Biostatistics (2017) **18**, 3, *pp.* 422–433
doi:10.1093/biostatistics/kxw053
Advance Access publication on 8 January 2017

Structural zeros in high-dimensional data with applications to microbiome studies

ABHISHEK KAUL*

Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, RTP, NC 27709, USA

abhishek.kaul@nih.gov

ORI DAVIDOV

Department of Statistics, University of Haifa, Haifa 31905, Israel

SHYAMAL D. PEDDADA

Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, RTP, NC 27709, USA

Summary

This paper is motivated by the recent interest in the analysis of high-dimensional microbiome data. A key feature of these data is the presence of "structural zeros" which are microbes missing from an observation vector due to an underlying biological process and not due to error in measurement. Typical notions of missingness are unable to model these structural zeros. We define a general framework which allows for structural zeros in the model and propose methods of estimating sparse high-dimensional covariance and precision matrices under this setup. We establish error bounds in the spectral and Frobenius norms for the proposed estimators and empirically verify them with a simulation study. The proposed methodology is illustrated by applying it to the global gut microbiome data of Yatsunenko *and others* (2012. Human gut microbiome viewed across age and geography. Nature 486, 222–227). Using our methodology we classify subjects according to the geographical location on the basis of their gut microbiome.

Keywords: Classification; High dimension; Microbiome data; Missing data; Sparsity.

1. Introduction

With the advancement of high-throughput technologies it is now common to encounter high-dimensional data where the number of parameters d exceeds the sample size n. One of many such examples is the human microbiome data obtained by the 16s rRNA sequencing technology. The resulting data, known as operational taxonomic units (OTUs), represent counts of thousands of microbial taxa (Mandal *and others*, 2015). In this setting it is often of interest to investigate relationships among the microbes to understand their effects on health outcomes. These relationships can in turn be used to predict the health status of an individual based on his/her microbial composition.

*To whom correspondence should be addressed.

Many such objectives can be achieved via the estimation of the covariance matrix ($\mathbf{\Sigma}$) or its inverse, the precision matrix ($\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$), which characterize the dependence or the conditional dependence structure between variables, respectively. In the high-dimensional setting, estimation of $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ has been discussed extensively in the literature and the existing literature can be broadly classified into two categories, the first approach involves estimation of the precision matrix by exploiting its natural sparsity in comparison to the covariance matrix (Friedman *and others*, 2008; Rothman, Bickel *and others*, 2009; Cai *and others*, 2011). A limitation of this approach is that it does not apply to low-rank matrices $\mathbf{\Sigma}$ since the precision matrix does not exist in this case. The second popular approach is to estimate $\mathbf{\Sigma}$ by assuming that $\mathbf{\Sigma}$ is itself sparse. One of several methods for this purpose is to threshold each element of the sample covariance matrix (Bickel and Levina, 2008; Rothman, Levina *and others*, 2009).

The current literature assumes the availability of independent and identically distributed (i.i.d.) copies of the vector $\mathbf{X} = (X_1, X_2, ..., X_d)^T$ whose distribution is Gaussian or more generally sub-Gaussian with $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ as the d-dimensional mean vector and covariance matrix, respectively. Note that a real-valued random variable (r.v.) X_1 is said to be sub-Gaussian if there exists a $b > 0$ such that for every $t \in \mathbb{R}$, one has $Ee^{tX_1} \leq e^{b^2t^2/2}$.

In contrast to typical high-dimensional data, not all variables (i.e. microbes) are observed in a sample of microbiome data. Thus if $\mathbf{X}$ represents a vector of abundances of d microbes in a specimen, then not all components of $\mathbf{X}$ may be observed. We refer to this missingness as structural zeros, and it is due to the underlying biology and not due to error in measurement or values below the minimum detection level. For example, it is known that the bacterial genus *Bacteroides* is prevalent in the human gut when the associated diet is high protein/fat diet, whereas it may not be present otherwise, e.g. carbohydrate rich diet. The total abundance of such bacteria is coded as 0 counts in the observational vector $\mathbf{X}$.

The missing structure required to model structural zeros is more general than typical notions of missingness in the literature. More precisely, in the classical notions of missingness, such as missing completely at random or missing at random (MAR), it is assumed that in place of $\mathbf{X}$ we observe a surrogate vector $\mathbf{U} = \mathbf{X} \oplus \mathbf{W}$, where $\oplus$ represents a component-wise product and $\mathbf{W}$ is a d-dimensional vector of independent Bernoulli r.v.'s. In effect, not all components of $\mathbf{X}$ are observed in $\mathbf{U}$. For example, $\mathbf{U} = (0, 0, X_3, .., X_d)^T$ corresponds to the case where the first two components of $\mathbf{X} = (X_1, ..., X_d)^T$ are not observed in $\mathbf{U}$ with $\mathbf{W} = (0, 0, 1, ..., 1)^T$. In this example, although X_1 and X_2 are absent in $\mathbf{U}$, they still influence the distribution of the remaining components $X_3, .., X_d$ through the underlying dependence structure of $\mathbf{\Sigma}$ and are only hidden by the corresponding multiplicative Bernoulli noise vector $\mathbf{W}$. In contrast, for the case of structural zeros the observed vector itself is $\mathbf{X} = (0, 0, X_3.., X_p)$, i.e. the first two components are truly absent from the observation and thus the missing components should not influence the distribution of the remaining components. It should also be noted from this example that in the latter case, implementing classical methods of imputation under such conditions would be a logical error due to the definition of a structural zero.

In this paper we introduce a framework which allows for structural zeros in the model and discuss consistent methods of estimating sparse high-dimensional covariance and precision matrices. We establish consistency in estimation of the proposed methodology and empirically support it with a simulation study. We also apply our methodology to classify observations to geographical locations based on the global human gut microbiome data of Yatsunenko *and others* (2012).

Some work related to ours in the literature is that of Kurkland and Heagerty (2005) who provide a regression setting for the analysis of longitudinal data truncated by deaths, here they make a similar distinction between zeros in response variable due to an individual dropping out of the study or due to death, in this context the zero due to death can be described as a structural zero in our definition. Estimation of covariance and precision matrices in the traditional missing values setting has also been discussed in the literature (Loh and Wainwright, 2012; Lounici, 2014; Kaul *and others*, 2016). As noted above and

as shall become more apparent in the following sections, our model allows for a more general notion of missingness while assuming weaker conditions in comparison to typical notions of missingness.

2. NOTATIONS AND FRAMEWORK

For any matrix $\mathbf{A} = [a_{ij}]$ define the *Sup*, *Frobenius* and *spectral* norms as $\|\mathbf{A}\|_\infty = \max_{i,j} |a_{ij}|$, $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$ and $\|\mathbf{A}\|_2 = \sup_{||x||_2 \leq 1} ||Ax||_E$, respectively, where $\|\cdot\|_E$ represents the usual Euclidean norm of a vector. We shall also require the matrix ℓ_1 norm, $\|A\|_1 = \max_j \sum_i |a_{ij}|$ and the elementwise ℓ_1 norm, $\|A\|_{L_1} = \sum_{i,j} |a_{ij}|$. The notation $\mathbf{A} \succ 0$ indicates the matrix $\mathbf{A}$ is positive definite and symbols c_0, c_1, and c_2 represent generic constants which may change according to the context but are independent of any model parameters. The notation $a_n = O(b_n)$ represents that $a_n \leq cb_n$ for some constant $c < \infty$ and n large enough. For any set of indices S, its cardinality is denoted by $|S|$ and for a subset $A \subseteq \{1, 2, \cdots, d\}$, $\mathbf{b}_A$ denotes the vector of components of $\mathbf{b}$ with indices in A. Finally we partition a $d \times d$ matrix $\mathbf{\Sigma}$ as follows:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{AA} & \mathbf{\Sigma}_{AA^c} \\ \mathbf{\Sigma}_{A^cA} & \mathbf{\Sigma}_{A^cA^c} \end{pmatrix}, \qquad \text{where } A^c \text{ denotes the compliment set of A.} \tag{2.1}$$

We begin by describing a framework that characterizes structural zeros. As stated in the previous section, these structural zeros represent components that are biologically absent in the specimen. Thus, the framework should allow for the distribution of the specimen to be determined by only the observed components. For this purpose, let $\mathcal{S}$ be the sample space of possible configurations of missing components be as follows.

$$\mathcal{S} = \begin{cases} (1, \ldots, 1), \\ (0, 1, \ldots, 1), (1, 0, \ldots, 1), \ldots, (1, \ldots, 1, 0) \\ (0, 0, 1, ...1), (0, 1, 0.., 1), \ldots, (1, \ldots, 0, 0) \\ . \\ . \\ (0, 0, \ldots, 1), (0, 0, \ldots, 1, 0), \ldots, (1, 0, ...0) \end{cases} . \tag{2.2}$$

Here $0, 1$ correspond to the cases where a component is unobserved or observed in the sample, respectively. We represent each of the above $2^d - 1$ events of the sample space by Configuration (Config.) (j), $j = 1, 2, \ldots, 2^d - 1$, in the order written in (2.2). For example, Config. (1) is the case where all components are observed and Config. $(2^d - 1)$ corresponds to the configuration where only the first component is observed. Assume for the ith sample, the missing structure is generated by independent r.v.'s $\mathbf{M}_i$, $1 \leq i \leq n$, with sample space described in (2.2).

In many applications, including the analysis of microbiome data, it may be unreasonable to assume that the missingness is generated by identically distributed r.v.'s since this distribution function may be influenced by factors or covariates such as geographical location, age, race, and gender of the subject. We allow for this flexibility by defining the distribution of the r.v.'s $\mathbf{M}_i$, $1 \leq i \leq n$ as follows,

$$P\Big(\mathbf{M}_i \text{ is in Config. } (j) \Big) = \delta^i_{(j)}, \quad 0 \leq \delta^i_{(j)} \leq 1, \quad 1 \leq j \leq 2^d - 1. \tag{2.3}$$

Under this general definition each sample may have a different probability of each configuration. This feature of allowing the missingness to be acting independently but not identically is reminiscent of the MAR structure. We now proceed to define the conditional distribution of the observed components of a specimen.

Let $\boldsymbol{\mu} = (\mu_1, ..., \mu_d)^T \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} = [\sigma_{ij}]_{d\times d}$ be a d-dimensional vector and positive definite matrix, respectively. For the ith subject, with missing configuration given by the r.v. $\mathbf{M}_i$, we denote the observed components by the index set

$$A_i = \{j,\, M_{ij} = 1\}. \tag{2.4}$$

Note that the index set A_i is a random set which is determined by the r.v. $\mathbf{M}_i$. Now assume that conditioned on $\mathbf{M}_i$, the components of $\mathbf{X}_i$ with indices in the index set A_i jointly follow a Gaussian distribution with mean and covariance being the corresponding sub-vector of $\boldsymbol{\mu}$ and sub-matrix of $\boldsymbol{\Sigma}$, respectively, i.e. for any $\mathbf{x} \in \mathbb{R}^d$,

$$P\Big(\mathbf{X}_{A_i} \le \mathbf{x}_{A_i}\Big|\mathbf{M}_i\Big) = \Phi_{A_i}(\mathbf{x}_{A_i}), \tag{2.5}$$

where Φ_{A_i} represents the Gaussian distribution function with mean $\boldsymbol{\mu}_{A_i}$ and covariance matrix $\boldsymbol{\Sigma}_{A_iA_i}$. For example let $\mathbf{M}_i = (1, 1, 0, \ldots, 0)$, then the observed vector is $\mathbf{X}_i = (X_{i1}, X_{i2}, 0, \ldots, 0)$ with the conditional distribution of the observed components as $P\Big(X_{i1} \le x_{i1}, X_{i2} \le x_{i2}\Big|\mathbf{M_i}\Big) = \Phi(x_{i1}, x_{i2})$. This completes the description of the distributional structure assumed in this paper. Under this definition, the zero components of an observation do not influence the distribution of the nonzero components, thus characterizing what we refer to as structural zeros.

To proceed further we shall require the following definitions. For all $1 \le l, m \le d$, let $n(l) = \{i \,:\, l \in A_i,\ 1 \le i \le n\}$, and $n(l, m) = \{i \,:\, l, m \in A_i\,, 1 \le i \le n\}$ be the number of subjects where lth component is observed and the number of subjects where both lth and mth components are observed, respectively. Note that these are random quantities determined by the random set A_i which in turn is determined by the r.v.'s $\mathbf{M_i}$, $1 \le i \le n$. Also define for each observation $i = 1, 2, \ldots, n$, and the indices $1 \le l, m \le d$, the collections,

$$C_i(l) = \Big\{1 \le j \le 2^d - 1,\ \text{component } l \text{ is present in Config. } (j) \text{ in r.v. } \mathbf{M}_i\Big\}$$
$$C_i(l, m) = \{1 \le j \le 2^d - 1,\ \text{components } l \text{ and } m \text{ are present in Config. } (j) \text{ in r.v. } \mathbf{M}_i\Big\}$$

In the sequel we make the following additional assumption over the missing structure.

(A1) There exists a constant $\delta_{\min} > 0$ such that for any $1 \le l,\, m \le d$,

$$\text{(i)}\ \frac{1}{n}\sum_{i=1}^{n}\sum_{j\in C_i(l)}\delta^i_{(j)} = \delta(l) > \delta_{\min} \quad \text{and} \quad \text{(ii)}\ \frac{1}{n}\sum_{i=1}^{n}\sum_{j\in C_i(l,m)}\delta^i_{(j)} = \delta(l,m) > \delta_{\min}.$$

The condition **(A1)** is a mild assumption on the missing structure. If the r.v.'s $\mathbf{M_i}$, $1 \le i \le n$ are assumed to be i.i.d., then **(A1)**(i) requires that each component is present in an observation with a nonzero probability, i.e. $\sum_{j\in C(l)}\delta_{(j)} > \delta_{\min}$, and **(A1)**(ii) requires that every pair of components are present in each observation vector with a nonzero probability, i.e. $\sum_{j\in C(l,m)}\delta_{(j)} > \delta_{\min}$.

3. Methodology

In this section we discuss methodologies to estimate the covariance and precision matrices. First we provide a ℓ_1 minimization approach to estimate the precision matrix $\boldsymbol{\Omega}$. Then a generalized thresholding procedure to estimate the covariance matrix $\boldsymbol{\Sigma}$. We shall provide error bounds for estimates obtained

by these methods that hold with asymptotic probability 1, unconditional on the missing structure. These error bounds shall also allow for the dimension d to increase exponentially with the sample size, thus allowing for high dimensions. To describe our methodology we require the following definitions. For each $1 \leq l, m \leq d$ let

$$\hat{\mu}_l = \frac{1}{|n(l)|} \sum_{i \in n(l)} X_{il}, \qquad 1 \leq l \leq d \tag{3.1}$$

and define a "re-normalized sample covariance" matrix $\hat{\mathbf{\Sigma}}$ as follows,

$$\hat{\sigma}_{lm} = \sum_{i \in n(l,m)} (X_{il} - \hat{\mu}_l)(X_{im} - \hat{\mu}_m) \Big/ |n(l,m)| \quad \text{and} \quad \hat{\mathbf{\Sigma}} = \left[\hat{\sigma}_{lm}\right]_{l,m=1,..,d}. \tag{3.2}$$

The matrix $\hat{\mathbf{\Sigma}}$ forms an initial estimator for obtaining consistent estimates of the covariance matrix and the precision matrices in the high-dimensional setting. The following lemma provides an approximation result between $\hat{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}$ in the *Sup* norm and shall be key to providing convergence rates of the estimators to follow later in this section.

Lemma 3.1 Suppose the observations $\mathbf{X}_i$, $1 \leq i \leq n$ follow the distribution (2.5) and that the missing structure satisfies condition **(A1)**. In addition assume that the variance components of the covariance matrix $\mathbf{\Sigma}$ are bounded above, i.e. $\sigma_{jj} \leq \sigma_x$, $1 \leq j \leq p$, for a constant $\sigma_x < \infty$. Then with probability at least $1 - c_1 \exp(-c_2 \log d)$,

$$\left\| \hat{\mathbf{\Sigma}} - \mathbf{\Sigma} \right\|_\infty \leq c_0 \frac{\sigma_x^2}{\delta_{\min}} \sqrt{\frac{\log d}{n}}, \tag{3.3}$$

for some universal constant $c_0 < \infty$.

To appreciate this result note that the re-normalized sample covariance matrix $\hat{\Sigma}$ is defined through the r.v.'s $\mathbf{X}_i$, $1 \leq i \leq n$, whose distribution is in turn defined conditionally of the missing structure $\mathbf{M}_i$. However, Lemma 3.1 provides an unconditional probability bound on the desired random quantity with only a mild assumption **(A1)** on the missing structure. The proof of this result relies on the observation that $|n(l,m)|$, $1 \leq l, m \leq d$ is a sum of independent r.v.'s, i.e. $|n(l,m)| = \sum_{i=1}^n I_{ilm}$ where $I_{ilm} = \mathbf{1}\left[M_{il} = 1 \;\&\; M_{im} = 1\right]$ for every $1 \leq l, m \leq d$. Here $\mathbf{1}$ represents the indicator function. This observation allows the applicability of the Hoeffding's inequality (Hoeffding, 1963) in combination with conditional expectation arguments. The details of the proof are provided in the supplementary material available at *Biostatistics* online. To proceed with the estimation of $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ we require these matrices to belong to the following class of approximately sparse matrices.

(A2) Assume that the covariance and precision matrices belong to the following classes of matrices, respectively: define for $0 \leq q < 1$,

$$\text{(i)}\ \ \mathcal{M}(q, s_o(d), K) = \left\{ \mathbf{\Sigma} \,:\, \sigma_{ii} \leq K,\ \max_{1 \leq i \leq d} \sum_{j=1}^d |\sigma_{ij}|^q \leq s_0(d) \right\} \text{ and}$$

$$\text{(ii)}\ \ \mathcal{U}(q, s_o(d), K) = \left\{ \mathbf{\Omega} \,:\, \mathbf{\Omega} \succ 0,\ \|\mathbf{\Omega}\|_{L_1} \leq K,\ \max_{1 \leq i \leq d} \sum_{j=1}^d |\omega_{ij}|^q \leq s_0(d) \right\}.$$

The quantity $s_0(d)$ is allowed to depend on d and thus is not and explicit restriction on sparsity. Two examples of matrices which satisfy this restrictions are, a p-diagonal matrix that satisfies this condition with any $0 \leq q < 1$ and $s_0(d) = K^q p$. Second, an $AR(1)$ covariance matrix where $\sigma_{ij} = \rho^{|i-j|}$, which satisfies the restriction with $s_0(d) = c_0$ for some constant $c_0 < \infty$.

3.1. *Estimation of the precision matrix* $\boldsymbol{\Omega}$

The problem of estimation of the precision matrix $\boldsymbol{\Omega}$ has received wide attention in the literature. Several solutions have been proposed in the context of i.i.d. sub-Gaussian observation and one such solution has been the penalized likelihood method for which exploits the i.i.d. Gaussian structure of observations (Friedman *and others*, 2008; Ravikumar *and others*, 2011). In our setup however, the observations $\mathbf{X}_i$, $1 \leq i \leq n$ are no longer identically distributed, and hence such likelihood-based approaches are no longer feasible. However this problem can be overcome by adapting penalized moment-based approaches such as the method "Clime" of Cai *and others* (2011) under our setting. Such moment-based approaches to the estimation of high-dimensional precision matrices do not rely on an explicit likelihood functions, instead only require probabilistic bounds on the quantity $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty$, and thus Lemma 3.1 stated earlier forms the connecting link between such approaches and the distributional structure of Section 1 without requiring the observations $\mathbf{X}_i$, $1 \leq i \leq n$ to be identically distributed.

Let $\hat{\boldsymbol{\Omega}}_1$ be the solution of the following convex program,

$$\min \|\boldsymbol{\Omega}\|_1 \quad \text{subject to} \quad \left|\hat{\boldsymbol{\Sigma}}_n \boldsymbol{\Omega} - \mathbf{I}\right|_\infty \leq \lambda_\Omega, \quad \boldsymbol{\Omega} \in \mathbb{R}^{p \times p}, \tag{3.4}$$

with a suitable choice of $\lambda_{\boldsymbol{\Omega}} > 0$. Here $\mathbf{I}$ represents the identity matrix and $\hat{\boldsymbol{\Sigma}}$ as defined in (3.2). Since the solution $\hat{\boldsymbol{\Omega}}_1$ may not be symmetric in general, the final estimate $\hat{\boldsymbol{\Omega}}$ is obtained by symmetrizing $\hat{\boldsymbol{\Omega}}_1 = [\omega^1_{ij}]_{d \times d}$ by choosing the smaller of $|\omega^1_{ij}|$ and $|\omega^1_{ji}|$ in the final estimate $\hat{\boldsymbol{\Omega}}$, i.e. $\hat{\boldsymbol{\Omega}} = (\hat{\omega}_{ij})$, with $\hat{\omega}_{ij} = \hat{\omega}_{ji} = \hat{\omega}^1_{ij}\mathbf{1}[|\omega^1_{ij}| \leq |\hat{\omega}^1_{ji}|] + \hat{\omega}^1_{ji}\mathbf{1}[|\omega^1_{ij}| > |\hat{\omega}^1_{ji}|]$.

The estimator $\hat{\boldsymbol{\Omega}}$ is not guaranteed to be positive definite; however, the following theorem shows that it converges to a positive definite limit with asymptotic probability 1. To ensure positive definiteness, in practice one may add a small constant to the diagonal elements of this matrix, or project the matrix onto the cone of positive definite matrices.

Theorem 3.1 Suppose $\mathbf{X}_i$, $1 \leq i \leq n$ follow the distribution (2.5) and that the missing structure satisfies condition **(A1)**. Also assume that $\boldsymbol{\Omega} \in \mathcal{U}$ and the regularizer is chosen $\lambda_{\boldsymbol{\Omega}} = c_0 K \sigma_x^2 \sqrt{\log d}\big/\delta_{\min}\sqrt{n}$, then the following bounds hold with probability at least $1 - c_1 \exp(-c_2 \log d)$,

$$(i) \quad \|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_\infty = O\Big[\sqrt{\frac{\log d}{n}}\Big]$$

$$(ii) \quad \|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_2 = O\Big[s_0(d)\Big(\sqrt{\frac{\log d}{n}}\Big)^{1-q}\Big], \quad \text{and,}$$

$$(iii) \quad \frac{1}{d}\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_F^2 = O\Big[s_0(d)\Big(\sqrt{\frac{\log d}{n}}\Big)^{2-q}\Big].$$

This methodology was introduced by Cai *and others* (2011) under the standard i.i.d. Gaussian setup, which is implemented using the sample covariance matrix $\hat{\boldsymbol{\Sigma}}^S = \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T/n$ as the initial estimate. The proof for the error bounds of Theorem 3.1 follows by deterministic arguments on the event where the inequality (3.3) holds and follows from the arguments of Cai *and others* (2011), a sketch of this proof has been provided in the supplementary material available at *Biostatistics* online.

3.2. *Estimation of the covariance matrix* $\mathbf{\Sigma}$

In this sub-section we discuss thresholding methods of estimating the covariance matrix $\mathbf{\Sigma}$ in our setting with structural zeros. These methods were first proposed in the standard i.i.d. setting by Wu and Pourahmadi (2003) and have also been studied by Bickel and Levina (2008) and Levina *and others* (2008) amongst several others. Although assuming sparsity on the covariance matrix is a stronger assumption than assuming the same on the precision matrix, however thresholding carries very little computational burden in comparison to the penalized methods such as the one described in the previous sub-section. Thus this procedure forms an attractive alternative to estimating the precision matrix, especially in very high-dimensional problems and real time applications. We adopt this methodology in our setup with structural zeros as follows.

Let $s_\lambda(x)$ be a generalized thresholding operator as defined by Rothman, Levina *and others* (2009). We restate the definition for the convenience of readers. A function $s_\lambda : \mathbb{R} \to \mathbb{R}$ satisfying

$$(i)\ |s_\lambda(x)| \le |x|, \quad (ii)\ s_\lambda(x) = 0 \text{ for } |x| \le \lambda\,, \text{ and } (iii)\ |s_\lambda(x) - x| \le \lambda \tag{3.5}$$

is said to be a generalized thresholding operator. In view of this definition, the covariance matrix $\mathbf{\Sigma}$ can be estimated by elementwise thresholding as, $s_\lambda(\hat{\mathbf{\Sigma}}) = \big[s_\lambda(\hat{\sigma}_{ij})\big]_{i,j=1,\ldots,d}$. The two common examples of these operators are the hard- and soft-thresholding operators defined as,

$$s^h_\lambda(x) = x\mathbf{1}(|x| > \lambda), \qquad s^s_\lambda(x) = \text{sign}(x)(|x| - \lambda)_+, \tag{3.6}$$

respectively. The soft-thresholding operator can also be defined as $s^s_\lambda(x) = \arg\min_\theta \Big\{(\theta - x)^2 + \lambda|\theta|\Big\}$, and has been studied by various authors, including Donoho *and others* (1995) and Tibshirani (1996). The hard-thresholding operator has been investigated by Bickel and Levina (2008) among others. Additional examples of thresholding operators include SCAD of Fan and Li (2001), adaptive Lasso of Zou (2006). The following result provides the consistency of this estimator.

Theorem 3.2 Suppose $\mathbf{X}_i$, $1 \le i \le n$ follow the distribution (2.5) and that the missing structure satisfies condition **(A1)**. Also assume that s_λ satisfies condition (3.5). Then uniformly on $\mathcal{M}(q, s_0(d), K)$ choosing the regularizer $\lambda = c_0\sigma_x^2\sqrt{\log d}\big/\delta_{\min}\sqrt{n}$ we obtain

$$\big\|s_\lambda(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\big\|_2 = O\Big[s_0(d)\Big(\sqrt{\frac{\log d}{n}}\Big)^{1-q}\Big], \tag{3.7}$$

with probability at least $1 - c_1\exp(-c_2\log d)$.

In the standard i.i.d. setting, Rothman, Levina *and others* (2009) introduced this generalized thresholding methodology based on usual sample covariance matrix $\hat{\mathbf{\Sigma}}^S$. The proof of this theorem relies on deterministic arguments of Rothman, Levina *and others* (2009) on the set where the bound provided in Lemma 3.1 holds and a sketch is provided in the supplementary material available at *Biostatistics* online.

4. SIMULATION STUDY

In this section we numerically illustrate that the methodology of Section 3 provides consistent estimates of Σ and Ω. We confirm that our method provides a significant improvement over the typical method of using the sample covariance $\hat{\mathbf{\Sigma}}^S = \sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T/n$ as the initial estimate in the methods of Sections 3.1 and 3.2, note that using the sample covariance matrix ignores the presence of these structural zeros.

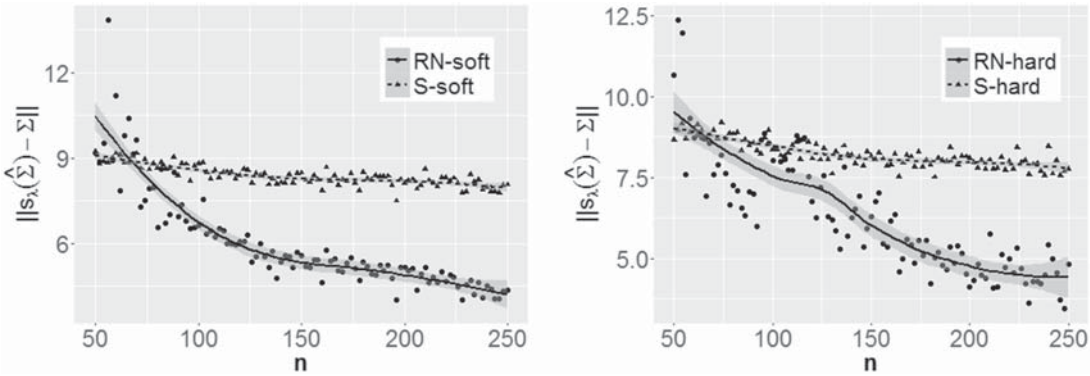

Fig. 1. Comparisons of Frobenius norms of error in estimation due to soft (left) and hard (right) thresholding, x-axis: n, y-axis: $\|s_\lambda(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\|_F$, results at d=75.

4.1. *Simulation setup and results*

The missingness is generated by r.v.'s $\mathbf{M}_i = \left(M_{i1}, ..., M_{id}\right)$ where each $M_{ij} \sim^{i.i.d.}$ Bernoulli$(1 - \rho_j)$, $1 \le i \le n$, $1 \le j \le d$. Here ρ_j denotes the probability of jth component being a structural zero. These ρ_j's $1 \le j \le d$ are chosen by uniformly between $(0, 0.35)$. Finally, for each $1 \le i \le n$, the components of $\mathbf{X}_i$ with indices in the set A_i defined in (2.5) are assumed to be zero mean Gaussian r.v.'s with the covariance being the corresponding sub-block of the matrix $\mathbf{\Sigma}_{A_iA_i}$.

We set the covariance matrix as $\mathbf{\Sigma} = [\sigma_{ij}]_{i,j=1,..,p}$ where $\sigma_{ij} = 0.5^{|i-j|}$ and all entries <0.01 are set to zero. Under this setting we estimate $\mathbf{\Sigma}$ as described in Section 3.2 by the hard- and soft-thresholding operations of (3.6) on the initial re-normalized (RN) estimate (3.2), we shall refer to these estimates as "RN-hard" and "RN-soft," respectively. Similarly we estimate $\mathbf{\Omega}$ as described in Section 3.1 and refer to the estimates obtained as "RN-clime." We also illustrate that our method provides a significant improvement over the standard method of using the sample covariance $\hat{\mathbf{\Sigma}}^{\mathrm{S}} = \sum_{i=1}^{n} \mathbf{X}_i\mathbf{X}_i^T/n$ as the initial estimate for the hard/soft thresholding and clime procedures, we refer to these as "S-hard," "S-soft," and "S-clime," respectively.

We repeat simulations on $\sim$ 100 independent data sets that are generated under the following settings. (i) d is fixed at 75 and n is allowed to vary from 50 to 250 with increments of 2, leading to 101 independent data sets; and (ii) d is allowed to vary from 10 to 100 with increments of 15 and n is allowed to vary from 50 to 200 with increments of 10 leading to 112 independent data sets.

Figures 1 and 2 illustrate the results of our simulation. Here each dot (triangle) represents Frobenius norm of estimation error for each independent data set. To measure the average performance over the simulated models, nonparametric regression lines are fit via the Loess method with its smoothing parameter set at 0.75. The solid line represents the average performance of our methods and the dotted line represents the average performance of the standard method which ignores the presence of structural zeros. For the soft-/hard-thresholding case, the tuning parameter λ is chosen via cross-validation under the Frobenius loss, and for the "clime" method, the tuning parameter is chosen by cross-validating with the loss function $\mathrm{Tr}(\hat{\mathbf{\Sigma}}\hat{\mathbf{\Omega}} - \mathbf{I})^2$.

Results of thresholding procedures are provided in Figure 1. This figure plots $||s_\lambda(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}||_F$ for different sample sizes n for the case $d = 75$. From this figure, it is clear that the error corresponding to the proposed estimator under both soft and hard thresholding tends to get smaller with sample size faster than the standard estimator that ignores the structural zeros. The left panel of Figure 2 plots of the errors for different values of scaled sample sizes $n/\log d$, for the soft-thresholding operator. This figure seems to confirm the result of Theorem 3.2 regarding the rate of convergence of the estimator. Similarly, the

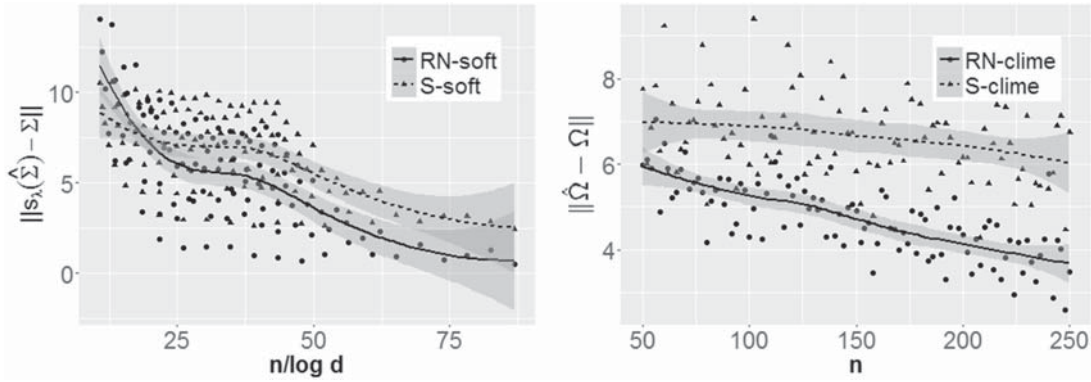

Fig. 2. Left panel: Comparison of Frobenius norms of error in estimation due to soft thresholding the renormalized sample covariance at scaled sample size, x-axis: $n/\log d$, $n \in [50, 200]$, $d \in [10, 100]$; y-axis: $\|s^s_\lambda(\hat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\|_F$. Right panel: Comparison of spectral norms of error in estimation due to Clime procedure on the renormalized sample covariance, $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_F$ and Clime procedure on the standard sample covariance $\|\hat{\mathbf{\Omega}^S} - \mathbf{\Omega}\|_F$. Results at $d = 75$.

right panel of Figure 2 describes the results of the simulation for the "clime" methodology described in Section 3.1.

5. Analysis of global human gut microbiome data

In this section we analyze the global human gut microbiome data of Yatsunenko *and others* (2012) available at the repository MG-RAST (http://metagenomics.anl.gov/) under accession numbers qiime:621 for fecal microbiome shotgun sequencing data sets. Here we estimate the precision matrix $\mathbf{\Omega}$ and use this estimate to classify observations according to geographical locations. Further details on the availability of data have been provided in the supplementary material available at *Biostatistics* online.

The data consist of microbial OTU counts obtained from individuals of United States (US), Venezuela (VE), and Malawi (MA). Our analysis is based on the "genus" level of bacterial taxonomy. We subdivide the data into two age categories "under 2 years" and "at least 2 years." This stratification is done since it is known that microbial composition of infants changes drastically when they switch over from breast milk (or formula milk) to solid food (Lozupone *and others*, 2013). The sample sizes in the two strata (under 2 years, at least 2 years) for US, VE and MA samples are (70, 225), (15, 74) and (32, 72), respectively. After several pre-processing steps of the raw OTU data we obtain for each age group and each location, a matrix of observations with $d = 75$ microbes and inherent structural zeros. Each row of these matrices is assumed to follow the conditional Gaussian distribution of (2.5). Note that for the "under 2 years" category for the pair VE–MA, the total number of observations is 47 and thus we have a high-dimensional scenario. The pre-processing of data is described in detail in the supplementary material available at *Biostatistics* online. For each pair of locations, we randomly split 5/6-th data into training set and the remaining one-sixth as the test set. The training set is used to estimate the common precision matrix using the procedure described in Section 3.

5.1. *Tuning parameter*

The regularizer $\lambda_{\mathbf{\Omega}}$ is chosen via 5-fold cross-validation within the combined training data for each pair of locations. The loss function used to evaluate cross-validation error is $\mathrm{Tr}(\hat{\mathbf{\Sigma}}\hat{\mathbf{\Omega}} - \mathbf{I})^2$. Also, in the construction of $\hat{\mathbf{\Sigma}}$ in (3.2) if a pair (l, m) does not occur then we set the pairwise covariance to zero.

Table 1. *Percentages of correct classification between locations*

	US–MA			US–VE			VE–MA		
Age/ % Correct	**US**	**MA**	**Total**	**US**	**VE**	**Total**	**VE**	**MA**	**Total**
Age $<$ 2 years	87.5	80.9	85.6	79.6	81.2	79.8	76.1	60.5	69.4
Age $\geq$ 2 years	92.8	95.0	93.3	80.9	73.3	73.3	62.1	57.5	60.0

5.2. *Classification of subjects to geographical locations*

In this section we exploit the assumed conditional Gaussian structure of observations to classify subjects of the test set to one of two geographical locations by using estimates of the corresponding precision matrices.We perform pairwise classifications of samples into (i) US and MA, (ii) US and VE, and finally (iii) VE and MA. Let $\hat{\boldsymbol{\mu}}_r, r = 1, 2$ be the estimated d-dimensional mean vector obtained as in (3.1) for each of the two populations under consideration, and $\boldsymbol{\Omega}$ be the common precision matrix of the two locations.

Let $\mathbf{X} = (X_1, .., X_d)^T$ represent the observation to be classified and let $A = \{j \ ; \ X_j \neq 0\}$ denote the collection of indices of the non-zero components of $\mathbf{X}$. For location $r = 1, 2$, We implement the following classification rule: let

$$\delta_r(\mathbf{X}_A) = \mathbf{X}_A^T \hat{\boldsymbol{\Sigma}}_{AA}^{-1} \hat{\boldsymbol{\mu}}_{rA} - \frac{1}{2} \hat{\boldsymbol{\mu}}_{rA}^T \hat{\boldsymbol{\Sigma}}_{AA}^{-1} \hat{\boldsymbol{\mu}}_{rA} \tag{5.1}$$

denote the linear discriminant function where an observation $\mathbf{X}_A$ is classified into population 1 if $\delta_1(\mathbf{X}_A) > \delta_2(\mathbf{X}_A)$, otherwise classified into population 2. The percentage of correctly classified observations from the test set is computed and the above process is repeated 20 times to obtain average correct classification percentages. The results for the two age groups and for every pair of locations are summarized in Table 1.

The classification results depend on the two populations under consideration. The trends in the correct classification rates (column "Total" in Table 1) are the same in both age categories, the best being US–MA and the worst VE–MA. A possible reason for this is that the microbial composition between US and MA is more different relative to the other two pairs. This is illustrated in Figure 3, which plots the largest 25 absolute differences in the signal-to-noise ratios (SNR) for each of the three pairs of location for the "at least 2 years" age group. More precisely, for each $j, j = 1, 2, \ldots, d$, we plot $\left|(\hat{\mu}_j^{\mathrm{US}}/\hat{\sigma}_j^{\mathrm{US}}) - (\hat{\mu}_j^{\mathrm{MA}}/\hat{\sigma}_j^{\mathrm{MA}})\right|$, $\left|(\hat{\mu}_j^{\mathrm{US}}/\hat{\sigma}_j^{\mathrm{US}}) - (\hat{\mu}_j^{\mathrm{VE}}/\hat{\sigma}_j^{\mathrm{VE}})\right|$, and $\left|(\hat{\mu}_j^{\mathrm{VE}}/\hat{\sigma}_j^{\mathrm{VE}}) - (\hat{\mu}_j^{\mathrm{MA}}/\hat{\sigma}_j^{\mathrm{MA}})\right|$. This decreasing trend in the absolute difference of SNRs is a possible reason for the lower correct classification rate between MA and VE than between US and MA.

6. Discussion

New technologies such as the 16s RNA sequencing have yielded high-dimensional data with characteristics that cannot be modeled by standard i.i.d. formulations of multivariate data. In this paper we describe one such characteristic, namely "structural zeros," which are encountered in microbiome studies. We proposed a conditional Gaussian distributional structure that characterizes these zeros and provide methods to estimate covariance and precision matrices in this context. We show that in spite of the distribution being conditional, it is indeed possible to obtain results that are unconditional. As future work, we believe that the conditional Gaussian distributional structure proposed in this paper can be used to carry forward the work of Kurkland and Heagerty (2005) in the high-dimensional setting where the covariates are subjected to structural zeros.

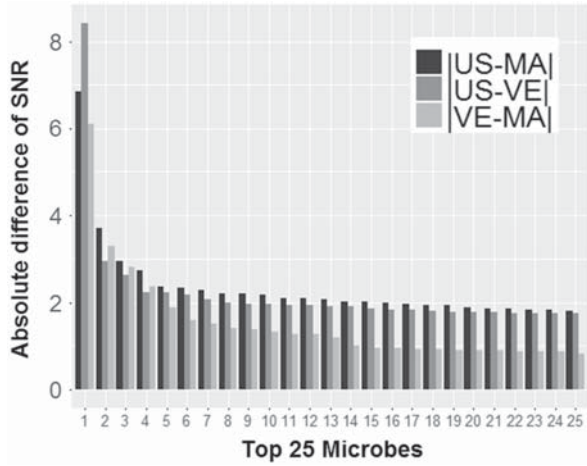

Fig. 3. Comparison of absolute difference in SNR for each pair of locations in descending order for the top 25 microbes. Results for the "at least 2 years age" category.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

We thank a reviewer and the associate editor for several useful suggestions that helped improve this manuscript.

Conflict of Interest: None declared.

FUNDING

Intramural Research Program of the NIH, NIEHS (Z01 ES101744-04) to S.D.P. and A.K.; Israeli Science Foundation (1256/13) to O.D.

REFERENCES

BICKEL, P. AND LEVINA, E. (2008). Covariance regularization by thresholding. *Annals of Statistics* 36, 2577–2604.

CAI, T., LIU, W. AND LUO, X. (2011). A constrained l_1 minimization approach to sparse precision matrix estimation. *Journal of American Statistical Association*, 106, 594–607.

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. AND PICKARD, D. (1995). Wavelet shrinakge: asymptopia? *Journal of the Royal Statistical Society*, Series. B, 57, 301–369.

FAN, J. AND LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360.

FRIEDMAN, J., HASTIE, T., TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9, 432–441.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded variables. *Journal of American Statistical Association*, 58, 13–30.

KAUL, A., KOUL, H. L., CHAWLA, A. AND LAHIRI, S. N. (2016). Two stage non-penalized corrected least squares for high dimensional linear models with measurement error or missing covariates. arXiv preprint arXiv:1605.03154v1.

KURKLAND, B. F. AND HEAGERTY, P. J. (2005). Directly parameterized regression conditioning on being alive: analysis of longitudinal data truncated by deaths. *Biostatistics* 6, 241–258.

LEVINA, E., ROTHMAN, A., AND ZHU, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. *Annals of Applied Statistics*, 1, 245–263.

LOH, P., AND WAINWRIGHT, M. J. (2012). High-dimensional regression with noisy and missing data: provable guarantees with non-convexity. *Annals of Statistics* 40, 1637–1664.

LOUNICI, K. (2014). High dimensional covariance matrix estimation with missing observations. *Bernoulli* 20, 1029–1058.

LOZUPONE, C. A., STOMBAUGH, J., GONZALEZ, A., ACKERMAN, G., WENDEL, D., VAZQUEZ-BAEZA, Y., JANSSON, J. K., GORDON, J.I., KNIGHT R. (2013). Meta-analyses of studies of the human microbiota. *Genome Research* 23(10), 1704–1714.

MANDAL, S., TEUREN, W. V., WHITE, R. A., EGGESBO, M., KNIGHT, R. AND PEDDADA, S. (2015) Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease* 26 1651–2235.

RAVIKUMAR, P., WAINWRIGHT, M., RASKUTTI, G., AND YU, B. (2011). High dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics* 5, 935–980.

ROTHMAN, A., BICKEL, P., LEVINA, E., AND ZHU, J. (2009). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* 2, 494–515.

ROTHMAN, A., LEVINA, E., AND ZHU, J. (2009). Generalized thresholding of large covariance matrices. *Journal of American Statistical Association* 104, 177–186.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58, 267–288.

WU, W. B. AND POURAHMADI, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* 90, 831–844.

YATSUNENKO, T., REY, F. E., MANARY, M. J., TREHAN, I., DOMINGUEZ-BELLO, M. G., CONTRERAS, M., MAGRIS, M., HIDALGO, G., BALDASSANO, R. N., ANOKHIN, A. P., *and others*. (2012). Human gut microbiome viewed across age and geography. *Nature*, 486, 222–227.

ZOU, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429.

[*Received May 17, 2016; revised August 30, 2016; accepted for publication November 8, 2016*]