Computational Statistics and Data Analysis 187 (2023) 107800

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

On graphical models and convex geometry

Haim Bar [a,*,1], Martin T. Wells [b]

[a] Department of Statistics, University of Connecticut, Room 315, Philip E. Austin Building, Storrs, 06269-4120, CT, USA
[b] Department of Statistics and Data Science, Cornell University, 1190 Comstock Hall, Ithaca, 14853, NY, USA

ARTICLE INFO

Article history:
Received 2 June 2022
Received in revised form 6 June 2023
Accepted 7 June 2023
Available online 14 June 2023

Keywords:
Convex geometry
Correlation matrix estimation
Expectation Maximization (EM) algorithm
Graphical models
Grassmann manifold
High-dimensional inference
Network models
Phase transition
Quasi-orthogonality
Two-group model

ABSTRACT

A mixture-model of beta distributions framework is introduced to identify significant correlations among P features when P is large. The method relies on theorems in convex geometry, which are used to show how to control the error rate of edge detection in graphical models. The proposed 'betaMix' method does not require any assumptions about the network structure, nor does it assume that the network is sparse. The results hold for a wide class of data-generating distributions that include light-tailed and heavy-tailed spherically symmetric distributions. The results are robust for sufficiently large sample sizes and hold for non-elliptically-symmetric distributions.

1. Introduction

Even in the age of 'big data' linear regression remains one of the most useful tools available to scientists in all disciplines. The linear regression framework has been built upon the sturdy foundation of the normal theory, and thus offers powerful inferential and prediction tools. In many applications the underlying assumptions of regression appear to be reasonable, namely, that the relationship between the predictors and the outcome is linear, and the measurement errors are independently and identically normally distributed and are uncorrelated with the predictors. However, in order for this theoretical result to be applicable the number of predictors, P, cannot exceed the sample size, n. Using conventional notation, where the outcome (response) variable is y, and we assume that it is a linear function of P predictors, x_j, plus some random (Gaussian) noise, $\epsilon \sim N(0, \sigma^2)$, that is,

$$y = \beta_0 + \sum_{j=1}^{P} \beta_j x_j + \epsilon \,.$$

[*] Corresponding author.
E-mail address: haim.bar@uconn.edu (H. Bar).

[1] The R package which implements the beta-mixture model is available from github, at https://github.com/haimbar/betaMix and data and code files used in this paper can be found at https://github.com/haimbar/betaMixFiles.

https://doi.org/10.1016/j.csda.2023.107800

Using matrix notation, the parameter vector is estimated by the ordinary least squares formula $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y$. If $P > n$, routine estimation of the regression parameters is not possible since the inverse of the matrix $X'X$ does not exist, and we say that $\boldsymbol{\beta}$ is unidentifiable. Even if $n > P$, inference about $\boldsymbol{\beta}$ may be impractical when P is sufficiently large because standard errors are often large and the width of the confidence interval grows with P. For example, Hotelling's T^2 yields confidence intervals with width which is proportional to $\sqrt{[P(n-1)F_{P,n-P,\alpha}]/[n(n-P)]}$.

To deal with the fact that many modern applications involve a large number of putative predictors and often a modest sample size, statisticians had to develop variable selection methods capable of identifying the true predictors, while limiting the number of irrelevant predictors from being included in the regression model. Arguably, most famous among such methods is the LASSO (Tibshirani, 1996). Such methods assume sparsity, and require that $\log(P)/n = o(1)$. For example, van de Geer et al. (2014) denote the active set of variables by $S_0 = \{j : \beta_j \neq 0\}$ and its cardinality by s_0. To prove their main result (Theorem 2.2) they further require that the n samples are i.i.d. Gaussian, $X'X$ has a strictly positive smallest eigenvalue, and that $(X'X)^{-1}$ is row-wise sparse: $\max_j s_j = |\{k \neq j : (X'X)^{-1} \neq 0\}| = o(n/\log(P))$. Under these assumptions van de Geer et al. (2014) derive asymptotic confidence intervals for the LASSO estimator, $\hat{b}_{LASSO}$ and obtain an $o_{\mathbb{P}}(1)$ estimator for the precision matrix, Σ^{-1}. We elaborate on other related methods and results in Section 5.

While applying the linear model with Gaussian errors to high-dimensional problems has been a natural step which yielded extraordinary advances, both in theory and in applications, it also required making strong assumptions, including sparsity of the mean vector and rows of the covariance (or precision) matrix. In the original, small-P setting, the assumption that the predictors are uncorrelated seemed reasonable, but correlation between columns of X is inevitable when P is large, and for the most part, the approaches to deal with such correlations have relied on a somewhat ad-hoc two-stage approach, where in the first step a dimension reduction is performed (e.g., by clustering) in order to restore at least in part the validity of the requirement that $X'X$ is invertible.

Another motivation for extending the linear model framework to the large-P setting has been the interpretability of the results, namely, that 'a unit change in some predictor, x_j is associated with β_j units increase in y'. However, in many cases involving a large number of predictors there is no reason to think that the relationship between Y and X is linear. For example, a quantitative trait may depend on the expression of many genes in an intricate way, so that we cannot use statements like 'holding all other variables constant', and we cannot draw conclusions like 'an increase in expression of gene j is associated with β_j units increase in the trait', because a change in the expression of that gene may not occur without a simultaneous change in many other genes. For the same reason the $\boldsymbol{\beta}$-sparsity assumption may not be valid, and the covariance matrix may not be sparse, either. It is quite possible, and in fact common, that a trait is associated with hundreds or even thousands of genes. Such is the case if genes form a highly connected network, which may be necessary because the trait requires the production of many different proteins or it may be evolutionary beneficial as a way to protect against mutations. It may also be the case that the assumption of underlying low dimensionality is not valid. For example, if the predictors have an auto-regressive structure, $AR(m)$, as is the case if the predictors represent repeated measurements (for example, daily log-returns of stocks). It is possible to reduce the dimensionality by taking representative predictors, but doing so results in loss of information about the most prominent feature of the data, namely, its $AR(m)$ structure.

Since in high-dimensional setting neither a linear relationship between X and Y, nor the uniqueness of $\boldsymbol{\beta}$, nor its sparsity is some laws of Nature, but rather, mathematically convenient assumptions, we propose to change the perspective and consider obtaining the whole network structure as the main objective, where nodes represent variables and edges represent strong associations between pairs of variables. While the graphical models in the literature aim to do just that, they almost always rely on sparsity assumptions. Detecting edges in a network via the graphical LASSO involves designating each predictor in its turn as the response and regressing on the other $P-1$ variables. Our approach, which is described in the next section, does not require sparsity assumptions, and attains the estimated network structure in a single step. We may, however, choose to treat one variable as a 'response', and use our method to perform variable selection by identifying all the nodes (variables) connected via an edge to the response node in the graph.

Like LARS (Efron et al., 2004), our method uses partial correlations, but in a very different way. LARS is a stepwise process in which each step involves updating the coefficients of the regression and the residuals, and recalculating the correlations between the remaining predictors and the residuals. Our method calculates all the pairwise correlations just once. LARS is a variable selection method, used when the goal is to find which predictors are associated with a response variable, while our method finds all the connections between all variables simultaneously. Furthermore, LARS relies on $\boldsymbol{\beta}$-sparsity, while our method does not. There is, however, a more subtle difference between our method and LARS. LARS uses an inclusion criterion, adding variables to the model while a cumulative threshold has not been exceeded, and that threshold depends on a tuning parameter, usually obtained via cross validation. In contrast, our method is based on an exclusion criterion which uses the distribution of pairwise correlations under the null hypothesis (which is discussed in the next section). Thus, our method provides an inferential framework, which is used to control the error rate even in the presence of massive multiple testing.

Our method is based on ideas and results from convex geometry, some of which may seem counter intuitive at first glance. The relevant theorems are stated in Section 2, but for a comprehensive (and very enjoyable) introduction to convex geometry see Ball (1997) and Blum et al. 2020, Chapter 2. The latter reference contains applications to modern data science challenges. The key to our method is 'flipping' the roles of variables and observations and treat the data as P points in $\mathbb{R}^n$, so that each predictor is characterized by n samples. The classical approach views data as n points in $\mathbb{R}^P$, and in the high dimensional setting where $P > n$, all the n points lie on a low-dimensional hyperplane in $\mathbb{R}^P$. This degeneracy causes difficulties for classical statistical methods. However, if we view the data as P vectors in $\mathbb{R}^n$, then such degeneracy problem no longer exists. This is the mathematical structure that underlies the asymptotics in the high-dimensional-low-sample-size framework developed by Hall et al. (2005).

Using a key distribution result (see Theorem 1 in Sec. 2) we know that pairs of uncorrelated predictors will be nearly perpendicular with high probability if n is sufficiently large. The null distribution of the squared sine of angles between random pairs is $Beta(\frac{n-1}{2}, \frac{1}{2})$ and we could use this fact to detect edges in a graph, while controlling the error rate as frequentists. However, we propose an empirical Bayes approach, and use a mixture of two beta distributions where the nonnull component is $Beta(a, b)$ and a, b are estimated from the data. This allows us to get more power, and check model adequacy. It also allows us to adapt the model to situations where the samples are not i.i.d., and we do so by replacing the $\frac{n-1}{2}$ parameter in the null component by $\frac{\nu-1}{2}$, where $\nu \in (1, n)$ is the 'effective sample size'. In many cases, the i.i.d. assumption is unrealistic, and using $Beta(\frac{n-1}{2}, \frac{1}{2})$ as the null distribution will lead to many false discoveries.

Furthermore, the convex geometry literature establishes a remarkable asymptotic relationship between n and P. Specifically, Theorem 2 in Sec. 2 states that the number of random lines that go through the origin in $\mathbb{R}^n$ and are approximately perpendicular, grows exponentially with n. In the context of graphical models this means that as the sample size grows, the number of possible null edges grows exponentially with n (equivalently, $\log(P) = O(n)$). Combined with Theorem 1, this means that our method can detect null edges in large graphs with $P \gg n$, while controlling the false detection rate.

Dempster (1972) first studied concentration graphical models as a tractable approach to multivariate dependence structures. These models reflect conditional dependencies in a multivariate probability distribution. In the Gaussian case, these models induce zeros in the inverse covariance matrix and correspond to natural exponential families. For example, Meinshausen and Bühlmann (2006); Friedman et al. (2008); Yuan and Lin (2007); Peng et al. (2009); Khare et al. (2015) develop estimation methods via a penalized Gaussian log-likelihood for a large, sparse precision matrix. In this article we consider the setting of estimating covariance graphical model that represents variables as nodes and marginal dependencies between variables as edges. A covariance graph is the corresponding graphical model for marginal independencies, not the conditional independence in concentration graphical models. Although the specification of covariance graphical models is intuitive, their analysis and estimation are somewhat complex (Cox and Wermuth, 1996; Drton and Richardson, 2008; Khare and Rajaratnam, 2011). The primary difficulty is that these models give rise to curved exponential families. The zero entries in the covariance matrix Σ translate into complicated restrictions on the entries of the natural parameter Σ^{-1} matrix, so a covariance graphical model cannot be viewed as a concentration graphical model.

In Section 2 we describe some mathematical background for convex geometry, and our mixture model. In Section 3 we demonstrate the betaMix model in simulations, and in Section 4 we show how the method can be applied to a wide range of big-data applications. In Section 5 we review related work, and we conclude with a brief discussion in Section 6.

2. Method

2.1. Background – convex geometry results

Most intuition about geometry which is based on two and three dimensions can often be misleading in high dimensions. Specifically, the field of statistics leans heavily on various orthogonal decompositions through the notion of Pythagorean right angles (in Greek, *ortho gonia*). From classical linear algebra perspective, the minimal number of orthogonal basis vectors needed to specify an object in a Euclidean space defines its orthogonal dimension. Recent work in convex geometry (Kainen and Krková, 2020) seeks to extend the notion of dimension to ϵ-quasi-orthogonal dimension of $\mathbb{R}^n$. The concept of quasi-orthogonal dimension is obtained by relaxing exact orthogonality so that angular distances between unit vectors are constrained to a closed symmetric interval about $\pi/2$. For $\epsilon \in [0, 1]$ a subset of $A \subset \mathcal{S}^{n-1}$ is a ϵ-quasi-orthogonal subset if $x \neq y \in A \Rightarrow |\langle x, y\rangle| \leq \epsilon$. The ϵ-quasi-orthogonal dimension of $\mathbb{R}^n$ is defined as $\dim_\epsilon(n) := \max\{|X| : X \subset \mathcal{S}^{n-1}, x \neq y \in X \Rightarrow |\langle x, y\rangle| \leq \epsilon\}$, where $|X|$ denotes the cardinality of the set X. Equivalently, the maximum number of nonzero vectors whose pairwise angles lie in the interval $[\arccos(\epsilon), \arccos(-\epsilon)]$ centered at $\pi/2$ or the maximum cardinality of an ϵ-quasi-orthogonal subset of $\mathbb{R}^n$.

Kainen and Krková (2020) showed that an exponential number of such quasi-orthogonal vectors exist as the Euclidean dimension increases, specifically $\dim_\epsilon(n) \geq \exp(n\epsilon^2/2)$. The argument for the existence of such large quasi-orthogonal sets comes from packing spherical caps into the surface of $\mathcal{S}^{n-1}$. The spherical caps consist of all points on the sphere within a fixed angular distance from some point, that for $y \in \mathcal{S}^{n-1}$ and $\epsilon > 0$, $\mathcal{C}(y, \epsilon) := \{x \in \mathcal{S}^{n-1}, |\langle x, y\rangle| \geq \epsilon\}$. It is known (Ball, 1997, p. 11) that the Lebesgue measure of $\mathcal{C}(y, \epsilon)$ is bounded above by $\exp(-n\epsilon^2/2)$. The proof of the $\dim_\epsilon(n)$ bound then follows from a maximum packing argument. It is of special note that the two bounds are reciprocals. The $\exp(-n\epsilon^2/2)$ upper bound on the Lebesgue measure of $\mathcal{C}(y, \epsilon)$ is quite counter-intuitive since for any fixed ϵ, the bound becomes very small as n increases. Hence, in high dimension, most of the area of the sphere lies very close to its equator. This is an incidence of the concentration of measure phenomena. Donoho (2000) notes that increases in dimensionality can often be helpful to the asymptotic analysis. The *blessings of dimensionality* include the concentration of measure phenomenon where certain random fluctuations are very well controlled in high dimensions.

A random vector $V \in \mathbb{R}^P$ is spherically symmetric if, for any orthogonal transformation $\mathcal{O}$, $\mathcal{O}V$ has the same distribution as V. The class of spherical distributions generalize the multivariate normal distribution and includes the Laplace,

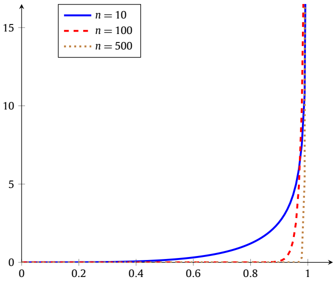

Fig. 1. $X \sim Beta\left(\frac{n-1}{2}, \frac{1}{2}\right)$, for $n = 10$ (solid), 100 (dashed), and 500 (dotted).

logistic, symmetric stable, and t- distributions as well as the family of scale mixtures of normal distributions. The connection between spherical symmetry and uniform distributions on the unit sphere is through the equivalence: a random vector $V \in \mathbb{R}^P$ has a spherically symmetric distribution if and only if V has the stochastic representation $V \overset{\mathrm{D}}{=} \|V\| U$ where $\|V\| = \langle V, V\rangle^{1/2}$ is the Euclidean norm, $Pr[\|V\| = 0] = 0$, U and V are independent, and U is uniformly distributed on the unit sphere. Note that $V/\|V\|$ has the same distribution for the entire family of spherically symmetric distributions. If $\mu \in \mathbb{R}^P$, Σ is a positive definite $P \times P$ matrix, and V is spherically symmetric, then $X \overset{\mathrm{D}}{=} \mu + \Sigma^{1/2} V \in \mathbb{R}^P$ has an elliptically symmetric distribution, $\mathcal{E}_P(\mu, \Sigma)$. See Chapter 4 of Fourdrinier et al. (2018) for further details on spherical and elliptical distributions. We first assume the data $\{X_i\}_{i=1}^n$ follow an elliptically symmetric distribution. Then, in Section 2.4 we discuss an extension of our method when the elliptical symmetry assumption does not hold.

Our method relies on the following theorem from the convex geometry literature which has appeared in many places (e.g., Muirhead (1982) and Watson (1983)) but we cite Theorem 1.1 in Frankl and Maehara (1990):

Theorem 1. *Let K be a fixed 1-space (line) in $\mathbb{R}^n$, and let L be a random l-space in $\mathbb{R}^n$. Let θ be the angle between K and L. The random variables $\cos^2\theta$ and $\sin^2\theta$ have the beta distributions $Beta(1/2, (n-l)/2)$ and $Beta((n-l)/2, 1/2)$, respectively.*

In the following we let L also be a line ($l = 1$), in which case Theorem 1 yields

$$Z \overset{\mathrm{def}}{=} \sin^2\theta \sim Beta\left(\frac{n-1}{2}, \frac{1}{2}\right). \tag{1}$$

Using an asymptotic approximation for the beta function and bounding the beta probability density function in (1), Theorem 3.1 in Frankl and Maehara (1990) gives an approximation for the cumulative distribution function of θ

$$Pr[\theta \le \alpha] = Pr[Z \le \sin^2\alpha] = [(\pi(n-1)/2)^{1/2} \cos\alpha]^{-1} (\sin\alpha)^{n-1} + o(1), \tag{2}$$

which holds for any $\alpha \in (0, \pi/2)$, where $o(1) \to 0$ as $n \to \infty$.

A consequence of (2) is that two random vectors are approximately perpendicular with high probability, if the dimension of the space is sufficiently large. If n is large enough, the distribution of Z is highly concentrated close to 1. This phenomena is what is anticipated from the concentration of measure results in (Ball, 1997, p. 11). For example, Fig. 1 shows the distribution of Z for $n = 10$, 100 and 500. Even with $n = 10$ it is very unlikely that z will be less than, say, 0.6, which means that the probability of two *random* vectors in $\mathbb{R}^{10}$ will be correlated by chance, is very small. In other words, even for moderate values of n, $\mathbb{R}^n$ is a pretty big space which allows for a lot of random vectors to be sufficiently far from one another (in terms of chordal-based distance). Formally, citing Theorem 3.2 from Frankl and Maehara (1990) with a slight change in notation, we have the following:

Theorem 2. *For any $\alpha \in (0, \frac{\pi}{2})$, there exist more than $m_\alpha(n) = \sqrt{\frac{\pi(n-1)}{2}} \cdot \frac{\cos\alpha}{(\sin\alpha)^{n-1}}$ lines in $\mathbb{R}^n$ going through the origin such that any two of them determine an angle greater than α.*

The number of lines going through the origin that can be drawn randomly in $\mathbb{R}^n$ such that the angle between each pair is at least α grows exponentially with n. This phenomena is what is foreseen from quasi-orthogonality results in Kainen and Krková (2020). Theorem 2 is illustrated in Fig. 2, which shows the asymptotically linear relationship between $\log_{10} m_\alpha(n)$ and n for three angles: $\alpha = \pi/4, \pi/3$ and $0.8\pi/2$.

We emphasize that Theorem 2 refers to *any* pair among at least $m_\alpha(n)$ randomly chosen lines. For example, with $n = 100$ there are at least 12,713,167 lines that can be drawn randomly so that the angle between any two lines is at least 60°.

In Appendix A we provide a more general convex geometry theoretical framework which can serve as the basis for interesting extensions to our model. For example, using the proper definition for an angle between planes it may be possible to develop similar estimation methods for repeated measurement models. We leave such possible extensions for future work.

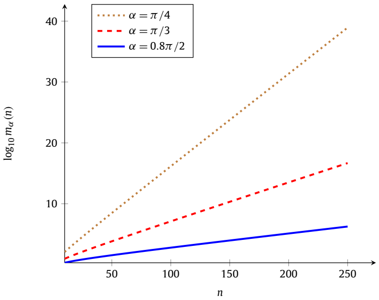

Fig. 2. The logarithm of the number of lines in $\mathbb{R}^n$ going through the origin such that any two of them determine an angle greater than $\frac{\pi}{4}$ (dotted), $\frac{\pi}{3}$ (dashed), and $\frac{0.8\pi}{2}$ (solid).

In the following subsection, we apply the distribution theory from Theorem 1 to graphical model construction. The underlying rationale of the methodology is that the assumption of elliptical symmetry reduces (by scale and rotational invariance) the sampling model assumption to that of a uniform distribution on the unit sphere (Fourdrinier et al., 2018; Muirhead, 1982). Once on the unit sphere, one can use the geometric results in Theorem 1 to deduce the distribution of inner products (i.e., correlations), which are in turn connected to the $\sin^2$ (or $\cos^2$) of the angle between vectors on the unit sphere (Frankl and Maehara, 1990). This distribution theory based on these angular measures is then applied to assess the magnitude of the inner products to define when predictors are sufficiently non-orthogonal under some probabilistic constraints that control the error rate of edge detection.

2.2. The beta-mixture method

Consider a situation in which we obtain P quantitative characteristics (predictors) of n random subjects, and assume that P is large, possibly much larger than n. Such datasets have become common in recent years due to advances in high-throughput technologies which allow researchers to obtain, for example, RNA sequencing data for tens of thousands of genes. Many graphical model methods regard the data as n points in $\mathbb{R}^P$, and apply a variable selection method to a multivariate linear (normal) model in order to detect strong relationship between pairs of predictors.

Our approach is different in that we consider the data as P points in $\mathbb{R}^n$, and rather than representing each subject by P quantitative characteristics, we view each predictor as a point which is determined by a sample of n subjects. 'Null' predictors correspond to randomly drawn points in $\mathbb{R}^n$. Using Theorem 1, pairs of null predictors will be nearly perpendicular with high probability, if n is sufficiently large. Therefore, our approach to detecting edges in the graphical model is to exclude all edges corresponding to pairs of approximately perpendicular vectors (predictors) in $\mathbb{R}^n$. We use the distribution of randomly drawn vectors to establish statistical properties, and to control the probability of erroneously detecting an edge. The known distribution of null edges allows us to use either a frequentist or a Bayesian inferential procedure.

A frequentist method: Let (X_k, X_ℓ) be the jth pair of predictors where $k < \ell \in \{1, \ldots, P\}$. Let θ_j be the angle between them, $j = 1, \ldots, P(P-1)/2$ and let $z_j = \sin^2\theta_j$. The null hypothesis is: all the pairs of features are uncorrelated, and the alternative is *not all* the pairs are uncorrelated. Since we are interested not only in rejecting the null hypothesis, but also in detecting which pairs of features are correlated (that is, the nonnull edges) we use the quantile approximation for the rth order statistic, for which we have the following asymptotic approximation

$$E(Z_{r:K}) \sim Q\left(\frac{r}{K+1}\right), \tag{3}$$

where $K = P(P-1)/2$ is the total number of edges (e.g. David 1981, Chapter 4).

For example, suppose that $P = 500$ and $n = 70$. Denote the δ quantile of the $Beta((n-1)/2, 1/2)$ distribution by Q_δ. If we set $\delta = 1/(500 \cdot 499/2)$ we obtain $Q_\delta(34.5, 0.5) \approx 0.75$, and the quantile approximation in (3) implies that under the null we expect no more than one z_j to be smaller than 0.75. Translating this threshold to degrees and correlation coefficients, it means that we include an edge in the graph if the angle (modulus 90) between the corresponding pair of vectors is less than 60°, which is equivalent to a correlation coefficient of at least 0.5 between the two predictors.

An empirical Bayes method: An alternative to the screening methodology is to use an empirical Bayes two-group approach. We define a mixture model (hereafter called the *betaMix* model),

$$\ell(z_j) = p_0 f_0(z_j) + (1 - p_0) f_{a,b}(z_j),$$

where p_0 is the probability that the jth pair of vectors follows the null distribution according to Theorem 1, and the distributions of the null and nonnull components are, respectively:

$$f_0(z_j) = \frac{1}{B(\frac{n-1}{2}, \frac{1}{2})} z_j^{\frac{n-1}{2}-1} (1 - z_j)^{-\frac{1}{2}} \text{ and}$$

$$f_{a,b}(z_j) = \frac{1}{B(a,b)} z_j^{a-1}(1-z_j)^{b-1}\,,$$

for some unknown parameters a, b. Note that the alternative distribution, $f_{a,b}(z)$, is very flexible. We do not impose any restrictions on a and b, except that they have to be positive.

In order to estimate the model we define random indicator variables m_{0j}, so that $m_{0j} = 1$ if the jth pair of predictors is approximately perpendicular. We assume that

$$m_{0j} \sim Ber(p_0)\,. \tag{4}$$

Since the mixture indicator variables are latent we use the EM algorithm (Dempster et al., 1977) to estimate the model parameters. In the E-step, at the t-th iterate $m_{0j}^{(t)}$ is estimated by the posterior mean,

$$\hat{m}_{0j}^{(t)} = \frac{\hat{p}_0^{(t-1)} f_0(z_j)}{\hat{p}_0^{(t-1)} f_0(z_j) + (1-\hat{p}_0^{(t-1)}) f_{\hat{a}^{(t-1)},\hat{b}^{(t-1)}}(z_j)}\,,$$

where $\hat{p}^{(t-1)}, \hat{a}^{(t-1)}, \hat{b}^{(t-1)}$ are parameter estimates from the previous, $t-1$ iterate of the EM algorithm. In the M-step we obtain the maximum likelihood estimates for a and b numerically: we update the t-th iterate estimates $\hat{a}^{(t)}, \hat{b}^{(t)}$ by finding

$$\arg\max_{a,b} \sum_j \hat{m}_{0j}^{(t)} \log f_{a,b}(z_j)\,.$$

The t-th iterate (maximum likelihood) estimate of p_0 is the mean of the Bernoulli random variables: $\hat{p}_0^{(t)} = \bar{m}_{0.}^{(t)}$. This process is repeated iteratively until convergence is achieved.

We say that an edge in the graph exists if the posterior null probability (under f_0) is smaller than some threshold, $\hat{m}_{0j} < \tau$. Note that the posterior null probability (4) is identical to Efron's definition of local false discovery rate (Efron, 2008). Replacing the p.d.f.'s in (4) with the cumulative distribution functions, (Efron, 2008) gives a Bayesian interpretation to the frequentist (Benjamini-Hochberg) false discovery rate (FDR), which is equivalent to declaring an edge to be in the nonnull group if its tail-area posterior null probability is no greater than some q, the desired FDR threshold. Since the null distribution is determined by the sample size, we can set τ so that $Q_\tau((n-1)/2, 0.5) = q$.

2.3. Adaptation to dependent samples

Parameter estimates for the beta-mixture model in the previous sub-section were derived under the assumption that the n samples are independent, but it may not always be a reasonable assumption. When this assumption is invalid it is possible, at least conceptually, to incorporate a certain dependence structure into the model and derive a distribution for the null set. One conceivable way to achieve this is to employ a random effects model which accounts for within-cluster correlations, and estimate the intraclass correlation coefficient (ICC). The ICC is often used to determine the *effective* sample size (ESS) of an experiment, and when the ICC is large, the ESS is much smaller than the actual sample size.

Rather than specifying a possibly incorrect dependence structure we propose a different approach, and instead we model the *consequence* of dependence among observations, namely, a *smaller effective sample size*. Let $\nu \le n$ be the unknown ESS, and let the null distribution be

$$f_0(z_j) = \frac{1}{B(\frac{\nu-1}{2}, \frac{1}{2})} z_j^{\frac{\nu-1}{2}-1}(1-z_j)^{-\frac{1}{2}}\,.$$

Note that the second parameter of the null distribution is still 1/2 because the null hypothesis is still that the P vectors are drawn randomly, which means that the results from Frankl and Maehara (1990) apply, but because observations may be dependent the dimension may be less than the sample size. The ESS ν is estimated from the data in the M-step of the EM algorithm. The estimating equations for the three parameters, a, b, and ν are given in the Supplementary Material.

2.4. Non elliptically-symmetric distributions

Although the proof for Theorem 1 relies on the data being drawn from a elliptically symmetric distribution, the betaMix method is much more broadly applicable if the sample size is sufficiently large. The rationale behind the applicability of betaMix to elliptically symmetric distributions was discussed briefly in the Introduction. Using rotation invariance properties it can be shown that the elliptically symmetric distribution assumption can be reduced to the isotropic Gaussian distribution.

However, as a reviewer of this paper pointed out, even when the elliptical symmetry assumption does not hold, the distribution of the Z statistics under the null is still, asymptotically, approximately $Beta((n-1)/2, 1/2)$, or equivalently, $1-Z$ is approximately $Beta(1/2, (n-1)/2)$. The result below shows that both the squared cosine of the angle between general random vectors in $\mathbb{R}^n$ and $1-Z$ have a limiting χ_1^2 distribution. In Section 3 we give some simulation evidence that when the sample size is large the betaMix methodology is applicable in cases where the elliptical symmetry assumption does not hold.

Table 1
Simulation results.

	Corr. Structure	ρ	N	P	Settings	TPR	FDR
1	Clusters	0.3	200	500	Cluster size 25	0.59	1.5e-3
2	Clusters	0.9	200	500	Cluster size 25	1.00	4.4e-5
3	Clusters	0.3	200	1000	Cluster size 100	0.66	1.2e-3
4	Clusters	0.9	200	1000	Cluster size 100	1.00	0.00
5	Clusters	0.3	200	1000	Cluster size 500	0.83	5e-4
6	Clusters	0.9	200	1000	Cluster size 500	1.00	0.00
7	Clusters	0.3	500	1000	Cluster size 500	0.99	1.5e-5
8	Clusters	0.9	500	1000	Cluster size 500	1.00	0.00
9	Band	0.3	200	500	Width 150	0.34	5.8e-4
10	Band	0.9	200	500	Width 150	0.92	2.7e-3
11	Band	0.3	200	1000	Width 150	0.38	1.5e-3
12	Band	0.9	200	1000	Width 150	0.93	1.4e-3
13	Band	0.3	200	1000	Width 30	0.38	1.7e-3
14	Band	0.9	200	1000	Width 30	0.98	5.7e-4
15	Band	0.3	500	1000	Width 150	0.86	8.6e-4
16	Band	0.9	500	1000	Width 150	1.00	1.6e-3
17	Band	0.3	500	1000	Width 30	0.91	3.8e-4
18	Band	0.9	500	1000	Width 30	1.00	1.6e-3
19	Cycle	0.3	200	500	Length 25	0.43	4.8e-3
20	Cycle	0.9	200	500	Length 25	1.00	2.2e-3
21	Cycle	0.3	200	1000	Length 50	0.32	1.9e-3
22	Cycle	0.9	200	1000	Length 50	1.00	1.1e-3
23	Cycle	0.3	500	1000	Length 50	0.98	1.2e-3
24	Cycle	0.9	500	1000	Length 50	1.00	4.7e-3

Proposition 1. *Let V_1 and V_2 be two random uncorrelated identically distributed vectors in $\mathbb{R}^n$ with finite variance.*

(i) Let θ be the random angle between V_1 and V_2 and $Y_n = \cos^2\theta$. Then when n is large, the distribution of $nY_n \xrightarrow{D} \chi_1^2$.

(ii) Suppose W_n has a $Beta(1/2, (n-1)/2)$ then when n is large, the distribution of $nW_n \xrightarrow{D} \chi_1^2$.

Proof. For part (i) note that for two uncorrelated random variables V_1, V_2 it can be shown by using Slutsky's theorem and the delta method (Serfling, 2009) that

$$nY_n = n\cos^2\theta = n\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{V_{1i}V_{2i}}{\|V_1\|\|V_2\|}\right\}^2 \xrightarrow{D} \chi_1^2 .$$

For part (ii), recall a random variable $W_n \sim Beta(1/2, (n-1)/2)$ can be written as

$$W_n \overset{D}{=} \frac{G}{G+H_n},$$

for independent $G \sim Gamma(1/2, 1)$ and $H_n \sim Gamma((n-1)/2, 1)$. From the additive property of the gamma distribution $H_n' = G + H_n \sim Gamma(n/2, 1)$. By the weak law of large numbers, for i.i.d. $G_i \sim Gamma(1/2, 1)$,

$$\frac{1}{n}H_n' = \frac{1}{n}\sum_{i=1}^{n} G_i \xrightarrow{\mathbb{P}} E(G_i) = \frac{1}{2}.$$

Therefore by Slutsky's theorem, in the limit, $nW_n \xrightarrow{D} 2G \sim Gamma(1/2, 1/2) = \chi_1^2$. □

3. Simulations

The method described in the previous section has been implemented as an R package called `betaMix`. We simulated data with different numbers of predictors, sample sizes, and correlation structures, and in each configuration we ran the `betaMix` function and evaluated the goodness of fit of the mixture-model and the ability of the method to recover the true correlation structure in terms of true- and false-positive edges that it detects. Table 1 shows representative results from our simulations. In all cases, data were generated from a multivariate normal distribution with P variables and N samples, and we varied the correlation structure of the normal distribution. For example, in configurations 1-8 the correlation matrix had a block-diagonal clustered structure, with cluster sizes set to 25, 100, or 500. All the variables within a cluster were correlated, and the correlation coefficient between each pair in a cluster was set to either low (0.3) or high (0.9). Pairs not

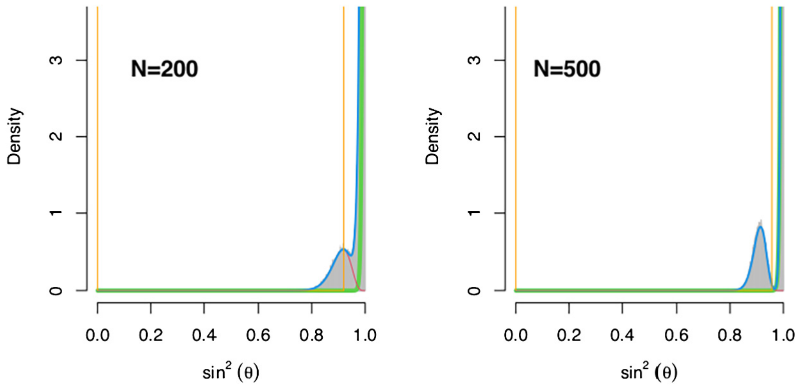

Fig. 3. Fitted distribution - simulated data: $P = 1000$, block-diagonal clustered correlation structure with cluster size 50, $\rho = 0.3$. Left $N = 200$, Right $N = 500$. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

belonging to the same cluster were set to be uncorrelated. In scenarios 9-18 we used a band matrix, with band-width set to either 30 or 150. Pairs of variables which correspond to cells within the band were set to be correlated (low or high), whereas outside the band the correlations were set to 0. The third family of examples in the table (lines 19-24) consists of correlation matrices with a block-diagonal cycle structure, with cycle length being 25 or 50. For example, in scenario 19 we have twenty 25 × 25 blocks along the diagonal so that the first variable is correlated with the second, the second with the third, etc., and the 25th variable is correlated with the first variable in the block. Six additional correlation structures are described in the Supplementary Material, for a total of 124 simulation configurations. Each configuration was used to generate 30 datasets, and we count the number of correctly and falsely detected edges in each run. An edge is correctly detected if and only if the corresponding cell in the actual correlation matrix was not zero.

The results show that as N increases the probability of detecting a true edge approaches 1. The number of false positive edges is very small, since our model allows to control the error rate. We set the FDR threshold to 0.01, and in all cases the observed error rate was lower. When the magnitude of the correlation coefficient is high, the power of our method increases, and even when N is much smaller than P, if ρ is 0.9 our method detects well over 90% of the edges in all scenarios. Even with a modest correlation coefficient our method has very good power, as a result of the mixture model which allows to borrow strength across pairs of variables. Note that our model does not assume sparsity, and performs just as well in the non-sparse cases. For instance, in scenarios 5-8 $P = 1000$ and there are two clusters, each with 500 variables.

In Fig. 3 we show the fitted model for two configurations in the block-diagonal clustered correlation structure with cluster size 50. In both cases, the correlation coefficient was set to $\rho = 0.3$. We used $N = 200$ (left) and $N = 500$ (right). The histograms in gray show the observed distribution of the z_j's. The green and red curves represent the fitted null and non-null distributions, respectively. The blue curve depicts the fitted mixture, which fits the data very well. The vertical orange lines show the range in which z_j is deemed small enough so that the corresponding pair is said to be strongly correlated. In this scenario, when $N = 200$ the threshold was found to be 0.92, and when $N = 500$ any pair with $z_j < 0.96$ is declared to be non-null.

In Fig. 4 we show a receiver operating characteristic (ROC) curve for three sample sizes: $N = 100$ (solid red line), 200 (dashed green), and 300 (dotted blue). We use $P = 500$, a block-diagonal clustered correlation structure with cluster size 25, and correlation coefficient $\rho = 0.3$. For each detection threshold between 0 to 1 in steps of 0.01 we calculated the true- and false-positive rates. The diamond-shaped points show the true- and false-positive rates when we use the threshold as defined in Section 2.2, with $q = 0.0001$. That is, we find τ such that $Q_\tau((N-1)/2, 0.5) = 0.0001$. These points show that the selected threshold controls the error rate very close to the desired level, with false positive rate being 0.0002, 0.0003, 0.0002 for $N = 100$, 200, 300, respectively. The numbers of falsely detected edges are 23, 45, and 27, while the numbers of true positives are 1,662, 3,269, and 4,427, out of 6,000 edges.

We also performed additional simulations in order to compare our method with other approaches. To do that, we used the R package ‘huge’ (Jiang et al., 2021) which includes the following three estimation methods: (1) the Meinshausen-Bühlmann method (mb) (Meinshausen and Bühlmann, 2006) (2) the graphical lasso (glasso) (Friedman et al., 2008), and (3) correlation thresholding (ct) (Mazumder and Hastie, 2012). In all cases (including when using our method) we used the default settings. The default threshold of the posterior probability of the null component with betaMix is 0.05 (that is, if $\hat{m}_{0j} < 0.05$ then the jth edge is determined to be in the graph). The methods in the huge package use tuning parameters to control sparsity, so we used the ‘huge.select’ function to automatically choose the best fit from the range of tuning parameters. The mb and glasso methods fit P regression models with the lasso, each time using a different variable as the response, which yields a directional graph because it is possible that node j will be found as a strong predictor for node i but not the other way around. The package offers two options – the *and* option considers i and j to be connected if both were found to be a strong predictor of the other, and the *or* option will detect an edge if either one is a predictor of the other. The latter (less conservative) is the default. In these simulations we used $P = 2,000$ variables, $\rho = 0.3$ (the correlation

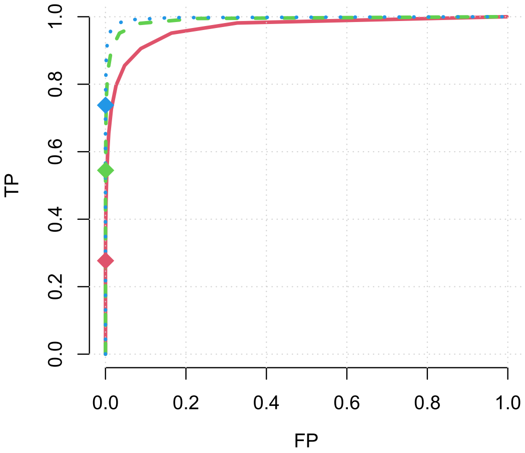

Fig. 4. ROC curve - simulated data: $P = 500$, block-diagonal clustered correlation structure with cluster size 25, $\rho = 0.3$. $N = 100$ (solid line), 200 (dashed), and 300 (dotted).

Table 2
The median true positives and false positives for the correlation thresholding (ct), Meinshausen-Bühlmann (mb), and betaMix (bx) methods with the cluster configuration (total number of edges in this configuration is 19,000, edge density of 0.01).

Method		$N = 100$	$N = 200$	$N = 400$
ct	TP	15,108	18,408	18,966
	FP	47,928	23,677	2,076
mb	TP	115	2,214	7,630
	FP	0	0	0
bx	TP	2,335	11,736	18,215
	FP	53	106	19

between variables which are not independent), and varied the sample size so that $N = 100, 200$, or 400. The relatively low correlation coefficient and the fact that $N < P$ make the estimation procedure quite challenging.

We used three different graph structures: 100 clusters each with 20 nodes, a band matrix with a width of 10, and a scale-free network. With each configuration we simulated ten datasets and estimated the network structure using betaMix and the competing methods. We measured the number of true- and false-positives, as well as the run-time.

Table 2 summarizes the median number of true- and false-positives for the cluster configuration. With the default settings, the correlation thresholding method yields too many false positives. The Meinshausen-Bühlmann method yields none in most simulations, but has a relatively low power to detect true edges. When $N = 100$ it finds only 115, and when $N = 400$ it detects 7,630 out of 19,000. Our method is more powerful and maintains a low number of false positive edges. The median run-times for these three methods were 122 s for correlation thresholding, 48.5 s for Meinshausen-Bühlmann, and 2 s for our method. The graphical lasso failed to yield results in some cases, and when it did converge it was considerably worse than the correlation thresholding method, both in terms of the number of false positives and the run-time.

Qualitatively similar result were obtained with the band configuration. With the scale-free structure the Meinshausen-Bühlmann method performed quite well in terms of the number of true- and false-positives when $N = 400$, but was much less powerful for smaller sample sizes, compared to our method. It was also slower than betaMix (a median of approximately 1 minute, versus 2 s for betaMix). Table 3 shows the median TP and FP for the correlation thresholding, Meinshausen-Bühlmann and betaMix methods. The glasso method was much slower with a median runtime of approximately 6 minutes, and yielded hundreds of thousands of false positives when $N = 100$ and $N = 200$. With $N = 400$ the glasso method found no edges at all.

Recall that in Section 2.2 we showed that the betaMix method is also applicable in cases where the elliptical symmetry assumption does not hold, when N is large enough. To demonstrate this point, we simulated $P = 2,000$ Bernoulli(0.5) random variables with a cluster configuration structure, as described above. The results shown here were obtained with $N = 300$ and $\rho = 0.33$ (the correlation between pairs of variables in the same cluster). Results are shown in Table 4 only for betaMix and the Meinshausen-Bühlmann method, which had the best performance among the competing methods in all the simulations. Again, betaMix recovers the true graph structure almost precisely, with 17,799 true edges and only 29 false ones. The mb method is much less powerful, and recovers only 2,715 edges from the 19,000 in the graph (with no false edges).

Table 3
The median true- and false-positives for the correlation thresholding (ct), Meinshausen-Bühlmann (mb), and betaMix (bx) methods with the scale-free configuration (total number of edges in this configuration is 1,999, edge density of 0.001).

Method		$N = 100$	$N = 200$	$N = 400$
ct	TP	1,485	1,876	1,970
	FP	45,861	24,428	3,292
mb	TP	17	332	1,581
	FP	0	0	1
bx	TP	26	674	1,682
	FP	0	8	3

Table 4
The median true- and false-positives and run-time for the Meinshausen-Bühlmann (mb) and betaMix (bx) methods with the cluster configuration of Bernoulli(0.5) variables. The total number of edges in this configuration is 19,000.

Method	True Positives	False Positives	Run-time
mb	2,715	0	49 s
bx	17,799	29	3 s

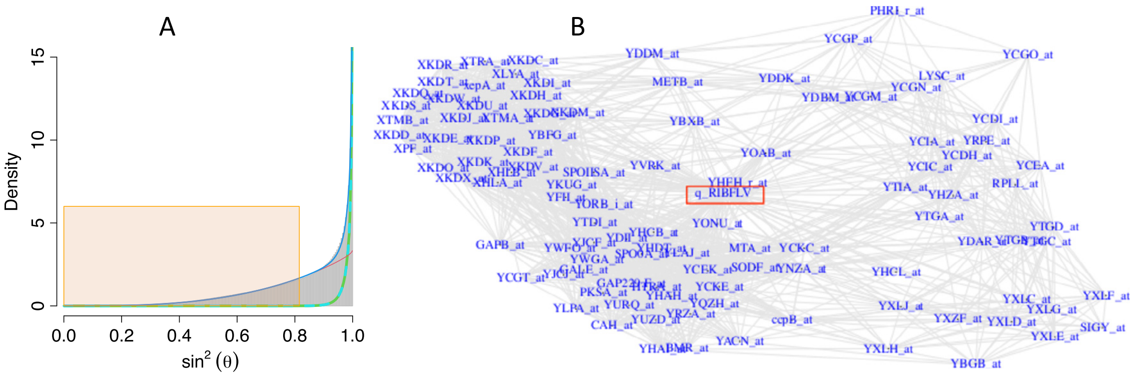

Fig. 5. A. The riboflavin data - fitted beta mixture model. B. 106 genes are selected as strong predictors for the production rate of riboflavin data.

4. Data analysis

We now demonstrate three applications of the `betaMix` method. Additional complex graphical model examples are provided in the Supplementary Material.

4.1. Variable selection and graphical models

The riboflavin data, which was introduced by Bühlmann et al. (2014), contains normalized expression data of 4,088 genes, and the objective here is to detect which of these genes is a predictor of riboflavin production rate (the response) in *Bacilluss subtilis*. There are $N = 71$ samples, which we assume to be independent. Variable selection with the beta-mixture method amounts to detecting the significant correlations or edges between the $P = 4088 + 1$ variables, and ultimately reporting the nodes which are found to be adjacent to the response variable's node. Fig. 5A shows the distribution of the z_j's and the fitted mixture model. The threshold for declaring a pair of variables as significantly correlated is found to be $\sin^2(\theta) > 0.815$ ($|r| > 0.43$). For the purpose of variable selection we are only interested in edges which connect to the riboflavin production rate variable (the highlighted node, q_RIBFLV in Fig. 5B) and the algorithm selects 106 variables, which form a highly interconnected network with two large clusters of genes.

The large number of selected predictors and the strong dependence among them suggests that riboflavin production is an intricate process which probably cannot be explained satisfactorily by a sparse, linear model. A change in one gene may cause a chain reaction in many other genes, quite possibly involving non-linear effects, thus making it complicated to predict the ultimate effect on the response variable.

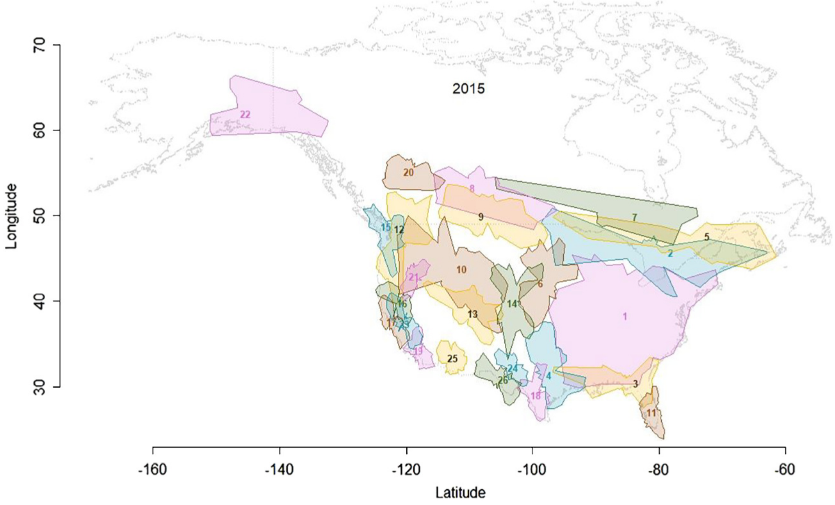

Fig. 6. The North American Breeding Bird Survey data – 2015.

4.2. Spatial models

One of the challenging steps in spatial modeling is to estimate the covariance matrix. Adjacent locations cannot be assumed to be independent, and the spatial correlation must be accounted for. We illustrate how one may use a data-driven approach with the beta mixture model to estimate the spatial covariance matrix (which may depend on covariates, so as to account for differences between regions). This can be useful for Kriging of spatial data. We use the 2020 release of the North American Breeding Bird Survey dataset, which contains bird species count for more than 700 North American bird taxa. The data is collected each June at thousands of random locations along routes in the United States and Canada. Each route is approximately 40 km long, with counting locations placed roughly every 800 meters (50 stops along the route). Counting is performed by a citizen scientist proficient in avian identification. A longitudinal study could be very interesting in order to detect trends in range, occurrence, and abundance of some birds, and perhaps help to establish ecological health indicators. However, since the number of routes has increased six-fold between 1966 and 2019, and also because conditions may vary significantly between years in some locations and there is only one observation per year, we use just one year (2015) to illustrate our approach.

Our dataset consists of the total number of species count per an entire route. Initially, we have 5,756 locations and 756 unique AOU's (the American Ornithologists' Union identification code for birds). We aggregate counts from routes which are close to each other (within 60 km). Birds which have not been observed in any location, and locations in which no birds have been observed, are eliminated, resulting in $N = 608$ birds and $P = 601$ locations. The counts are log-transformed in order to normalize the data (adding 1 to all counts, in order to avoid taking the logarithm of zero).

Since our objective is to obtain a spatial covariance matrix, we treat the locations as our nodes, and use the beta mixture model to find which pairs of locations are strongly correlated. We use the vectors of 608 bird species counts per location to calculate the z_i's which are used when fitting the model. In this type of analysis we must take into account that the observations (bird counts) are not independent, and use the adaptation mentioned in Section 2.3. This yields an effective sample size $\hat{\nu} \approx 32$ – much smaller than the actual N. Out of 180,300 possible pairs, 20,349 are detected as highly correlated, so the graph is only relatively sparse (with 11% of the possible edges). With these edges, we create 26 clusters (shown in Fig. 6) which consist of locations with similar bird abundance vectors. In the clustering step we first identify the cluster centers based on their degree and clustering coefficient, and then include a location in a cluster if its abundance vector is found to be strongly correlated with the central node by our beta mixture method. There are many possible clustering methods, and each can be configured with a set of parameters, thus yielding different cluster configurations, but since this is not the focus of this paper we refer the reader to the Supplementary Materials for details.

Some points are worth mentioning about Fig. 6. First, the algorithm uses only bird abundance data, and although the coordinates of points are *not* used in the network construction, the edges found by the beta mixture model yield clusters which correspond very nicely to geographical regions. For example, the Florida panhandle (11), the Sonora desert (25), along the Missouri River (9, 6, 4), and the subarctic region (22). The shape and location of the clusters correspond to common habitat conditions, such as climate, vegetation, water resources, and proximity to the shore. Second, the clusters have a fair amount of overlap (for example, clusters 12, 15, 16, 17, 19, and 23 along the west coast). The edges we find based on bird abundance correlations allow to capture subtle differences between similar clusters, and although we only have one time point, these overlapping clusters can be used to detect and account for migration paths. Third, notice that some regions are not associated with any cluster. This may very well be due to under-sampling, as is probably the case in the western deserts, and in northern Canada. In spatial data analysis this can be quite important. One may use nearby

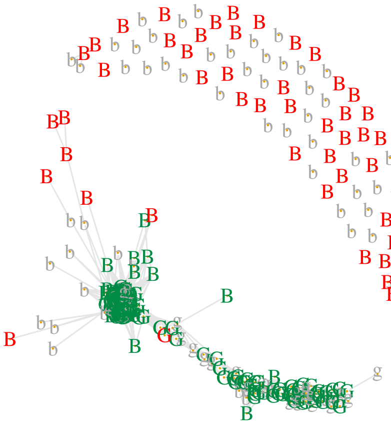

Fig. 7. The ionosphere network – lower case letters represent training data, and upper case letters represent test data. The letter specifies the true classification, and the color represents our algorithm's classification (red=bad, green=good). The training dataset has 60 good radar returns and 60 bad ones.

clusters to impute the covariance structure in such areas, or, possibly infer that the covariance matrix is indeed singular in certain regions. For example, a region may be inhabitable, and thus, imposing a non-singular covariance matrix may lead to incorrect predictions.

Obtaining data-driven spatial covariance matrices can be very useful in ecological studies. The approach demonstrated here with one time point from one year may be extended to longitudinal studies, and to consider species-environment interactions and include covariates such as temperature, precipitation, and major events such as hurricanes, wildfires, and volcanic activity.

4.3. Classification

To demonstrate how the betaMix approach may be used to perform classification we use the 'ionosphere' data (Sigillito et al., 1989) from the Machine Learning Repository (Dua and Graff, 2017). The data was collected from a radar system which aimed radio waves at the ionosphere in order to detect free electrons. If the returned signal does not show evidence of some type of structure, it means that the signal passed through the ionosphere and this radar return should be classified as 'bad'. If there is evidence of some type of structure in the radar return then this is classified as good. The objective of the original technical report was to show that the process of determining the quality of the return signal can be done automatically, accurately and efficiently by using a multilayer feed forward neural network. At the time of the report the radar in the experiment was producing data every 5 seconds, year round, and since the quality control process was in part manual it was very time-consuming.

The data has 34 continuous attributes, and a binary outcome – classification by an expert into 'good' or 'bad'. Two of the continuous variables have low variability so we exclude them from our analysis (the first variable has 38 zeroes and 313 ones, and the second contains only zeroes). There are 126 bad cases and 225 good ones in the dataset. For the training data we use 60 of each response type, leaving 66 bad and 165 good cases for the test data. Each observation in the test data is classified based on its similarity (in terms of the adjacency matrix obtained from betaMix) to points in the training data. We use a simple majority rule – if most of its training set neighbors are good then the point is classified as good. Otherwise it is classified as bad. To fit the model we set the frequentist error rate parameter to 1e-5 and the Bayesian posterior probability threshold to 0.001. The fitted plot is provided in the Supplementary Material, and the threshold for detecting an edge in the graph is found to be $\sin^2(\theta) < 0.56$. The network for one randomly drawn training data set is shown in Fig. 7. We use the igraph package (Csardi and Nepusz, 2006) for this graphical representation.

There are a few striking characteristics in the network plot. Most of the good cases are clustered very tightly (in the lower left quadrant). A smaller set of good cases forms a second cluster, with a high degree of connectivity, but less than the main cluster (in the lower right quadrant). Most of the bad cases are not connected to any other nodes, and among ones which are connected, most appear in the fringes of the clusters because they are similar to a relatively small number of nodes in the cluster. Still, in this particular training data set there are some bad cases which are very similar to the large and tight cluster of good cases. This suggests that correctly classifying all the bad cases is expected to be challenging.

With this typical training data we get an overall accuracy of 91.3% when we classify the test dataset. The sensitivity is 99.4% and the specificity is 71.2% (correctly classifying 47 of the 66 bad cases). From the plot we see that we can improve the results by using a stricter condition for the classification rule. For example, if we only classify a point as good if it has at least four training set neighbors, of which the majority are good, then the accuracy increases to 94.8%, the sensitivity is 98.7% and the specificity is 84.8%. To improve the specificity further we may consider using higher orders of the 32 variables in the input to betaMix. However, this is outside the scope of this article. Our goal here is to demonstrate that networks obtained from the betaMix model can be used to perform accurate classification, while also providing insights about the relationships among samples and classes. Additional complex graphical model examples are provided in the Supplementary Material.

5. Related work

5.1. Support discovery and covariance matrix estimation in high-dimensional settings

Wainwright (2009) establishes precise conditions on P, s_0 (adopting the notation from the introduction), and n, the sample size needed to recover the sparsity pattern using the LASSO. One consequence of his analysis is that under the assumptions of s_0-sparsity of the true $\boldsymbol{\beta}$ and invertibility of the matrix $X'_{S_0}X_{S_0}$, the LASSO estimator converges to $\boldsymbol{\beta}$ in the ℓ_2 norm if $\log(P)/n = o(1)$. In Corollary 2 of his main result Wainwright (2009) shows that for standard Gaussian designs (a) the LASSO can only recover $\boldsymbol{\beta}$ with support cardinality $s_0 \leq (1+o(1))n/2\log(P)$ where $n = \nu P$ for some $\nu \in (0,1)$, and it fails with probability converging to 1 if there exist $c_2 > c_1 > 0$ such that $s_0 \in (c_1 P, c_2 P)$; and (b) if $s_0 = \alpha P$ for some $\alpha \in (0,1)$ then the LASSO requires a sample size $n > 2\alpha P \log[(1-\alpha)P]$ in order to obtain an exact recovery of $\boldsymbol{\beta}$.

Reid and Tibshirani (2016) propose a sparse regression and marginal testing approach for data with correlated predictors. They first cluster the predictors, and then take the most informative predictor in a cluster as the ‘prototype’. They then apply either the LASSO or marginal significance testing to the much smaller set of predictors which were selected as prototypes. Efron et al. (2004) introduce the popular LARS method (least angle regression), which tackles the same problem. The idea is essentially similar to forward stepwise selection. Initially all $\beta_j = 0$ and in the first iteration the variable most correlated with the response is selected and its coefficient is set according to the sign of its correlation with the response. In each step, the current estimator is used to update the residuals, and the selected β_j is increased in the direction of the sign of its correlation with y, until some other predictor x_k is as correlated with the residual vector as x_j. The process is repeated – (β_j, β_k) are increased in their joint least squares direction, until another predictor x_k is as correlated with the residual, and so forth, until all the predictors are in the model. With a simple modification to the algorithm, Efron et al. (2004) show that the LARS algorithm yields all the LASSO solutions.

In the literature mentioned thus far the main objective was to recover the vector of regression parameters, $\boldsymbol{\beta}$. Many authors have extended the scope to the case in which the covariance matrix in the large-P setting also has to be estimated. This is important in a number of applications, including dimension reduction via principal component analysis (PCA) or singular value decomposition (SVD), spatial analysis, classification via discriminant analysis, and fitting graphical models. Bickel and Levina (2008) consider a method based on hard-thresholding and show that if $\log P/n \to 0$ and the true covariance matrix is sparse in the sense that in each row at most $c_0(P) \ll P$ elements are non-zero, then the threshold estimate is consistent. Furthermore, the rate of convergence of their estimator is shown to be $O_{\mathbb{P}}[c_0(P)(\log(P)/n)^{(1-q)/2}]$, for $q \in [0,1)$. Bickel and Yan (2008) consider the importance of sparsity in covariance matrix estimation. They define sparsity in terms of points lying on or near a low dimensional sub-manifold of a P-dimensional space. They consider sparsity in the covariance matrix or in the precision matrix, so that in either case each row in the matrix is sparse in the operator norm, that is, the number of non-zero elements in each row in the covariance matrix is small (less than some s). Bickel and Yan (2008) discuss properties of the estimator for the ‘true’ dimension of the data, which is assumed to be much smaller than P.

Sparse estimation of the covariance and covariance selection has been studied extensively in recent years. See, for example, Levina et al. (2008), Rothman et al. (2008), and Warton (2008). Cai et al. (2013a) consider testing equality of covariance matrices and support discovery in two-sample, high-dimensional and sparse settings, and Zhu et al. (2017) use high-dimensional covariance matrices tests to detect schizophrenia risk genes. In the context of graphical models, a common approach is to identify edges in a high-dimensional graph by using the LASSO P times, each time taking another variable as the response and performing variable selection on the other $P-1$. The l_1 penalty imposed on a Gaussian log-likelihood induces sparsity which is controlled by a tuning parameter. Other methods based on solving the estimation of a large, sparse precision matrix via l_1 penalized Gaussian log-likelihood include, for example, Meinshausen and Bühlmann (2006); Friedman et al. (2008); Yuan and Lin (2007); Peng et al. (2009); Khare et al. (2015). Some methods also deal with non-Gaussian data (e.g. Banerjee et al. (2008)). The computational complexity of these penalized approaches is polynomial in P. In contrast, correlation screening methods that are mentioned in Section 2.2 are non-iterative algorithms so the computational complexity is of the order $P\log P$ (Hero and Rajaratnam, 2015).

5.2. Approaches using convex geometry

Ideas from convex geometry have been applied in the statistics literature to developing thresholds for correlation screening. In correlation screening the objective is to select variables whose maximal correlation exceeds a given threshold. Hero and Rajaratnam (2011, 2015) developed a novel threshold for marginal correlation screening in the high-dimensional-low-sample-size setting using spherical cap calculations. Their threshold is derived as asymptotic expressions for the mean number of correct discoveries. These expressions depend on a Bhattacharyya measure (Basseville, 1989) of average pairwise dependency of the P multivariate scores defined on $\mathcal{S}^{n-2}$. Hero and Rajaratnam (2011) give $(1 - c_n(P-1)^{-\frac{2}{n-4}})^{\frac{1}{2}}$ (for c_n equal to the volume of $\mathcal{S}^{n-2}$) as useful correlation screening threshold. A similar threshold, $(1 - P^{-\frac{2}{n-2}})^{\frac{1}{2}}$, was also developed in Zhang (2017) for detecting spurious correlations and low rank correlation structure also by using a spherical cap packing perspective. Note that both of these thresholds are of the same order of $(1 - P^{-\frac{2}{n}})^{\frac{1}{2}} = (1 - \exp\{-\frac{2}{n}\log P\})^{\frac{1}{2}} \sim (1 - (1 - \frac{2}{n}\log P))^{\frac{1}{2}} = (\frac{2}{n}\log P)^{\frac{1}{2}}$ which is connected to the classical rate of convergence mentioned above.

Cai and Jiang (2011, 2012); Cai et al. (2013b) take an approach to screening correlations via the analysis of discovering minimal pairwise angles. These articles give a very careful analysis of the normalizing constants for the convergence to the particular Weibull-type extremal distribution using spherical cap calculations. They consider the different asymptotic phase transition regimes where $\log P/n \to \{0, \text{a constant}, \infty\}$. The phase transitions that are developed are similar in spirit to those in Hero and Rajaratnam (2015) and Zhang (2017). Hero et al. (2021) extend their previous work by making clever connections to random geometric graphs (Penrose, 2003) and demonstrate that the distribution of the number of significant correlations beyond a certain threshold is approximated by a compound Poisson distribution. The compound Poisson approximation rate parameter develop depends on a spherical cap comptutation. Hero et al. (2021) evaluate different asymptotic regimes when P is finite and ultra high-dimensional settings when P diverges to ∞ as well as dealing with previous technical block-sparse assumptions.

The screening rules in Hero and Rajaratnam (2011, 2015); Hero et al. (2021); Cai and Jiang (2011, 2012); Cai et al. (2013b); Zhang (2017) are based on the maximum correlation exceeding a threshold so these rules are defined by ℓ_∞ balls (hypercubes). There are interesting differences between the volume of a hypercube with unit length sides and the volume of a unit radius sphere in high dimensions (Blum et al., 2020). As the dimension of the unit cube increases, its volume is always one and the maximum possible distance between two points grows as the square root of the dimension. Cast in terms of a confidence set, the ℓ_∞ cube has fixed volume but a diverging diameter. In contrast, as the dimension of a unit sphere increases, its volume goes to zero exponentially (Ball, 1997) and the maximum possible distance between two points stays fixed. Efron (2006) pointed out that confidence regions should be constructed to minimize volume. Consequently, in high dimensions confidence regions based on hypercubes, as in screening, may be problematic as they have exponentially larger volume than those based on spherical rules.

Reverter and Chan (2008) also use partial correlations combined with information theory for the reconstruction of gene co-expression networks. Our approach is more similar to that of Bar and Bang (2021) who also considered a mixture model for detecting significant correlations, but their approach relies on Fisher's Z-transformed correlations and their asymptotic normal distribution under the null hypothesis. The nonnull edges in their edgefinder method are modeled as two lognormal distributions (for significantly positive/negative correlations). Their edgefinder method performs very well especially when n is sufficiently large, but the approach presented here does not require a normalizing transformation, and controlling the error rate relies on a general convex geometry theory.

The idea of flipping the roles of predictors and observations appeared in the context of variable selection via a method called SEMMS in Bar et al. (2020), although there the motivation was to improve computational efficiency via the Woodbury identity when the true number of predictors is much smaller than P. The representation into a higher dimensional space is also akin to the kernel trick that represents the data in a higher dimensional feature space. The approach presented here is more general in that it does not depend on sparsity, nor does it require to define one of the variables as the response. Therefore, the same convex geometry principles can be used to explain the excellent performance of SEMMS. It should be noted that as a variable selection method, SEMMS has been extended to the generalized linear models framework, as well as to quantile regression in the $\boldsymbol{\beta}$-sparse, high-dimensional setting. We briefly discuss these, and other extensions in Section 6.

6. Discussion

We have introduced a mixture-model of beta distributions to identify significant correlations among P predictors when P is large. The method relies on theorems in convex geometry, which were used here to show how to control the error rate of edge detection in graphical models. The betaMix method does not require any assumptions about the network structure, nor does it assume that the network is sparse. When the network is dense the null probability parameter may be estimated as $\hat{p}_0 = 0$ and a single beta component will be used to fit the data. However, in the applications discussed here, such as co-expression of genes or co-abundance of birds in spatial modeling, it is expected that many pairs (genes, locations) will be uncorrelated, and betaMix has been motivated by such applications. Applying betaMix to datasets in which there are no uncorrelated pairs of nodes in the graph is possible, but may require some modifications to the model. For example, it may be the case that while there are no uncorrelated pairs of nodes, there are still many 'mostly uncorrelated' ones. So, from the modeling perspective the null component may be allowed to be concentrated near 1, but not with $\frac{N-1}{2}$ and $\frac{1}{2}$ parameters as implied from the convex geometry theory. This will require a different geometric/probabilistic justification. Similarly, the correlated pairs may actually arise from two or more regimes, in which case it may be better for the non-null component to be a mixture of multiple beta distributions. These extensions are left for future research.

Another future avenue of research is to explore the possibility to extend betaMix for the analysis of causal models. In such models the edges need to be directional, while the method presented here only accommodates bi-directional ones, since it relies on correlations which are symmetric. It will be useful to incorporate covariates in the estimation of edges, since the existence of an edge in the graph may depend on time, location, temperature, and so on. It will also be interesting to explore the possibility to extend the approach beyond linear correlations. For example, the association between variables may be strong only when considering certain quantiles, but they may not be correlated in the usual sense (Pearson or Spearman).

Acknowledgements

We want to acknowledge the excellent suggestions of the referees and associate editor. Wells' research was partially supported by NIH grant R01GM135926-03.

Appendix A. Convex geometry

Despite the widespread and applicability in statistical modeling, linear subspaces suffer from the drawback that they cannot be analyzed using Euclidean geometry. Indeed, subspaces of $\mathbb{R}^n$ lie on a special type of Riemannian manifolds, the Grassmann manifold, which has a nonlinear structure. The Grassmann manifold $\mathbb{G}_{n,k}$ is used to study the geometry of the space of all k dimensional subspaces of $\mathbb{R}^n$. $\mathbb{G}_{n,k}$ is isomorphic to the quotient set $\mathbb{O}(n)/(\mathbb{O}(k) \times \mathbb{O}(n-k))$, where $\mathbb{O}(m)$ denotes the group of $m \times m$ orthogonal matrices (Absil et al., 2009; Ye and Lim, 2016).

The manifold $\mathbb{G}_{n,k}$ has an invariant measure (James, 1954; Lv, 2013) which can be used to calculate the volumes of sets which are specified in terms of the principal angles θ_i between k and l dimensional subspaces of $\mathbb{R}^n$. The principal angles between subspaces are the generalization of the concept of the angle between lines. Let $\mathcal{U}$ and $\mathcal{V}$ be two subspaces in $\mathbb{G}_{n,k}$ and $\mathbb{G}_{n,l}$ $(k \le l)$ having a set of principal angles $(\theta_1, \dots, \theta_k)$, with $\pi/2 \ge \theta_1 \ge \cdots \ge \theta_k \ge 0$ where $\theta_i = \max\{\frac{\langle u, v\rangle}{\|u\|\|v\|} : u \perp u_m, v \perp v_m, m = 1, 2, \dots, i-1\}$ for $u \in \mathcal{U}$ and $v \in \mathcal{V}$. The corresponding k pairs of orthogonal unit vectors are (u_i', v_i'). By setting $\rho_i = \cos\theta_i$ gives the canonical correlations $(\rho_1, \dots, \rho_k)$ and corresponding pairs of canonical variables $\{u_i', v_i'\}_{i=1}^k$ (Lv, 2013). The chordal distance between $\mathcal{U}$ and $\mathcal{V}$ is $(\sum_{i=1}^k \sin^2\theta_i)^{1/2}$ (Conway et al., 1996) and the maximum chordal distance is $\sin\theta_1$ (Absil et al., 2009; Ye and Lim, 2016). The invariant measure of the principal angles $(\theta_1, \dots, \theta_k)$ can be constructed by viewing $\mathbb{G}_{n,k}$ as $\mathbb{V}_{n,k}/\mathbb{O}(k)$, where $\mathbb{V}_{n,k}$ denotes the Stiefel manifold of all orthonormal k-frames in $\mathbb{R}^n$ (Lv, 2013). By deriving the exterior differential forms on those manifolds, James (1954) gave an expression for the invariant (uniform) measure of the principal angles $(\theta_1, \dots, \theta_k)$ for $(k \le l)$

$$d\mu_{k,l}^n = C_{k,l}^n \prod_{i=1}^k (\cos^2\theta_i)^{\frac{l-k}{2}} \prod_{i=1}^k (\sin^2\theta_i)^{\frac{n-l-k}{2}} \prod_{1\le i<j\le k} \left(\sin^2\theta_i - \sin^2\theta_j\right) d\theta_1 \cdots d\theta_k\,, \tag{A.1}$$

over $\Theta = \{(\theta_1, \dots, \theta_k) : \pi/2 > \theta_1 > \cdots > \theta_k > 0\}$. The normalization constant is given by

$$C_{k,l}^n = \prod_{i=1}^k \frac{A_{k-i+1} A_{k-i+1} A_{n-l-i+1}}{2A_{n-i+1}}, \tag{A.2}$$

where $A_j = 2\pi^{j/2}/\Gamma(j/2)$ is the area of the unit sphere S^{j-1}.

In the special case where $k = l = 1$ (which is the focus of this paper) the Grassmann manifold $\mathbb{G}_{n,1}$ is a generalization of the projective space $\mathbb{P}^{n-1}$ corresponding to the lines passing through the origin of the Euclidean space (Absil et al., 2009, pg.30). The chordal distance between two lines is the sine of their angle. On $\mathbb{G}_{n,1}$ the invariant measure $\mu_{1,1}^n$ has a simple expression for the density of the canonical angle θ (Absil et al., 2006; Lv, 2013)

$$\nu_1^n(\theta) = \frac{1}{B(\frac{1}{2}, \frac{n-1}{2})} (cos^2\theta)^{-1/2} (1 - cos^2\theta)^{(n-1)/2-1}, \tag{A.3}$$

where $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the beta function. This nice result implies that the $cos^2\theta$ has a $Beta(1/2, (n-1)/2)$ distribution or equivalently $sin^2\theta$ has a $Beta((n-1)/2, 1/2)$ distribution. These are precisely the measure the underlie the spherical cap calculations discussed in the previous section (Hero and Rajaratnam, 2011, 2015; Cai and Jiang, 2011, 2012; Cai et al., 2013b; Zhang, 2017) that was used in the development of correlation screening rules. Note that Theorem 1 can be expressed in terms of the invariant measure $\mu_{k,l}^n$ for the case $k = l = 1$ in (A.1).

To generalize our method we would like to consider the sine of the principal angles between two random k dimensional subspaces (rather than just lines). Absil et al. (2006) use the Gauss hypergeometric function ${}_2F_1$ with a matrix argument to give the density of the largest principal angle between two random subspaces. We give the necessary definitions of the Gauss hypergeometric function in the Supplementary Material. Using Theorem 1 in Absil et al. (2006) and the transformation $z = \sin^2\theta_1$ gives the multivariate analog of (1):

Theorem 3. *Let K and L be two k-planes in $\mathbb{R}^n$. Let θ_1 be the largest principal angle between K and L. Then the random variable $Z = sin^2\theta_1$ has a p.d.f.*

$$\frac{k(n-k)}{2} \frac{\Gamma(\frac{k+1}{2})\Gamma(\frac{n-k+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n+1}{2})} \frac{z^{\frac{k(n-k)}{2}-1}}{(1-z)^{\frac{1}{2}}} {}_2F_1\left(\frac{n-k-1}{2}, \frac{1}{2}; \frac{n+1}{2}; zI_{k-1}\right). \tag{A.4}$$

Appendix B. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2023.107800.

References

Absil, P.A., Edelman, A., Koev, P., 2006. On the largest principal angle between random subspaces. Linear Algebra Appl. 414, 288–294.

Absil, P.A., Mahony, R., Sepulchre, R., 2009. Optimization Algorithms on Matrix Manifolds. Princeton University Press.

Ball, K., 1997. An elementary introduction to modern convex geometry. In: Flavors of Geometry. Univ. Press, pp. 1–58.

Banerjee, O., Ghaoui, L.E., d'Aspremont, A., 2008. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. J. Mach. Learn. Res. 9, 485–516.

Bar, H., Bang, S., 2021. A mixture model to detect edges in sparse co-expression graphs with an application for comparing breast cancer subtypes. PLoS ONE 16, 1–20. https://doi.org/10.1371/journal.pone.0246945.

Bar, H.Y., Booth, J.G., Wells, M.T., 2020. A scalable empirical Bayes approach to variable selection in generalized linear models. J. Comput. Graph. Stat., 1–12. https://doi.org/10.1080/10618600.2019.1706542.

Basseville, M., 1989. Distance measures for signal processing and pattern recognition. Signal Process. 18, 349–369.

Bickel, P.J., Levina, E., 2008. Covariance regularization by thresholding. Ann. Stat. 36, 2577–2604. https://doi.org/10.1214/08-AOS600.

Bickel, P.J., Yan, D., 2008. Sparsity and the possibility of inference. Sankhya, Ser. A 2008 (70), 1–24. http://www.jstor.org/stable/41234399.

Blum, A., Hopcroft, J., Kannan, R., 2020. Foundations of Data Science. Cambridge University Press.

Bühlmann, P., Kalisch, M., Meier, L., 2014. High-dimensional statistics with a view toward applications in biology. Annu. Rev. Stat. Appl. 1, 255–278.

Cai, T., Liu, W., Xia, Y., 2013a. Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. J. Am. Stat. Assoc. 108, 265–277.

Cai, T.T., Fan, J., Jiang, T., 2013b. Distributions of angles in random packing on spheres. J. Mach. Learn. Res. 14, 1837.

Cai, T.T., Jiang, T., 2011. Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. Ann. Stat. 39, 1496–1525.

Cai, T.T., Jiang, T., 2012. Phase transition in limiting distributions of coherence of high-dimensional random matrices. J. Multivar. Anal. 107, 24–39.

Conway, J.H., Hardin, R.H., Sloane, N.J., 1996. Packing lines, planes, etc.: packings in grassmannian spaces. Exp. Math. 5, 139–159.

Cox, D., Wermuth, N., 1996. Multivariate Dependencies; Models, Analysis, Interpretation. Publisher Chapman and Hall, London.

Csardi, G., Nepusz, T., 2006. The igraph software package for complex network research. Int. J. Complex Syst., 1695. http://igraph.org.

David, H.A., 1981. Order Statistics. Wiley, New York.

Dempster, A.P., 1972. Covariance selection. Biometrics, 157–175.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B 39, 1–38.

Donoho, D.L., 2000. High-dimensional data analysis: the curses and blessings of dimensionality. In: AMS Math Challenges Lecture, vol. 1, p. 32.

Drton, M., Richardson, T.S., 2008. Graphical methods for efficient likelihood inference in Gaussian covariance models. J. Mach. Learn. Res. 9, 893–914.

Dua, D., Graff, C., 2017. UCI machine learning repository. http://archive.ics.uci.edu/ml.

Efron, B., 2006. Minimum volume confidence regions for a multivariate normal mean vector. J. R. Stat. Soc. B 68, 655–670.

Efron, B., 2008. Microarrays, empirical Bayes and the two-groups model. Stat. Sci. 23, 1–22. https://doi.org/10.1214/07-STS236.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Stat. 32, 407–499.

Fourdrinier, D., Strawderman, W.E., Wells, M.T., 2018. Shrinkage Estimation. Springer.

Frankl, P., Maehara, H., 1990. Some geometric applications of the beta distribution. Ann. Inst. Stat. Math. 42, 463–474.

Friedman, J., Hastie, T., Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9, 432–441.

van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Stat. 42, 1166–1202. https://doi.org/10.1214/14-AOS1221.

Hall, P., Marron, J.S., Neeman, A., 2005. Geometric representation of high dimension, low sample size data. J. R. Stat. Assoc., Ser B 67, 427–444.

Hero, A., Rajaratnam, B., 2011. Large-scale correlation screening. J. Am. Stat. Assoc. 106, 1540–1552.

Hero, A.O., Rajaratnam, B., 2015. Foundational principles for large-scale inference: illustrations through correlation mining. Proc. IEEE 104, 93–110.

Hero, A.O., Rajaratnam, B., Wei, Y., 2021. A unified framework for correlation mining in ultra-high dimension. arXiv preprint. arXiv:2101.04715.

James, A.T., 1954. Normal multivariate analysis and the orthogonal group. Ann. Math. Stat. 25, 40–75.

Jiang, H., Fei, X., Liu, H., Roeder, K., Lafferty, J., Wasserman, L., Li, X., Zhao, T., 2021. huge: high-dimensional undirected graph estimation. https://CRAN.R-project.org/package=huge. r package version 1.3.5.

Kainen, P.C., Krková, V., 2020. Quasiorthogonal dimension. In: Beyond Traditional Probabilistic Data Processing Techniques: Interval, Fuzzy etc. Methods and Their Applications. Springer, pp. 615–629.

Khare, K., Oh, S.Y., Rajaratnam, B., 2015. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. J. R. Stat. Soc. B, 803–825.

Khare, K., Rajaratnam, B., 2011. Wishart distributions for decomposable covariance graph models. Ann. Stat. 39, 514–555.

Levina, E., Rothman, A., Zhu, J., 2008. Sparse estimation of large covariance matrices via a nested lasso penalty. Ann. Appl. Stat., 245–263.

Lv, J., 2013. Impacts of high dimensionality in finite samples. Ann. Stat. 41, 2236–2262.

Mazumder, R., Hastie, T., 2012. Exact covariance thresholding into connected components for large-scale graphical lasso. J. Mach. Learn. Res. 13, 781–794. https://doi.org/10.1007/s11306-017-1284-x.

Meinshausen, N., Bühlmann, P., 2006. High-dimensional graphs and variable selection with the lasso. Ann. Stat., 1436–1462.

Muirhead, R.J., 1982. Aspects of Multivariate Statistical Theory. John Wiley & Sons.

Peng, J., Wang, P., Zhou, N., Zhu, J., 2009. Partial correlation estimation by joint sparse regression models. J. Am. Stat. Assoc. 104, 735–746.

Penrose, M., 2003. Random Geometric Graphs, vol. 5. OUP, Oxford.

Reid, S., Tibshirani, R., 2016. Sparse regression and marginal testing using cluster prototypes. Biostatistics 17, 364–376. https://doi.org/10.1093/biostatistics/kxv049. https://academic.oup.com/biostatistics/article-pdf/17/2/364/6692798/kxv049.pdf.

Reverter, A., Chan, E.K.F., 2008. Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. Bioinformatics 24, 2491–2497. https://doi.org/10.1093/bioinformatics/btn482. https://academic.oup.com/bioinformatics/article-pdf/24/21/2491/16884178/btn482.pdf.

Rothman, A.J., Bickel, P.J., Levina, E., Zhu, J., et al., 2008. Sparse permutation invariant covariance estimation. Electron. J. Stat. 2, 494–515.

Serfling, R.J., 2009. Approximation Theorems of Mathematical Statistics. John Wiley & Sons.

Sigillito, V.G., Wing, S.P., Hutton, L.V., Baker, K.B., 1989. Classification of radar returns from the ionosphere using neural networks. Technical Report 10. Johns Hopkins APL Technical Digest.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B 58, 267–288.

Wainwright, M., 2009. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). IEEE Trans. Inf. Theory 55, 2183–2202. https://doi.org/10.1109/TIT.2009.2016018.

Warton, D.I., 2008. Penalized normal likelihood and ridge regularization of correlation and covariance matrices. J. Am. Stat. Assoc. 103, 340–349.

Watson, G.S., 1983. Statistics on Spheres. Wiley-Interscience.

Ye, K., Lim, L.H., 2016. Schubert varieties and distances between subspaces of different dimensions. SIAM J. Matrix Anal. Appl. 37, 1176–1197.

Yuan, M., Lin, Y., 2007. Model selection and estimation in the Gaussian graphical model. Biometrika 94, 19–35.

Zhang, K., 2017. Spherical cap packing asymptotics and rank-extreme detection. IEEE Trans. Inf. Theory 63, 4572–4584.

Zhu, L., Lei, J., Devlin, B., Roeder, K., 2017. Testing high-dimensional covariance matrices, with application to detecting schizophrenia risk genes. Ann. Appl. Stat. 11, 1810.