed problem. The optimization of $\mathbf{Z}$, for fixed $\mathbf{X}$, can be performed on each column independently, and amounts to a Euclidean projection to the unit-ball. Likewise, the optimization $\mathbf{X}$, for fixed $\mathbf{Z}$, is also convex with an analytic solution given by application of the proximal operator $\mathbf{prox}_{\frac{\lambda}{\rho}h}(\mathbf{Z})$.

Because the aforementioned alternating optimization procedure is limited by slow convergence and sensitivity to the penalty, we instead conduct an optimization through an augmented Lagrangian formulation. Specifically, we employ the alternating direction method of multipliers algorithms (ADMM) (Gabay and Mercier, 1976; Eckstein and Bertsekas, 1992; Combettes and Pesquet, 2011; Boyd et al., 2011) to find a fixed point of the augmented Lagrangian

$$\mathcal{L}_\rho(\mathbf{Z},\mathbf{X},\mathbf{M}) = I(\mathbf{Z}) + \langle \mathbf{Z}, -\mathbf{T}\rangle + \lambda h(\mathbf{X}) + \tfrac{\rho}{2}\|\mathbf{Z}-\mathbf{X}\|_F^2 + \langle \mathbf{M}, \mathbf{Z}-\mathbf{X}\rangle, \tag{13}$$

where I is the indicator function for $\operatorname{conv}(\mathcal{OB})$: $f(\mathbf{Z}) = 0$ for $\mathbf{Z} \in \operatorname{conv}(\mathcal{OB})$ and $f(\mathbf{Z}) = +\infty$ otherwise. This leads to a minimization that alternates between two minimization steps on each non-fixed variables $\mathbf{Z}$ and $\mathbf{X}$. Using scaled dual variables $\mathbf{U} = \frac{1}{\rho}\mathbf{M}$, the steps are

$$\min_{\mathbf{Z}\in\operatorname{conv}(\mathcal{OB})} \left\|\mathbf{Z}-\mathbf{X}+\mathbf{U}-\tfrac{1}{\rho}\mathbf{T}\right\|_F^2,$$
$$\min_{\mathbf{X}\in\mathbb{R}^{b\times n}} \tfrac{\rho}{2}\|\mathbf{Z}-\mathbf{X}+\mathbf{U}\|_F^2 + \lambda h(\mathbf{X}),$$

and are followed by an update of the dual variables $\mathbf{U}$ with the residuals $\mathbf{Z}-\mathbf{X}$. The proposed optimization procedure is summarized in Algorithm 1. The computational complexity of each iteration is $\mathcal{O}(b^2 n)$, which corresponds to the singular value decomposition of the embedding matrix required by the proximal operator of h.

```
Algorithm 1: Correlation Matrix Denoising
input  : Initial embedding $\mathbf{T}$ of size $b \times n$, and parameters $\lambda > 0$, $\rho > 0$, and $\eta = \frac{\lambda}{\rho}$
$\mathbf{X} \leftarrow \mathbf{T}$, $\mathbf{U} \leftarrow \mathbf{0}$
while not converged do
    $\mathbf{W} \leftarrow \mathbf{X} - \mathbf{U} + \frac{1}{\rho}\mathbf{T}$
    $\mathbf{z}_i \leftarrow \frac{\mathbf{w}_i}{\max(1, \|\mathbf{w}_i\|_2)}$, to project each column to unit-ball
    $\mathbf{X} \leftarrow \mathbf{prox}_{\eta h}(\mathbf{Z}+\mathbf{U}) = \bar{\mathbf{X}} + \tilde{\mathbf{X}}$, as defined in Equation 10 with
        $\bar{\mathbf{X}} = \frac{n}{n+2\eta}(\mathbf{Z}+\mathbf{U})\frac{1}{n}\mathbf{J}$,
        $\tilde{\mathbf{X}} = \mathcal{Z}_\beta\left(\mathbf{ZH}+\mathbf{UH}\right)$, where $\beta = \frac{2\eta}{n(n-1)}$ and $\mathcal{Z}_\beta$ defined in Equation E.1.
    $\mathbf{U} \leftarrow \mathbf{U} + \mathbf{Z} - \mathbf{X}$, for the dual update
end
output: A denoised embedding $\mathbf{Z}$
```

In practice, we use $\rho = 1$, and find that if λ is chosen appropriately, the solution at convergence lies on the oblique manifold, i.e., $\mathbf{Z} \in \mathcal{OB}$. This means that the solution to the relaxed problem (12) satisfies the constraint of the original problem (11). To understand why the convex relaxation yields solutions to the original problem, we note that for any λ, the relaxed problem can be written as a linear program over the convex set defined by the intersection of conv($\mathcal{OB}$) and $\mathcal{I}_c = \{\mathbf{X} : \sqrt{h(\mathbf{X})} \leq c\}$, where small values of c correspond to large values of λ, and $\mathcal{I}_c$ corresponds to the set embeddings with informativeness above $1 - c$. As any linear program is optimized at extremal points of a convex set, an optimizing point of the relaxed solution will either lie on the oblique manifold and within the interior of $\mathcal{I}_c$, or if λ is too large, on the boundary of $\mathcal{I}_c$ and in the interior of conv($\mathcal{OB}$).

The convergence of the ADMM algorithm can be proven under mild conditions (Boyd et al., 2011), which are met for Algorithm 1. The conditions require the objective functions to be closed, proper, and convex; and the existence of a saddle point $(\mathbf{Z}^\star, \mathbf{X}^\star, \mathbf{M}^\star)$ of the unaugmented Lagrangian function $\mathcal{L}_0(\mathbf{Z}, \mathbf{X}, \mathbf{M})$, i.e., Equation 13 with $\rho = 0$, such that $\mathcal{L}_0(\mathbf{Z}^\star, \mathbf{X}^\star, \mathbf{M}) \leq \mathcal{L}_0(\mathbf{Z}^\star, \mathbf{X}^\star, \mathbf{M}^\star) \leq \mathcal{L}_0(\mathbf{Z}, \mathbf{X}, \mathbf{M}^\star)$ for all $\mathbf{Z}, \mathbf{X}, \mathbf{M}$. The saddle point $(\mathbf{Z}^\star, \mathbf{X}^\star, \mathbf{0})$ corresponds to the solution $(\mathbf{Z}^\star, \mathbf{X}^\star)$ of the relaxed optimization problem (12). This solution necessarily exists, since as mentioned, the relaxed problem is equivalent to a linear objective optimized over the convex set defined by the intersection of conv($\mathcal{OB}$) and $\mathcal{I}_c$. Furthermore, the convergence rate of the ADMM algorithm for convex objectives is $\mathcal{O}(1/t)$, where t is the number of iterations (He and Yuan, 2015).

To obtain embeddings which are progressively more informative (low-rank and nearly centered), the algorithm can be ran with increasing values of λ and warm restarts. We start with $\lambda = 1$, increase it by 125% at each restart, and stop the sequence of warm restarts if the solution at convergence no longer lies on the boundary of the convex hull, that is, $\mathbf{Z} \notin \mathcal{OB}$. The main computational burden of the algorithm is the singular value decomposition required in computing the proximal operator of $h(\cdot)$. This burden can somewhat be alleviated by using lower dimensional embeddings (i.e., a smaller b). Rather than choosing a small initial b, which cannot preserve as much information about the target matrix, we use an adaptive truncation heuristic. Specifically, we note that $h(\cdot)$ is invariant to unitary transformations of the embedding coordinates. Thus, we can use a singular value decomposition of the solution $\mathbf{Z} = \mathbf{\Psi}\mathbf{\Sigma}\mathbf{\Upsilon}^\top$ and transform $\mathbf{Z}$ and $\mathbf{T}$ with the left singular vectors, and then remove embedding dimensions (rows) from both $\mathbf{\Psi}^\top\mathbf{Z} = \mathbf{\Sigma}\mathbf{\Upsilon}^\top$ and $\mathbf{\Psi}^\top\mathbf{T}$, when the corresponding singular values of $\mathbf{Z}$ are less than $\frac{\sigma_1}{100}$, with $\sigma_1 = \|\mathbf{Z}\|_2$ being the maximum singular value. While truncation does distort the target matrix, this approach lowers the dimensions and decreases the processing requirements for the subsequent warm restarts.

5. Experiments

In this section we highlight example applications of the measures of informativeness[7] and compare the different measures listed in Table 2. Applications include selecting an appropriate kernel bandwidth, selecting the dimension of a low-rank kernel approximation, and detecting clustered graphs. We also demonstrate the correlation matrix denoising algorithm as a preprocessor for spectral clustering.

5.1 Kernel Bandwidth Selection

Gaussian kernel functions of the form $k_\sigma(\mathbf{x}, \mathbf{y}) = \exp\left(\frac{-1}{2\sigma^2}\|\mathbf{x} - \mathbf{y}\|_2^2\right)$ are widely used for kernel-based clustering and classification of vectors in Euclidean space, but an appropriate kernel bandwidth σ needs to be selected. For any choice of σ, evaluating the kernel function for a set of vectors yields a correlation matrix with entries $K_{i,j} = k_\sigma(\mathbf{x}_i, \mathbf{x}_j)$. If all of the vectors are distinct, the limit $\sigma \to 0$ yields the identity matrix $\mathbf{I}$, and the limit $\sigma \to \infty$ results in the constant matrix $\mathbf{J}$. Essentially, the kernel bandwidth parameterizes a curve within the elliptope that begins and finishes at non-informative correlation matrices. Since informativeness is a function of the distance to the set of non-informative correlation matrices, choosing the kernel bandwidth that maximizes the informativeness corresponds to finding an extremal point of this curve as illustrated in Figure 7. The example shows informativeness can be used to select a kernel bandwidth appropriate for the scale of the data. In comparison, measures such as the von Neumann entropy or distance from the identity matrix are maximized by the smallest and largest kernel bandwidths, respectively.

We compare the different informativeness measures on four synthetic examples in Figure 8. In these examples, the measures are maximized by kernel bandwidths dependent on the scale of the data. This holds for all measures except CKA, which is maximized by large kernel bandwidths that yield nearly constant matrices. We quantify the appropriateness of informativeness-based bandwidth selection for spectral clustering—specifically, the algorithm by Ng et al. (2002)—by using the normalized mutual information[8] between the labels corresponding to the true grouping and the clusters found when using different kernel bandwidths. For each example, 10 random instances are tested with the correct number of groups provided to the clustering algorithm. On the first example, each group is a convex shape and a large kernel bandwidth performs well. On the remaining examples, there is an intermediate kernel bandwidth that maximizes the normalized mutual information.

[7] MATLAB code that implements the informativeness measures and reproduces the figures and tables is available at `http://pcwww.liv.ac.uk/~goulerma/software/brockmeier17a-code.zip`.

[8] Normalized mutual information is the mutual information between two variables divided by the maximum entropy of the two (Kvålseth, 1987).

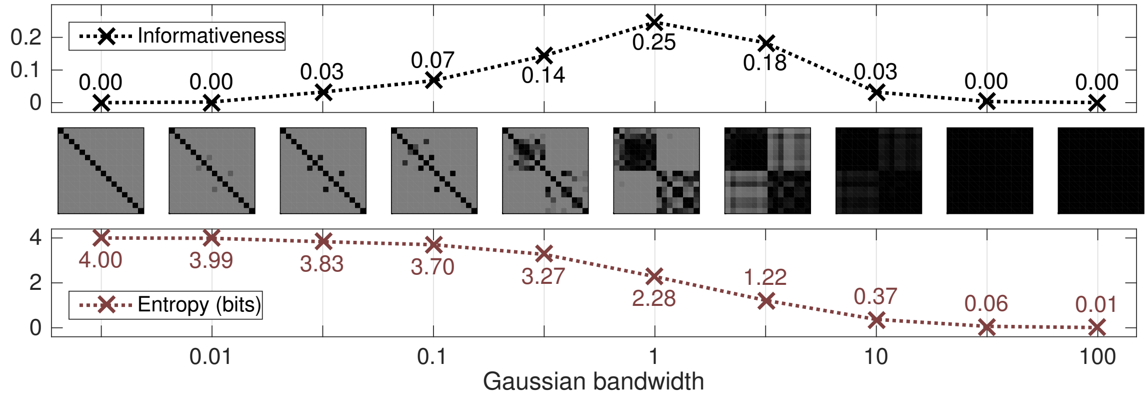

Figure 7: Informativeness versus von Neumann entropy for correlation matrices obtained from applying a Gaussian kernel with varying bandwidths to a sample drawn from a univariate mixture of two normal distributions separated by a distance of 4. (The Bures-based informativeness is used in this example.)

For each measure of informativeness and for Bartlett's and Lawley's test statistics, we collect the values of normalized mutual information corresponding to the bandwidths that maximize the measures. The results are reported as box and whisker plots in Figure 8. From the results, there appear to be four subsets of measures that select similar bandwidths: the cosine measure, which chooses the smallest kernel bandwidth, has the worst median performance on the first data set, but performs the best on the remaining; the Chernoff, QH, and Bures-based measures select a larger bandwidth, tie for the best performance on the first data set, achieve the second-best performance across the remaining; the Euclidean, HSIC, and sub-Bures measures select larger bandwidths and perform worse; and the CKA measure along with Bartlett's and Lawley's test statistics are maximized by the largest kernel bandwidth within the range, which yields a poor clustering on three of the four data sets. This example shows that the informativeness measures (except CKA) can adaptively select kernel bandwidths appropriate for the structure of the data.

5.2 Selecting an Embedding Dimension

Kernel principal component analysis (Schölkopf et al., 1998) approximates a kernel matrix using a lower-dimensional embedding obtained by truncating the eigenvalue decomposition of the centered kernel matrix. Choosing the embedding dimension (how many eigenvectors to retain) requires a criterion. We show that choosing the dimension that maximizes informativeness is an effective approach.

Specifically, kernel principal component analysis uses the top-k eigenvectors of the centered kernel matrix $\tilde{\mathbf{K}} = \mathbf{HKH}$. Let denote $\tilde{\mathbf{K}}^{(k)}$ the rank-k approximation of $\tilde{\mathbf{K}}$. In general $\tilde{\mathbf{K}}^{(k)}$ is positive definite but not a correlation matrix, and requires symmetric normalization as defined in Equation 1. However, if the original kernel matrix is far from being centered then applying normalization after centering and truncation can introduce severe distortion.

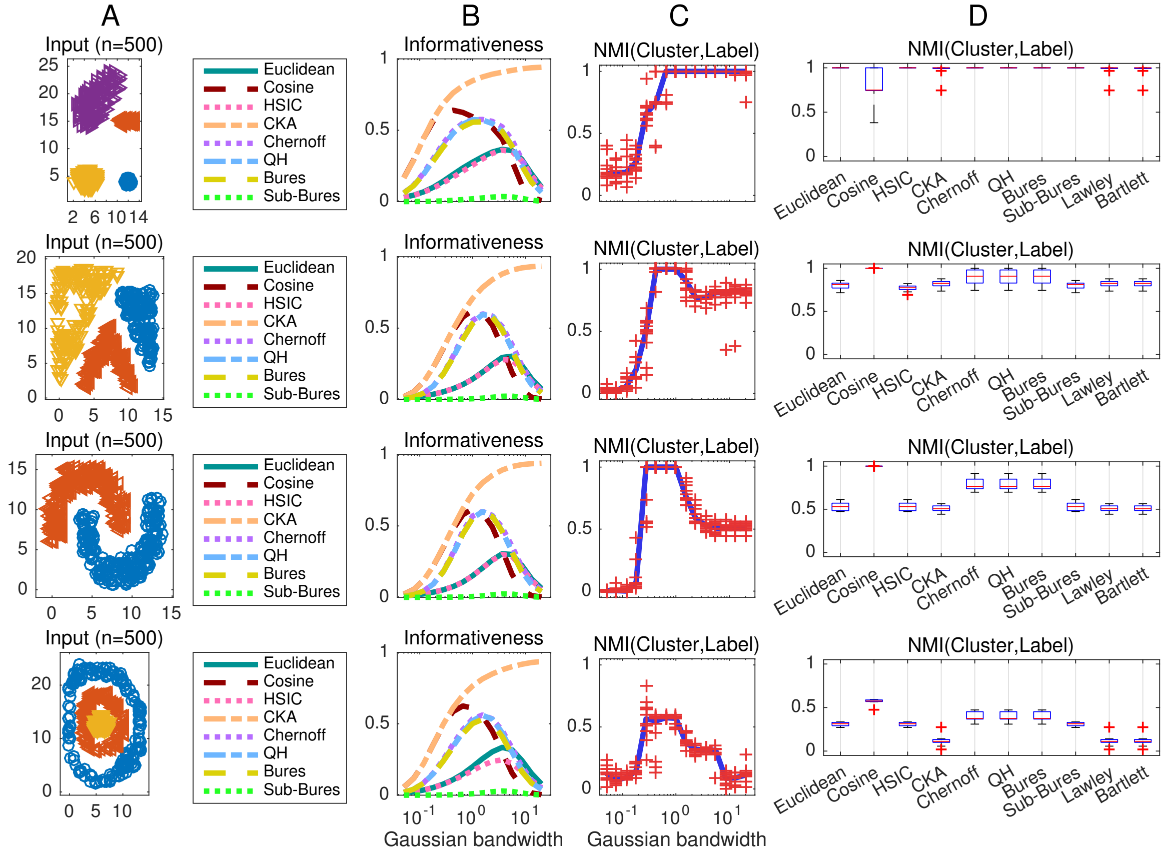

Figure 8: Selecting a Gaussian kernel bandwidth for spectral clustering by maximizing informativeness. (A) Samples are drawn from two-dimensional distributions with distinct groups. (B) Informativeness is plotted versus the kernel bandwidth. (C) For each kernel bandwidth, spectral clustering is applied with the correct number of clusters and the normalized mutual information (NMI) between the cluster label and the true label is recorded (crosses indicate values across 10 random samples; solid curve for the median). (D) Box and whisker plots of the NMI values for the clusterings obtained when the kernel bandwidth is selected by informativeness or by Lawley's and Bartlett's test statistics.

As an alternative, we truncate the original, uncentered correlation matrix and then perform renormalization.

We test the dimensionality selection method on a set of UCI data sets (Lichman, 2013). A Gaussian kernel function with a heuristic bandwidth is used for all cases. Specifically, the bandwidth is a linear combination of the minimum and maximum Euclidean distances $\sigma = 2d_{min} + \frac{2}{9}(d_{max} - 2d_{min})$, where d_{min} and d_{max} are the minimum and maximum pairwise distances (a similar heuristic was used by Shi and Malik, 2000).

Table 6: Nearest neighbor classification performance for kernel-based dimensionality reduction across 7 UCI data sets. Each entry lists the classification accuracy as a percentage, followed by the average dimensionality in parentheses. The last two rows list the average difference in performance versus using the original space and the average loss margin (absolute difference to the best method on the run) across 20 Monte Carlo runs and data sets (standard deviation across data sets).

	Original	CKA[1]	CKA[2]	Euclid.	Cosine	QH	Bures
breast cancer	95 (569)	90 (2.0)	87 (1.0)	91 (3)	95 (328)	95 (52)	95 (41)
sonar	83 (208)	83 (126.3)	82 (130.1)	62 (3)	83 (113)	83 (28)	83 (26)
ionosphere	85 (351)	74 (1.0)	88 (69.5)	84 (2)	78 (176)	94 (38)	94 (33)
parkinson	92 (195)	84 (21.4)	80 (39.5)	82 (2)	93 (100)	91 (19)	91 (19)
iris	94 (150)	85 (3.0)	84 (2.2)	80 (2)	94 (21)	93 (10)	94 (11)
glass	68 (214)	65 (181.0)	68 (211.6)	60 (3)	68 (112)	68 (20)	68 (19)
ecoli	80 (336)	78 (31.6)	77 (48.2)	80 (6)	80 (158)	80 (24)	80 (24)
vs. original	—	5.5±4	4.9±6	8.4±7	0.9±2	-0.9±4	-1.1±4
Loss margin	2.2±3	7.7±6	7.1±4	10.6±7	3.2±6	1.3±0.6	1.2±0.5

[1] Centered kernel alignment using the training portion of the uncentered kernel matrix

[2] Centered kernel alignment using the training portion of the centered kernel matrix

	HSIC	CKA	Chernoff	Sub-Bures	Bartlett	Lawley
breast cancer	90 (2)	70 (1)	95 (33)	91 (3)	70 (1)	70 (1)
sonar	62 (3)	53 (1)	84 (30)	62 (3)	57 (2)	57 (2)
ionosphere	84 (2)	74 (1)	94 (38)	84 (2)	84 (2)	84 (2)
parkinson	82 (2)	67 (1)	91 (15)	82 (2)	82 (2)	82 (2)
iris	80 (2)	66 (1)	93 (10)	80 (2)	80 (2)	80 (2)
glass	60 (3)	38 (1)	68 (18)	60 (3)	38 (1)	38 (1)
ecoli	74 (3)	40 (1)	80 (18)	80 (6)	66 (2)	66 (2)
vs. original	9.5±6	27.2±9	-1.0±4	8.4±7	17.3±10	17.3±10
Loss margin	11.8±6	29.4±7	1.3±0.3	10.6±7	19.5±9	19.5±9

For comparison, we use supervised selection of the dimensionality based on centered kernel alignment (Cortes et al., 2012). We test both the centered and uncentered truncations, and select the dimension that maximizes the centered kernel alignment to a target matrix obtained from the classification labels in the training set. In all cases, we use a first nearest neighbor classifier with half of the instances for training; the average classification accuracy is recorded in Table 6.

Across these data sets, the Bures, Chernoff, and QH-based informativeness measures perform the best. They select a lower dimensional embedding and improve the classification performance versus using the original distances. The cosine-based informativeness measure also outperforms using centered kernel alignment with training labels. In comparison, the Euclidean and HSIC-based informativeness measures select very low-dimensional embeddings that do not perform well, and the CKA-based informativeness and Bartlett's and

Lawley's test statistics are inappropriate for this task as they often select one-dimensional embeddings, which is an inadequate representation for a nearest neighbor classifier.

5.3 Informativeness as an Indicator of Structure in Graphs

We show that informativeness is indicative of structure in undirected graphs. We consider binary graphs whose edges are defined by a symmetric adjacency matrix $\mathbf{G} \in \{0,1\}^{n \times n}$, where $G_{i,j} = 1$ indicates an edge between vertices i and j. To enable informativeness to be applied directly, we represent a graph by its normalized graph Laplacian matrix $\mathbf{L}$, which is a correlation matrix under the following definition

$$L_{i,j} = \begin{cases} 1 & i = j, \\ \frac{-1}{\sqrt{d_i d_j}} & G_{i,j} = 1, \\ 0 & \text{otherwise}, \end{cases}$$

where d_i is the degree of the ith vertex. When the graph has unconnected vertices, the corresponding rows and columns of the normalized graph Laplacian are all zeros except at the diagonal entry. This is different than the standard convention (Chung, 1997), where the diagonal entry for an unconnected vertex would be zero.

Informativeness can be used to make relative comparisons between graphs with the same number of vertices. In particular, informativeness can be used to rank graphs and identify graphs that are more informative than a regular graph with the same number of edges. Firstly, we consider the informativeness for small graphs, $n = 5$, for which we evaluate all 34 non-isomorphic graphs. Figure 9 shows the Bures-based informativeness for each graph plotted against the number of edges; there is a clear separation of the values for the three most informative graphs, which correspond to clustered graphs. Of the remaining graphs, the path graph (a tree with maximum degree of 2) has the next highest informativeness. This ordering held for all informativeness measures. We then test the set of non-isomorphic trees for $n = 6$. The trees are shown in Figure 10 ordered by their informativeness. Once again, the path graph has the highest informativeness of all trees; the ordering is consistent for all informativeness measures.

For larger graphs, it becomes infeasible to enumerate all of the non-isomorphic graphs. As a baseline we consider regular graphs, which are defined as graphs whose vertices have the same degree. To further sample from the universe of possible graphs with $n = 100$ vertices, we consider the following random graph models:

- Erdős-Rényi model described by an edge-link probability (Erdős and Rényi, 1959),
- Barabási-Albert scale-free model with a power law degree distribution (Barabási and Albert, 1999),
- Watts and Strogatz small-world model (Watts and Strogatz, 1998), which randomly rewires each edge of a regular graphs with a certain probability (we use 5%),
- a clustered regular graph model, generated by removing all edges between vertices assigned to different clusters (the number of clusters is varied between 2 and 20).

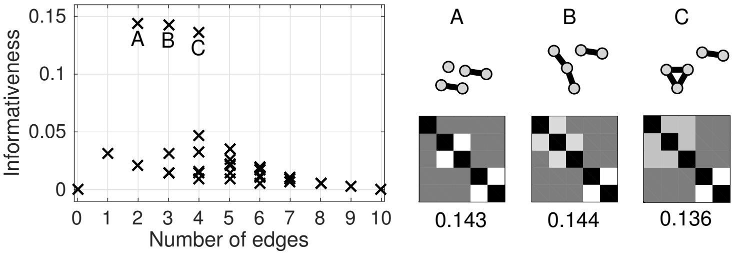

Figure 9: Bures-based informativeness of non-isomorphic graphs with $n = 5$ vertices. The three graphs (labeled A, B, C) with highest informativeness correspond to graphs with two or three clusters.

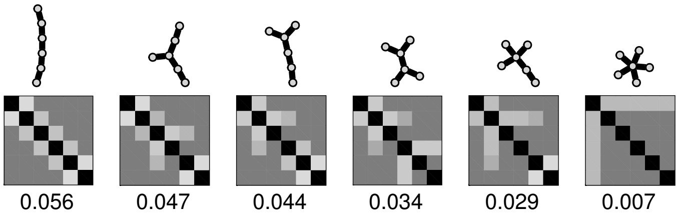

Figure 10: Non-isomorphic trees with 6 vertices sorted by Bures-based informativeness.

Erdős-Rényi graphs and Barabási-Albert scale-free graphs are both unstructured graph models that differ in their degree distributions. The degree distribution of the former is a Poisson distribution, whereas the scale-free graphs have a long-tailed degree distribution, meaning that a small number of vertices have a large number of edges.

We evaluate the informativeness of regular graphs and realizations of each random graph model, scanning any free parameters to vary the number of edges. Example graphs and the Bures-based informativeness as a function of the average degree[9] are shown in Figure 11. For any given number of edges, the informativeness is maximized by clustered graphs, and minimized by the Erdős-Rényi graphs and Barabási-Albert scale-free graphs. The random rewiring involved with the Watts and Strogatz small-world model does not drastically change the informativeness from that of regular graphs.

We compare informativeness to two existing test statistics for clustering in graphs: the modularity of two-way partitions identified using an eigenvector-based method (Newman, 2006) and the statistic defined by Bickel and Sarkar (2016) on the adjacency matrix. The latter statistic is used for testing the hypothesis that the graph is drawn from an Erdős-Rényi model; the test statistic is based on the largest eigenvalue of the difference between the adjacency matrix and the expectation over adjacency matrices under an Erdős-Rényi model, where the normalized graph Laplacian of this expectation is a non-informative correlation matrix. Bickel and Sarkar (2016) used the test statistic to identify cluster structure. Bickel and Sarkar's statistic is normalized, such that its distribution for Erdős-Rényi graphs is invariant to the number of edges. Figure 11 also shows the values obtained from these test

[9] The average degree is given as $\bar{d} = \frac{2e}{n}$, where e is the number of edges.

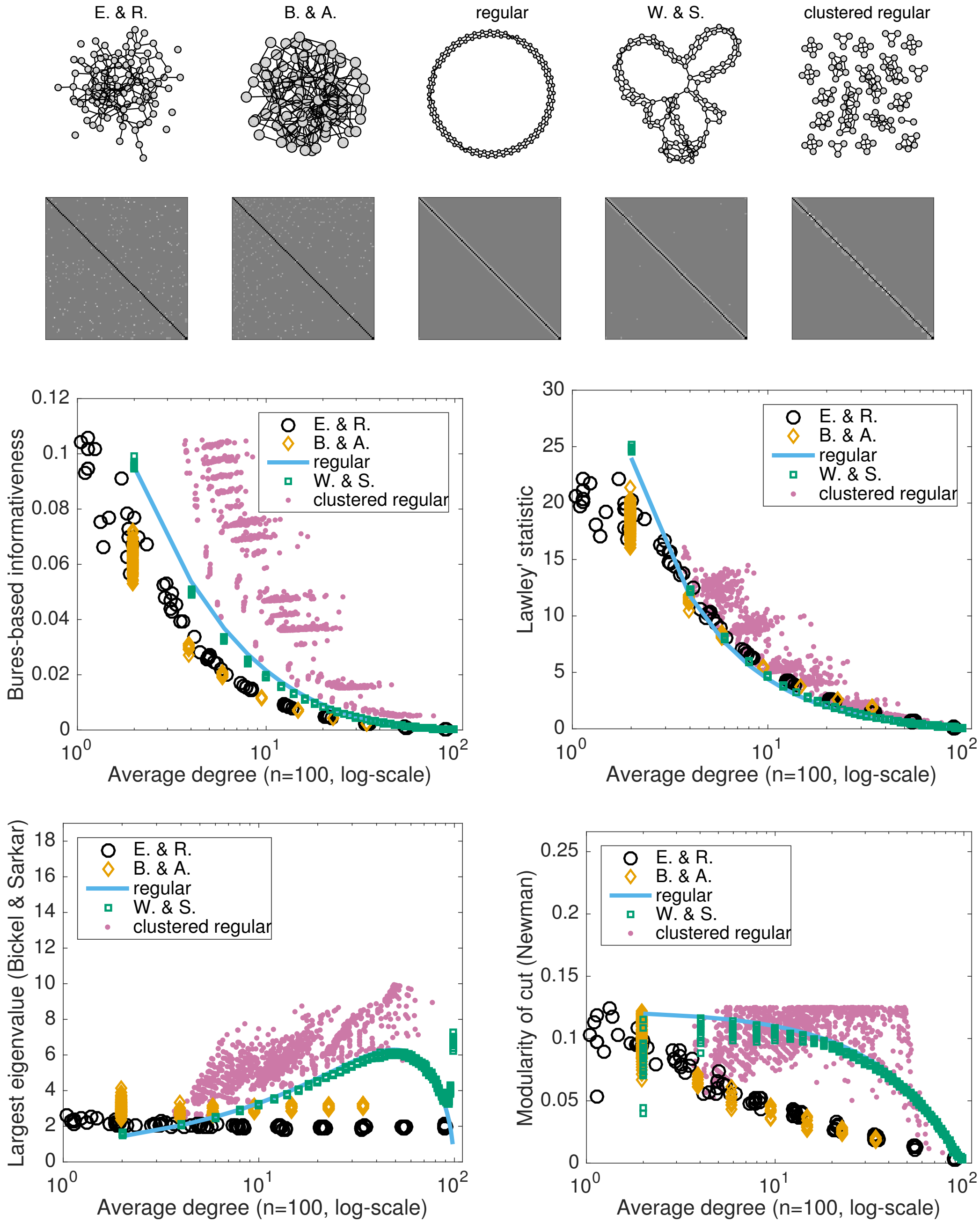

Figure 11: Comparison of informativeness and other statistics on different types of random graphs. Bures-based informativeness and Lawley's test statistic for homogeneous correlation coefficients are applied to the normalized graph Laplacians, and compared to the graph-clustering test statistics proposed by Bickel and Sarkar (2016) and Newman (2006). The random graphs were drawn from Erdős-Rényi (E. & R.), Barabási-Albert scale-free (B. & A.), Watts-Strogatz small-world (W. & S.), and clustered regular models.

Table 7: Proportion of graphs with informativeness, or other measure, greater than a threshold set as the value of a nearly-regular graph with the same number of edges. Graphs have 100 vertices. The random graphs were drawn from Erdős-Rényi (E. & R.), Barabási-Albert scale-free (B. & A.), Watts-Strogatz small-world (W. & S.), and clustered regular models.

	E. & R.	B. & A.	W. & S.	Clustered regular
Number of graphs:	200	200	990	1200
Euclidean	0.241	0.195	0.986	0.968
Sub-Bures	0.241	0.195	0.986	0.968
Cosine	0.241	0.195	0.987	0.968
HSIC	0.134	**0**	0.969	0.957
CKA	0.134	**0**	0.974	0.957
Chernoff	0.053	**0**	0.550	**1**
QH	0.053	**0**	0.558	**1**
Bures	0.118	**0**	0.553	**1**
Bartlett	**0.016**	**0**	0.974	0.957
Lawley	0.364	0.340	1	0.996
Bickel and Sarkar	0.460	0.700	0.232	0.962
Newman	0.102	**0**	**0.099**	0.601

statistics on the various graphs. The test statistics have a different range as compared to the measures of informativeness; nonetheless, for a given statistic or measure of informativeness, the values can be used to make relative comparisons between the different graphs.

The value of informativeness is dependent on the number of edges, with graphs with fewer edges being more informative. As a threshold for identifying structure, we use the informativeness of a nearly-regular graph[10] with the same number of edges. The thresholds for the other statistics are formed in the same way. Ideally, the value of a clustered graph will be above the threshold, and the value will be below the threshold for an unstructured graph. Table 7 details the proportion of random graphs above the threshold for each measure.

The QH, Chernoff, and Bures-based measures appear to be the best measures to distinguish clustered graphs from unstructured graphs. The HSIC and CKA-based measures and Bartlett's statistic perform nearly as well. The Euclidean, sub-Bures, and cosine-based measures are not as specific—a larger proportion of unstructured graphs are evaluated above the threshold. Bickel and Sarkar's and Lawley's statistics are even less specific. Newman's modularity index (designed to detect binary clustering) evaluates fewer clustered graphs above the threshold; but, it identifies the lowest proportion of Watts-Strogatz small-world model, which are not clustered. To summarize, a subset of the informativeness measures and Bartlett's statistic are specific indicators of regular structure and clustering in graphs, and Newman's modularity is useful for identifying clustered graphs specifically.

[10] A graph is regular if the number of edges is divisible by the number of vertices, otherwise a nearly-regular graph is constructed by adding edges to a regular graph, where the additional edges are evenly distributed, such that the degrees of any two vertices differ by at most 1.

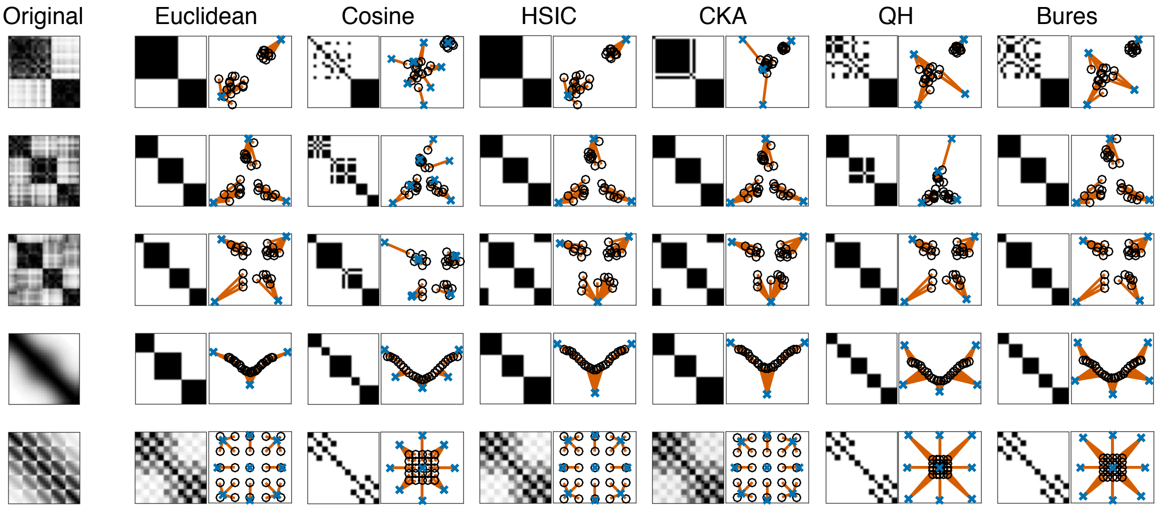

Figure 12: Sample denoising based on optimizing the informativeness of a Gaussian kernel matrix with respect to the sample's two-dimensional coordinates. Each row shows the original Gaussian kernel matrix ($n = 25$), the kernel matrices corresponding to the optimized coordinates, the location of the optimized coordinates, denoted as x's, and the original coordinates, denoted as o's.

5.4 Sample Denoising via Informativeness

To illustrate that informativeness can be used as an unsupervised objective function, we explore first-order optimization of informativeness with respect to a sample of points $\{\mathbf{x}_i\}_{i=1}^{n}$, defining a correlation matrix through a Gaussian function as $K_{i,j} = \kappa(\|\mathbf{x}_i - \mathbf{x}_j\|_2^2) = \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x}_i - \mathbf{x}_j\|_2^2\right)$. We first find the bandwidth σ that maximizes the informativeness using a golden search over $\theta \in [-5, 5]$, where $2\sigma^2 = 10^{-\theta}$. Then, we optimize the sample coordinates using the conjugate gradient method implemented in minFunc (Schmidt, 2012) with the gradients given in Section 4.1. For all methods, the optimization is performed in terms of $\acute{\mathbf{K}} = (1 - \eta)\mathbf{K} + \eta\mathbf{I}$, where $\eta = 10^{-6}$, to ensure the correlation matrix is positive definite. A log-barrier term of $-\log(1 - \frac{1}{n^2}\mathbf{1}^\top\mathbf{K}\mathbf{1})$ is added to the cost function for the CKA measure. For the Bures-based measure, a smoothing parameter of $\gamma = 10^{-9}$ is used.

We apply the first-order optimization on five synthetic data sets consisting of $n = 25$ two-dimensional points. As shown in Figure 12, the optimization exhibits mode-seeking—i.e., groups of nearby points cluster together, which enhances the cluster structure seen in the correlation matrices.

We then apply the denoising process to a set of grayscale images of handwritten digits in the USPS data set (Hull, 1994).[11] Taking a random set of 40 images for each digit yields a sample with $n = 400$. The resulting correlation matrices and the images corresponding to the optimized point locations are shown in Figure 13. There is a clear difference in the results for the different measures: With the Euclidean measure, different digits are

[11] The USPS handwritten digits data set (Hull, 1994) is available from `http://web.stanford.edu/~hastie/ElemStatLearn/data.html`, where it is labeled 'ZIP code'.

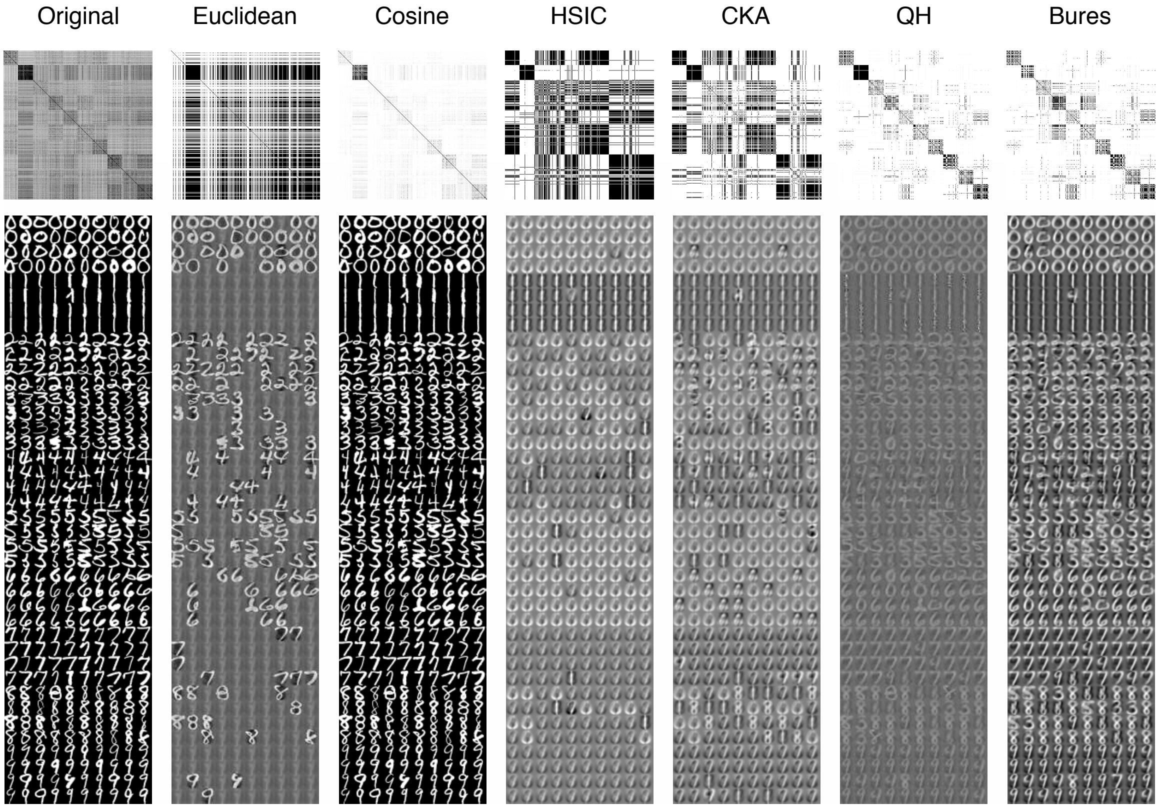

Figure 13: Sample denoising on 200 images from the USPS handwritten digits data set. The correlation matrix is formed from the Gaussian kernel applied to vectors of the pixel values. The first row shows the correlation matrices and the second row shows the corresponding images based on the optimized pixel values.

grouped together into indistinct images, while the remaining images remain unchanged. The cosine measure only slightly modifies the images. The HSIC and CKA measures seem to map most of the digits to a subset of stereotyped digits, but besides 0's, 1's, and 7's/9's the correspondence to the original digits is lost. On the other hand, the QH and Bures measures produce correlation matrices that correspond much more closely to the true class structure. Furthermore, the images appear to be denoised with the majority of the digits mapped to stereotyped versions of the original digits. To quantify the enhanced class structure, we compute the centered kernel alignment between the correlation matrix for the ground-truth labels and the optimized correlation matrix, and record the values in Table 8. As a baseline, we use the original coordinates (that is the original images) for the sample and select a kernel bandwidth that maximizes the centered kernel alignment. Only the Bures measure yields a correlation matrix with higher centered kernel alignment than this baseline.

The QH and Bures measures are also distinguished by their computation time as shown in Table 8, as the gradient calculations for these two measures have higher computational

Table 8: Performance results for sample denoising on the subset of 200 USPS digit images shown in Figure 13. The centered kernel alignment (CA) is calculated between the ground-truth label matrix (a binary correlation matrix) and the correlation matrix for the original sample and the denoised sample for each measure. Computation time logged in MATLAB R2015b on a 2.8 GHz Intel Core i7 with 16 GB RAM.

	Original	Euclidean	Cosine	HSIC	CKA	QH	Bures
CA to labels:	0.53	0.11	0.45	0.27	0.28	0.47	0.55
Time (s):	—	0.99	0.17	1.56	1.10	503.85	37.13

complexity $\mathcal{O}(n^3)$ versus $\mathcal{O}(n^2)$ for the other measures. The QH measure requires solving two Sylvester equations, and the Bures measure requires a single eigendecomposition. This makes the running time with the QH measure much longer in practice.

5.5 Correlation Matrix Denoising

We now demonstrate Algorithm 1 for correlation matrix denoising. The algorithm finds an embedding that maximizes the Bures-based informativeness within a neighborhood of a target embedding. The algorithm is completely unsupervised and can be used to obtain an arbitrary rank embedding. We use it to reduce the rank of an $n \times n$ kernel matrix to $\lfloor \sqrt{n} \rfloor$ as a preprocessing for clustering.

The initial correlation matrix is formed by using a Gaussian kernel between images for five thumbnail image data sets: ORL, MNIST, UMIST, USPS, and COIL-20.[12] The kernel bandwidth heuristic described in Section 5.2 is again used to select this parameter. The initial target embedding was given as $\mathbf{T} = \sqrt{\mathbf{\Sigma}}\mathbf{U}^\top$, where $\mathbf{U\Sigma U}^\top = \mathbf{K}$ is the eigendecomposition of the Gaussian kernel matrix. Figure 14 shows a visualization of the various matrices and also plots the informativeness and centered kernel alignment to the label matrix as a function of rank. As designed, the denoised matrices are more informative than the original matrix and more informative than truncated and renormalized kernel matrices (centered or uncentered) of equal rank. Additionally, except on the ORL data set, the denoised matrices have higher centered kernel alignment.

To demonstrate that the denoising preserves and enhances task-relevant structure, we perform spectral clustering with and without denoising. Starting from an embedding of the original correlation matrix, we obtain a denoised embedding $\mathbf{Z}$ with a rank of $\lfloor \sqrt{n} \rfloor$. Since spectral clustering methods assume a non-negative affinity matrix as input, we apply non-negative thresholding on the entries of the denoised correlation matrix $\mathbf{K}^\star = \mathbf{Z}^\top\mathbf{Z}$ and treat the resulting non-negative, symmetric matrix $[\mathbf{K}^\star]_+$ as the input for Ng et al.'s normalized

[12] The Olivetti Research Laboratory's (ORL) face image data set was previously hosted by AT&T Cambridge and is now hosted by Cambridge University Computer Laboratory: `http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html`. The MNIST handwritten digits test set is available from `http://yann.lecun.com/exdb/mnist/`. The UMIST data set (Graham and Allinson, 1998) is now the Sheffield Face Database `https://www.sheffield.ac.uk/eee/research/iel/research/face`. The COIL-20 data set consists of images of rotated objects (Nene et al., 1996), available at `http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php`.

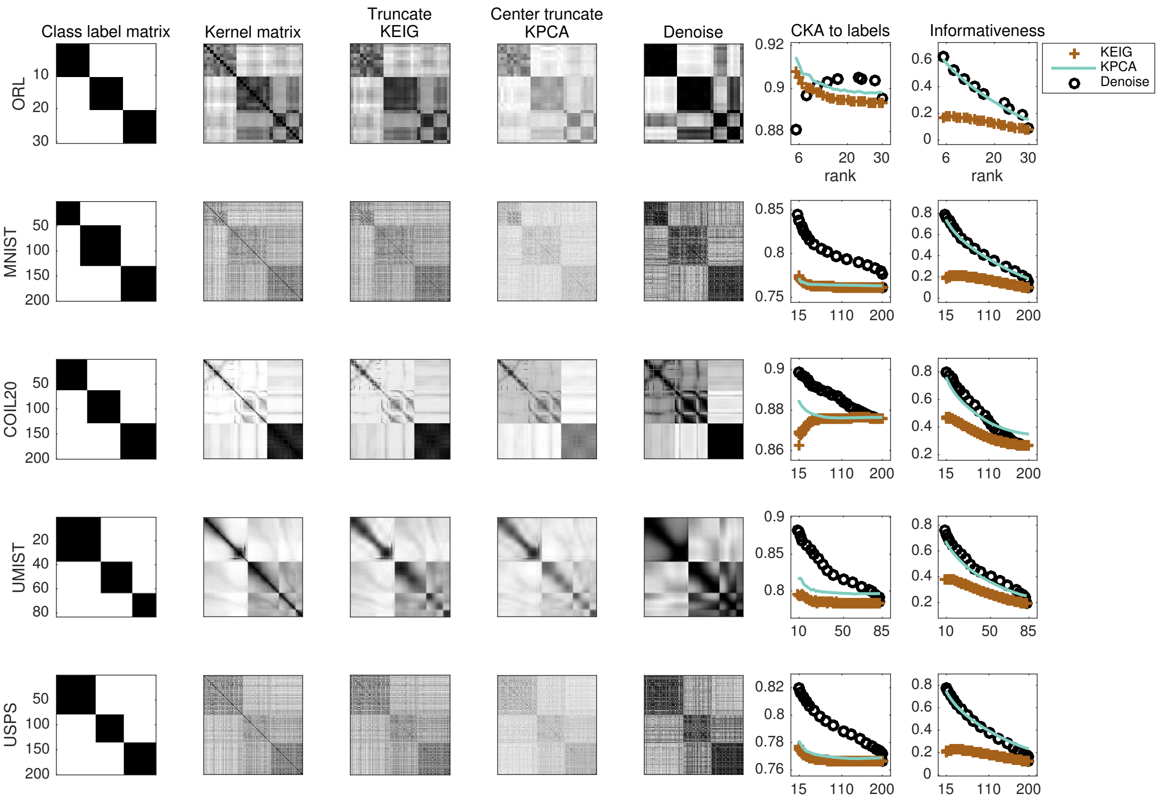

Figure 14: Examples of correlation matrix denoising on Gaussian kernel matrices obtained from samples of the ORL, MNIST, COIL-20, UMIST, and USPS data sets. Each sample contains three different classes. Each row shows the kernel matrix for the ground truth labels, the original Gaussian kernel matrix, the truncated and renormalized matrices corresponding to the uncentered kernel matrix (KEIG) and centered kernel matrix (KPCA), and the denoised correlation matrix. The latter three matrices have rank $\lfloor\sqrt{n}\rfloor$ and the symmetric normalization described in Equation 1 is applied to ensure they are correlation matrices. To the right of the matrices, the centered kernel alignment (CKA) to the ground-truth label matrix and the Bures-based informativeness is shown as a function of rank.

spectral clustering algorithm (Ng et al., 2002). For comparison we also use the spectral clustering method by Shi and Malik (2000) without denoising.

We compare clustering performance on the 20 Newsgroup text data set[13] and the five image data sets used in the previous example. Random subsets of different number of classes (2, 3, 4, and 5) are drawn with 200 samples in each Monte Carlo draw. For the text data set, each document is represented as a bag-of-words vector (a sparse vector of word counts for each document), and in each Monte Carlo division, words that are used in less than 4 documents are dropped. The remaining word counts are multiplied by the

[13] A MATLAB/Octave version is provided by Jason Rennie `http://qwone.com/~jason/20Newsgroups/`.

Table 9: Clustering performance using normalized spectral clustering (NCut), the algorithm by Ng, Jordan, and Weiss (NJW), and the NJW algorithm after denoising the correlation matrix. In each case, k-means with correct number of classes is used to determine the cluster assignment from the spectral embedding. Values correspond to average accuracy (higher is better) or variation in information (lower is better) across 20 Monte Carlo runs; the average loss margin to the best performing method is shown parenthesized. Boldfaced entries indicate significantly better median performance (Wilcoxon signed-rank test with a significance threshold of 0.05).

Accuracy (after mapping clusters to classes)					Variation in information		
	k	NCut	NJW	denoise+NJW	NCut	NJW	denoise+NJW
20 News	2	0.64 (0.29)	0.92 (0.01)	0.93 (0.01)	1.09 (0.46)	0.67 (0.04)	0.68 (0.06)
	3	0.60 (0.22)	0.73 (0.09)	**0.79 (0.03)**	1.70 (0.29)	1.64 (0.24)	1.51 (0.11)
	4	0.54 (0.21)	0.68 (0.07)	**0.75 (0.01)**	2.17 (0.42)	1.99 (0.25)	**1.82 (0.08)**
	5	0.53 (0.17)	0.62 (0.08)	**0.69 (0.01)**	2.34 (0.30)	2.30 (0.25)	**2.16 (0.11)**
COIL-20	2	0.70 (0.26)	0.91 (0.05)	**0.96 (0.00)**	1.20 (1.00)	0.50 (0.29)	**0.22 (0.01)**
	3	0.64 (0.26)	0.86 (0.04)	**0.88 (0.02)**	1.27 (0.77)	0.65 (0.15)	0.60 (0.11)
	4	0.60 (0.27)	0.77 (0.10)	**0.87 (0.00)**	1.50 (0.80)	0.94 (0.25)	**0.70 (0.01)**
	5	0.58 (0.20)	0.72 (0.06)	0.76 (0.02)	1.79 (0.66)	1.37 (0.24)	1.23 (0.10)
MNIST	2	0.78 (0.16)	0.92 (0.02)	**0.94 (0.00)**	1.17 (0.63)	0.70 (0.16)	**0.56 (0.01)**
	3	0.77 (0.09)	0.78 (0.08)	**0.85 (0.01)**	1.43 (0.35)	1.47 (0.40)	**1.14 (0.06)**
	4	0.70 (0.07)	0.71 (0.06)	**0.75 (0.01)**	1.89 (0.23)	1.86 (0.21)	**1.71 (0.05)**
	5	0.64 (0.07)	0.66 (0.05)	0.69 (0.02)	2.30 (0.20)	2.32 (0.22)	**2.14 (0.04)**
USPS	2	0.85 (0.09)	0.91 (0.03)	0.93 (0.01)	0.88 (0.37)	0.67 (0.16)	0.61 (0.10)
	3	0.75 (0.10)	0.76 (0.09)	0.80 (0.05)	1.33 (0.33)	1.30 (0.29)	**1.13 (0.13)**
	4	0.66 (0.13)	0.74 (0.05)	0.76 (0.03)	1.81 (0.57)	1.39 (0.15)	1.32 (0.08)
	5	0.66 (0.09)	0.71 (0.04)	0.73 (0.02)	1.93 (0.28)	1.77 (0.11)	1.79 (0.13)
UMIST	2	0.75 (0.15)	0.72 (0.18)	**0.87 (0.03)**	1.17 (0.62)	1.34 (0.80)	**0.66 (0.11)**
	3	0.64 (0.20)	0.67 (0.17)	**0.81 (0.03)**	1.40 (0.65)	1.31 (0.56)	**0.92 (0.17)**
	4	0.59 (0.15)	0.63 (0.11)	**0.69 (0.05)**	1.56 (0.42)	1.52 (0.38)	1.37 (0.23)
	5	0.54 (0.11)	0.59 (0.06)	0.61 (0.04)	1.81 (0.32)	1.82 (0.33)	1.69 (0.20)
ORL	2	0.78 (0.22)	0.98 (0.02)	1.00 (0.00)	1.00 (1.00)	0.11 (0.11)	0.00 (0.00)
	3	0.81 (0.17)	0.97 (0.01)	0.97 (0.01)	0.66 (0.54)	0.18 (0.06)	0.15 (0.04)
	4	0.77 (0.20)	0.92 (0.04)	0.95 (0.02)	0.84 (0.64)	0.41 (0.21)	0.29 (0.09)
	5	0.81 (0.16)	0.94 (0.03)	0.94 (0.03)	0.75 (0.56)	0.31 (0.12)	0.31 (0.12)

logarithm of the inverse of each word's document frequency (this is the standard TF-IDF weighting), and cosine similarity is used to compute the correlation matrix. For the image data sets, the Gaussian kernel (the kernel bandwidth heuristic described in Section 5.2 is used once again). We fix the number of clusters to the number of classes. The results in terms of accuracy and variation of information (Meilă, 2003) are recorded in Table 9. The clustering obtained when using the denoised and thresholded matrix better matches the ground truth (in terms of both performance metrics) than the clustering obtained using the original kernel matrix across almost every data set and number of classes.

6. Discussion

We have introduced a framework for identifying nontrivial structure in correlation matrices using informativeness, which is a distance-based framework we have proposed for measuring the equality of correlation coefficients (Bartlett, 1954; Anderson, 1963; Lawley, 1963; Gleser, 1968; Aitkin et al., 1968; Steiger, 1980; Brien et al., 1984). Correlation matrices appear in various contexts within machine learning and statistical analysis. Measures which yield correlation matrices include the normalized inner product; Pearson's, Spearman's, and Kendall's correlation coefficients between sets of vectors; kernel matrices formed from non-linear kernel functions on data of various types (since any positive semidefinite kernel function can be normalized); and the normalized graph Laplacian, which can represent any set of non-negative similarity measurements on an undirected graph.

We defined the informativeness of a correlation matrix to be proportional to the distance between it and the closest correlation matrix whose off-diagonal entries are all equal. Motivated by the fact that a scaled correlation matrix is a valid quantum density matrix, we explored quantum distance metrics, including the Bures distance. For specific distance metrics, we derived closed-form expressions of the minimal distance, and a lower-bound for the Bures distance. In particular, we proved the lower-bound, referred to as the Bures-based informativeness, is maximized by centered, rank-1 correlation matrices, which means a maximally informative matrix corresponds to a balanced cut.

The measures of informativeness are unsupervised and can be used as objective functions for machine learning applications. In particular, informativeness can be used for selecting kernel parameters or the embedding dimension for kernel-based dimensionality reduction, testing for structure in graphs, and first-order optimizations of correlation matrices. In the first case, we explored informativeness as a criterion for automatically selecting an appropriate kernel bandwidth for spectral clustering. For this task, the proposed informativeness measures (except the CKA-based measure) consistently outperform Bartlett's and Lawley's test statistics for equality of correlation coefficients, which like the CKA measure always select the largest possible kernel bandwidth. The results also indicate that there is a clear grouping of the different informativeness measures that behave similarly. The cosine measure performs best on three data sets, and the three measures applicable to quantum density matrices—Chernoff bound, QH, and Bures—have the second best performance on three data sets.

In Section 5.2, we investigated using informativeness to select the dimensionality of a kernel-based embedding in a semi-supervised case, where the kernel matrix is formed from both labeled and unlabeled samples, and the performance is assessed by the first nearest-neighbor classification error rates in the reduced dimension space. On this task, Bartlett's and Lawley's test statistics fail to select a meaningful dimensionality (selecting one or two dimensions), which yields an average classification error rate 17.3 percentage points (ppts) higher than the original, as detailed in Table 6. In comparison, the Chernoff, QH, and Bures-based measures have the best average performance with an error rate >0.9 ppts lower than using the distances in the original space. Furthermore, these measures of informativeness outperform centered kernel alignment (Cortes et al., 2012) on this task, which has an error rate of 4.9 ppts higher than the original, even though centered kernel alignment uses the training set labels.

Informativeness is also useful for quantifying the amount of structure in undirected graphs via their normalized graph Laplacian representation, as discussed in Section 5.3. Empirically, the measures of informativeness are all able to distinguish regular and clustered graphs from unstructured graphs. The quantum-based distances (Chernoff bound, QH, and Bures) perform well in distinguishing structured graphs from random graphs (identifying <15% of random graphs as structured and 100% of clustered regular graphs). These informativeness measures are better at identifying structured graphs than the test statistic for determining if the graph is a realization of an Erdős-Rényi graph (Bickel and Sarkar, 2016), which identifies 46% of random graphs and 96.2% of clustered regular graphs. For this task, Bartlett's test statistic also performs well (identifying >2% of random graphs as structured and 95.7% of clustered regular graphs), but Lawley's test statistics evaluates a larger proportion of random graphs higher than the regular graphs (36.4%).

Across these experiments, the three quantum-based distances (Chernoff bound, QH, and Bures) perform consistently well in identifying correlation matrices that are structured and more suitable for clustering or classification. One drawback of these measures is that their evaluations have a higher computation cost than the other measures of informativeness that are based on the Frobenius norm, e.g., the Euclidean distance, cosine similarity, HSIC, and CKA. Specifically, the Bures-based measure requires all of the eigenvalues of an $n \times n$ positive semidefinite matrix, the Chernoff-bound requires a full eigendecomposition, and the QH measures requires the matrix square root—a complexity of $\mathcal{O}(n^3)$ in each case. In comparison, the measures based on the Frobenius norm can be computed directly from the elements of the correlation matrix (or its centered version), with a complexity of $\mathcal{O}(n^2)$.

Closed-form measures of informativeness can also be used as optimization criteria for applications like sample denoising and metric learning through positive definite kernel functions. Using the gradients of smooth cost functions derived from the informativeness measures, we discussed first-order optimizations in Section 4.1. The experimental results for sample denoising in Section 5.4 show that when the coordinates of a sample of points are updated the points form local clusters to maximize informativeness. In this way, informativeness functions as an objective for finding local modes in a sample. This behavior is scale-invariant, since when a Gaussian or other applicable kernel is used, the scale can be automatically set by choosing the bandwidth that maximizes the informativeness. When the informativeness-based optimizations are applied to a thumbnail image data set, the optimization based on the Bures measures enhances the class structure, yielding a higher centered kernel alignment than the correlation matrix for the original sample.

Along similar lines, we investigated the optimization problem of finding maximally informative correlation matrices in terms of their lower-dimensional embeddings (Section 4.2). We proved that the Bures-based measure is especially suited for this task since it corresponds to a convex cost function on Euclidean embeddings. This cost function is a novel matrix norm, which can be used to simultaneously penalize high-rank and uncentered embeddings. Based on the proximal operator of this norm (Appendix E), we proposed a correlation matrix denoising algorithm using an ADMM scheme. In Section 5.5, we tested the proposed correlation matrix denoising algorithm on data sets with known cluster structure. On thumbnail image data sets, the denoising method is able to preserve information relevant to ground-truth classes. Furthermore, it consistently increases the clustering performance when used as a preprocessing for spectral clustering (Ng et al., 2002).

In conclusion, informativeness appears to be a well-suited criterion for a variety of unsupervised data analysis tasks that can be formulated in terms of correlation matrices, and offers a unique measure for assessing the organizational properties of a sample based on pairwise similarity measurements.

Acknowledgments

This work was supported by the Medical Research Council grant MR/L01078X/1 "Supporting Evidence-based Public Health Interventions using Text Mining."

Appendix A. Bures Distance

Bures (1969) defined a distance metric between two non-commutative, positive semidefinite Hermitian matrices that generalizes the Hellinger-Kakutani probability distance to non-commuting spaces. The Bures distance is a measure of the statistical distance between two quantum states, i.e., a symmetric measure of how likely it is for one state to transition to the other (Uhlmann, 1976). In particular, as shown later, when two matrices commute the Bures distance corresponds to the Hellinger-Kakutani distance between the eigenvalues.

Let $\mathbf{A}, \mathbf{B}$ denote two positive-semidefinite matrices, which are trace-normalized so that $\text{tr}\mathbf{A} = \text{tr}\mathbf{B} = 1$. The squared Bures distance between them can be computed (Uhlmann, 1976) as $d_B^2(\mathbf{A}, \mathbf{B}) = 2 - 2\text{tr}\sqrt{\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}}$, where all square roots denote matrix roots. The Bures distance is a chordal distance on the space of positive semidefinite matrices with trace one. The trace term in the previous expression, denoted here as $C(\mathbf{A}, \mathbf{B})$, is equal to the trace norm of the matrix $\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}$, that is

$$C(\mathbf{A}, \mathbf{B}) = \text{tr}\sqrt{\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}} = \left\|\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\right\|_* .$$

This quantity is a symmetric similarity measure between positive definite matrices. From the properties of trace-normalized positive semidefinite matrices, we have $0 \leq C(\mathbf{A}, \mathbf{B}) \leq 1$; $C(\mathbf{A}, \mathbf{B}) = 0$ iff $\langle \mathbf{A}, \mathbf{B} \rangle = 0$ (since $\mathbf{AB}$ and $\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}$ have the same eigenvalues), and $C(\mathbf{A}, \mathbf{B}) = 1$ iff $\mathbf{A} = \mathbf{B}$. In quantum information theory, the quantity $C(\mathbf{A}, \mathbf{B})$, or the concave quantity $C^2(\mathbf{A}, \mathbf{B})$ as originally defined, is known as fidelity (Jozsa, 1994). Fidelity is a measure of how well $\mathbf{A}$ preserves the information of $\mathbf{B}$. It can also be used to directly define a geodesic distance, which is known as the statistical distance or quantum angle, and can be computed using the standard chordal to geodesic conversion, as $\arccos\left(C(\mathbf{A}, \mathbf{B})\right) = \arccos\left(1 - \frac{1}{2}d_B^2(\mathbf{A}, \mathbf{B})\right)$.

If $\mathbf{A}$ and $\mathbf{B}$ commute, then they are simultaneously diagonalizable by an orthogonal matrix $\mathbf{M}$, that is $\mathbf{M}\mathbf{\Lambda}_A\mathbf{M}^\top = \mathbf{A}$ and $\mathbf{M}\mathbf{\Lambda}_B\mathbf{M}^\top = \mathbf{B}$, with $\mathbf{\Lambda}_A$ and $\mathbf{\Lambda}_B$ being the corresponding spectral matrices. Therefore, $\text{tr}\sqrt{\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}} = \text{tr}\big(\mathbf{M}\sqrt{\mathbf{\Lambda}_A\mathbf{\Lambda}_B}\mathbf{M}^\top\big) = \text{tr}(\sqrt{\mathbf{\Lambda}_A}\sqrt{\mathbf{\Lambda}_B})$ and $\text{tr}\big(\mathbf{M}\sqrt{\mathbf{\Lambda}_A\mathbf{\Lambda}_B}\mathbf{M}^\top\big) = \langle\sqrt{\mathbf{A}}, \sqrt{\mathbf{B}}\rangle$, which makes the fidelity $C(\mathbf{A}, \mathbf{B})$ equal to the affinity measure in Table 1, or equivalently, the Bures and the quantum Hellinger distance to coincide because $d_B^2(\mathbf{A}, \mathbf{B}) = \|\sqrt{\mathbf{A}} - \sqrt{\mathbf{B}}\|_F^2$. In this case, by using $\mathbf{p}_A = \text{diag}\left(\mathbf{\Lambda}_A\right)$ and $\mathbf{p}_B = \text{diag}\left(\mathbf{\Lambda}_B\right)$ to denote vectors containing the eigenvalues, $C(\mathbf{A}, \mathbf{B})$ is also equal to $\langle\sqrt{\mathbf{p}_A}, \sqrt{\mathbf{p}_B}\rangle$. This last quantity is the Bhattacharyya coefficient between the eigenspectra of $\mathbf{A}$ and $\mathbf{B}$, and hence, $d_B^2(\mathbf{A},\mathbf{B}) = \|\sqrt{\mathbf{p}_A} - \sqrt{\mathbf{p}_B}\|_2^2$ is the squared Euclidean distance between them.

A further point to note about the Bures distance is that it serves as a lower bound for the trace distance. Specifically, as shown by Fuchs and van de Graaf (1999), we have $\frac{1}{2}d_B^2(\mathbf{A},\mathbf{B}) = 1 - C(\mathbf{A},\mathbf{B}) \leq \frac{1}{2}\|\mathbf{A}-\mathbf{B}\|_*$. It is also a lower bound on quantum Hellinger distance (Luo and Zhang, 2004). This can also be directly seen from the fact that for a given matrix $\mathbf{Q}$, we have $\operatorname{tr}(\mathbf{Q}) \leq \max_{\|\mathbf{U}\|_2\leq 1}\operatorname{tr}(\mathbf{U}^\top\mathbf{Q}) = \|\mathbf{Q}\|_*$ from the duality of the trace and spectral norms. By setting $\mathbf{Q} = \sqrt{\mathbf{A}}\sqrt{\mathbf{B}}$, we have the fidelity measure $C(\mathbf{A},\mathbf{B})$ dominating the affinity measure, $\langle\sqrt{\mathbf{A}},\sqrt{\mathbf{B}}\rangle$.

For the context of this work, we will show that a useful property of the Bures distance is that in its computation, the matrix square roots $\sqrt{\mathbf{A}}$ and $\sqrt{\mathbf{B}}$ can be substituted with any arbitrary embeddings in a Hilbert space that represent the correlation structure. Assume we have two arbitrary embeddings $\mathbf{Z}_A \in \mathbb{R}^{b_A\times n}$ and $\mathbf{Z}_B \in \mathbb{R}^{b_B\times n}$ such that $\mathbf{Z}_A^\top\mathbf{Z}_A = \mathbf{A}$ and $\mathbf{Z}_B^\top\mathbf{Z}_B = \mathbf{B}$. Since $\|\mathbf{Q}\|_* = \left\|\mathbf{Q}^\top\right\|_* = \operatorname{tr}\sqrt{\mathbf{Q}^\top\mathbf{Q}}$, we have

$$\begin{aligned}\left\|\mathbf{Z}_A\mathbf{Z}_B^\top\right\|_* &= \operatorname{tr}\sqrt{\mathbf{Z}_A\mathbf{Z}_B^\top\mathbf{Z}_B\mathbf{Z}_A^\top} = \operatorname{tr}\sqrt{\mathbf{Z}_A\mathbf{B}\mathbf{Z}_A^\top} = \left\|\mathbf{Z}_A\mathbf{B}^{\frac{1}{2}}\right\|_* \\ &= \operatorname{tr}\sqrt{\mathbf{B}^{\frac{1}{2}}\mathbf{Z}_A^\top\mathbf{Z}_A\mathbf{B}^{\frac{1}{2}}} = \operatorname{tr}\sqrt{\mathbf{B}^{\frac{1}{2}}\mathbf{A}\mathbf{B}^{\frac{1}{2}}} = \left\|\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\right\|_*.\end{aligned} \tag{A.1}$$

A critical conclusion from these equations, is that the fidelity $\|\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\|_*$ and the Bures distance d_B can potentially use any of the forms in Equation A.1 for calculating the constituent trace norm, according to computational convenience and embeddings dimensionality, even though the embeddings $\mathbf{Z}_A$ and $\mathbf{Z}_B$ are not unique representations of $\sqrt{\mathbf{A}}$ and $\sqrt{\mathbf{B}}$. For instance, the constraint $\mathbf{Z}_A^\top\mathbf{Z}_A = \mathbf{A}$ cannot imply a unique $\mathbf{Z}_A$, as a row permutation, sign change, or any orthogonal transformation of the rows of $\mathbf{Z}_A$ would not alter the underlying similarity matrix. Furthermore, one can choose a transformation matrix $\mathbf{U}_A$, which is not necessarily square, with orthonormal columns (and, by definition, orthonormal rows if $\mathbf{U}_A$ is square) that introduces the embedding into a higher dimensional space. For example, if $\mathbf{U}_A \in \mathbb{R}^{q\times b_A}$, the projection $\mathbf{U}_A\mathbf{Z}_A$ is equivalent to applying an orthogonal transformation to $\binom{\mathbf{Z}_A}{\mathbf{0}} \in \mathbb{R}^{q\times n}$. These invariances lead to an alternative formulation of fidelity via the following theorem, which is known as Uhlmann's theorem in the more general case for quantum density matrices (Uhlmann, 1976; Jozsa, 1994; Nielsen and Chuang, 2000).

Theorem 1 *For positive semidefinite matrices* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *and* $\mathbf{B} \in \mathbb{R}^{n\times n}$*,*

$$C(\mathbf{A},\mathbf{B}) = \left\|\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\right\|_* = \max_{\substack{\mathbf{Z}_A,\mathbf{Z}_B \\ \mathbf{Z}_A^\top\mathbf{Z}_A=\mathbf{A} \\ \mathbf{Z}_B^\top\mathbf{Z}_B=\mathbf{B}}} \langle\mathbf{Z}_A,\mathbf{Z}_B\rangle.$$

Proof Let $\bar{\mathbf{Z}}_A \in \mathbb{R}^{b_A\times n}$ and $\bar{\mathbf{Z}}_B \in \mathbb{R}^{b_B\times n}$ be arbitrary embeddings such that $\bar{\mathbf{Z}}_A^\top\bar{\mathbf{Z}}_A = \mathbf{A}$ and $\bar{\mathbf{Z}}_B^\top\bar{\mathbf{Z}}_B = \mathbf{B}$ (without loss of generality, we assume $b_A \leq b_B$). The set of valid embeddings for $\mathbf{A}$ and $\mathbf{B}$ can then be expressed as transformations of $\bar{\mathbf{Z}}_A$ and $\bar{\mathbf{Z}}_B$, i.e.,

$$\left\{(\mathbf{Z}_A,\mathbf{Z}_B) : \mathbf{Z}_A^\top\mathbf{Z}_A = \mathbf{A},\ \mathbf{Z}_B^\top\mathbf{Z}_B = \mathbf{B}\right\} \quad = \quad \left\{(\mathbf{U}_A\bar{\mathbf{Z}}_A,\mathbf{U}_B\bar{\mathbf{Z}}_B) : \mathbf{U}_A^\top\mathbf{U}_A = \mathbf{I},\ \mathbf{U}_B^\top\mathbf{U}_B = \mathbf{I}\right\},$$

where $\mathbf{U}_A \in \mathbb{R}^{q \times b_A}$ and $\mathbf{U}_B \in \mathbb{R}^{q \times b_A}$ are matrices with orthonormal columns and the same number of rows. From the equivalence of these two sets, we pose the equivalent maximization problem

$$\max_{\substack{\mathbf{U}_A, \mathbf{U}_B \\ \mathbf{U}_A^\top \mathbf{U}_A = \mathbf{I} \\ \mathbf{U}_B^\top \mathbf{U}_B = \mathbf{I}}} \left\{ \langle \mathbf{U}_A \bar{\mathbf{Z}}_A, \mathbf{U}_B \bar{\mathbf{Z}}_B \rangle = \operatorname{tr}\Big(\bar{\mathbf{Z}}_A^\top \mathbf{U}_A^\top \mathbf{U}_B \bar{\mathbf{Z}}_B \Big) = \operatorname{tr}\Big(\mathbf{U}_B^\top \mathbf{U}_A \bar{\mathbf{Z}}_A \bar{\mathbf{Z}}_B^\top \Big) \right\}.$$

Since $\mathbf{U}_A$ and $\mathbf{U}_B$ are semi-orthogonal, they both have all their non-zero singular values equal to one, and from the sub-multiplicative property it follows that

$$\|\mathbf{U}_B^\top \mathbf{U}_A\|_2 \leq \|\mathbf{U}_A\|_2 \cdot \|\mathbf{U}_B\|_2 = 1,$$

and furthermore, the set of pairs of semi-orthogonal matrices is a strict subset of the set of pairs with product within the spectral norm unit ball:

$$\Big\{ (\mathbf{U}_A, \mathbf{U}_B) : \mathbf{U}_A^\top \mathbf{U}_A = \mathbf{I},\ \mathbf{U}_B^\top \mathbf{U}_B = \mathbf{I} \Big\} \ \subset \ \Big\{ (\mathbf{U}_A, \mathbf{U}_B) : \|\mathbf{U}_B^\top \mathbf{U}_A\|_2 \leq 1 \Big\}.$$

Letting $\mathbf{U}^\top = \mathbf{U}_B^\top \mathbf{U}_A$, we have

$$\max_{\substack{\mathbf{U}_A, \mathbf{U}_B \\ \mathbf{U}_A^\top \mathbf{U}_A = \mathbf{I} \\ \mathbf{U}_B^\top \mathbf{U}_B = \mathbf{I}}} \operatorname{tr}\Big(\mathbf{U}_B^\top \mathbf{U}_A \bar{\mathbf{Z}}_A \bar{\mathbf{Z}}_B^\top \Big) \ \leq \ \max_{\substack{\mathbf{U} \\ \|\mathbf{U}\|_2 \leq 1}} \operatorname{tr}\Big(\mathbf{U}^\top \bar{\mathbf{Z}}_A \bar{\mathbf{Z}}_B^\top \Big) = \|\bar{\mathbf{Z}}_A \bar{\mathbf{Z}}_B^\top\|_*,$$

where the equality follows from the duality of the trace and spectral norms. Furthermore, the optimization in terms of $\mathbf{U}_A, \mathbf{U}_B$ achieves the maximum when $\mathbf{U}_A = [\mathbf{R},\ \mathbf{0}_{b_A \times (b_B - b_A)}]^\top$ and $\mathbf{U}_B = \mathbf{V}$, where $\mathbf{R}\mathbf{S}\mathbf{V}^\top$ is a singular value decomposition of $\bar{\mathbf{Z}}_A \bar{\mathbf{Z}}_B^\top$, since in this case[14]

$$\begin{aligned} \operatorname{tr}\Big(\mathbf{U}_B^\top \mathbf{U}_A \bar{\mathbf{Z}}_A \bar{\mathbf{Z}}_B^\top \Big) &= \operatorname{tr}\Big(\mathbf{U}_B^\top \mathbf{U}_A \mathbf{R}\mathbf{S}\mathbf{V}^\top \Big) = \operatorname{tr}\Big(\mathbf{V}^\top \mathbf{U}_B^\top \mathbf{U}_A \mathbf{R}\mathbf{S} \Big) \\ &= \operatorname{tr}\bigg(\mathbf{V}^\top \mathbf{V} \Big[\mathbf{R}^\top \mathbf{R},\ \mathbf{0}_{b_A \times (b_B - b_A)} \Big]^\top \mathbf{S} \bigg) = \operatorname{tr}\Big(\big[\mathbf{I}_{b_A \times b_A},\ \mathbf{0}_{b_A \times (b_B - b_A)} \big]^\top \mathbf{S} \Big) = \|\bar{\mathbf{Z}}_A \bar{\mathbf{Z}}_B^\top\|_*. \end{aligned}$$

By the properties of the trace norm (A.1), $\|\bar{\mathbf{Z}}_A \bar{\mathbf{Z}}_B^\top\|_* = \|\mathbf{A}^{\frac{1}{2}} \mathbf{B}^{\frac{1}{2}}\|_*$, which completes the proof. ■

It is a significant consequence of this theorem that $d_B(\mathbf{A}, \mathbf{B})$ corresponds to the minimal distance $\|\mathbf{Z}_A - \mathbf{Z}_B\|_F$ between all possible embeddings represented by $\mathbf{A}$ and $\mathbf{B}$. This is very desirable, as it makes d_B independent to the choice of their embeddings used to measure it with any of the forms in Equation A.1.

We note that Theorem 1 can also be used to show that the symmetric normalization in Equation 1, discussed in Section 2.1, which transforms a positive semidefinite matrix to a correlation matrix, corresponds to finding the closest elliptope element in the Bures distance sense.

[14] Letting $\mathbf{Q} = \mathbf{V}^\top \mathbf{U}_B^\top \mathbf{U}_A \mathbf{R} = (\mathbf{U}_B \mathbf{V})^\top \mathbf{U}_A \mathbf{R}$ we have $\operatorname{tr}\big(\mathbf{U}_B^\top \mathbf{U}_A \bar{\mathbf{Z}}_A \bar{\mathbf{Z}}_B^\top \big) = \operatorname{tr}(\mathbf{Q}\mathbf{S})$. We note that no element of $\mathbf{Q}$ can exceed unity by the Cauchy-Bunyakovsky-Schwarz inequality since $\mathbf{U}_A \mathbf{R}$ and $\mathbf{U}_B \mathbf{V}$ have orthonormal columns. From this, it follows that $\operatorname{tr}\big(\mathbf{U}_B^\top \mathbf{U}_A \bar{\mathbf{Z}}_A \bar{\mathbf{Z}}_B^\top \big) = \operatorname{tr}(\mathbf{Q}\mathbf{S}) \leq \operatorname{tr}(\mathbf{S}) = \|\bar{\mathbf{Z}}_A \bar{\mathbf{Z}}_B^\top\|_*$.

Lemma 2 *For a positive semidefinite matrix* $\tilde{\mathbf{K}} \in \mathbb{R}^{n\times n}$ *with* $\mathrm{tr}(\tilde{\mathbf{K}}) = n$*,*

$$\operatorname*{arg\,min}_{\mathbf{K}\in\mathcal{E}} \; d_B\left(\tfrac{1}{n}\mathbf{K}, \tfrac{1}{n}\tilde{\mathbf{K}}\right) \;=\; \mathbf{D}\tilde{\mathbf{K}}\mathbf{D},$$

where $\mathbf{D}$ *is a diagonal matrix with entries* $D_{i,i} = \frac{1}{\sqrt{\tilde{K}_{i,i}}}$*. Furthermore,*

$$\min_{\mathbf{K}\in\mathcal{E}} \; d_B\left(\tfrac{1}{n}\mathbf{K}, \tfrac{1}{n}\tilde{\mathbf{K}}\right) = \sqrt{\frac{2}{n}\sum_{i\in[n]}\left(1-(\tilde{K}_{i,i})^{\frac{1}{2}}\right)},$$

i.e., the distance depends on how far the roots of the diagonal entries of $\tilde{\mathbf{K}}$ *are from unity.*

Proof Let $\tilde{\mathbf{Z}}$ be an embedding of $\tilde{\mathbf{K}}$ such that $\tilde{\mathbf{Z}}^\top\tilde{\mathbf{Z}} = \tilde{\mathbf{K}}$. Then for any choice of $\tilde{\mathbf{Z}}$,

$$\min_{\mathbf{K}\in\mathcal{E}} d_B\left(\tfrac{1}{n}\mathbf{K}, \tfrac{1}{n}\tilde{\mathbf{K}}\right) = \min_{\mathbf{Z}\in\mathcal{OB}} \left\| \tfrac{1}{\sqrt{n}}\mathbf{Z} - \tfrac{1}{\sqrt{n}}\tilde{\mathbf{Z}} \right\|_F = \sqrt{\frac{1}{n}\sum_{i\in[n]} \min_{\mathbf{x}:\|\mathbf{x}\|_2=1} \|\mathbf{x}-\tilde{\mathbf{z}}_i\|_2^2}\;,$$

where the second equality follows from the fact that both the oblique manifold constraint and the Frobenius norm are amenable to column-wise separation. The solution for the ith column is $\mathbf{z}_i = \frac{\tilde{\mathbf{z}}_i}{\|\tilde{\mathbf{z}}_i\|_2}$, which corresponds to finding the point on the unit sphere closest to the vector $\tilde{\mathbf{z}}_i$. This is equivalent to setting $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] = \tilde{\mathbf{Z}}\mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix with entries $D_{i,i} = \frac{1}{\|\tilde{\mathbf{z}}_i\|_2} = \frac{1}{\sqrt{\tilde{K}_{i,i}}}$, which yields the closest matrix in the elliptope as $\mathbf{D}\tilde{\mathbf{Z}}^\top\tilde{\mathbf{Z}}\mathbf{D} = \mathbf{D}\tilde{\mathbf{K}}\mathbf{D} \in \mathcal{E}$.

The distance itself is given as

$$\begin{aligned} d_B\left(\tfrac{1}{n}\mathbf{D}\tilde{\mathbf{K}}\mathbf{D}, \tfrac{1}{n}\tilde{\mathbf{K}}\right) &= \sqrt{2-2\left\|\tfrac{1}{n}\tilde{\mathbf{Z}}\mathbf{D}\tilde{\mathbf{Z}}^\top\right\|_*} = \sqrt{2-\frac{2}{n}\mathrm{tr}\left(\mathbf{D}\tilde{\mathbf{Z}}^\top\tilde{\mathbf{Z}}\right)} \\ &= \sqrt{2-\frac{2}{n}\sum_{i\in[n]}\|\tilde{\mathbf{z}}_i\|_2} = \sqrt{\frac{2}{n}\sum_{i\in[n]}\left(1-(\tilde{K}_{i,i})^{\frac{1}{2}}\right)}\;. \end{aligned}$$

■

This simple normalization can be contrasted with the alternating projection algorithm required to find the closest correlation matrix in terms of the Euclidean distance (Higham, 2002). The latter could be considered more general, since it can be applied to any square matrix, whereas the Bures distance is only applicable to positive semidefinite matrices. Nonetheless, the symmetric normalization can be applied after first finding the nearest positive semidefinite matrix (Higham, 1988).

Another property of the Bures distance, is that its square is a convex function, which follows from the following theorem for real-valued positive-semidefinite matrices, which is a special case of the proof by Uhlmann (1976).

Theorem 3 *The squared fidelity measure* $C^2(\mathbf{A},\mathbf{B}) = \|\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\|_*^2$ *is concave for positive-semidefinite matrices.*

Proof Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \in \mathbb{R}^{n\times n}$ be three positive-semidefinite matrices. Let $\Psi_Z = \bigoplus_{i=1}^n \psi_{z_i} \in \mathcal{H}_X$ and $\Psi_X = \bigoplus_{i=1}^n \psi_{x_i} \in \mathcal{H}_X$ denote Hilbert-space embeddings defined via the direct sum $\mathcal{H}_X = \mathcal{H}_x \oplus \cdots \oplus \mathcal{H}_x = \bigoplus_{i=1}^n \mathcal{H}_x$ such that $Z_{i,j} = \langle \psi_{z_i}, \psi_{z_j}\rangle, X_{i,j} = \langle \psi_{x_i}, \psi_{x_j}\rangle$, and $\|\mathbf{Z}^{\frac{1}{2}}\mathbf{X}^{\frac{1}{2}}\|_* = \langle \Psi_Z, \Psi_X\rangle$. Likewise, let $\Phi_Z = \bigoplus_{i=1}^n \phi_{z_i} \in \mathcal{H}_Y$ and $\Psi_Y = \bigoplus_{i=1}^n \phi_{y_i} \in \mathcal{H}_Y$ denote Hilbert-space embeddings such that $Z_{i,j} = \langle \phi_{z_i}, \phi_{z_j}\rangle, Y_{i,j} = \langle \phi_{y_i}, \phi_{y_j}\rangle$, and $\|\mathbf{Z}^{\frac{1}{2}}\mathbf{Y}^{\frac{1}{2}}\|_* = \langle \Phi_Z, \Phi_Y\rangle$.

For any λ_X, λ_Y such that $\lambda_X^2 + \lambda_Y^2 = 1$, define

$$\Omega_Z = \bigoplus_{i=1}^n \omega_{z_i} = (\lambda_X \Psi_Z) \oplus (\lambda_Y \Phi_Z) \in \mathcal{H}_X \oplus \mathcal{H}_Y \equiv \mathcal{H}_W.$$

Then $Z_{i,j} = \langle \omega_{z_i}, \omega_{z_j}\rangle = \lambda_X^2 \langle \psi_{z_i}, \psi_{z_j}\rangle + \lambda_Y^2 \langle \phi_{z_i}, \phi_{z_j}\rangle$. Let $\mathbf{W} = \alpha\mathbf{X} + (1-\alpha)\mathbf{Y}$ denote a convex combination, where $0 \le \alpha \le 1$. Then $\Omega_W = (\sqrt{\alpha}\Psi_X) \oplus (\sqrt{1-\alpha}\Phi_Y) \in \mathcal{H}_W$ is a Hilbert-space embedding of $\mathbf{W}$. Consequently, $\langle \Omega_Z, \Omega_W\rangle = \lambda_X \sqrt{\alpha}\langle \Psi_Z, \Psi_X\rangle + \lambda_Y \sqrt{1-\alpha}\langle \Phi_Z, \Phi_Y\rangle$.

From Theorem 1 it follows that

$$\|\mathbf{Z}^{\frac{1}{2}}\mathbf{W}^{\frac{1}{2}}\|_*^2 \ge |\langle \Omega_Z, \Omega_W\rangle|^2 = \left|\lambda_X \sqrt{\alpha}\langle \Psi_Z, \Psi_X\rangle + \lambda_Y \sqrt{1-\alpha}\langle \Phi_Z, \Phi_Y\rangle\right|^2.$$

Maximizing over λ_X, λ_Y on the right hand side (which corresponds to an optimization along the positive quadrant of the circle parametrically defined by coordinates λ_X and λ_Y) we obtain

$$\|\mathbf{Z}^{\frac{1}{2}}\mathbf{W}^{\frac{1}{2}}\|_*^2 \ge \left|\sqrt{\alpha}\langle \Psi_Z, \Psi_X\rangle\right|^2 + \left|\sqrt{1-\alpha}\langle \Phi_Z, \Phi_Y\rangle\right|^2 = \alpha\|\mathbf{Z}^{\frac{1}{2}}\mathbf{X}^{\frac{1}{2}}\|_*^2 + (1-\alpha)\|\mathbf{Z}^{\frac{1}{2}}\mathbf{Y}^{\frac{1}{2}}\|_*^2.$$

■

Appendix B. Derivation of Informativeness Measures

In this appendix, we detail the derivation of the analytic expression of each measure of informativeness found in Table 2.

B.1 Euclidean Distance

We define the measure of informativeness for any correlation matrix $\mathbf{K} \in \mathcal{E}$ using the Euclidean distance as $d_{\mathcal{N}}(\mathbf{K}) = \frac{1}{n}\min_{\mathbf{N}_a \in \mathcal{N}} \|\mathbf{K} - \mathbf{N}_a\|_F$. Expanding the Frobenius norm term yields

$$\begin{aligned}
\|\mathbf{K} - \mathbf{N}_a\|_F^2 &= \|\mathbf{K}\|_F^2 + \|\mathbf{N}_a\|_F^2 - 2\langle \mathbf{K}, \mathbf{N}_a\rangle, \\
\|\mathbf{N}_a\|_F^2 &= (an)^2 + (1-a)^2\frac{n^2}{n-1}, \\
\langle \mathbf{K}, \mathbf{N}_a\rangle &= a\mathbf{1}^\top\mathbf{K}\mathbf{1} + (1-a)\frac{n}{n-1}\mathrm{tr}(\mathbf{H}\mathbf{K}) = a\bar{k}n^2 + (1-a)(1-\bar{k})\frac{n^2}{n-1},
\end{aligned}$$

since $\mathrm{tr}(\mathbf{H}\mathbf{K}) = n - \frac{1}{n}\mathrm{tr}(\mathbf{J}\mathbf{K}) = n(1-\bar{k})$, where $\bar{k} = \frac{1}{n^2}\mathbf{1}^\top\mathbf{K}\mathbf{1}$. Taking the derivative with respect to a yields a linear function of a whose root is $\bar{k}$. That is, the value of the parameter

a that defines the closest non-informative matrix $\mathbf{N}_{a^\star}$ is

$$a^\star = \underset{0\leq a\leq 1}{\arg\min}\ \|\mathbf{K}-\mathbf{N}_a\|_F^2 = \bar{k}, \tag{B.1}$$

and

$$\|\mathbf{K}-\mathbf{N}_{\bar{k}}\|_F^2 = \|\mathbf{K}\|_F^2 - (n\bar{k})^2 - (1-\bar{k})^2\frac{n^2}{n-1} = \|\mathbf{K}\|_F^2 - \frac{n^2}{n-1}(n\bar{k}^2 - 2\bar{k}+1).$$

The informativeness measure is

$$d_{\mathcal{N}}(\mathbf{K}) = \sqrt{\frac{1}{n^2}\|\mathbf{K}\|_F^2 - \frac{1}{n-1}(n\bar{k}^2-2\bar{k}+1)}. \tag{B.2}$$

This expression is a specific case of the formula for the Euclidean distance of an arbitrary matrix to the closest correlation matrix with uniform off-diagonal elements (Borsdorf et al., 2010).

B.2 Cosine Distance

Using the cosine similarity, one can define a chordal distance, which amounts to Euclidean distance after normalizing the matrices by their Frobenius norm. By squaring the chordal distance and dividing by two, we obtain a measure of informativeness as $1 - \max_{0\leq a\leq 1}\frac{\langle \mathbf{K},\mathbf{N}_a\rangle}{\|\mathbf{K}\|_F\|\mathbf{N}_a\|_F}$. For a specific value of a and $\mathbf{K}\in\mathcal{E}$ we have

$$\frac{\langle \mathbf{K},\mathbf{N}_a\rangle}{\|\mathbf{K}\|_F\|\mathbf{N}_a\|_F} = \frac{a\mathbf{1}^\top\mathbf{K}\mathbf{1} + (1-a)\frac{n}{n-1}\mathrm{tr}(\mathbf{HK})}{\|\mathbf{K}\|_F\sqrt{(an)^2+(1-a)^2\frac{n^2}{n-1}}} = \frac{a\bar{k}n + (1-a)(1-\bar{k})\frac{n}{n-1}}{\|\mathbf{K}\|_F\sqrt{(a^2n-2a+1)\frac{1}{n-1}}}.$$

Taking the derivative of this expression and solving for its root, yields

$$a^\star = \underset{0\leq a\leq 1}{\arg\max}\ \frac{\langle \mathbf{K},\mathbf{N}_a\rangle}{\|\mathbf{K}\|_F\|\mathbf{N}_a\|_F} = \bar{k}. \tag{B.3}$$

Substituting $a^\star$ back into the cosine similarity, leads to

$$\max_{0\leq a\leq 1}\frac{\langle \mathbf{K},\mathbf{N}_a\rangle}{\|\mathbf{K}\|_F\|\mathbf{N}_a\|_F} = \frac{\langle \mathbf{K},\mathbf{N}_{a^\star}\rangle}{\|\mathbf{K}\|_F\|\mathbf{N}_{a^\star}\|_F} = \frac{n\bar{k}^2 + (1-\bar{k})^2\frac{n}{n-1}}{\|\mathbf{K}\|_F\sqrt{(\bar{k}^2n-2\bar{k}+1)\frac{1}{n-1}}} = \frac{n}{\|\mathbf{K}\|_F}\sqrt{\frac{n\bar{k}^2-2\bar{k}+1}{n-1}}.$$

The measure of informativeness based on the chordal distance is

$$\frac{1}{2}d_{\mathcal{N}}^2(\mathbf{K}) = 1 - \frac{n}{\|\mathbf{K}\|_F}\sqrt{\frac{n\bar{k}^2-2\bar{k}+1}{n-1}}. \tag{B.4}$$

B.3 Hilbert-Schmidt Independence Criterion (HSIC)

The use of HSIC as a measure of similarity between a correlation matrix and the closest non-informative one leads to a trivial function. Specifically, for any a we have

$$\mathrm{HSIC}(\mathbf{K},\mathbf{N}_a) = \langle \mathbf{HKH},\mathbf{N}_a\rangle = \mathrm{tr}(\mathbf{HKHN}_a) = \frac{(1-a)n}{n-1}\mathrm{tr}(\mathbf{HK}).$$

This similarity is maximized for $a = 0$ and the informativeness would be inversely proportional to $\mathrm{tr}(\mathbf{HK}) = n(1-\bar{k})$ for $\mathbf{K} \in \mathcal{E}$, where $\bar{k}$ is as defined previously to be the average of the entries of $\mathbf{K}$.

A more meaningful measure of informativeness, however, is based on the Euclidean distance after centering (see Table 1). This is defined as

$$d_{\mathcal{N}}(\mathbf{K}) = \frac{1}{n}\min_{\mathbf{N}_a \in \mathcal{N}} \|\mathbf{HKH} - \mathbf{HN}_a\mathbf{H}\|_F = \frac{1}{n}\min_{\mathbf{N}_a \in \mathcal{N}} \left\| \mathbf{HKH} - \frac{(1-a)n}{n-1}\mathbf{H} \right\|_F .$$

Expanding the norm we have

$$\left\| \mathbf{HKH} - \frac{(1-a)n}{n-1}\mathbf{H} \right\|_F^2 = \|\mathbf{HKH}\|_F^2 + \frac{(1-a)^2n^2}{n-1} - 2\frac{(1-a)n}{n-1}\mathrm{tr}(\mathbf{HK}).$$

Taking the derivative of this expression and solving for its root yields

$$a^\star = \operatorname*{arg\,min}_{0\le a\le 1} \|\mathbf{HKH} - \mathbf{HN}_a\mathbf{H}\|_F^2 = \bar{k}, \tag{B.5}$$

and the measure of informativeness is

$$d_{\mathcal{N}}(\mathbf{K}) = \sqrt{\frac{1}{n^2}\|\mathbf{HKH}\|_F^2 - \frac{1}{n-1}(1-\bar{k})^2}. \tag{B.6}$$

B.4 Centered Kernel Alignment (CKA)

The CKA similarity measurement between kernel matrices can be used to define a chordal distance between centered kernel matrices. By squaring the chordal distance and dividing by two, we define the measure of informativeness as

$$\begin{aligned}\frac{1}{2}d_{\mathcal{N}}^2(\mathbf{K}) &= 1 - \max_{0\le a\le 1} \frac{\langle \mathbf{HKH}, \mathbf{N}_a\rangle}{\|\mathbf{HKH}\|_F\|\mathbf{HN}_a\mathbf{H}\|_F} = 1 - \max_{0\le a\le 1}\frac{\frac{(1-a)n}{n-1}\mathrm{tr}(\mathbf{HK})}{\|\mathbf{HKH}\|_F\frac{(1-a)n}{n-1}\|\mathbf{H}\|_F} \\ &= 1 - \frac{\mathrm{tr}(\mathbf{HK})}{\|\mathbf{HKH}\|_F\sqrt{n-1}} = 1 - \frac{n(1-\bar{k})}{\|\mathbf{HKH}\|_F\sqrt{n-1}},\end{aligned} \tag{B.7}$$

where $\bar{k}$ is as defined before, and $\mathrm{tr}(\mathbf{HK}) = n(1-\bar{k})$ for $\mathbf{K} \in \mathcal{E}$. The measure is undefined for the constant matrix $\mathbf{K} = \mathbf{N}_1 = \mathbf{J}$, because $\|\mathbf{HKH}\|_F = 0$, but since $\mathbf{J}$ is non-informative, the informativeness measure in this case should be defined to be zero. Furthermore, for a rank-1 matrix $\mathbf{K} = \mathbf{v}\mathbf{v}^\top$, this measure is constant with value of $1 - \frac{1}{\sqrt{n-1}}$, since $\mathrm{tr}(\mathbf{HK}) = \mathbf{v}^\top\mathbf{Hv} = \sqrt{\mathrm{tr}(\mathbf{v}^\top\mathbf{Hvv}^\top\mathbf{Hv})} = \|\mathbf{Hvv}^\top\mathbf{H}\|_F = \|\mathbf{HKH}\|_F$.

B.5 Chernoff Bound

The quantum Chernoff bound is a lower bound on both the affinity and fidelity, and it can thus be used to define an upper bound on a function of the quantum Hellinger and Bures distances (Audenaert et al., 2007). We define a measure of informativeness based on the Chernoff bound $Q(\frac{1}{n}\mathbf{K}, \frac{1}{n}\mathbf{N}_a) = \min_{0\le s\le 1}\frac{1}{n}\langle \mathbf{K}^s, \mathbf{N}_a^{1-s}\rangle$, for $\mathbf{K} \in \mathcal{E}$ as

$$d_{\mathcal{N}}(\mathbf{K}) = 1 - \max_{0\le a\le 1} Q(\tfrac{1}{n}\mathbf{K}, \tfrac{1}{n}\mathbf{N}_a) = 1 - \max_{0\le a\le 1}\min_{0\le s\le 1} f(\mathbf{K}, a, s),$$

where

$$f(\mathbf{K}, a, s) \equiv \frac{(an)^{1-s}}{n} \operatorname{tr}(\mathbf{K}^s \mathbf{J}) + \left(\frac{(1-a)n}{n-1}\right)^{1-s} \operatorname{tr}(\mathbf{K}^s \mathbf{H}). \tag{B.8}$$

For a fixed s, the value of a that maximizes $f(\mathbf{K}, a, s)$ can be found analytically. When $\operatorname{tr}(\mathbf{KJ}) = 0$, letting $a = 0$ maximizes the quantity $f(\mathbf{K}, a, s)$. Otherwise, taking the derivative with respect to a, yields

$$\frac{\partial f(\mathbf{K}, s, a)}{\partial a} = \frac{1-s}{n^{s+1}} a^{-s} \operatorname{tr}(\mathbf{K}^s \mathbf{J}) - \frac{1-s}{n-1} \left(\frac{(1-a)n}{n-1}\right)^{-s} \operatorname{tr}(\mathbf{K}^s \mathbf{H}).$$

Equating this quantity to zero, rearranging, and taking the sth root of both sides, gives the maximizing value defined as

$$a_s^* \equiv \left(1 + \sqrt[s]{\frac{\operatorname{tr}(\mathbf{K}^s \mathbf{H})}{\operatorname{tr}(\mathbf{K}^s \mathbf{J})} \frac{n}{(n-1)^{1-s}}}\right)^{-1}, \tag{B.9}$$

for $\operatorname{tr}(\mathbf{K}^s \mathbf{J}) > 0$. Now defining

$$q(\mathbf{K}) \equiv \min_{0 \le s \le 1} f(\mathbf{K}, s, a_s^*) = \min_{0 \le s \le 1} \max_{0 \le a \le 1} \langle \mathbf{K}^s, \mathbf{N}_a^{1-s} \rangle, \tag{B.10}$$

the measure of informativeness using the Chernoff bound is

$$d_{\mathcal{N}}(\mathbf{K}) = 1 - q(\mathbf{K}). \tag{B.11}$$

By precomputing the eigenvalue decomposition of $\mathbf{K} = \mathbf{U}\Lambda\mathbf{U}^\top$, each evaluation of $f(\mathbf{K}, s, a_s^*)$ requires computations on the order of $\mathcal{O}(n)$ (Mendonça et al., 2008). Specifically, $\operatorname{tr}(\mathbf{K}^s \mathbf{J}) = \operatorname{tr}(\Lambda^s \mathbf{U}^\top \mathbf{1}\mathbf{1}^\top \mathbf{U}) = \sum_i \lambda_i^s \langle \mathbf{u}_i, \mathbf{1} \rangle^2$, and $\operatorname{tr}(\mathbf{K}^s \mathbf{H}) = \sum_i \lambda_i^s (1 - \frac{1}{n} \langle \mathbf{u}_i, \mathbf{1} \rangle^2)$. By consequence, the complexity of computing $q(\mathbf{K})$ via a line search on s, is the same as computing the Chernoff bound.

B.6 Quantum Hellinger (QH)

The quantum version of Hellinger distance shares some of the same properties as the Bures distance (Luo and Zhang, 2004). Squaring the quantum Hellinger distance (a chordal distance on the space of trace-normalized positive semidefinite matrices) and dividing by two yields a measure of informativeness for correlation matrices $\mathbf{K} \in \mathcal{E}$ as

$$\frac{1}{2} d_{\mathcal{N}}^2(\mathbf{K}) = \min_{0 \le a \le 1} 1 - \frac{1}{n} \langle \sqrt{\mathbf{K}}, \sqrt{\mathbf{N}_a} \rangle.$$

Firstly, we expand the affinity term as

$$\frac{1}{\sqrt{n}} \langle \sqrt{\mathbf{K}}, \sqrt{\mathbf{N}_a} \rangle = \langle \sqrt{\mathbf{K}}, \tfrac{\sqrt{a}}{n} \mathbf{J} \rangle + \langle \sqrt{\mathbf{K}}, \sqrt{\tfrac{1-a}{n-1}} \mathbf{H} \rangle = \frac{\sqrt{a}}{n} \mathbf{1}^\top \sqrt{\mathbf{K}} \mathbf{1} + \frac{\sqrt{1-a}}{\sqrt{n-1}} \operatorname{tr}(\mathbf{H}\sqrt{\mathbf{K}}).$$

Maximizing the latter, corresponds to an optimization along the positive quadrant of the circle parametrically defined by coordinates $\sqrt{a}$ and $\sqrt{1-a}$. The maximizing value is

$$a^\star = \operatorname*{arg\,max}_{0 \le a \le 1} \langle \sqrt{\mathbf{K}}, \sqrt{\mathbf{N}_a} \rangle = \frac{\left(\frac{1}{n} \mathbf{1}^\top \sqrt{\mathbf{K}} \mathbf{1}\right)^2}{\left(\frac{1}{n} \mathbf{1}^\top \sqrt{\mathbf{K}} \mathbf{1}\right)^2 + \frac{1}{n-1} \operatorname{tr}^2(\mathbf{H}\sqrt{\mathbf{K}})}. \tag{B.12}$$

Substituting $a^\star$ back into the expression of affinity leads to

$$\max_{0\le a\le 1}\langle\sqrt{\mathbf{K}},\sqrt{\mathbf{N}_a}\rangle = \langle\sqrt{\mathbf{K}},\sqrt{\mathbf{N}_{a^\star}}\rangle = \sqrt{\left(\tfrac{1}{n}\mathbf{1}^\top\sqrt{\mathbf{K}}\mathbf{1}\right)^2 + \tfrac{1}{n-1}\mathrm{tr}^2(\mathbf{H}\sqrt{\mathbf{K}})},$$

and the measure of informativeness is

$$\frac{1}{2}d^2_{\mathcal{N}}(\mathbf{K}) = 1 - \sqrt{\tfrac{1}{n^3}(\mathbf{1}^\top\sqrt{\mathbf{K}}\mathbf{1})^2 + \tfrac{1}{n(n-1)}\mathrm{tr}^2(\mathbf{H}\sqrt{\mathbf{K}})}. \tag{B.13}$$

B.7 Bures

By employing the Bures distance d_B for trace-normalized matrices in Equation 3, we can define a measure of informativeness of an arbitrary correlation matrix $\mathbf{K}\in\mathcal{E}$ as

$$\frac{1}{2}d^2_{\mathcal{N}}(\mathbf{K}) = \frac{1}{2}\min_{\mathbf{N}\in\mathcal{N}} d^2_B\left(\tfrac{1}{n}\mathbf{K},\tfrac{1}{n}\mathbf{N}\right) = 1 - 1\max_{0\le a\le 1}\tfrac{1}{n}\left\|\sqrt{\mathbf{K}}\sqrt{\mathbf{N}_a}\right\|_* . \tag{B.14}$$

Since $0 \le \frac{1}{n}\left\|\sqrt{\mathbf{K}}\sqrt{\mathbf{N}_a}\right\|_* \le 1$ (discussed in Appendix A), this distance is bounded as $0 \le d^2_{\mathcal{N}}(\mathbf{K}) \le 2$, with $d_{\mathcal{N}}(\mathbf{K}) = 0$ when $\mathbf{K}\in\mathcal{N}$.

We now define for convenience the function

$$f(\mathbf{K},a) \equiv \frac{1}{\sqrt{n}}\left\|\sqrt{\mathbf{K}}\sqrt{\mathbf{N}_a}\right\|_* = \left\|\tfrac{\sqrt{a}}{n}\sqrt{\mathbf{K}}\mathbf{J} + \sqrt{\tfrac{1-a}{n-1}}\sqrt{\mathbf{K}}\mathbf{H}\right\|_* ,$$

where the right hand side is obtained by using the equation $\sqrt{\mathbf{N}_a} = \sqrt{\frac{a}{n}}\mathbf{J} + \sqrt{\frac{(1-a)n}{n-1}}\mathbf{H}$ introduced in Section 3.1. We note that $\frac{1}{n}f^2(\mathbf{K},a) = C^2(\frac{1}{n}\mathbf{K},\frac{1}{n}\mathbf{N}_a)$, where $C^2(\cdot,\cdot)$ is the concave fidelity measure (Jozsa, 1994) discussed in Appendix A. The function simplifies for $a=0$ and $a=1$, yielding the following two quantities that will be useful for subsequent calculations

$$\begin{aligned} f^2(\mathbf{K},1) &= \frac{1}{n^2}\left\|\sqrt{\mathbf{K}}\mathbf{J}\right\|_*^2 = \frac{1}{n^2}\mathrm{tr}^2\sqrt{\mathbf{JKJ}} = \frac{1}{n}\mathbf{1}^\top\mathbf{K}\mathbf{1}\ , \\ f^2(\mathbf{K},0) &= \frac{1}{n-1}\left\|\sqrt{\mathbf{K}}\mathbf{H}\right\|_*^2 = \frac{1}{n-1}\mathrm{tr}^2\sqrt{\mathbf{HKH}}\ . \end{aligned}$$

For general correlation matrices $\mathbf{K}$, exact computation of the informativeness $d_{\mathcal{N}}(\mathbf{K})$ requires finding the value of a that maximizes $f(\mathbf{K},a)$, or equivalently $f^2(\mathbf{K},a)$. Since the latter expression is concave, this concave maximization is tractable and can be done using line search. However, each evaluation of $f(\mathbf{K},a)$ requires the calculation of a trace norm, and therefore the singular values of an $n\times n$ matrix; for large matrices, this would be a costly search. Consequently, as a surrogate, we seek a lower bound on $d_{\mathcal{N}}(\mathbf{K})$, or equivalently an upper bound on $\max_a\|\sqrt{\mathbf{K}}\sqrt{\mathbf{N}_a}\|_*$. One such useful bound can be found by simply employing the sub-additive property $\|\mathbf{A}+\mathbf{B}\|_* \le \|\mathbf{A}\|_* + \|\mathbf{B}\|_*$ for matrices $\mathbf{A}$ and $\mathbf{B}$ of equal size. Thus, for any $a\in[0,1]$, we have

$$\begin{aligned} f(\mathbf{K},a) &\le \frac{\sqrt{a}}{n}\left\|\sqrt{\mathbf{K}}\mathbf{J}\right\|_* + \sqrt{\frac{1-a}{n-1}}\left\|\sqrt{\mathbf{K}}\mathbf{H}\right\|_* = \sqrt{\frac{a}{n}\mathbf{1}^\top\mathbf{K}\mathbf{1}} + \sqrt{\frac{1-a}{n-1}}\left\|\sqrt{\mathbf{K}}\mathbf{H}\right\|_* \\ &= \sqrt{a}\,f(\mathbf{K},1) + \sqrt{1-a}\,f(\mathbf{K},0)\ \equiv\ \overline{f}(\mathbf{K},a). \end{aligned} \tag{B.15}$$

Additionally, we have $\max_a f(\mathbf{K}, a) \leq \max_a \overline{f}(\mathbf{K}, a)$. As the maximization of $\overline{f}(\mathbf{K}, a)$ corresponds to a linear program constrained to a circle parameterized by coordinates $\sqrt{a}$ and $\sqrt{1-a}$, the unique optimizing value of a is given as

$$a^* = \underset{0 \leq a \leq 1}{\arg\max}\, \overline{f}(\mathbf{K}, a) = \frac{f^2(\mathbf{K}, 1)}{f^2(\mathbf{K}, 1) + f^2(\mathbf{K}, 0)} = \frac{\mathbf{1}^\top \mathbf{K} \mathbf{1}}{\mathbf{1}^\top \mathbf{K} \mathbf{1} + \frac{n}{n-1} \mathrm{tr}^2 \sqrt{\mathbf{HKH}}}. \tag{B.16}$$

Substituting a^* into Equation B.15 yields $\overline{f}(\mathbf{K}, a^*) = \sqrt{f^2(\mathbf{K}, 1) + f^2(\mathbf{K}, 0)}$. This finally provides a lower bound for $d^2_{\mathcal{N}}(\mathbf{K})$ as

$$\begin{aligned} \frac{1}{2} d^2_{\mathcal{N}}(\mathbf{K}) &\geq 1 - \frac{1}{\sqrt{n}} \sqrt{f^2(\mathbf{K}, 1) + f^2(\mathbf{K}, 0)} \\ &= 1 - \sqrt{\tfrac{1}{n^2} \mathbf{1}^\top \mathbf{K} \mathbf{1} + \tfrac{1}{n(n-1)} \mathrm{tr}^2 \sqrt{\mathbf{HKH}}} \;\equiv\; i(\mathbf{K}). \end{aligned} \tag{B.17}$$

We refer to $i(\mathbf{K})$ as the *Bures-based informativeness*, since $\sqrt{2i(\mathbf{K})}$ is a lower bound on the Bures distance between $\mathbf{K}$ and the closest member in the set of non-informative correlation matrices $\mathcal{N}$.

Theorem 4 *The Bures-based informativeness $i(\mathbf{K})$ is convex. For $\mathbf{K}_1, \mathbf{K}_0 \in \mathcal{E}$ and $0 \leq \alpha \leq 1$, $i(\alpha \mathbf{K}_1 + (1-\alpha)\mathbf{K}_0) \leq \alpha i(\mathbf{K}_1) + (1-\alpha) i(\mathbf{K}_0)$.*

Proof We write the Bures-based informativeness as the composition $i(\mathbf{K}) = g(f(\mathbf{K}))$ where $f(\mathbf{K}) = \frac{1}{n^2} \mathbf{1}^\top \mathbf{K} \mathbf{1} + \frac{1}{n(n-1)} \mathrm{tr}^2 \sqrt{\mathbf{HKH}}$ and $g(u) = 1 - \sqrt{u}, \quad 0 \leq u$. The function $f(\mathbf{K}) = \langle \mathbf{K}, \frac{1}{n^2} \mathbf{J} \rangle + \frac{1}{n(n-1)} \| \sqrt{\mathbf{K}} \sqrt{\mathbf{H}} \|_*^2$ is concave as it is the sum of a linear function and the concave function $\frac{1}{n(n-1)} \| \sqrt{\mathbf{K}} \sqrt{\mathbf{H}} \|_*^2$, which is itself proportional to the concave fidelity function (Uhlmann, 1976; Jozsa, 1994) (see Theorem 3 in Appendix A). Furthermore, the function g is convex and nonincreasing. From these properties, it follows that

$$\begin{aligned} i(\alpha \mathbf{K}_1 + (1-\alpha)\mathbf{K}_0) = g(f(\alpha \mathbf{K}_1 + (1-\alpha)\mathbf{K}_0)) &\leq g(\alpha f(\mathbf{K}_1) + (1-\alpha) f(\mathbf{K}_0)) \\ &\leq \alpha g(f(\mathbf{K}_1)) + (1-\alpha) g(f(\mathbf{K}_0)) \\ &= \alpha i(\mathbf{K}_1) + (1-\alpha) i(\mathbf{K}_0), \end{aligned}$$

where the first inequality follows from the concavity of f and the monotonicity of g, and the second inequality follows from the convexity of g. ■

Special case 1: If $\mathbf{K}$ is centered, that is $\mathbf{K}\mathbf{1} = \mathbf{0}$, we have $\sqrt{\mathbf{K}}\mathbf{1} = \mathbf{0}$ and $f(\mathbf{K}, a) = \sqrt{\frac{1-a}{n-1}} \mathrm{tr} \sqrt{\mathbf{K}}$. Therefore, in this case we obtain a closed-form for Equation B.14, given as

$$\frac{1}{2} d^2_{\mathcal{N}}(\mathbf{K}) = 1 - \frac{1}{\sqrt{n(n-1)}} \mathrm{tr} \sqrt{\mathbf{K}}. \tag{B.18}$$

Thus, Equation B.17 yields an equality with $\frac{1}{2} d^2_{\mathcal{N}}(\mathbf{K}) = i(\mathbf{K})$. Centered correlation matrices have zero-mean embeddings, that is, $\mathbf{Z}\mathbf{1} = \mathbf{0}$, for any $\mathbf{Z}$ such that $\mathbf{Z}^\top \mathbf{Z} = \mathbf{K}$.

Special case 2: If $\mathbf{K}$ is a rank-1, then $\left\|\sqrt{\mathbf{K}}\sqrt{\mathbf{N}_a}\right\|_* = \left\|\sqrt{\mathbf{K}}\sqrt{\mathbf{N}_a}\right\|_F = \sqrt{\operatorname{tr}(\mathbf{N}_a\mathbf{K})}$. In this case, we have

$$f^2(\mathbf{K},a) = \frac{a}{n}\mathbf{1}^\top\mathbf{K}\mathbf{1} + \frac{1-a}{n-1}\operatorname{tr}(\mathbf{H}\mathbf{K}) = \frac{an^2(\bar{k}-\frac{1}{n})}{n-1} + \frac{n(1-\bar{k})}{n-1},$$

where $\bar{k} = \frac{1}{n^2}\mathbf{1}^\top\mathbf{K}\mathbf{1}$ is the average of all the entries of $\mathbf{K}$, and $\operatorname{tr}(\mathbf{H}\mathbf{K}) = n(1-\bar{k})$. This shows that $f^2(\mathbf{K},a)$ is linear with respect to a, and will be maximized by the extreme values 1 or 0 of a (depending on whether or not, $\bar{k}$ is greater than $\frac{1}{n}$). Substituting the expression of $f(\mathbf{K},a)$ into Equation B.14 and simplifying yields

$$\frac{1}{2}d^2_{\mathcal{N}}(\mathbf{K}) = 1 - \sqrt{\max\left(\bar{k}, \frac{1-\bar{k}}{n-1}\right)}. \tag{B.19}$$

Similarly, when $\mathbf{K}$ is rank-1, then $\operatorname{tr}^2\sqrt{\mathbf{HKH}} = \operatorname{tr}(\mathbf{HKH}) = n(1-\bar{k})$, and Equation B.17 becomes

$$i(\mathbf{K}) = 1 - \sqrt{\bar{k} + \frac{1-\bar{k}}{n-1}} = 1 - \sqrt{\frac{\bar{k}(n-2)+1}{n-1}}. \tag{B.20}$$

Rank-1 correlation matrices $\mathbf{K} = \mathbf{v}\mathbf{v}^\top$ with $\mathbf{v} \in \{\pm 1\}^n$ correspond to vertices of the elliptope (discussed in Section 2.1). Using the cut vector $\mathbf{v}$, it is easy to see that $\mathbf{1}^\top\mathbf{K}\mathbf{1} = (\mathbf{v}^\top\mathbf{1})^2 = (2c-n)^2 = 4(c-\frac{n}{2})^2$, where c is the number of positive entries of $\mathbf{v}$. Thus, Equation B.19 can be directly expressed in terms of c as

$$\frac{1}{2}d^2_{\mathcal{N}}(\mathbf{K}) = \begin{cases} 1 - \frac{2}{n}\sqrt{\frac{c(n-c)}{n-1}}, & \text{if } c \in \left[\frac{n-\sqrt{n}}{2}, \frac{n+\sqrt{n}}{2}\right], \\ 1 - \frac{2}{n}\left|c - \frac{n}{2}\right|, & \text{otherwise,} \end{cases}$$

and likewise Equation B.20 can be expressed as

$$i(\mathbf{K}) = 1 - \frac{2}{n}\sqrt{\left(c - \frac{n}{2}\right)^2 \frac{n-2}{n-1} + \frac{n^2}{4(n-1)}}.$$

Figure 15 compares $\sqrt{2i(\mathbf{K})}$ and $d_{\mathcal{N}}(\mathbf{K})$ for rank-1 matrices. For this case, $i(\mathbf{K})$ provides a tight lower bound except near the balanced cut.

Finally, we note the following result for the Bures-based informativeness.

Theorem 5 *For $n > 2$ and even, the Bures-based informativeness is maximized by rank-1 correlation matrices that are centered. That is, if $\mathbf{K}^\star \in \mathbb{R}^{n\times n}$ is a rank-1 correlation matrix with $\bar{k}^\star = \frac{1}{n^2}\mathbf{1}^\top\mathbf{K}^\star\mathbf{1} = 0$, then $i(\mathbf{K}^\star) = \max_{\mathbf{K}\in\mathcal{E}} i(\mathbf{K})$.*

Proof From Equation B.20, it follows that $i(\mathbf{K}^\star) = 1 - \sqrt{\frac{1}{n-1}}$. Maximizing $i(\mathbf{K})$ over $\mathbf{K} \in \mathcal{E}$ corresponds to minimizing the expression $\bar{k} + \frac{1}{n(n-1)}\|\mathbf{H}\sqrt{\mathbf{K}}\|_*^2$. Using the matrix norm inequality $\|\mathbf{A}\|_* \geq \|\mathbf{A}\|_F$, it follows that

$$\frac{1}{n(n-1)}\left\|\mathbf{H}\sqrt{\mathbf{K}}\right\|_*^2 \geq \frac{1}{n(n-1)}\left\|\mathbf{H}\sqrt{\mathbf{K}}\right\|_F^2 = \frac{1}{n(n-1)}\operatorname{tr}\left(\mathbf{H}\mathbf{K}\right) = \frac{1-\bar{k}}{n-1}.$$

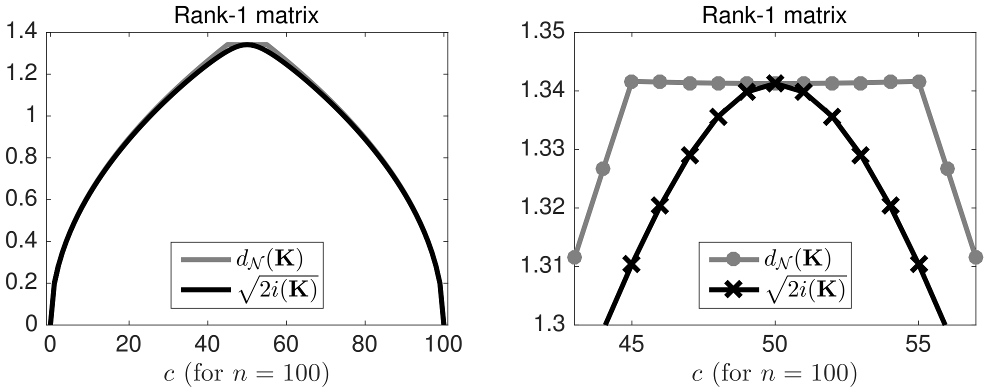

Figure 15: Comparison of the distance to the non-informative matrix family $d_{\mathcal{N}}(\mathbf{K})$ using the exact Bures distance d_B versus the lower bound $\sqrt{2i(\mathbf{K})}$ for rank-1 matrices $\mathbf{K} = \mathbf{v}\mathbf{v}^\top$, where c is the number of positive elements in the cut vector $\mathbf{v}$. The plot on the right magnifies the area around a balanced cut.

Substituting this expression back into the informativeness measure and maximizing over the possible values of $\bar{k}$, it follows that

$$i(\mathbf{K}) \leq 1 - \sqrt{\bar{k} + \frac{1-\bar{k}}{n-1}} \leq 1 - \sqrt{\frac{1}{n-1}} = i(\mathbf{K}^\star).$$

■

B.8 Sub-Bures Dissimilarity via Super-Fidelity

The modified fidelity or super-fidelity provides an upper bound of fidelity used in the Bures distance, and its computation has a lower complexity (Mendonça et al., 2008; Miszczak et al., 2009). Using super-fidelity, a lower bound on the Bures distance can be formed, which we refer to as the sub-Bures dissimilarity. (We use the term dissimilarity since it is not a metric, although a distance metric version can be obtained by removing the square root in the corresponding expression.) Using the sub-Bures dissimilarity, the informativeness is

$$\frac{1}{2} d_{\mathcal{N}}^2(\mathbf{K}) = 1 - \max_{0 \leq a \leq 1} \sqrt{G(\tfrac{1}{n}\mathbf{K}, \tfrac{1}{n}\mathbf{N}_a)}.$$

The super-fidelity between $\mathbf{K}$ and a non-informative matrix $\mathbf{N}_a$ is given by

$$\begin{aligned} G(\tfrac{1}{n}\mathbf{K}, \tfrac{1}{n}\mathbf{N}_a) &= \frac{1}{n^2}\langle \mathbf{K}, \mathbf{N}_a \rangle + \sqrt{\left(1 - \left\|\tfrac{1}{n}\mathbf{K}\right\|_F^2\right)\left(1 - \left\|\tfrac{1}{n}\mathbf{N}_a\right\|_F^2\right)} \\ &= \frac{a(n\bar{k}-1)+1-\bar{k}}{n-1} + \sqrt{\tfrac{n(1-a^2)+2(1-a)}{n-1}\left(1 - \left\|\tfrac{1}{n}\mathbf{K}\right\|_F^2\right)}, \end{aligned} \quad \text{(B.21)}$$

where $\bar{k}$ is as defined before, and $\left\|\frac{1}{n}\mathbf{N}_a\right\|_F^2 = a^2 + \frac{(1-a)^2}{n-1}$ was used to simplify the right hand side.

If $\mathbf{K}$ is rank-1, then $\|\frac{1}{n}\mathbf{K}\|_F = 1$ and the super-fidelity is proportional to $\langle \mathbf{K}, \mathbf{N}_a \rangle$. The inner product is maximized for $a = 1$ if $\bar{k} > \frac{1}{n}$, and $a = 0$ otherwise. In these two cases, super-fidelity obtains the values of $\bar{k}$ and $\frac{1-\bar{k}}{n-1}$, respectively. For more general cases, the maximizing value of a can be found using the derivative of super-fidelity given by

$$\frac{\partial G(\frac{1}{n}\mathbf{K}, \frac{1}{n}\mathbf{N}_a)}{\partial a} = \frac{\bar{k}n - 1}{n-1} + \frac{1 - an}{\sqrt{n(1-a^2) + 2(1-a)}}\sqrt{\frac{1 - \|\frac{1}{n}\mathbf{K}\|_F^2}{n-1}},$$

which vanishes when

$$\frac{\bar{k}n - 1}{n-1} = \frac{an - 1}{\sqrt{n(1-a^2) + 2(1-a)}}\sqrt{\frac{1 - \|\frac{1}{n}\mathbf{K}\|_F^2}{n-1}}.$$

Squaring both sides yields a quadratic expression whose roots are given by

$$a^{\pm} = \frac{b + cn \pm \sqrt{(cn+b)^2 - (cn^2 + bn)(c + b(2-n))}}{cn^2 + bn},$$

where $b = \frac{(\bar{k}n-1)^2}{(n-1)^2}$ and $c = \frac{1 - \left\|\frac{1}{n}\mathbf{K}\right\|_F^2}{n-1}$.

The super-fidelity is maximized when a is at one of its limits or by the root that satisfies $\operatorname{sign}(\bar{k} - \frac{1}{n}) = \operatorname{sign}(a - \frac{1}{n})$. Now, if we define $a_1 \equiv \min(1, a^+)$ and $a_0 \equiv \max(0, a^-)$, we have

$$a^\star = \underset{0 \le a \le 1}{\arg\max}\, G(\tfrac{1}{n}\mathbf{K}, \tfrac{1}{n}\mathbf{N}_a) = \begin{cases} a_1 & \text{if } G(\frac{1}{n}\mathbf{K}, \frac{1}{n}\mathbf{N}_{a_1}) \ge G(\frac{1}{n}\mathbf{K}, \frac{1}{n}\mathbf{N}_{a_0}) \\ a_0 & \text{otherwise.} \end{cases} \tag{B.22}$$

The maximal super-fidelity is

$$\max_{0 \le a \le 1} G(\tfrac{1}{n}\mathbf{K}, \tfrac{1}{n}\mathbf{N}_a) = \max\left(G(\tfrac{1}{n}\mathbf{K}, \tfrac{1}{n}\mathbf{N}_{a_1}),\ G(\tfrac{1}{n}\mathbf{K}, \tfrac{1}{n}\mathbf{N}_{a_0}) \right) \equiv g(\mathbf{K}), \tag{B.23}$$

and, hence, the sub-Bures measure of informativeness is

$$\frac{1}{2} d_{\mathcal{N}}^2(\mathbf{K}) = 1 - \sqrt{g(\mathbf{K})}. \tag{B.24}$$

Although this closed-form expression seems algebraically complex, it only depends on $\bar{k}$ and the norm of $\mathbf{K}$.

Appendix C. Statistics for Equality of Correlation Coefficients

In this appendix, we present Bartlett's and Lawley's test statistics for equality of correlation coefficients (Lawley, 1963) using notation consistent with the proposed measures of informativeness.

C.1 Lawley's Test Statistic

Lawley proposed a test statistic for equality of correlation coefficients (Lawley, 1963), which is a function of the Frobenius norm, element-wise mean, and variance of the row sums of a correlation matrix $\mathbf{K}$. The test assumes the correlation matrix is estimated from m observations of n-dimensional vectors. Under the null hypothesis the true correlation matrix is $\mathbf{N}_{a^*}$ for some value $0 \leq a^* \leq 1$. Letting $\lambda = \frac{n(1-a^*)}{n-1}$ and $\bar{k} = \frac{1}{n^2}\mathbf{1}^\top\mathbf{K}\mathbf{1}$, the test statistic can be written as

$$\begin{aligned} T_1(\mathbf{K}) &= \frac{m}{\lambda^2}\left(\frac{1}{2}\left\|\mathbf{K}-\mathbf{N}_{\bar{k}}\right\|_F^2 - \frac{1-\lambda^2}{n-\lambda^2(n-2)}\left\|\mathbf{K}\mathbf{1}-\mathbf{N}_{\bar{k}}\mathbf{1}\right\|_2^2\right) \\ &= \frac{m}{\lambda^2}\left(\frac{1}{2}\left(\|\mathbf{K}\|_F^2 - \frac{n^2}{n-1}(\bar{k}^2 n - 2\bar{k}+1)\right) - \frac{1-\lambda^2}{n-\lambda^2(n-2)}\left\|\mathbf{K}\mathbf{1}-n\bar{k}\mathbf{1}\right\|_2^2\right), \end{aligned} \quad \text{(C.1)}$$

where the simplification comes from $\|\mathbf{K}-\mathbf{N}_{\bar{k}}\|_F^2 = \|\mathbf{K}\|_F^2 - (n\bar{k})^2 - (1-\bar{k})^2\frac{n^2}{n-1} = \|\mathbf{K}\|_F^2 - \frac{n^2}{n-1}(n\bar{k}^2 - 2\bar{k}+1)$.

Asymptotically (as $m \to \infty$), the test statistic under the null hypothesis has a χ^2 distribution with $\frac{1}{2}(n+1)(n-2)$ degrees of freedom. In practice, a^* is unknown, but $\bar{k}$ can be used with $\lambda = \frac{n(1-\bar{k})}{n-1}$. This substitution does not affect the χ^2 limiting distribution (Lawley, 1963). Given only $\mathbf{K}$, m is unknown, and while it must be equal or greater than the rank of $\mathbf{K}$, we use the quantity $\frac{1}{m}T_1(\mathbf{K})$ for comparisons.

C.2 Bartlett's Asymptotic Test Statistic

Similarly, Bartlett's asymptotic test statistic (Lawley, 1963) is also a function of the Frobenius norm, element-wise mean, and variance of the rows sum of the given correlation matrix. Letting $\bar{k}$ and λ be defined as before, this test statistic is

$$\begin{aligned} T_2(\mathbf{K}) &= \frac{m}{\lambda^2}\left(\frac{1}{2}\left\|\mathbf{K}-\mathbf{N}_{\bar{k}}\right\|_F^2 - \frac{1}{n}\left\|\mathbf{K}\mathbf{1}-\mathbf{N}_{\bar{k}}\mathbf{1}\right\|_2^2\right) \\ &= \frac{m}{\lambda^2}\left(\frac{1}{2}\left(\|\mathbf{K}\|_F^2 - \frac{n^2}{n-1}(\bar{k}^2 n - 2\bar{k}+1)\right) - \frac{1}{n}\left\|\mathbf{K}\mathbf{1}-\bar{k}n\mathbf{1}\right\|_2^2\right). \end{aligned} \quad \text{(C.2)}$$

For $n = 3$ the asymptotic distribution of the test statistic under the null hypothesis is a scaled χ^2 distribution with two degrees of freedom (Anderson, 1963). For $n > 3$ the test statistic under the null hypothesis has a distribution described by the conic combination of two χ^2 distributions; however, this combination depends on the unknown value of a^* in terms of λ (Lawley, 1963). Again, we use the quantity $\frac{1}{m}T_2(\mathbf{K})$ since we assume m is unknown.

Appendix D. Matrix Norm Properties of $\sqrt{h(\cdot)}$

Theorem 6 *The cost function* $\sqrt{h(\mathbf{X})} = \sqrt{\frac{1}{n}\|\mathbf{X}\frac{1}{n}\mathbf{J}\|_*^2 + \frac{1}{n(n-1)}\|\mathbf{X}\mathbf{H}\|_*^2}$*, derived from the Bures-based informativeness measure, is a matrix norm for* $\mathbf{X} \in \mathbb{R}^{b\times n}$ *that satisfies the* *following properties:*

$$\sqrt{h(\mathbf{X})} \geq 0, \tag{D.1}$$

$$\sqrt{h(c\mathbf{X})} = |c| \sqrt{h(\mathbf{X})} \quad \forall c, \tag{D.2}$$

$$\sqrt{h(\mathbf{X}+\mathbf{Y})} \leq \sqrt{h(\mathbf{X})} + \sqrt{h(\mathbf{Y})}, \tag{D.3}$$

$$\sqrt{h(\mathbf{X})} = 0 \textit{ iff } \mathbf{X} = \mathbf{0}. \tag{D.4}$$

Furthermore, $\sqrt{h(\cdot)}$ *is convex.*

Proof The first two properties of non-negativity and scalability are readily apparent. We proceed to show that $\sqrt{h(\cdot)}$ satisfies the triangle inequality (D.3). For compactness, we set $\alpha_1 = \frac{1}{n}$, $\alpha_2 = \frac{1}{n(n-1)}$, and $\acute{\mathbf{J}} = \frac{1}{n}\mathbf{J}$. For all conformable matrices $\mathbf{X}$ and $\mathbf{Y}$, we have

$$\begin{aligned}
h(\mathbf{X}+\mathbf{Y}) &= \alpha_1 \|(\mathbf{X}+\mathbf{Y})\acute{\mathbf{J}}\|_*^2 + \alpha_2 \|(\mathbf{X}+\mathbf{Y})\mathbf{H}\|_*^2 \\
&\leq \alpha_1 (\|\mathbf{X}\acute{\mathbf{J}}\|_* + \|\mathbf{Y}\acute{\mathbf{J}}\|_*)^2 + \alpha_2 (\|\mathbf{X}\mathbf{H}\|_* + \|\mathbf{Y}\mathbf{H}\|_*)^2 \\
&= h(\mathbf{X}) + h(\mathbf{Y}) + 2\sqrt{\left(\alpha_1 \|\mathbf{X}\acute{\mathbf{J}}\|_* \|\mathbf{Y}\acute{\mathbf{J}}\|_* + \alpha_2 \|\mathbf{X}\mathbf{H}\|_* \|\mathbf{Y}\mathbf{H}\|_*\right)^2} \\
&= h(\mathbf{X}) + h(\mathbf{Y}) \\
&\quad + 2\sqrt{\alpha_1^2 \|\mathbf{X}\acute{\mathbf{J}}\|_*^2 \|\mathbf{Y}\acute{\mathbf{J}}\|_*^2 + \alpha_2^2 \|\mathbf{X}\mathbf{H}\|_*^2 \|\mathbf{Y}\mathbf{H}\|_*^2 + 2\alpha_1\alpha_2 \|\mathbf{X}\acute{\mathbf{J}}\|_* \|\mathbf{Y}\acute{\mathbf{J}}\|_* \|\mathbf{X}\mathbf{H}\|_* \|\mathbf{Y}\mathbf{H}\|_*} \\
&\leq h(\mathbf{X}) + h(\mathbf{Y}) \\
&\quad + 2\sqrt{\alpha_1^2 \|\mathbf{X}\acute{\mathbf{J}}\|_*^2 \|\mathbf{Y}\acute{\mathbf{J}}\|_*^2 + \alpha_2^2 \|\mathbf{X}\mathbf{H}\|_*^2 \|\mathbf{Y}\mathbf{H}\|_*^2 + \alpha_1\alpha_2 \left(\|\mathbf{X}\acute{\mathbf{J}}\|_*^2 \|\mathbf{Y}\mathbf{H}\|_*^2 + \|\mathbf{Y}\acute{\mathbf{J}}\|_*^2 \|\mathbf{X}\mathbf{H}\|_*^2\right)} \\
&= h(\mathbf{X}) + h(\mathbf{Y}) + 2\sqrt{h(\mathbf{X})h(\mathbf{Y})} = \left(\sqrt{h(\mathbf{X})} + \sqrt{h(\mathbf{Y})}\right)^2,
\end{aligned}$$

where the first inequality follows from the sub-additivity and non-negativity of the trace norm, and the second one from the arithmetic and geometric means inequality. Therefore, taking the square root yields $\sqrt{h(\mathbf{X}+\mathbf{Y})} \leq \sqrt{h(\mathbf{X})} + \sqrt{h(\mathbf{Y})}$.

To ensure that $\sqrt{h(\mathbf{X})}$ is a norm, rather than a seminorm, we now show that property (D.4) holds. Firstly, if $\mathbf{X} = \mathbf{0}$, then $h(\mathbf{X}) = \alpha_1 \|\mathbf{0}\acute{\mathbf{J}}\|_*^2 + \alpha_2 \|\mathbf{0}\mathbf{H}\|_*^2 = 0$. We prove the converse by contradiction for any $\alpha_1, \alpha_2 > 0$. Suppose that $h(\mathbf{X}) = 0$ and $\mathbf{X} \neq \mathbf{0}$. For the trace norm, $\|\mathbf{X}\acute{\mathbf{J}}\|_*^2 = 0$ iff $\mathbf{X}\acute{\mathbf{J}} = \mathbf{0}$, and $\|\mathbf{X}\mathbf{H}\|_*^2 = 0$ iff $\mathbf{X}\mathbf{H} = \mathbf{0}$. Since the trace-norm is non-negative and $\alpha_1, \alpha_2 > 0$, $h(\mathbf{X}) = 0$ implies that $\mathbf{X}\acute{\mathbf{J}} = \mathbf{0}$ and $\mathbf{X}\mathbf{H} = \mathbf{0}$. Summing these yields $\mathbf{X}\acute{\mathbf{J}} + \mathbf{X}\mathbf{H} = \mathbf{0}$, which, since $\acute{\mathbf{J}} + \mathbf{H} = \mathbf{I}$, yields $\mathbf{X} = \mathbf{0}$, but this contradicts the assumption that $\mathbf{X} \neq \mathbf{0}$.

The convexity of $\sqrt{h(\cdot)}$ follows from the norm properties. For $\mathbf{X}_1, \mathbf{X}_0$ and $0 \leq \alpha \leq 1$, $\sqrt{h(\alpha\mathbf{X}_1 + (1-\alpha)\mathbf{X}_0)} \leq \sqrt{h(\alpha\mathbf{X}_1)} + \sqrt{h((1-\alpha)\mathbf{X}_0)} = \alpha\sqrt{h(\mathbf{X}_1)} + (1-\alpha)\sqrt{h(\mathbf{X}_0)}$. ■

Appendix E. Second-Order Spectral Soft-Thresholding Operator

The trace norm is well known as the tightest convex surrogate for the rank of a matrix (Fazel et al., 2001; Recht et al., 2010) and its proximal operator, the spectral soft-thresholding operator, has been used in the context of matrix completion (Cai et al., 2010; Mazumder et al., 2010; Ma et al., 2011). The generalization of proximal operators on vectors to matrices holds for any vector functional that is absolutely symmetric (invariant to permutation and change of sign), or equivalently any matrix functional that is unitarily invariant, since in these cases Lewis (1995) proved that the subdifferential of the matrix functional can be written as a composition of the subdifferential of the vector functional and the singular value functional (Lewis's result generalizes the classical theorem by von Neumann (1937) relating symmetric gauge functions to unitary invariant norms). For example, the trace norm can be written as the ℓ_1-norm of the singular values of a matrix, and thus, the spectral soft-thresholding operator is the matrix analog of the soft-thresholding operator Donoho et al. (1995), which is the proximal operator of the ℓ_1-norm (Wright et al., 2009).

By the same reasoning, the squared trace-norm is the matrix analog to the squared ℓ_1-norm. The squared ℓ_1-norm has been used in the context of multiple kernel learning (Lanckriet et al., 2004; Bach, 2008; Kowalski and Torrésani, 2009; Martins et al., 2011), and its proximal operator uses the soft-thresholding operator (Donoho et al., 1995) with a data-dependent penalty (Bach, 2008; Kowalski and Torrésani, 2009; Martins et al., 2011). Since the squared ℓ_1-norm is absolutely symmetric, the subdifferential of the squared trace norm is found by composition of the subdifferential of the squared ℓ_1-norm and the singular value functional (Lewis, 1995). The proximal operator for the squared trace norm applies the proximal operator of the squared ℓ_1-norm to the singular values of the target matrix.

Let $\mathbf{X} \in \mathbb{R}^{m \times n}$ denote an arbitrary matrix (without loss of generality we assume $m \leq n$), and let $\mathbf{U\Sigma V}^\top$ denote the singular value decomposition of $\mathbf{X}$, such that $\mathbf{U}, \mathbf{V}$ are unitary and $\mathbf{\Sigma}$ is a diagonal matrix with its first m entries $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m \geq 0$ and any remaining entries being 0. The spectral soft-thresholding operator is defined as

$$\mathcal{S}_\tau(\mathbf{X}) = \arg\min_{\mathbf{Y}} \frac{1}{2} \|\mathbf{X} - \mathbf{Y}\|_F^2 + \tau \|\mathbf{Y}\|_* = \mathbf{U}[\mathbf{\Sigma} - \tau \mathbf{I}]_+ \mathbf{V}^\top,$$

where $[\cdot]_+$ denotes the element-wise non-negative thresholding operation that sets any negative values to zero. Using the proximal operator for the squared ℓ_1-norm (Kowalski and Torrésani, 2009; Martins et al., 2011), we define the second-order spectral soft-thresholding operator $\mathcal{Z}_\beta(\cdot)$ as

$$\begin{aligned} \mathcal{Z}_\beta(\mathbf{X}) &= \mathcal{S}_{\tau^\star}(\mathbf{X}), \\ \tau^\star &= \tfrac{\beta}{1+k^\star\beta} \|\mathbf{X}\|_{k^\star}, \\ k^\star &= \max\{k \in [m] \ : \ \sigma_k - \tfrac{\beta}{1+k\beta}\|\mathbf{X}\|_k > 0\}, \\ \|\mathbf{X}\|_k &= \sigma_1 + \cdots + \sigma_k \quad \text{(Ky-Fan } k\text{-norm).} \end{aligned} \tag{E.1}$$

Theorem 7 *$\mathcal{Z}_\beta(\cdot)$ is the proximal operator of the function $\frac{\beta}{2}\|\cdot\|_*^2$ such that*

$$\mathcal{Z}_\beta(\mathbf{X}) = \arg\min_{\mathbf{Y}} \ \tfrac{1}{2} \|\mathbf{X} - \mathbf{Y}\|_F^2 + \tfrac{\beta}{2} \|\mathbf{Y}\|_*^2.$$

Proof

Since $\frac{1}{2} \|\mathbf{X} - \mathbf{Y}\|_F^2 + \frac{\beta}{2}\|\mathbf{Y}\|_*^2$ is convex, to prove the optimality of the second-order spectral thresholding operator it is sufficient to show that

$$\mathbf{0} \in \partial \tfrac{1}{2} \|\mathbf{X} - \mathcal{Z}_\beta(\mathbf{X})\|_F^2 + \partial \tfrac{\beta}{2} \|\mathcal{Z}_\beta(\mathbf{X})\|_*^2,$$

where $\partial f(\mathbf{X})$ denotes the subdifferential of the function f at $\mathbf{X}$, defined as

$$\partial f(\mathbf{X}) = \{\mathbf{G} : \forall \mathbf{Y} \in \mathbb{R}^{m \times n}, f(\mathbf{Y}) \geq f(\mathbf{X}) + \langle \mathbf{Y} - \mathbf{X}, \mathbf{G} \rangle\}.$$

To do this we use the optimality of the spectral thresholding operator (Cai et al., 2010; Mazumder et al., 2010; Ma et al., 2011), which implies that for any $\tau \geq 0$

$$\mathbf{0} \in \partial \tfrac{1}{2} \left\| \mathbf{X} - \mathcal{S}_\tau(\mathbf{X}) \right\|_F^2 + \partial \tau \| \mathcal{S}_\tau(\mathbf{X}) \|_*,$$

where the subdifferential of the trace norm (Watson, 1992) is

$$\partial \|\mathbf{X}\|_* = \{\mathbf{U}_1 \mathbf{V}_1^\top + \mathbf{M} : \mathbf{U}_1^\top \mathbf{M} = \mathbf{0}, \mathbf{M}\mathbf{V}_1 = \mathbf{0}, \|\mathbf{M}\|_2 \leq 1\},$$

where the columns of $\mathbf{U}_1$ and $\mathbf{V}_1$ are the left and right singular vectors associated to non-zero singular values.

We note that any subgradient of squared trace-norm is a scaled version of the subgradient of the trace-norm with location-dependent scaling, i.e., $(2\|\mathbf{X}\|_*)\mathbf{G} \in \partial \|\mathbf{X}\|_*^2$ iff $\mathbf{G} \in \partial \|\mathbf{X}\|_*$.

Given $\mathcal{Z}_\beta(\mathbf{X}) = \mathcal{S}_{\tau^\star}(\mathbf{X})$, $\tau^\star = \frac{\beta}{1+k^\star\beta}\|\mathbf{X}\|_{k^\star}$, and $\|\mathcal{Z}_\beta(\mathbf{X})\|_* = \|\mathbf{X}\|_{k^\star} - k^\star\tau^\star = \frac{\tau^\star}{\beta}$, it follows that

$$\partial \tfrac{\beta}{2} \|\mathcal{Z}_\beta(\mathbf{X})\|_*^2 = \tfrac{\beta}{2}(2\|\mathcal{Z}_\beta(\mathbf{X})\|_*)\partial \|\mathcal{Z}_\beta(\mathbf{X})\|_* = \tau^\star \partial \|\mathcal{S}_{\tau^\star}(\mathbf{X})\|_*.$$

Thus, $\partial \tfrac{1}{2} \left\| \mathbf{X} - \mathcal{Z}_\beta(\mathbf{X}) \right\|_F^2 + \partial \tfrac{\beta}{2} \|\mathcal{Z}_\beta(\mathbf{X})\|_*^2 = \partial \tfrac{1}{2} \left\| \mathbf{X} - \mathcal{S}_{\tau^\star}(\mathbf{X}) \right\|_F^2 + \partial \tau^\star \|\mathcal{S}_{\tau^\star}(\mathbf{X})\|_*$. ■

References

P.-A. Absil and K. A. Gallivan. Joint diagonalization on the oblique manifold for independent component analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages V–945–V–948, 2006.

M. A. Aitkin, W. C. Nelson, and Karen H. Reinfurt. Tests for correlation matrices. *Biometrika*, 55(2):327–334, 1968.

T. W. Anderson. Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, 34(1):122–148, 1963.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

K. M. R. Audenaert, J. Calsamiglia, R. Munoz-Tapia, E. Bagan, Ll. Masanes, A. Acin, and F. Verstraete. Discriminating states: the quantum Chernoff bound. *Physical Review Letters*, 98(16):160501, 2007.

Francis R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

Richard H. Bartels and G. W. Stewart. Solution of the matrix equation $AX + XB = C$. *Communications of the ACM*, 15(9):820–826, 1972.

M. S. Bartlett. A note on the multiplying factors for various χ^2 approximations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 16(2):296–298, 1954.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Peter J. Bickel and Purnamrita Sarkar. Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):253–273, 2016.

Rüdiger Borsdorf, Nicholas J. Higham, and Marcos Raydan. Computing a nearest correlation matrix with factor structure. *SIAM Journal on Matrix Analysis and Applications*, 31(5):2603–2622, 2010.

Jonathan M. Borwein and Adrian S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. CMS Books in Mathematics. Springer New York, 2010.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

Samuel L. Braunstein and Carlton M. Caves. Statistical distance and the geometry of quantum states. *Physical Review Letters*, 72(22):3439–3443, 1994.

C. J. Brien, W. N. Venables, A. T. James, and O. Mayo. An analysis of correlation matrices: equal correlations. *Biometrika*, 71(3):545–554, 1984.

Thomas R. Bromley, Marco Cianciaruso, Rosario Lo Franco, and Gerardo Adesso. Unifying approach to the quantification of bipartite correlations by Bures distance. *Journal of Physics A: Mathematical and Theoretical*, 47(40):405302, 2014.

Arne Brøndsted. *An Introduction to Convex Polytopes*, volume 90 of *Graduate Texts in Mathematics*. Springer New York, 2012.

Samuel Burer and Renato D.C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

Donald Bures. An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w*-algebras. *Transactions of the American Mathematical Society*, 135:199–212, 1969.

Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

Zeineb Chebbi and Maher Moakher. Means of Hermitian positive-definite matrices based on the log-determinant α-divergence function. *Linear Algebra and its Applications*, 436 (7):1872–1889, 2012.

Anoop Cherian, Suvrit Sra, Arindam Banerjee, and Nikolaos Papanikolopoulos. Efficient similarity search for covariance matrices via the Jensen-Bregman logdet divergence. In *2011 International Conference on Computer Vision*, pages 2399–2406, 2011.

Fan R. K. Chung. *Spectral Graph Theory.* Number 92 in CBMS regional conference series in mathematics. American Mathematical Society, 1997.

Andrzej Cichocki, Sergio Cruces, and Shun-ichi Amari. Log-determinant divergences revisited: Alpha-beta and gamma log-det divergences. *Entropy*, 17(5):2988–3034, 2015.

Patrick L. Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In Heinz H. Bauschke, Regina S. Burachik, Patrick L. Combettes, Veit Elser, D. Russell Luke, and Henry Wolkowicz, editors, *Fixed-point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer, New York, NY, 2011.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13:795–828, 2012.

Nello Cristianini, John Shawe-Taylor, André Elisseeff, and Jaz Kandola. On kernel-target alignment. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.

David L. Donoho, Iain M. Johnstone, Gérard Kerkyacharian, and Dominique Picard. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(2):301–369, 1995.

Jonathan Eckstein and Dimitri P. Bertsekas. On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1):293–318, 1992.

P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 6: 290–297, 1959.

Maryam Fazel, Haitham Hindi, and Stephen P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the American Control Conference*, pages 4734–4739, 2001.

Christopher Fuchs and Jeroen van de Graaf. Cryptographic distinguishability measures for quantum-mechanical states. *IEEE Transactions on Information Theory*, 45(4):1216–1227, 1999.

Kenji Fukumizu, Francis R. Bach, and Michael I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.

Leon Jay Gleser. On testing a set of correlation coefficients for equality: Some asymptotic results. *Biometrika*, 55(3):513–517, 1968.

Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.

Daniel B. Graham and Nigel M. Allinson. Characterising virtual eigensignatures for general purpose face recognition. In Harry Wechsler, P. Jonathon Phillips, Vicki Bruce, Françoise Fogelman Soulié, and Thomas S. Huang, editors, *Face Recognition: From Theory to Applications*, pages 446–456. Springer, Berlin, Heidelberg, 1998.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita, editors, *Algorithmic Learning Theory*, volume 3734 of *Lecture Notes in Computer Science*, pages 63–77. Springer, Berlin, Heidelberg, 2005.

Igor Grubišić and Raoul Pietersz. Efficient rank reduction of correlation matrices. *Linear Algebra and its Applications*, 422(2):629–653, 2007.

Masahito Hayashi. *Quantum Information.* Springer, 2006.

Bingsheng He and Xiaoming Yuan. On non-ergodic convergence rate of Douglas–Rachford alternating direction method of multipliers. *Numerische Mathematik*, 130(3):567–577, 2015.

Nicholas J. Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*, 103:103–118, 1988.

Nicholas J. Higham. Computing the nearest correlation matrix—a problem from finance. *IMA Journal of Numerical Analysis*, 22(3):329–343, 2002.

Michael Hintermüller and Tao Wu. A smoothing descent method for nonconvex TV^q-models. In Andrés Bruhn, Thomas Pock, and Xue-Cheng Tai, editors, *Efficient Algorithms for Global Optimization Methods in Computer Vision*, volume 8293 of *Lecture Notes in Computer Science*, pages 119–133. Springer, Berlin, Heidelberg, 2014.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994.

M. Journée, F. Bach, P.-A. Absil, and R. Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20(5):2327–2351, 2010.

Richard Jozsa. Fidelity for mixed quantum states. *Journal of Modern Optics*, 41(12): 2315–2323, 1994.

Vladimir Koltchinskii and Dong Xia. Optimal estimation of low rank density matrices. *Journal of Machine Learning Research*, 16:1757–1792, 2015.

Risi Kondor and Tony Jebara. A kernel between sets of vectors. In *Proceedings of the 20th International Conference on Machine Learning*, pages 361–368, 2003.

Matthieu Kowalski and Bruno Torrésani. Structured sparsity: from mixed norms to structured shrinkage. In *SPARS'09 - Signal Processing with Adaptive Sparse Structured Representations*, 2009.

Tarald O. Kvålseth. Entropy and correlation: some comments. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(17):517–519, 1987.

Gert R. G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

Monique Laurent and Svatopluk Poljak. On a positive semidefinite relaxation of the cut polytope. *Linear Algebra and its Applications*, 223–224:439–461, 1995.

Monique Laurent and Svatopluk Poljak. On the facial structure of the set of correlation matrices. *SIAM Journal on Matrix Analysis and Applications*, 17(3):530–547, 1996.

D. N. Lawley. On testing a set of correlation coefficients for equality. *The Annals of Mathematical Statistics*, 34(1):149–151, 1963.

Hosoo Lee and Yongdo Lim. Invariant metrics, contractions and nonlinear matrix equations. *Nonlinearity*, 21(4):857–878, 2008.

A. S. Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1/2):173–183, 1995.

A. S. Lewis. Derivatives of spectral functions. *Mathematics of Operations Research*, 21(3): 576–588, 1996.

Chi-Kwong Li and Bit-Shun Tam. A note on extreme correlation matrices. *SIAM Journal on Matrix Analysis and Applications*, 15(3):903–908, 1994.

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, 2002.

David G. Lowe. Similarity metric learning for a variable-kernel classifier. *Neural Computation*, 7(1):72–85, 1995.

Shunlong Luo and Qiang Zhang. Informational distance on quantum-state space. *Physical Review A*, 69(3):032106, 2004.

Shiqian Ma, Donald Goldfarb, and Lifeng Chen. Fixed point and Bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 128(1-2):321–353, 2011.

A. P. Majtey, P. W. Lamberti, and D. P. Prato. Jensen-Shannon divergence as a measure of distinguishability between mixed quantum states. *Physical Review A*, 72(5):052310, 2005.

Jérôme Malick. A dual approach to semidefinite least-squares problems. *SIAM Journal on Matrix Analysis and Applications*, 26(1):272–284, 2004.

André F. T. Martins, Noah A. Smith, Eric P. Xing, Pedro M. Q. Aguiar, and Mário A. T. Figueiredo. Online learning of structured predictors with multiple kernels. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 507–515, 2011.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.

Marina Meilă. Comparing clusterings by the variation of information. In B. Schölkopf and M. K. Warmuth, editors, *Learning Theory and Kernel Machines*, pages 173–187. Springer, Berlin, Heidelberg, 2003.

Paulo E. M. F. Mendonça, Reginaldo d. J. Napolitano, Marcelo A. Marchiolli, Christopher J. Foster, and Yeong-Cherng Liang. Alternative fidelity measure between quantum states. *Physical Review A*, 78(5):052330, 2008.

Jarosław Adam Miszczak, Zbigniew Puchała, Paweł Horodecki, Armin Uhlmann, and Karol Życzkowski. Sub-and super-fidelity as bounds for quantum fidelity. *Quantum Information & Computation*, 9(1&2):103–130, 2009.

Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

Sameer A. Nene, Shree K. Nayar, and Hiroshi Murase. Columbia object image library (COIL-20). Technical report, CUCS-005-96, 1996.

Yu. Nesterov. Gradient methods for minimizing composite objective function. Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, 2007.

M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104, 2006.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 849–856, Cambridge, MA, 2002. MIT Press.

Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information.* Cambridge University Press, 2000.

Peder A. Olsen, Steven J. Rennie, and Vaibhava Goel. Efficient automatic differentiation of matrix functions. In Shaun Forth, Paul Hovland, Eric Phipps, Jean Utke, and Andrea Walther, editors, *Recent Advances in Algorithmic Differentiation*, volume 87 of *Lecture Notes in Computational Science and Engineering*, pages 71–81. Springer, Berlin, Heidelberg, 2012.

Neal Parikh and Stephen P. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

Il Memming Park, Sohan Seth, Antonio R. C. Paiva, Lin Li, and Jose C. Principe. Kernel methods on spike train space for neuroscience: a tutorial. *IEEE Signal Processing Magazine*, 30(4):149–160, 2013.

Houduo Qi and Defeng Sun. A quadratically convergent Newton method for computing the nearest correlation matrix. *SIAM Journal on Matrix Analysis and Applications*, 28(2): 360–385, 2006.

Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

Luis Gonzalo Sanchez Giraldo, Murali Rao, and Jose C. Principe. Measures of entropy from data using infinitely divisible kernels. *IEEE Transactions on Information Theory*, 61(1): 535–548, 2015.

J. Saunderson, V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1395–1416, 2012.

Mark Schmidt. minFunc, 2012. Software available at `http://www.di.ens.fr/~mschmidt/Software/minFunc.html`.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

James H. Steiger. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245–251, 1980.

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

Nickolay T. Trendafilov and Ross A. Lippert. The multimode procrustes problem. *Linear Algebra and its Applications*, 349(1):245–264, 2002.

A. Uhlmann. The transition probability in the state space of a*-algebra. *Reports on Mathematical Physics*, 9(2):273–279, 1976.

John von Neumann. Some matrix inequalities and metrization of matric-space. *Tomsk University Review*, 1:286–300, 1937.

John von Neumann. *Mathematical Foundations of Quantum Mechanics*. Princeton University Press, 1955.

G. A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33–45, 1992.

Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

Stephen J. Wright, Robert D. Nowak, and Mário A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. Distance metric learning with application to clustering with side-information. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 505–512, Cambridge, MA, 2003. MIT Press.