Claremont Colleges
Scholarship @ Claremont

Pomona Faculty Publications and Research

Pomona Faculty Scholarship

1-1-2013

A method for generating realistic correlation matrices

Johanna S. Hardin
Pomona College

Stephan Ramon Garcia
Pomona College

David Golan
Tel Aviv University

Recommended Citation

A METHOD FOR GENERATING REALISTIC CORRELATION MATRICES

BY JOHANNA HARDIN[1], STEPHAN RAMON GARCIA[2] AND DAVID GOLAN[1,3]

Pomona College, Pomona College and Tel Aviv University

Simulating sample correlation matrices is important in many areas of statistics. Approaches such as generating Gaussian data and finding their sample correlation matrix or generating random uniform $[-1, 1]$ deviates as pairwise correlations both have drawbacks. We develop an algorithm for adding noise, in a highly controlled manner, to general correlation matrices. In many instances, our method yields results which are superior to those obtained by simply simulating Gaussian data. Moreover, we demonstrate how our general algorithm can be tailored to a number of different correlation models. Using our results with a few different applications, we show that simulating correlation matrices can help assess statistical methodology.

1. Introduction. As computational resources continue to improve, researchers can take advantage of simulation studies to investigate properties and results associated with novel statistical methodology. In particular, simulating correlation matrices with or without a given structure can provide insight into the sensitivity of a model. There has been extensive work on simulating correlation matrices with random entries, that is, generating positive-semidefinite matrices with all entries bounded by $[-1, 1]$ and having ones along the diagonal. Seminal work by Marsaglia and Olkin (1984) discusses distributional characteristics and eigenvalues of simulated random correlation matrices. Although there has been additional work expanding the ideas associated with generating random correlation matrices [Joe (2006), Lewandowski, Kurowicka and Joe (2009), Holmes (1991), Davies and Higham (2000), Rae (1997)] and even randomly generating correlation matrices within particular settings [Ng and Joe (2010), Holmes (1989)], to our knowledge there is no literature devoted to the problem of adding noise to given template correlation structures.

We discuss the need to simulate realistic correlation matrices in a specific context. By *realistic* we mean not only that the correlation matrix has some prescribed structure (dependent upon the requirements of the particular application), but also

that it is *noisy*. Below, we discuss the importance of simulating correlation matrices in probit analysis, Bayesian shrinkage estimation, meta-analysis, multiple comparisons, management science, factor analysis, heritability estimation, network analysis and classification.

In order to ensure identifiability of model parameters, covariance matrices in probit analysis on longitudinal data are often constrained to be correlation matrices. Liu and Daniels (2006) and Zhang, Boscardin and Belin (2006) discuss advantages and disadvantages of different prior distributions used with a Metropolis Hastings algorithm to sample correlation matrices from a posterior distribution.

Barnard, McCulloch and Meng (2000) use hierarchical Bayesian models to describe the covariance between parameters in a regression model with shrinkage. In practice, joint uniform and inverse-Wishart priors are used to simulate correlation matrices.

One important task in meta-analysis is to combine correlation matrices from different studies. Different methods for combining such matrices are used to make inferences about marginal correlations. Hafdahl (2007) ran a Monte Carlo study generating sample correlation matrices using Gaussian deviates from a given population correlation matrix.

One of the big challenges in developing accurate multiple comparisons techniques is knowing the underlying correlation structure of the many items being compared. Simply knowing the rate of null versus alternative tests in a given simulation does not provide enough information for realistic application to studies with possibly strongly correlated data and hypotheses. In order to model false discovery rates (FDR) in settings more realistic than i.i.d. (independent and identically distributed), Kim and van de Wiel (2008) use Gaussian deviates to simulate nested correlation matrices with constrained correlation strengths.

In order to maximize expected performance, Nelson and Goldsman (2001) use a stochastic simulation to compare a variety of management systems (e.g., queues). The authors use a modification of the method of Marsaglia and Olkin (1984) to simulate correlation matrices. Instead of generating random points on the k-dimensional unit sphere, they constrain their search to the part of the unit sphere with all nonnegative coordinates (inducing nonnegative correlations).

In factor analysis, sample correlation matrices based on population correlation matrices are typically used in simulation studies. Methods which incorporate model error as well as sampling error create more realistic structures from which to model data. Hong (1999) recommends using the eigenstructure of the population correlation matrix along with random chi-square deviates to directly obtain a random sample covariance matrix, from which the sample correlation matrix can be computed.

Additionally, as we detail in Section 4, simulated correlation matrices are used to estimate heritability in Genome Wide Association Studies (GWASs) and to assess network and classification algorithms. We present work done by Lee et al.

(2011) on estimating heritability without considering unknown noise in the correlation structure. Alternatively, some clustering and classification methods simulate correlations (or covariances) using uniform distributions [Kraus and Kestler (2010), Tai and Pan (2007), Huang, Tong and Zhao (2010), Kraj et al. (2008)]. However, randomly simulated univariate correlations do not typically produce matrices that are positive semidefinite. We argue that neither the no-noise strategy nor the uniform-noise strategy is realistic for assessing methodology. Our work is applicable to any context where simulating realistic correlation matrices is important.

Suppose that we are given a $N \times N$ correlation matrix $\Sigma = (\Sigma_{ij})_{i,j=1}^{N}$. Generating a noisy correlation matrix $S = (S_{ij})_{i,j=1}^{N}$ based upon the template Σ can be difficult since noise must be added to Σ in such a way that S remains positive semidefinite and satisfies $S_{ii} = 1$ and $-1 \leq S_{ij} \leq 1$ for $1 \leq i, j \leq N$. Moreover, for numerical purposes (e.g., generating data from S) one might also require an explicit upper bound on the condition number $\kappa(S)$ of S (see Section A.1) to ensure its numerical stability (e.g., for matrix inversion). Unfortunately, naively adding random noise to a correlation matrix can result in matrices which violate any or all of the above constraints.

1.1. *Simulating data for evaluating algorithms.* In certain applications, it is important to have a known (or assumed) structure based on a covariance or correlation matrix. Many authors use a particular structure and simulate Gaussian data from that matrix. For example, in a recent paper, Tritchler, Parkhomenko and Beyene (2009) simulate Gaussian data to assess a method for filtering genes prior to, for example, network analysis. Their structure consists of within group correlations of 0.4 and between group correlations of 0. Using clustering to find differentially expressed genes, Hu, Qiu and Glazko (2010) generate Gaussian deviates in a two-group setting with one group of 100 observations correlated at 0.94, another group of 608 observations clustered at 0.9, and observations from different groups correlated at 0.697.

We appreciate the difficulty in generating realistic data with known structure. However, we believe that using Gaussian deviates often adds an additional unnecessary layer of assumptions. Indeed, much recent work has been applied to high-throughput data and, for example, we do not believe that microarray data have Gaussian distributions [Hardin and Wilson (2009)]. In Section 3.1 we demonstrate that our method produces matrices that are more general than the class of matrices produced by finding the sample correlation of Gaussian data. In particular, our method is able to produce Gaussian-like deviates or other distributional deviates (e.g., uniform-like deviates). Instead of simulating Gaussian data from a known correlation structure, we argue in favor of simulating correlation matrices directly based on a known correlation structure. The random correlation matrices can then be used to assess the algorithm at hand.

1.2. *Three existing models.* The goal of our work is to provide an algorithm for simulating correlation structures that can be used to evaluate statistical methodology in a realistic context. Instead of relying on a known structure, noise is added to the matrix to represent variability across different components of the entries. Additionally, the noise is added in a way that can represent any underlying data structure.

Below we have outlined three methods for generating correlation matrices, each of which describes different dependence structures for simulating data. Each of the three methods is taken from a different area of application (estimating heritability from GWAS, classification and network analysis). Our paper offers a flexible way to generate correlation structures given a reasonable model of what we would expect across observational units.

1.2.1. *Constant correlation model.* Heritability is the proportion of variability in a phenotypic characteristic which is due to genetic differences between individuals. The estimation procedure for heritability is based on a mixture model specified by a large correlation structure defining the correlations between the genetic effects of individuals in a study. These correlations are typically referred to as genetic correlations. The genetic correlation structure is then used to decompose the phenotypic variance to genetic and environmental components, resulting in an estimate of heritability.

Recent work has assumed that the genetic correlation structure is known [Lee et al. (2011)], despite estimating it from genetic data. Simulations using a known correlation structure are used to evaluate the heritability estimation algorithm. In the actual data analysis, the estimate of the correlation matrix is plugged into the algorithm as if it were the true value.

The simulation study generates genetic relationships between 10,000 individuals in the following manner: Groups of size 100 are simulated to have genetic correlations of 0.05. Uncorrelated environmental effects are added to the genetic effects. The variances of the genetic and environmental effects are predetermined by the value of heritability used in the simulation. According to the liability threshold model used by Lee et al. (2011), individuals for which the sum of effects crosses a pre-defined threshold are considered to be cases, while the rest are considered controls.

Since the prevalence of most interesting phenotypes is small, the threshold is set such that only a small fraction of the individuals in each group are considered cases. To simulate realistic case-control studies, the cases are kept along with the same number of randomly selected controls from the group, while the rest of the controls are discarded. The process is repeated until 5000 cases and 5000 controls are obtained.

Depending on the number of cases in each group, the resulting groups of genetic correlation 0.05 are as small as a few individuals or as large as 100 individuals.

Genetic correlations between different groups are assumed to be zero [Lee et al. (2011)]. For future reference, we let

$$(1) \qquad \Sigma_k = \begin{pmatrix} 1 & 0.05 & 0.05 & 0.05 & \cdots & 0.05 \\ 0.05 & 1 & 0.05 & 0.05 & \cdots & 0.05 \\ 0.05 & 0.05 & 1 & 0.05 & \cdots & 0.05 \\ 0.05 & 0.05 & 0.05 & 1 & \cdots & 0.05 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0.05 & 0.05 & 0.05 & 0.05 & \cdots & 1 \end{pmatrix}$$

be the genetic correlation matrix for the kth group of individuals, where the size of the kth block matrix is a random variable (i.e., the group size), with a distribution which is defined by the parameter settings.

1.2.2. *Toeplitz model.* Another structure is one that models high correlation for observations which are close together in the correlation matrix and models decreasing correlation values for observations which are increasingly far away. In building a classification model, Guo, Hastie and Tibshirani (2007) describe a Toeplitz structure (sometimes referred to as an auto-regressive structure) to the correlation matrix, where adjacent pairs of observations are highly correlated, and those further away are less correlated. For future reference we let

$$(2) \qquad \Sigma_k = \begin{pmatrix} 1 & \rho_k & \rho_k^2 & \rho_k^3 & \cdots & \rho_k^{g_k-1} \\ \rho_k & 1 & \rho_k & \rho_k^2 & \cdots & \rho_k^{g_k-2} \\ \rho_k^2 & \rho_k & 1 & \rho_k & \cdots & \rho_k^{g_k-3} \\ \rho_k^3 & \rho_k^2 & \rho_k & 1 & \cdots & \rho_k^{g_k-4} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_k^{g_k-1} & \rho_k^{g_k-2} & \rho_k^{g_k-3} & \rho_k^{g_k-4} & \cdots & 1 \end{pmatrix}$$

be the correlation matrix for the kth class, given by the base correlation value ρ_k. In this model, the between group correlations are set to zero. Additional classification models have used similar Toeplitz structure for simulating data from a correlation matrix [Guo, Hastie and Tibshirani (2007), Dabney and Storey (2007), Witten and Tibshirani (2009), Zuber and Strimmer (2009), Pang, Tong and Zhao (2009), Huang, Tong and Zhao (2010)]. In fact, Huang et al. use a $U[0.5, 1.5]$ distribution to simulate the variance components in order to add noise to the above prescribed structure. Further, this Toeplitz correlation structure is seen in time series models where simulating correlation matrices is also important [Joe (2006), Ng and Joe (2010)]. The Toeplitz structure has been used extensively in classification and discriminant analysis as a model for group correlations

1.2.3. *Hub observation model.* The last model which we consider is one that is hierarchical in nature based on a single hub-observation and the relationship of each observation to that original hub. Within the context of network analysis,

Horvath et al. [Zhang and Horvath (2005), Langfelder, Zhang and Horvath (2008), Langfelder and Horvath (2008)] define a structure with respect to a particular profile (or hub-observation). Each observation in a group is correlated with the hub-observation with decreasing strength (from a user supplied maximum correlation to a given minimum correlation). Additionally, groups are generated independently (i.e., with correlation zero between groups). Letting observation 1 correspond to the hub, for the ith observation ($i = 2, 3, \ldots, g$), the correlation between it and the hub-observation is given by

$$\Sigma_{i,1} = \rho_{\max} - \left(\frac{i-2}{g-2}\right)^{\gamma}(\rho_{\max} - \rho_{\min}).$$

Note that the correlation between the ith observation and the hub will range from $\rho_{\max}$ to $\rho_{\min}$; the rate at which the correlations decay is controlled by the exponent γ (where $\gamma = 1$ would indicate a linear decay).

1.2.4. *Overview.* Motivated by the models above, we provide algorithms for adding noise to prescribed correlation matrices. We begin in Section 2 detailing algorithms for the three specific models of correlation matrices discussed above. In Section 3.1 we demonstrate the benefits of generating random deviates from the correlation matrix instead of using random deviates from a particular distribution. Section 4 gives applications of how our method can be used to assess new and standard statistical procedures. Following a brief conclusion in Section 5, we present the theoretical justifications of our algorithms in the Appendix.

2. Recipes. Using a single basic procedure (Algorithm 4 in Section A.2) for adding noise to a given correlation matrix, we can take advantage of our theoretical understanding of certain known correlation structures to yield stronger results. This is carried out for the constant correlation structure (Algorithm 1 in Section 2.1), the Toeplitz correlation structure (Algorithm 2 in Section 2.2) and the hub correlation structure (Algorithm 3 in Section 2.3). Each model describes a population based on multiple groups with the same underlying structure (with different sizes and parameter values). Since the justifications of these procedures are rather involved, we defer the technical details until Appendix.

2.1. *Constant correlation structure.* Our first correlation structure is based on constant correlations within each group and between each group (values of the correlation differ for each relationship). In particular, observe that the approach below yields a noisy correlation matrix which has a significant amount of noise on the *off-diagonal* blocks. This is clearly more realistic than simply assuming that all of these entries are zero. A detailed justification of the following algorithm can be found in Section A.3 of Appendix.

ALGORITHM 1.   *Let*

- *K denote a positive integer* (*the number of groups*) *and* $k = 1, 2, \ldots, K$,
- g_k *be a positive integer* (*the size of the kth group*),
- $N = \sum_{k=1}^{K} g_k$ (*size of the desired matrix*),
- ρ_k *such that* $0 \le \rho_k < 1$ (*baseline correlation in the kth group*),
- $\rho_{\min} = \min\{\rho_1, \rho_2, \ldots, \rho_K\}$ (*minimum correlation in any group*),
- $\rho_{\max} = \max\{\rho_1, \rho_2, \ldots, \rho_K\}$ (*maximum correlation in any group*),
- δ *such that* $0 \le \delta < \rho_{\min}$ (*baseline noise between group*),
- Σ_k *be the* $g_k \times g_k$ *matrix*

$$(3) \qquad \Sigma_k = \begin{pmatrix} 1 & \rho_k & \cdots & \rho_k \\ \rho_k & 1 & \cdots & \rho_k \\ \vdots & \vdots & \ddots & \vdots \\ \rho_k & \rho_k & \cdots & 1 \end{pmatrix}$$

  (*correlation matrix for kth group*),
- Σ *be the* $N \times N$ *matrix having the blocks* $\Sigma_1, \Sigma_2, \ldots, \Sigma_k$ *along the diagonal and zeros elsewhere,*
- ε *such that* $0 \le \varepsilon < 1 - \rho_{\max}$ (*maximum entry-wise random noise*),
- *M be a positive integer* (*the dimension of the noise space*).

Select N unit vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$ *randomly from* $\mathbb{R}^M$. *The* $N \times N$ *matrix* $S = (S_{ij})_{i,j=1}^{N}$ *defined by*

$$(4) \qquad S_{ij} = \begin{cases} 1, & \text{if } i = j, \\ \rho_k + \varepsilon \mathbf{u}_i^T \mathbf{u}_j, & \text{if } i, j \text{ are in the kth group and } i \ne j, \\ \delta + \varepsilon \mathbf{u}_i^T \mathbf{u}_j, & \text{if } i, j \text{ are in different groups,} \end{cases}$$

is a correlation matrix whose condition number satisfies

$$(5) \qquad \boxed{\kappa(S) \le \frac{N(1 + \varepsilon) + 1}{1 - \rho_{\max} - \varepsilon}.}$$

2.2. *Toeplitz correlation structure.*   The Toeplitz structure has been used extensively in classification, discriminant analysis and in the time series literature as a model for group correlations. In particular, the model we follow assumes that each pair of adjacent observations is highly correlated and that the correlations between the ith and jth observations decay exponentially with respect to $|i - j|$. The following algorithm, whose justification can be found in Section A.4 of Appendix, produces noisy correlation matrices based upon the Toeplitz template.

ALGORITHM 2.   *Let*

- *K denote a positive integer* (*the number of clusters*) *and* $k = 1, 2, \ldots, K$,
- g_k *be a positive integer* (*the size of the kth group*),

- $N = \sum_{k=1}^{K} g_k$ (*size of the desired matrix*),
- ρ_k *be such that* $0 \le \rho_k < 1$ (*correlation factor in the kth group*),
- $\rho_{\max} = \max\{\rho_1, \rho_2, \ldots, \rho_K\}$ (*maximum correlation factor*),
- Σ_k *be the* $g_k \times g_k$ *Toeplitz correlation matrix*

$$
(6) \qquad \Sigma_k = \begin{pmatrix}
1 & \rho_k & \rho_k^2 & \rho_k^3 & \cdots & \rho_k^{g_k-1} \\
\rho_k & 1 & \rho_k & \rho_k^2 & \cdots & \rho_k^{g_k-2} \\
\rho_k^2 & \rho_k & 1 & \rho_k & \cdots & \rho_k^{g_k-3} \\
\rho_k^3 & \rho_k^2 & \rho_k & 1 & \cdots & \rho_k^{g_k-4} \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\rho_k^{g_k-1} & \rho_k^{g_k-2} & \rho_k^{g_k-3} & \rho_k^{g_k-4} & \cdots & 1
\end{pmatrix}
$$

(*correlation matrix for kth group*),
- Σ *be the* $N \times N$ *matrix having the blocks* $\Sigma_1, \Sigma_2, \ldots, \Sigma_k$ *along the diagonal and zeros elsewhere,*
- $0 < \varepsilon < \dfrac{1 - \rho_{\max}}{1 + \rho_{\max}}$ (*maximum entry-wise random noise*),
- M *be a positive integer* (*the dimension of the noise space*).

Select N *unit vectors* $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$ *from* $\mathbb{R}^M$ *and form the* $M \times N$ *matrix* $U = (\mathbf{u}_1|\mathbf{u}_2|\cdots|\mathbf{u}_N)$ *whose columns are the* $\mathbf{u}_i$. *The* $N \times N$ *matrix*

$$
(7) \qquad S = \Sigma + \varepsilon(U^T U - I)
$$

is a correlation matrix whose entries satisfy $|S_{ij} - \Sigma_{ij}| \le \varepsilon$ *and whose condition number satisfies*

$$
(8) \qquad \boxed{\kappa(S) \le \frac{(1 + \rho_{\max})/(1 - \rho_{\max}) + (N - 1)\varepsilon}{(1 - \rho_{\max})/(1 + \rho_{\max}) - \varepsilon}.}
$$

Among other things, let us remark that for typical values of ρ [e.g., Guo, Hastie and Tibshirani (2007) let $\rho = 0.9$] the noise level ε can be made quite large compared to most of the entries in each Σ_k. This occurs because the eigenvalue estimates (24) obtained in Section A.4 are remarkably strong and because the off-diagonal entries of each submatrix Σ_k are small (due to exponential decay) if one is far away from the main diagonal. Thus, the approach outlined above yields a flexible method for introducing noise into the Toeplitz model. In fact, one can introduce so much noise (while still obtaining a correlation matrix with controlled condition number) that the original block-Toeplitz structure becomes difficult to discern.

2.3. *Hub correlation structure.* The hub correlation structure assumes a known correlation between a *hub observation* (typically the first observation) and each of the other observations. Moreover, one typically assumes that the correlation between the 1st and the ith observation decays as i increases.

Let us describe a typical example which has been considered frequently in the literature. Suppose that the first row (and hence column) of a $g \times g$ correlation matrix A is to consist of the prescribed values

$$A_{11} = 1, \qquad A_{1i} = \rho_{\max} - (\rho_{\max} - \rho_{\min})\left(\frac{i-2}{g-2}\right)^{\gamma}$$

which decrease (linearly if $\gamma = 1$) from $A_{12} = \rho_{\max}$ to $A_{1g} = \rho_{\min}$ for $2 \le i \le g$. For instance, this model is considered in Horvath et al. [Langfelder and Horvath (2008), Langfelder, Zhang and Horvath (2008), Zhang and Horvath (2005)]. For the sake of simplicity, we consider the linear case $\gamma = 1$ and adopt a more convenient notation. Rather than specifying $\rho_{\max}$ and $\rho_{\min}$, we specify only $\rho_{\max}$ and work instead with the step size $\tau = (\rho_{\max} - \rho_{\min})/(g-2)$.

After specifying the first row, there are a variety of ways to generate the remainder of such a correlation matrix. Using any hub structure correlation matrix, we can find the smallest resulting eigenvalue which can be fed into Algorithm 4 of Section A.2. For example, we can use a Toeplitz structure to fill out the remainder of the hub correlation matrix and, using the well-developed theory of truncated Toeplitz matrices [Böttcher and Silbermann (1999)], obtain eigenvalue bounds which can be fed directly into Algorithm 4. This approach yields the following algorithm, whose justification can be found in Section A.5 of Appendix.

ALGORITHM 3. *Let*

- *K denote a positive integer (the number of groups) and $k = 1, 2, \ldots, K$,*
- *g_k be a positive integer (the size of the kth group),*
- *$N = \sum_{k=1}^{K} g_k$ (size of the desired matrix),*
- *ρ_k (maximum correlation in the first row of kth group),*
- *τ_k (step size in first row/column of kth group),*
- *$\alpha_{k,1} = 1$ and $\alpha_{k,i} = \rho_k - \tau_k(i-2)$ (correlations between hub and observations),*
- *Σ_k be the $g_k \times g_k$ hub-Toeplitz correlation matrix*

$$(9) \qquad \Sigma_k = \begin{pmatrix} 1 & \alpha_{k,2} & \alpha_{k,3} & \alpha_{k,4} & \cdots & \alpha_{k,g_k} \\ \alpha_{k,2} & 1 & \alpha_{k,2} & \alpha_{k,3} & \cdots & \alpha_{k,g_k-1} \\ \alpha_{k,3} & \alpha_{k,2} & 1 & \alpha_{k,2} & \cdots & \alpha_{k,g_k-2} \\ \alpha_{k,4} & \alpha_{k,3} & \alpha_{k,2} & 1 & \cdots & \alpha_{k,g_k-3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{k,g_k} & \alpha_{k,g_k-1} & \alpha_{k,g_k-2} & \alpha_{k,g_k-3} & \cdots & 1 \end{pmatrix}$$

(correlation matrix for kth group),
- *Σ be the $N \times N$ matrix having the blocks $\Sigma_1, \Sigma_2, \ldots, \Sigma_k$ along the diagonal and zeros elsewhere,*
- *$0 < \varepsilon < \min\{1 - \rho_k - \frac{3}{4}\tau_k : 1 \le k \le K\}$ (ε is the maximum noise level),*
- *M be a positive integer (the dimension of the noise space).*

Select N unit vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$ from $\mathbb{R}^M$ and form the $M \times N$ matrix $U = (\mathbf{u}_1 | \mathbf{u}_2 | \cdots | \mathbf{u}_N)$ whose columns are the $\mathbf{u}_i$. The $N \times N$ matrix

$$(10) \qquad\qquad S = \Sigma + \varepsilon(U^T U - I)$$

is a correlation matrix whose entries satisfy $|S_{ij} - \Sigma_{ij}| \leq \varepsilon$ and whose condition number satisfies

$$(11) \qquad\qquad \boxed{\kappa(S) \leq \frac{\lambda_1(\Sigma) + (N-1)\varepsilon}{\lambda_N(\Sigma) - \varepsilon}}$$

where

$$(12) \quad \lambda_1(\Sigma) \leq \max\left\{ 1 + (g_k - 1)\rho_k - \tau_k \frac{(g_k - 2)(g_k - 1)}{2} : 1 \leq k \leq K \right\},$$

$$(13) \quad \lambda_N(\Sigma) \geq \min\left\{ 1 - \rho_k - \tfrac{3}{4}\tau_k : 1 \leq k \leq K \right\}.$$

2.4. *Extensions.* Before proceeding, let us remark that our general Algorithm 4, which can be found in Section A.2 of Appendix, is applicable to any given positive-definite correlation matrix. The amount of noise which can be added to the original matrix is determined by its smallest eigenvalue. For several specific classes of correlation matrices, one can obtain simple, but powerful, lower bounds on this lowest eigenvalue. For such correlation matrices, we have provided explicit, specialized algorithms which provide a significant amount of noise while also maintaining quantitative control over the condition number of the resulting matrix.

3. Distribution of error terms. As described above, our method uses the dot product of normalized vectors as the error terms which are added to a given correlation matrix. Below we discuss three methods for generating normalized vectors with given distributions.

1. *Random uniform vectors on the M-dimensional unit sphere*: Consider

$$x_i \sim \text{i.i.d. } N(0, 1), \qquad i = 1, 2, \ldots, M,$$

$$\mathbf{x} = (x_1, x_2, \ldots, x_M),$$

$$\mathbf{v} = \frac{\mathbf{x}}{\|\mathbf{x}\|}.$$

It is known that $\mathbf{v}$ will be uniformly distributed on the M-dimensional unit sphere [Muller (1959)]. Additionally, for vectors distributed uniformly on the unit sphere, the distribution of their dot product is well characterized [Cho (2009)]:

$$\mathbf{v}, \mathbf{w} \sim \text{uniformly on the } M\text{-dimensional unit sphere,}$$

$$Z = \mathbf{v}^T \mathbf{w},$$

$$f_Z(z; M) = \frac{\Gamma(M/2)}{\Gamma((M-1)/2)\sqrt{\pi}} \left(\sqrt{1 - z^2}\right)^{M-3}, \qquad -1 \leq z \leq 1,$$

is the probability density function for the dot product of $\mathbf{v}$ and $\mathbf{w}$. Note that if $M = 2$, the distribution of Z is of the form $1/\pi\sqrt{1 - z^2}$ which gives a U-shaped distribution favoring values of Z closer to -1 and 1. If $M = 3$, the distribution is uniform across -1 to 1. For $M > 3$, the distribution function is mound shaped and converges to a Gaussian distribution for large M (see below).

2. *Random independent and identically distributed vectors*: Consider two vectors generated independently from identical distributions in $\mathbb{R}^M$ with mean zero,

$$\mathbf{v}, \mathbf{w} \sim \text{i.i.d. } F_M(\mu = \mathbf{0}),$$

$$Z = \frac{\mathbf{v}^T\mathbf{w}}{\|\mathbf{v}\|\|\mathbf{w}\|},$$

$$\sqrt{M}Z \xrightarrow{D} N(0, 1).$$

The asymptotic distribution of Z is a straightforward application of the Central Limit theorem and Slutsky's theorem.

3. *Arbitrary distribution*: Some situations may call for a particular distribution of error noise. The distribution can be controlled through the α parameter as seen in equation (21).

Note that typically the error terms added to the correlation entries are of the form

$$\text{error} = \varepsilon \cdot \frac{\mathbf{v}^T\mathbf{w}}{\|\mathbf{v}\|\|\mathbf{w}\|}.$$

If the dot product is approximately distributed with a variance of $1/M$, then the variance of the error term is ε^2/M, resulting in a standard error of the correlation values,

$$(14) \qquad\qquad \text{SE(correlation)} \approx \frac{\varepsilon}{\sqrt{M}}.$$

The distribution of error terms will necessarily depend on the application. For some problems, uniform error terms may be most appropriate; for other problems, Gaussian errors will be preferable. In fact, for Gaussian data, correlations between vectors are approximately Gaussian, which may motivate a user to want to add Gaussian noise to the given correlation structure.

3.1. *Comparison to a correlation matrix from Gaussian vectors*. One method for generating a noisy correlation matrix is to simulate Gaussian data from an original template and then find the sample correlation matrix from the data. Varying the sample size of the generated data can create correlation matrices which are more or less variable (in magnitude). However, from Gaussian data the nature of the variability (distribution) of the resulting correlations is similar across different sample sizes—uniform or U-shaped distributions of error terms are not possible

given correlations from Gaussian data. In addition, the majority of the entries in a given sample correlation matrix generated from Gaussian data are quite close to the template matrix. Only a handful of observations deviate from the template substantially. In fact, the sample size needed in order to get a large amount of variability could be smaller than the dimension of the correlation matrix (thus producing sample correlation matrices which are not positive definite).

To demonstrate the restriction associated with simulating Gaussian data as a way to find sample correlation matrices, we generate multiple correlation matrices using both Gaussian samples and our method. The Gaussian noise is created by simulating data of a particular sample size (25, 250, or 1000) from a template correlation matrix. We then compute the sample correlation matrix and find the difference between the estimate and the template; histograms of those differences describe the distribution of the correlation error terms. For example, Gauss25 was created by simulating 25 observations from a 230×230 template correlation matrix. The difference between the correlations of the 25 observations and the template matrix are computed; the histograms of the differences are given in Figure 1.

The three Gaussian structures (sample sizes 25, 250, 1000) show the same tendencies with more spread for smaller sample sizes (see Figure 1: Gauss25, Gauss250, Gauss1000). The three simulations using our method are based on uniform random vectors on the unit sphere in $\mathbb{R}^2$, $\mathbb{R}^3$ and $\mathbb{R}^{25}$ (see Table 1 for different generating scenarios). For each simulation we used a constant correlation structure with three groups of sizes $g_1 = 100$, $g_2 = 50$, $g_3 = 80$; within group correlations of $\rho_1 = 0.7$, $\rho_2 = 0.7$, $\rho_3 = 0.4$; $\varepsilon = 0.29$; and between group correlations of $\delta = 0.25$ (see Algorithm 1 in Section A.3).

Note that in our simulations below, the distribution of differences from uniform vectors in $\mathbb{R}^{25}$ (S25) is similar to that obtained from the correlation matrix generated by sampling 250 random Gaussian vectors (Gauss250). In fact, not only can our method create correlations with univariate distributions that are similar to Gaussian deviations, but the eigenvalues of the respective matrices (ours compared with correlations from Gaussian data) are indistinguishable (results not shown).

We see that our method is able to add larger noise terms than the Gaussian simulation. Figure 1 shows the distribution of the differences. Depending on the application, one might prefer large noise components, uniform noise components or bell-shaped noise components. Our work provides a template for generating multiple different structures depending on the problem at hand.

4. Applications. To demonstrate the effectiveness of our method, we simulate data from two applications to show that noise added to a known correlation structure can be useful in practice. It may not always be obvious which format to use to incorporate the noise; the format of the noise will be situation dependent and should be based on the underlying data structure. In Section 3.1 we have provided more details about the different noise models.
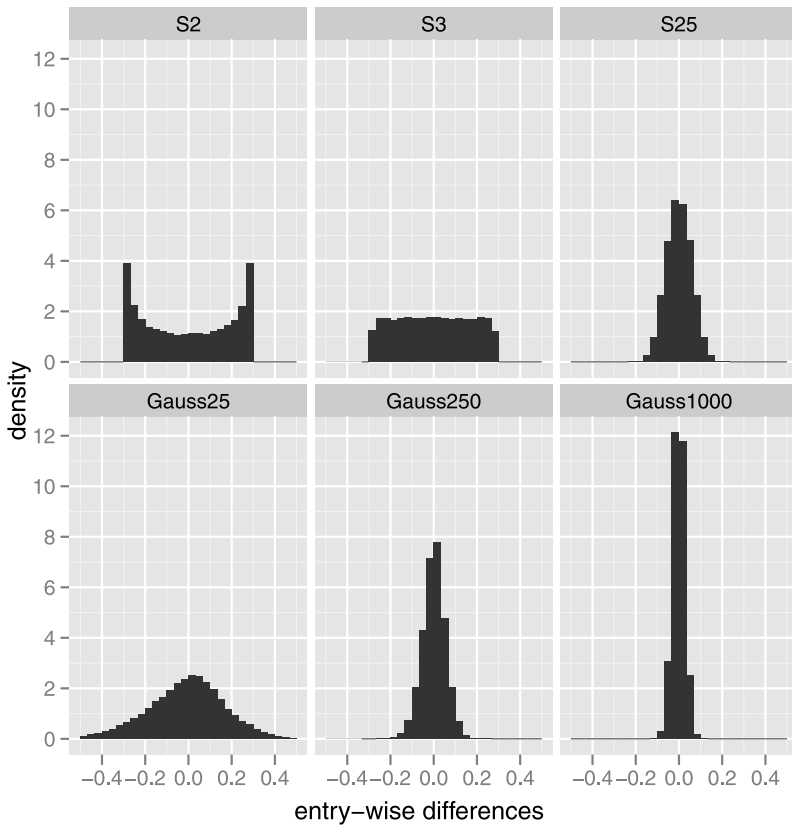
Fig. 1. *Each histogram represents the distribution of entry-wise differences between the generated matrix and the template. The distribution of differences for random vectors in $\mathbb{R}^{25}$ (S25) is similar to that obtained from the correlation matrix generated by sampling 250 random Gaussian vectors (Gauss250).*

TABLE 1
Six different correlation matrix generating scenarios. S2, S3 and S25 use the algorithms given in the paper for constant correlation; to simulate the error terms, we generate vectors on the unit sphere using standard Gaussian deviates. The Gaussian simulations use the template matrix with the given sample size of random vectors. Each correlation matrix is based on a setting of 3 groups with sizes (100, 50, 80)

Template	$\rho_1 = 0.7$	$\rho_2 = 0.7$	$\rho_3 = 0.4$	$\delta = 0.25$	$\varepsilon = 0$
S2	$\rho_1 = 0.7$	$\rho_2 = 0.7$	$\rho_3 = 0.4$	$\delta = 0.25$	$\mathbf{u}_i \in S^2, \varepsilon = 0.29$
S3	$\rho_1 = 0.7$	$\rho_2 = 0.7$	$\rho_3 = 0.4$	$\delta = 0.25$	$\mathbf{u}_i \in S^3, \varepsilon = 0.29$
S25	$\rho_1 = 0.7$	$\rho_2 = 0.7$	$\rho_3 = 0.4$	$\delta = 0.25$	$\mathbf{u}_i \in S^{25}, \varepsilon = 0.29$
Gauss25	$\rho_1 = 0.7$	$\rho_2 = 0.7$	$\rho_3 = 0.4$	$\delta = 0.25$	25 vectors
Gauss250	$\rho_1 = 0.7$	$\rho_2 = 0.7$	$\rho_3 = 0.4$	$\delta = 0.25$	250 vectors
Gauss1000	$\rho_1 = 0.7$	$\rho_2 = 0.7$	$\rho_3 = 0.4$	$\delta = 0.25$	1000 vectors

4.1. *Heritability.* Heritability is the proportion of variability in a phenotypic characteristic which is due to genetic differences. The understanding and estimation of heritability is of great importance in directing future studies as well as understanding the architecture of human genetic diseases such as type-1 and type-2 diabetes, Crohn's disease, schizophrenia and bipolar disorder. The study of heritability in human disease presents the so-called "mystery" of the missing heritability [Maher (2008)]: a considerable gap between the estimated heritability as obtained from family studies and the estimated heritability as obtained from genetic studies (known as genome-wide association studies, or GWASs), with the latter estimate of heritability being considerably smaller than the former for a wide range of phenotypes.

Recently, a novel method for estimation of heritability from the genotypes of unrelated individuals was introduced by Yang et al. (2010). The method was first applied for random samples from a population [Yang et al. (2010)] and later adapted for the more relevant scenario of case-control studies [Lee et al. (2011)]. These works presented genome-based estimates of heritability that were considerably higher than previous estimates, thus bridging, at least in part, the gap between family-based and genome-based estimates of heritability. As expected, these works attracted attention and are the focus of recent research and debate [see, e.g., Golan and Rosset (2011), Lee et al. (2012)].

The central idea behind these methods is to estimate a population-wise correlation structure from the genotypes of individuals and use this estimated structure in a Restricted Maximum Likelihood (REML) estimation of the heritability. However, the REML estimation does not account for the fact that the correlation structure is estimated rather than known. Moreover, the simulations in Lee et al. (2011) use a known correlation structure to demonstrate the validity of the method, which in turn uses an estimated correlation structure. Such simulations might produce an overly optimistic evaluation of the method used to estimate heritability in terms of both bias and variance.

To demonstrate the sensitivity of the heritability estimate to the known correlation structure, we reran the simulations in Lee et al. (2011) with and without noise. As expected, adding noise to the matrix introduces bias to the estimators. Our methods provide a mechanism for understanding the behavior of heritability estimates under different correlation and error structures.

For our investigation, we are interested in estimating heritability in the setting of a binary response, in particular, we want to estimate heritability for case control studies. As done in Lee et al. (2011), we assume there is an underlying liability continuous variable (e.g., glucose level) determining the binary measured phenotype (e.g., diabetes). We can find the heritability on the observed scale (with respect to the binary disease trait) and transform it back to the value of interest, the heritability on the liability scale. The transformation considers the disease prevalence in the population [Lee et al. (2011)].

TABLE 2

The average heritability from 100 *simulations for the given population prevalence and heritability of liability. In parentheses we provide the estimate given by* Lee et al. (2011) *with no error. The error is added as described in Algorithm* 1 *with* $\varepsilon = 0.001$, 0.01 *and* 0.02, *corresponding to a SE of the noise terms of* 0.0002, 0.002 *and* 0.004, *respectively*; *see equation* (14).

			True heritability of liability				
		Prevalence of disease in pop.	0.1	0.3	0.5	0.7	0.9
$\varepsilon = 0.001$ SE $= 0.0002$		0.5	0.10 (0.09)	0.28 (0.28)	0.47 (0.51)	0.64 (0.70)	0.80 (0.90)
		0.1	0.11 (0.11)	0.29 (0.30)	0.49 (0.49)	0.70 (0.71)	0.87 (0.89)
		0.001	0.05 (0.17)	0.23 (0.31)	0.37 (0.56)	0.56 (0.75)	0.77 (0.94)
$\varepsilon = 0.01$ SE $= 0.002$		0.5	0.08 (0.09)	0.23 (0.28)	0.37 (0.51)	0.54 (0.70)	0.69 (0.90)
		0.1	0.06 (0.11)	0.23 (0.30)	0.42 (0.49)	0.60 (0.71)	0.78 (0.89)
		0.001	0.00 (0.17)	0.01 (0.31)	0.01 (0.56)	0.02 (0.75)	0.02 (0.94)
$\varepsilon = 0.02$ SE $= 0.004$		0.5	0.07 (0.09)	0.19 (0.28)	0.35 (0.51)	0.51 (0.70)	0.65 (0.90)
		0.1	0.03 (0.11)	0.16 (0.30)	0.22 (0.49)	0.53 (0.71)	0.73 (0.89)
		0.001	0.01 (0.17)	0.01 (0.31)	0.01 (0.56)	0.01 (0.75)	0.01 (0.94)

We followed the simulation procedure of Lee et al. (2011), which is outlined in Section 1.2.1. See Lee et al. (2011) for a more detailed description of the simulation procedure. Subsequently, we added noise to the correlation matrices using Algorithm 1 in Section 2.1. We used the software Genome-wide Complex Trait Analysis (GCTA) to estimate heritability and standard errors of the estimate [Yang et al. (2011)].

Results are presented in Table 2. Each table entry contains the estimated heritability of liability from our simulations with noise as well as the corresponding estimate given by Lee et al. (2011)—calculated using the known correlation structure. As expected, the more noise added to the relationship matrix, the more bias in estimating the heritability. Additionally, we see that there is a strong interaction: for low prevalence, even a small amount of error can have a large impact on the estimate of heritability. With high prevalence, moderate amounts of error can bias the estimate.

We run a second set of simulations to see how often we can capture the true heritability in a confidence interval using ± 2 SE (provided from the GCTA software) when noisy correlation matrices are used; we expect roughly 95% of the confidence intervals to capture the true heritability value. We simulated 100 heritability values from a uniform $(0.1, 0.9)$ distribution. We then simulated phenotypes and a corresponding correlation matrix with error (as described previously) for each heritability. We constructed corrected CIs, using a multiplicative factor correction obtained from our first set of simulations as well as uncorrected confidence intervals, and counted the number of times these CIs contained the true heritability.

Out of 100 *simulations, the number of true heritabilities captured in the interval two standard errors around the estimated heritability. We would expect* 95% *confidence intervals to capture the true heritability value* 95 *times out of* 100. *For each parameter setting, we calculated two intervals, thus, we report two coverage rates. The first number uses a correction factor (for both the heritability and the SE of the heritability) calculated from the bias estimated from Table* 2 *above. The second number uses no correction. The error is added as described in Algorithm* 1 *with* $\varepsilon = 0.001, 0.01$ *and* 0.02, *corresponding to a SE of the noise terms of* 0.0002, 0.002 *and* 0.004, *respectively; see equation* (14).

Prevalence of disease in population	Amount of error added to the correlation structure			
	$\varepsilon = 0$ SE = 0	$\varepsilon = 0.001$ SE = 0.0002	$\varepsilon = 0.01$ SE = 0.002	$\varepsilon = 0.02$ SE = 0.004
0.5	93/89	94/88	94/76	93/46
0.1	89/90	94/94	91/72	68/57

Results for the CIs are given in Table 3. Each entry gives the number of true heritabilities captured in the interval two standard errors around the adjusted estimated heritability (expected to be 95 when CIs are accurate). Our results show that neglecting to account for the effects of noise on heritability estimation resulted in problematic confidence intervals. Correcting the bias using the method described above resolved the issue for low and moderate levels of noise.

We note that for the case of estimating heritability, the bias suggests an immediate way of choosing ε for proper inference. The bias is a result of the noisy estimation of the true underlying genetic correlation structure, and so can be estimated using the actual genotypes, as done by Yang et al. (2010). We thus suggest choosing ε such that the bias in simulations of correlation structures equals the bias estimated from the genetic data as in Yang et al. (2010). Since the bias increases with ε (as the signal-to-noise ratio decreases), finding the appropriate value of ε should be straightforward using simulations.

The example provided on estimating heritability shows that a sensitivity analysis can uncover obstacles in applying methods—which work for simulated data—to actual data collected with error. In the following section we provide an additional sensitivity analysis applied to a clustering algorithm.

4.2. *Clustering.* As an additional example, we consider the PAM algorithm [Kaufman and Rousseeuw (1990)] to cluster data which has known structure but different levels of noise. We use the adjusted Rand statistic [Rand (1971), Yeung and Ruzzo (2001)] to measure the degree of concordance between the clustering output and the truth. Using silhouette width, the unsupervised PAM algorithm will give the optimal number of clusters. The adjusted Rand statistic models the degree of concordance between the PAM results and the truth. An adjusted Rand of 1

indicates perfect concordance; an adjusted Rand of zero indicates a random partition. For each of the models we tested, we created the correlation matrix (including noise) using an appropriately customized algorithm.

Clustering results from simulation. Using the hub correlation structure discussed in Section 1.2.3, we are required to choose a method to fill out the rest of the correlation matrix. We use the Toeplitz structure as discussed in Algorithm 3 with the parameter settings below. All simulations were done in the three cluster setting with groups of the following size: $g_1 = 100$, $g_2 = 50$, $g_3 = 80$. Recall that with the hub-Toeplitz correlation, the correlation values descend according to some power (here linearly) from a specified maximum to a specified minimum correlation (see Figure 2):

(a) hTC1 $\quad \rho_1 \in (0.7 \to 0) \quad \rho_2 \in (0.7 \to 0) \quad \rho_3 \in (0.4 \to 0) \quad \mathbf{u}_i \in \mathbb{R}^2 \quad \varepsilon = 0.23$

(b) hTC2 $\quad \rho_1 \in (0.7 \to 0.5) \quad \rho_2 \in (0.7 \to 0.6) \quad \rho_3 \in (0.4 \to 0.2) \quad \mathbf{u}_i \in \mathbb{R}^2 \quad \varepsilon = 0.29$

(c) hTC3 $\quad \rho_1 \in (0.7 \to 0.5) \quad \rho_2 \in (0.7 \to 0.6) \quad \rho_3 \in (0.4 \to 0.2) \quad \mathbf{u}_i \in \mathbb{R}^{25} \quad \varepsilon = 0.29$

(d) hTC4 $\quad \rho_1 \in (0.7 \to 0.5) \quad \rho_2 \in (0.7 \to 0.6) \quad \rho_3 \in (0.4 \to 0.2) \quad \mathbf{u}_i \in \mathbb{R}^2 \quad \varepsilon = 0.1$

(e) hTC5 $\quad \rho_1 \in (0.7 \to 0.5) \quad \rho_2 \in (0.7 \to 0.6) \quad \rho_3 \in (0.4 \to 0.2) \quad \mathbf{u}_i \in \mathbb{R}^2 \quad \varepsilon = 0.25$

(f) hTC6 $\quad \rho_1 \in (0.8 \to 0) \quad \rho_2 \in (0.75 \to 0) \quad \rho_3 \in (0.7 \to 0) \quad \mathbf{u}_i \in \mathbb{R}^2 \quad \varepsilon = 0.19$

For each of the scenarios above, we simulated 1000 correlation matrices. We then clustered the data using PAM; the clustering results were assessed by determining the number of clusters the algorithm produced (truth was 3 clusters) as well as the concordance between the clustering results and the truth (1 gives perfect concordance).

Our results show that adding noise can create scenarios about which the algorithm is unable to determine the true structure (hTC1 and hTC2) and scenarios where the noise is not sufficient to decrease the performance of the algorithm (hTC3, hTC4 and hTC6), as well as situations that work only sometimes (hTC5) (see Table 4). For correlation structures that degrade all the way to zero (hTC1 and hTC6), the algorithm is able to discern the structure if the original correlations are

TABLE 4
Results from optimal number of clusters as well as the adjusted Rand. The original correlation structure had 3 *clusters. A perfect allocation of points gives an adjusted Rand of* 1.

Scenario	hTC1	hTC2	hTC3	hTC4	hTC5	hTC6
min # clusters	3	3	3	3	3	3
median # clusters	11	8	3	3	3	3
max # clusters	20	13	3	3	10	3
median adj Rand	0.320	0.414	1	1	0.770	1

(a) hTC1

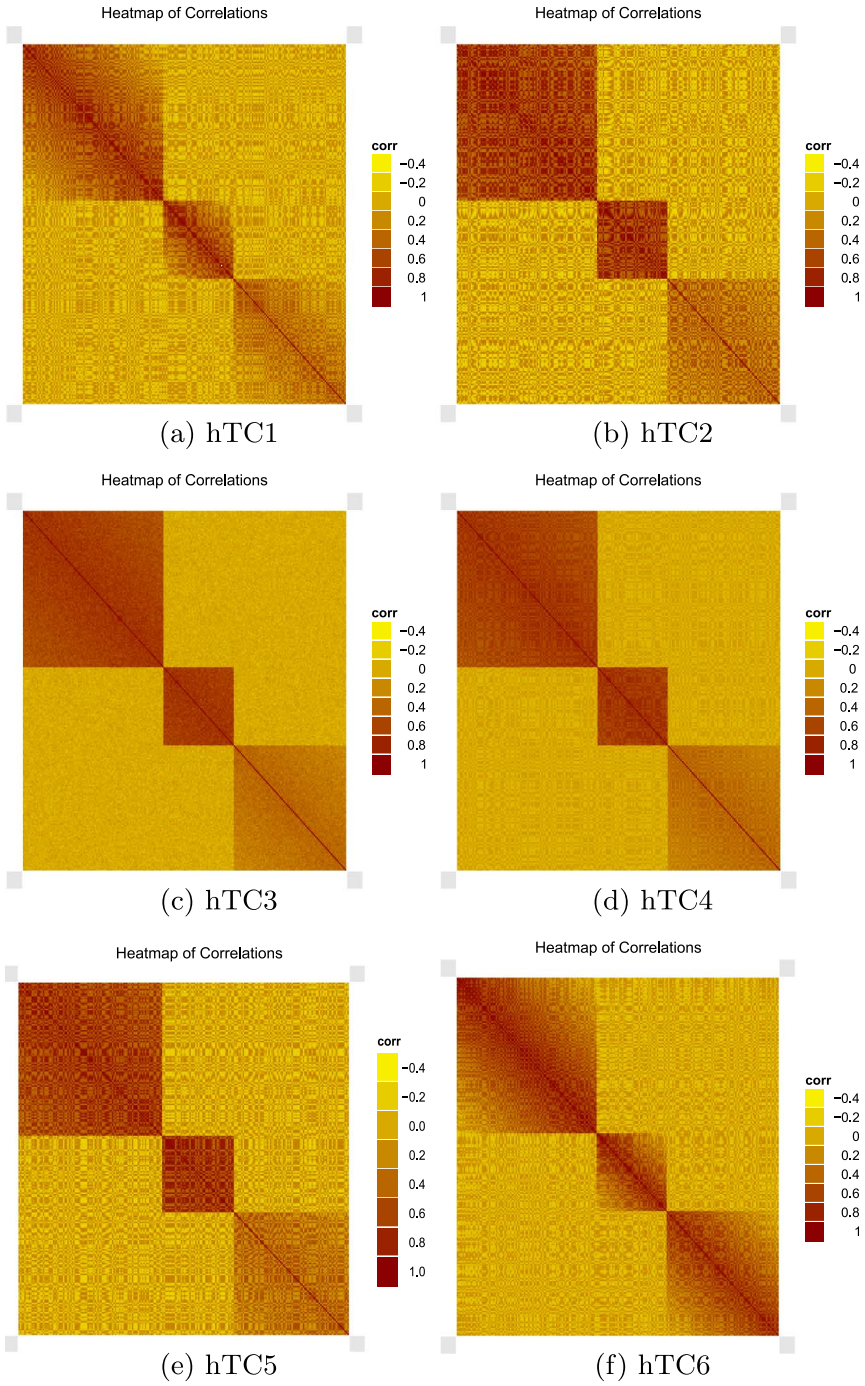(b) hTC2

(c) hTC3

(d) hTC4

(e) hTC5

(f) hTC6

FIG. 2.   *Each heatmap represents the correlation matrix from the scenarios given above.*

TABLE 5
Results for Fisher's Iris data from optimal number of clusters as well as the adjusted Rand. The original data have three species, but I. virginica *and* I. versicolor *are remarkably similar according to the measurements given in the data set and are typically grouped together. A perfect allocation of points gives an adjusted Rand of* 1.

Scenario	$\varepsilon = 0$ (no error)	$\mathbb{R}^{25}, \varepsilon = 0.05$	$\mathbb{R}^{10}, \varepsilon = 0.2$	$\mathbb{R}^5, \varepsilon = 0.2$	$\mathbb{R}^2, \varepsilon = 0.15$
min # clusters	2	2	2	2	2
median # clusters	2	2	2	3	4
max # clusters	2	2	3	5	5
median adj Rand (2 species)	1	1	0.973	0.570	0.469

large (hTC6). For correlation structures that degrade only a small amount (hTC2, hTC3, hTC4), the results are based on the amount of error and the dimension from which the noise vectors from Algorithm 3 are selected.

Clustering results on Fisher's Iris data. As an application to real data, we consider Fisher's Iris data [Fisher (1936)] which have been used extensively to asses discriminant and cluster analysis methods. For 50 iris specimens in each of three species, *Iris setosa*, *I. versicolor* and *I. virginica*, the sepal length, sepal width, petal length and petal width are measured in millimeters (see Figure 3). Though there are three species measured, *I. versicolor* and *I. virginica* are typically quite difficult to differentiate with unsupervised clustering methods [Mezzich and Solomon (1980), page 85]. Indeed, when applying the PAM algorithm to the iris data, we get a perfect separation into two groups (with the three group silhouette width being slightly smaller).

In order to assess the sensitivity of the output, we can add noise consistent with the variability of the observations. In this case, we assume that the correlations within a group are constant, and we estimate the standard error of the correlations to be approximately 0.01. Such errors might be generated by using vectors from $\mathbb{R}^{25}$, with $\varepsilon = 0.05$. [Refer to the discussion preceding equation (14) for choice of ε.] Even without noise, the unsupervised PAM algorithm separates the flowers into two distinct groups (see Table 5). As would be expected with the iris data, substantially more noise is needed before the two-group structure is lost.

5. Conclusion. We have developed an algorithm for adding noise, in a highly controlled manner, to a template correlation matrix in order to obtain a more realistic correlation matrix. Moreover, we have demonstrated how our general procedure can be tailored to a number of different correlation models (e.g., constant correlation, Toeplitz structure).

Our method allows for noisy correlation matrices which differ more from the initial template than the estimated correlation matrix based on simulated Gaussian
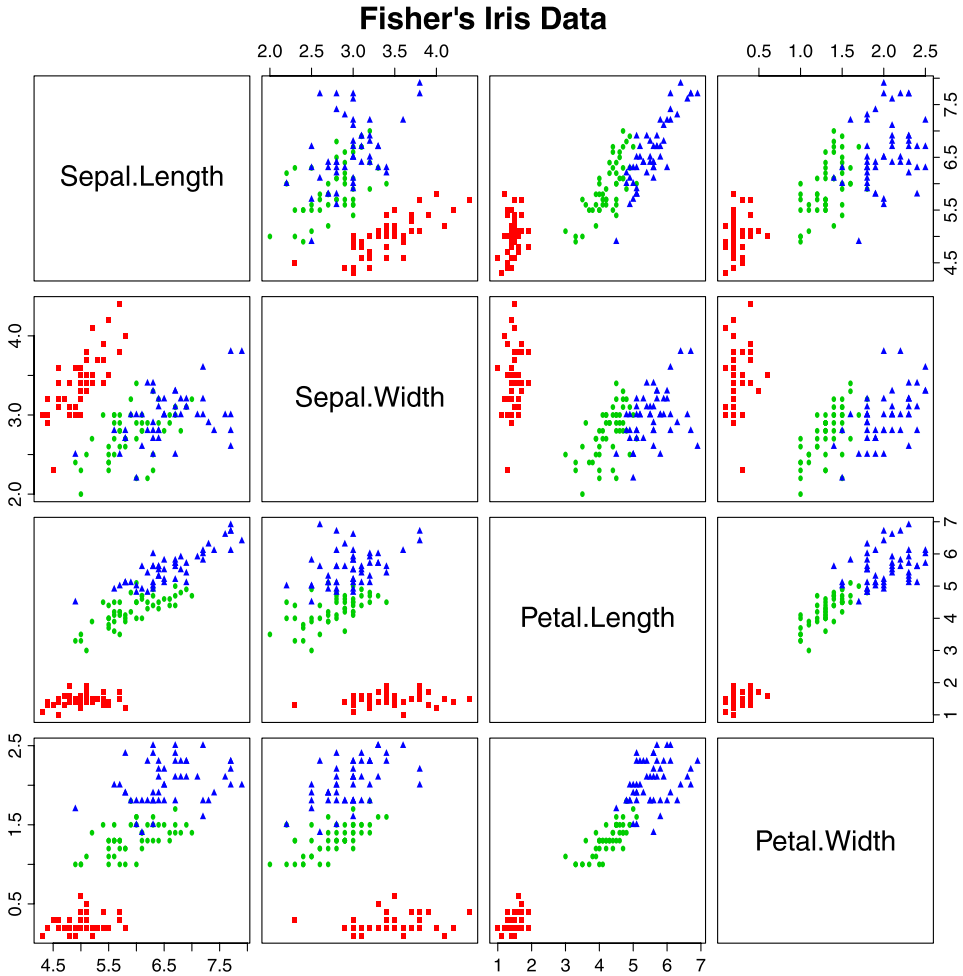
FIG. 3. *The* I. setosa *flowers are clearly distinct*, *while* I. versicolor *and* I. virginica *are difficult to differentiate*, *especially with respect to sepal measurements*.

data. Using Gaussian data produces a sample correlation matrix with limited and well-behaved (possibly unrealistic) differences from the original template correlation if the generated sample is large. If the generated sample is small, then the sample correlation matrix is not positive definite (i.e., most of the eigenvalues will be zero). Using uniform $[-1, 1]$ deviates as random correlation values produces a matrix that is in general not even positive semidefinite. It can also create relationships between observations that are meaningless (e.g., a trio of observations where the first is highly correlated to the other two, but the other two are negatively correlated).

Using a recent—though already influential—model for heritability and a standard clustering algorithm, we have shown that simulated correlation matrices can

be used to assess new or existing statistical methodology. We provide the user with detailed algorithms to use on several standard clustering structures, as well as a general algorithm to apply to any correlation matrix for which the smallest eigenvalue can be reasonably estimated.

APPENDIX

A.1. Preliminaries. Recall that if A is a $N \times N$ symmetric matrix, then each of its eigenvalues is real and, hence, we may list them in descending order

$$\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_N(A)$$

where each eigenvalue is repeated according to its multiplicity. According to this convention, A is positive semidefinite if and only if $\lambda_N(A) \geq 0$ and A is positive definite if and only if $\lambda_N(A) > 0$.

The *norm* of a $N \times N$ matrix A is defined to be

$$(15) \qquad \|A\| = \max_{\|\mathbf{v}\|=1} \|A\mathbf{v}\|$$

which equals $\lambda_1(A)$ if A is positive semidefinite. To be more specific, the expression (15) is often called the *operator norm* to distinguish it from other frequently used matrix norms (e.g., the *Frobenius norm*). The *condition number* [Horn and Johnson (1990), page 336] of a symmetric matrix A is defined to be

$$\kappa(A) = \begin{cases} \|A^{-1}\|\|A\|, & \text{if } A \text{ is nonsingular,} \\ \infty, & \text{if } A \text{ is singular.} \end{cases}$$

In particular, if A is positive semidefinite, then we have

$$\kappa(A) = \begin{cases} \dfrac{\lambda_1(A)}{\lambda_N(A)}, & \text{if } \lambda_N(A) > 0, \\ \infty, & \text{if } \lambda_N(A) = 0. \end{cases}$$

In the following, we let I_g denote the $g \times g$ identity matrix and $\mathbb{1}_g$ denote the $g \times g$ matrix whose entries are all equal to 1.

A.2. The basic algorithm. Given an $N \times N$ prototype correlation matrix $\Sigma = (\Sigma_{ij})_{i,j=1}^N$, we might wish to add noise to Σ in a computationally efficient way such that the resulting matrix S is also a correlation matrix. Furthermore, we might also require effective bounds on the condition number $\kappa(S)$ of S to ensure that S is a suitable candidate for certain numerical procedures (e.g., matrix inversion). For example, in the statistical software R, the default tolerance for detecting linear dependencies in the columns of a matrix is a condition number $\leq 10^{15}$. The following simple procedure accomplishes this task.

ALGORITHM 4. *Let*

1. Σ *be a given $N \times N$ correlation matrix*,
2. $0 < \varepsilon < \lambda_N(\Sigma)$ (ε *is the maximum noise level*),
3. *M be a positive integer* (*the dimension of the noise space*).

Select N unit vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$ from $\mathbb{R}^M$ and form the $M \times N$ matrix $U = (\mathbf{u}_1|\mathbf{u}_2|\cdots|\mathbf{u}_N)$ whose columns are the $\mathbf{u}_i$. The $N \times N$ matrix

$$(16) \qquad S = \Sigma + \varepsilon(U^T U - I)$$

is a correlation matrix whose entries satisfy $|S_{ij} - \Sigma_{ij}| \leq \varepsilon$ for $1 \leq i, j \leq N$ and whose condition number $\kappa(S)$ satisfies

$$(17) \qquad \boxed{\kappa(S) \leq \frac{\lambda_1(\Sigma) + (N-1)\varepsilon}{\lambda_N(\Sigma) - \varepsilon}.}$$

We might also desire that $\kappa(S) \leq \kappa_{\max}$ for some fixed $\kappa_{\max}$, which depends upon the particular requirements of the software being employed. From (17), it is easy to see that any $\varepsilon > 0$ satisfying the additional constraint

$$(18) \qquad \boxed{\varepsilon \leq \frac{\kappa_{\max}\lambda_N(\Sigma) - \lambda_1(\Sigma)}{\kappa_{\max} + (N-1)}}$$

yields an S such that $\kappa(S) \leq \kappa_{\max}$.

JUSTIFICATION OF ALGORITHM 4.   Let $E = U^T U$ so that

$$E = \begin{pmatrix} 1 & \mathbf{u}_1^T\mathbf{u}_2 & \cdots & \mathbf{u}_1^T\mathbf{u}_N \\ \mathbf{u}_2^T\mathbf{u}_1 & 1 & \cdots & \mathbf{u}_2^T\mathbf{u}_N \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{u}_N^T\mathbf{u}_1 & \mathbf{u}_N^T\mathbf{u}_2 & \cdots & 1 \end{pmatrix}$$

and note that E is symmetric and positive semidefinite [i.e., $\lambda_N(E) \geq 0$]. Moreover, E is positive definite if and only if the $\mathbf{u}_i$ are linearly independent [Horn and Johnson (1990), Theorem 7.2.10].

Now recall that Geršgorin's Disk theorem [Horn and Johnson (1990), Theorem 6.1.1] asserts that if $A = (A_{ij})_{i,j=1}^N$ is a $N \times N$ matrix, then for each eigenvalue λ of A there exists a corresponding index i such that

$$|\lambda - A_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^N |A_{ij}|.$$

By Geršgorin's theorem and Cauchy–Schwarz, it follows that every eigenvalue λ of E satisfies

$$|\lambda - 1| \leq \sum_{\substack{j=1 \\ j \neq i}}^{N} |\mathbf{u}_i^T \mathbf{u}_j| \leq (N-1),$$

whence $0 \leq \lambda_i(E) \leq N$ for $i = 1, 2, \ldots, N$.

We next define S by (16) and observe that S is of the form

$$(19) \qquad S = \begin{pmatrix} 1 & \Sigma_{12} + \varepsilon \mathbf{u}_1^T \mathbf{u}_2 & \cdots & \Sigma_{1N} + \varepsilon \mathbf{u}_1^T \mathbf{u}_N \\ \Sigma_{21} + \varepsilon \mathbf{u}_2^T \mathbf{u}_1 & 1 & \cdots & \Sigma_{2N} + \varepsilon \mathbf{u}_2^T \mathbf{u}_N \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{N1} + \varepsilon \mathbf{u}_N^T \mathbf{u}_1 & \Sigma_{N2} + \varepsilon \mathbf{u}_N^T \mathbf{u}_2 & \cdots & 1 \end{pmatrix}.$$

In particular, S is our original matrix Σ with "noise" terms $\varepsilon \mathbf{u}_i^T \mathbf{u}_j$ of magnitude at most ε added to the off-diagonal entries. To analyze the impact of adding this noise, we require Weyl's Inequalities [Horn and Johnson (1990), Theorem 4.3.1], which assert that if A and B are $N \times N$ symmetric matrices, then

$$(20) \qquad \lambda_j(A) + \lambda_N(B) \leq \lambda_j(A + B) \leq \lambda_j(A) + \lambda_1(B)$$

for $j = 1, 2, \ldots, N$. Applying the lower inequality in (20) with $j = N$, $A = \Sigma - \varepsilon I_N$ and $B = \varepsilon E$, we obtain

$$0 < \lambda_N(\Sigma) - \varepsilon = \lambda_N(\Sigma - \varepsilon I_N) \leq \lambda_N(\Sigma - \varepsilon I_N) + \lambda_N(\varepsilon E) \leq \lambda_N(S),$$

from which we conclude that S is positive definite. Next, we apply the upper inequality in (20) with $j = 1$, which yields

$$\lambda_1(S) \leq \lambda_1(\Sigma - \varepsilon I) + \lambda_1(\varepsilon E) \leq (\lambda_1(\Sigma) - \varepsilon) + N\varepsilon = \lambda_1(\Sigma) + (N-1)\varepsilon.$$

Putting this all together, we obtain the estimates

$$0 < \lambda_N(\Sigma) - \varepsilon \leq \lambda_N(S) \leq \lambda_1(S) \leq \lambda_1(\Sigma) + (N-1)\varepsilon.$$

The inequality (17) follows since $\kappa(S) = \lambda_1(S)/\lambda_N(S)$. $\quad \square$

There are several arguments which can be made in favor of adding noise in this manner. First of all, the procedure described above is easy to implement numerically, and it can be rapidly executed. Moreover, it offers a great deal of flexibility since the dimension M of the ambient space that the vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$ are drawn from and the manner in which these vectors are selected is arbitrary and can be tailored to the particular application at hand. Finally, our method is completely general in the sense that any positive-definite $N \times N$ matrix E having constant diagonal 1 can be factored as $E = U^T U$ where U is some matrix whose columns are unit vectors (e.g., let U be the positive-semidefinite square root of E). In other

words, regardless of the method one employs to produce a positive-semidefinite matrix $E = U^T U$ for use in (16), the same E can in principle be generated using our approach.

Let us now say a few words about the manner in which the vectors $\mathbf{u}_i$ are selected. If M is very small (e.g., $2 \le M \le 5$), then many of the dot products $\mathbf{u}_i^T \mathbf{u}_j$ will be large in magnitude. For many purposes, this yields a very noisy coefficient matrix S based upon the original template Σ. Moreover, even if M is relatively large, then the matrix $E = U^T U$ can be computed extremely rapidly since generating the unit vectors $\mathbf{u}_i$ and computing the dot products $\mathbf{u}_i^T \mathbf{u}_j$ involve straightforward computations (e.g., no eigenvalue calculations).

There are of course many other ways which one could select the $\mathbf{u}_i$. If one wishes the $\mathbf{u}_i^T \mathbf{u}_j$ to be consistently large in magnitude while also ensuring that E has full rank, one lets $M \ge N$ and then selects numbers $\alpha_1, \alpha_2, \ldots, \alpha_N$ at random from $[-1, 1]$ using a continuous probability density function $f(x)$ on $[-1, 1]$ which favors extreme values (e.g., $f(x) = |x|$, $f(x) = \frac{2 - 2\sqrt{1 - x^2}}{4 - \pi}$ or a Beta distribution transformed to exist on the range $[-1, 1]$). One then replaces the numbers $\mathbf{u}_i^T \mathbf{u}_j$ in (19) by

$$(21) \qquad \alpha_i \alpha_j + \sqrt{(1 - |\alpha_i|^2)(1 - |\alpha_j|^2)} \mathbf{u}_i^T \mathbf{u}_j.$$

In effect, one is replacing the $\mathbf{u}_i \in \mathbb{R}^M$ with the unit vectors $(\alpha_i, \sqrt{1 - |\alpha_i|^2} \mathbf{u}_i) \in \mathbb{R}^{M+1}$. These vectors tend to have high negative or positive correlations (but they are linearly independent) since the numbers α_i favor extreme values in the interval $[-1, 1]$.

A.3. Justification of Algorithm 1. In order to introduce a significant amount of noise to the off-diagonal blocks, we work instead with the modified correlation matrix

$$(22) \qquad \Sigma' = \underbrace{\begin{pmatrix} \Sigma_1 - \delta \mathbb{1}_{g_1} & & & \\ & \Sigma_2 - \delta \mathbb{1}_{g_2} & & \\ & & \ddots & \\ & & & \Sigma_K - \delta \mathbb{1}_{g_K} \end{pmatrix}}_{A} + \delta \mathbb{1}_N$$

where $\mathbb{1}_g$ denotes the $g \times g$ matrix whose entries are all 1. Since

$$\Sigma_k - \delta \mathbb{1}_{g_k} = (1 - \rho_k) I_{g_k} + (\rho_k - \delta) \mathbb{1}_{g_k},$$

it follows that

$$(23) \quad \lambda_j(\Sigma_k - \delta \mathbb{1}_{g_k}) = \begin{cases} g_k(\rho_k - \delta) + (1 - \rho_k), & \text{if } j = 1, \\ 1 - \rho_k, & \text{if } j = 2, 3, \ldots, g_k, \end{cases}$$

and that the eigenspace corresponding to the largest eigenvalue of $\Sigma_k - \delta\mathbb{1}_{g_k}$ is spanned by the vector $\mathbf{1}_{g_k} = (1, 1, \ldots, 1) \in \mathbb{R}^{g_k}$. In particular, the eigenspace corresponding to the eigenvalue $1 - \rho_k$ is $(g_k - 1)$-dimensional and any eigenvector $\mathbf{v} = (v_1, v_2, \ldots, v_{g_k})$ belonging to this eigenspace is orthogonal to $\mathbf{1}_{g_k}$ (i.e., satisfies $\sum_{i=1}^{g_k} v_i = 0$).

If we augment $\mathbf{v}$ by placing $N - g_k$ zeros appropriately, we obtain a vector

$$\mathbf{v}' = (\underbrace{0, 0, \ldots, 0}_{g_1 + \cdots + g_{k-1}}, v_1, v_2, \ldots, v_{g_k}, \underbrace{0, 0, \ldots, 0}_{g_{k+1} + \cdots + g_K}) \in \mathbb{R}^N$$

which is an eigenvector of Σ' corresponding to the eigenvalue $1 - \rho_k$ since $A\mathbf{v}' = (1 - \rho_k)\mathbf{v}'$ and $\mathbb{1}_N\mathbf{v}' = \mathbf{0}$. It follows that the lowest $N - K$ eigenvalues of Σ are the numbers $1 - \rho_k$, each repeated $g_k - 1$ times. In particular,

$$\lambda_N(\Sigma') = 1 - \rho_{\max}.$$

An upper bound on the eigenvalues of Σ follows from (20) and (23):

$$\begin{aligned}
\lambda_1(\Sigma') &\leq \lambda_1(A) + \lambda_1(\delta\mathbb{1}_N) \\
&\leq \max_{1 \leq k \leq K} \{g_k(\rho_k - \delta) + (1 - \rho_k)\} + N\delta \\
&\leq N(1 - \delta) + 1 + N\delta \\
&= N + 1.
\end{aligned}$$

Plugging the matrix Σ' into Algorithm 4 and using the preceding estimates for $\lambda_1(\Sigma')$ and $\lambda_N(\Sigma')$ into (17), we obtain the desired estimate (5) for $\kappa(S)$. $\quad\square$

A.4. Justification of Algorithm 2. Using the spectral theory of self-adjoint Toeplitz operators, it is possible to show that T_g is positive definite and that its eigenvalues satisfy

$$(24) \qquad \frac{1 - \rho}{1 + \rho} \leq \lambda_j(T_g) \leq \frac{1 + \rho}{1 - \rho}$$

for $j = 1, 2, \ldots, g$. We also remark that the preceding bounds are quite sharp in the sense that

$$(25) \qquad \lim_{g \to \infty} \lambda_1(T_g) = \frac{1 + \rho}{1 - \rho}, \qquad \lim_{g \to \infty} \lambda_g(T_g) = \frac{1 - \rho}{1 + \rho}$$

as the size g of the matrix tends to infinity. In light of the explicit bounds (24), a straightforward application of Algorithm 4 yields the following procedure.

To justify the crucial inequalities (24) and the limits (25), first observe that the Toeplitz matrix

$$
(26) \qquad T_g = \begin{pmatrix}
1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{g-1} \\
\rho & 1 & \rho & \rho^2 & \cdots & \rho^{g-2} \\
\rho^2 & \rho & 1 & \rho & \cdots & \rho^{g-3} \\
\rho^3 & \rho^2 & \rho & 1 & \cdots & \rho^{g-4} \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\rho^{g-1} & \rho^{g-2} & \rho^{g-3} & \rho^{g-4} & \cdots & 1
\end{pmatrix}
$$

is simply the upper-left corner of the infinite Toeplitz matrix

$$
(27) \qquad T = \begin{pmatrix}
1 & \rho & \rho^2 & \rho^3 & \cdots \\
\rho & 1 & \rho & \rho^2 & \cdots \\
\rho^2 & \rho & 1 & \rho & \cdots \\
\rho^3 & \rho^2 & \rho & 1 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \ddots
\end{pmatrix}
$$

which induces a linear operator T on the Hilbert space ℓ^2 of all square-summable infinite sequences. Since the ijth entry of T is the $(i-j)$th complex Fourier coefficient of the function $P_\rho(\theta) : [-\pi, \pi] \to \mathbb{R}$ defined by

$$
P_\rho(\theta) = \sum_{n=-\infty}^{\infty} \rho^{|n|} e^{in\theta} = \frac{1-\rho^2}{1 - \rho\cos\theta + \rho^2},
$$

we conclude from Böttcher and Silbermann (1999), Theorem 1.9, that T is a bounded self-adjoint operator whose spectrum equals the range of P_ρ [Halmos (1982), Problem 250] [note that $P_\rho(\theta)$ is the so-called *Poisson kernel* from the study of harmonic functions]. A short calculus exercise reveals that $P_\rho(\theta)$ achieves its maximum value $\frac{1+\rho}{1-\rho}$ at $\theta = 0$ and its minimum value $\frac{1-\rho}{1+\rho}$ at $\theta = \pm\pi$ (see Figure 4), from which we conclude that the spectrum of T is precisely the closed interval $[\frac{1-\rho}{1+\rho}, \frac{1+\rho}{1-\rho}]$. By Böttcher and Silbermann (1999), Proposition 2.17, it follows that the eigenvalues of T_n are also contained in this interval. This establishes the inequalities (24). The limiting behavior (25) follows immediately from Böttcher and Silbermann (1999), Theorem 5.14. □

A.5. Justification of Algorithm 3. By Geršgorin's Disk theorem Horn and Johnson (1990), Theorem 6.11, the largest eigenvalue $\lambda_1(\Sigma_k)$ of Σ_k satisfies

$$
\lambda_1(\Sigma_k) \le 1 + \rho_k + (\rho_k - \tau_k) + \cdots + \big(\rho_k - (g_k - 2)\tau_k\big)
$$

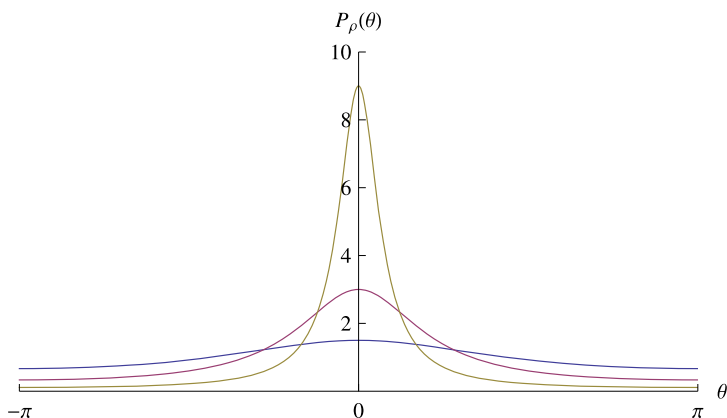$$
= 1 + (g_k - 1)\rho_k - \tau_k \frac{(g_k - 2)(g_k - 1)}{2}.
$$

FIG. 4. *The Poisson kernel $P_\rho(\theta)$ for $\rho = 0.2, 0.5, 0.8$. As $\rho \to 1^-$, the graphs spike sharply at $\theta = 0$ while tending rapidly to zero for θ away from 0. Intuitively, the functions $P_\rho(\theta)$ approximate a point mass (i.e., Dirac δ-function) at $\theta = 0$ as $\rho \to 1^-$.*

This immediately yields (12). On the other hand, it is possible to show that the smallest eigenvalue of Σ_k satisfies

$$(28) \qquad \lambda_{g_k}(\Sigma_k) \geq 1 - \rho_k - \tfrac{3}{4}\tau_k.$$

To be brief, one regards the original $g_k \times g_k$ Toeplitz matrix Σ_k as the upper-left principal submatrix of a $(2g_k - 1) \times (2g_k - 1)$ symmetric circulant matrix, the eigenvalues of which can be exactly computed using well-known techniques [Böttcher and Grudsky (2005), page 32]. A series of elementary but tedious algebraic manipulations and a standard eigenvalue interlacing result [Böttcher and Grudsky (2005), Theorem 9.19] yield the desired inequality (28), from which (13) follows. We thank A. Böttcher, the author of Böttcher and Grudsky (2005), Böttcher and Silbermann (1999), for suggesting this approach to us.

SUPPLEMENTARY MATERIAL

R code (DOI: 10.1214/13-AOAS638SUPP; .r). R code for functions available at http://pages.pomona.edu/~jsh04747/research/simcor.r.

REFERENCES

BARNARD, J., MCCULLOCH, R. and MENG, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statist. Sinica* **10** 1281–1311. MR1804544

BÖTTCHER, A. and GRUDSKY, S. M. (2005). *Spectral Properties of Banded Toeplitz Matrices.* SIAM, Philadelphia, PA. MR2179973

BÖTTCHER, A. and SILBERMANN, B. (1999). *Introduction to Large Truncated Toeplitz Matrices.* Springer, New York. MR1724795

CHO, E. (2009). Inner product of random vectors. *Int. J. Pure Appl. Math.* **56** 217–221. MR2571302

DABNEY, A. R. and STOREY, J. D. (2007). Optimality driven nearest centroid classification from genomic data. *PLoS ONE* **2** e1002.

DAVIES, P. I. and HIGHAM, N. J. (2000). Numerically stable generation of correlation matrices and their factors. *BIT* **40** 640–651. MR1799307

FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7** 179–188.

GOLAN, D. and ROSSET, S. (2011). Accurate estimation of heritability in genome wide studies using random effects models. *Bioinformatics* **27** i317–i323.

GUO, Y., HASTIE, T. and TIBSHIRANI, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* **8** 86–100.

HAFDAHL, A. (2007). Combing correlation matrices: Simulation analysis of improved fixed-effects methods. *Journal of Educational and Behavioral Statistics* **32** 180–205.

HALMOS, P. R. (1982). *A Hilbert Space Problem Book*, 2nd ed. *Graduate Texts in Mathematics* **19**. Springer, New York. MR0675952

HARDIN, J. and WILSON, J. (2009). A note on oligonucleotide expression values not being normally distributed. *Biostatistics* **10** 446–450.

HARDIN, J., GARCIA, S. R. and GOLAN, D. (2013). Supplement to "A method for generating realistic correlation matrices." DOI:10.1214/13-AOAS638SUPP.

HOLMES, R. B. (1989). On random correlation matrices. II. The Toeplitz case. *Comm. Statist. Simulation Comput.* **18** 1511–1537. MR1046545

HOLMES, R. B. (1991). On random correlation matrices. *SIAM J. Matrix Anal. Appl.* **12** 239–272. MR1089158

HONG, S. (1999). Generating correlation matrices with model error for simulation studies in factor analysis: A combination of the Tucker–Koopman–Linn model and Wijsman's algorithm. *Behavior Research Methods, Instruments & Computers* **31** 727–730.

HORN, R. A. and JOHNSON, C. R. (1990). *Matrix Analysis*. Cambridge Univ. Press, Cambridge. MR1084815

HU, R., QIU, X. and GLAZKO, G. (2010). A new gene selection procedure based on the covariance distance. *Bioinformatics* **25** 348–354.

HUANG, S., TONG, T. and ZHAO, H. (2010). Bias-corrected diagonal discriminant rules for high-dimensional classification. *Biometrics* **66** 1096–1106. MR2758497

JOE, H. (2006). Generating random correlation matrices based on partial correlations. *J. Multivariate Anal.* **97** 2177–2189. MR2301633

KAUFMAN, L. and ROUSSEEUW, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York. MR1044997

KIM, K. I. and VAN DE WIEL, M. A. (2008). Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics* **9** 114.

KRAJ, P., SHARMA, A., GARGE, N., PODOLSKY, R. and MCINDOE, R. A. (2008). ParaKMeans: Implementation of a parallelized K-means algorithm suitable for general laboratory use. *BMC Bioinformatics* **9** 200.

KRAUS, J. M. and KESTLER, H. A. (2010). A highly efficient multi-core algorithm for clustering extremely large datasets. *BMC Bioinformatics* **11** 169.

LANGFELDER, P. and HORVATH, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9** 559.

LANGFELDER, P., ZHANG, B. and HORVATH, S. (2008). Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. *Bioinformatics* **24** 719–720.

LEE, S. H., WRAY, N. R., GODDARD, M. E. and VISSCHER, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88** 294–305.

LEE, S. H., DECANDIA, T. R., RIPKE, S., YANG, J., SCHIZOPHRENIA PSYCHIATRIC GENOME-WIDE ASSOCIATION STUDY CONSORTIUM (PGC-SCZ), INTERNATIONAL SCHIZOPHRENIA CONSORTIUM (ISC), MOLECULAR GENETICS OF SCHIZOPHRENIA COLLABORATION (MGS), SULLIVAN, P. F., GODDARD, M. E., KELLER, M. C., VISSCHER, P. M. and WRAY, N. R. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* **44** 247–250.

LEWANDOWSKI, D., KUROWICKA, D. and JOE, H. (2009). Generating random correlation matrices based on vines and extended onion method. *J. Multivariate Anal.* **100** 1989–2001. MR2543081

LIU, X. and DANIELS, M. J. (2006). A new algorithm for simulating a correlation matrix based on parameter expansion and reparameterization. *J. Comput. Graph. Statist.* **15** 897–914. MR2297634

MAHER, B. (2008). Personal genomes: The case of the missing heritability. *Nature* **456** 18–21.

MARSAGLIA, G. and OLKIN, I. (1984). Generating correlation matrices. *SIAM J. Sci. Statist. Comput.* **5** 470–475. MR0740861

MEZZICH, J. E. and SOLOMON, H. (1980). *Taxonomy and Behavioral Science*. Academic Press, San Diego, CA.

MULLER, M. (1959). A note on a method for generating points uniformly on N-dimensional spheres. *Communications of the ACM* **2** 19–20.

NELSON, B. L. and GOLDSMAN, D. (2001). Comparisons with a standard in simulation experiments. *Management Science* **47** 449–463.

NG, C. T. and JOE, H. (2010). Generating random AR(p) and MA(q) Toeplitz correlation matrices. *J. Multivariate Anal.* **101** 1532–1545. MR2609512

PANG, H., TONG, T. and ZHAO, H. (2009). Shrinkage-based diagonal discriminant analysis and its applications in high-dimensional data. *Biometrics* **65** 1021–1029. MR2756489

RAE, G. (1997). A FORTRAN 77 program for generating sample correlation matrices. *Educ. Psychol. Meas.* **57** 189–192.

RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.* **66** 846–850.

TAI, F. and PAN, W. (2007). Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. *Bioinformatics* **23** 3170–3177.

TRITCHLER, D., PARKHOMENKO, E. and BEYENE, J. (2009). Filtering genes for cluster and network analysis. *BMC Bioinformatics* **10** 193.

WITTEN, D. M. and TIBSHIRANI, R. (2009). Covariance-regularized regression and classification for high dimensional problems. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 615–636. MR2749910

YANG, J., BENYAMIN, B., MCEVOY, B. P., GORDON, S., HENDERS, A. K., NYHOLT, D. R., MADDEN, P. A., HEATH, A. C., MARTIN, N. G., MONTGOMERY, G. W., GODDARD, M. E. and VISSCHER, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42** 565–569.

YANG, J., LEE, S. H., GODDARD, M. E. and VISSCHER, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics* **88** 76–82.

YEUNG, K. Y. and RUZZO, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics* **17** 763–774.

ZHANG, X., BOSCARDIN, W. J. and BELIN, T. R. (2006). Sampling correlation matrices in Bayesian models with correlated latent variables. *J. Comput. Graph. Statist.* **15** 880–896. MR2297633

ZHANG, B. and HORVATH, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4** Art. 17, 45 pp. (electronic). MR2170433

ZUBER, V. and STRIMMER, K. (2009). Gene ranking and biomarker discovery under correlation. *Bioinformatics* **25** 2700–2707.

J. HARDIN
S. R. GARCIA
DEPARTMENT OF MATHEMATICS
POMONA COLLEGE
610 N. COLLEGE AVE.
CLAREMONT, CALIFORNIA 91711
USA
E-MAIL: jo.hardin@pomona.edu
          Stephan.Garcia@pomona.edu

D. GOLAN
DEPARTMENT OF STATISTICS AND OR
SCHOOL OF MATHEMATICAL SCIENCES
TEL AVIV UNIVERSITY
69975 TEL AVIV
ISRAEL
E-MAIL: davidgo5@post.tau.ac.il