High-dimensional principal component analysis with heterogeneous missingness

Ziwei Zhu[*,‡], Tengyao Wang[*,†] and Richard J. Samworth[*]

[*] *Statistical Laboratory, University of Cambridge*
[†] *Department of Statistics, London School of Economics*
[‡] *Department of Statistics, University of Michigan, Ann Arbor*

E-mail: ziweiz@umich.edu, t.wang59@lse.ac.uk, r.samworth@statslab.cam.ac.uk

Summary. We study the problem of high-dimensional Principal Component Analysis (PCA) with missing observations. In a simple, homogeneous observation model, we show that an existing observed-proportion weighted (OPW) estimator of the leading principal components can (nearly) attain the minimax optimal rate of convergence, which exhibits an interesting phase transition. However, deeper investigation reveals that, particularly in more realistic settings where the observation probabilities are heterogeneous, the empirical performance of the OPW estimator can be unsatisfactory; moreover, in the noiseless case, it fails to provide exact recovery of the principal components. Our main contribution, then, is to introduce a new method, which we call `primePCA`, that is designed to cope with situations where observations may be missing in a heterogeneous manner. Starting from the OPW estimator, `primePCA` iteratively projects the observed entries of the data matrix onto the column space of our current estimate to impute the missing entries, and then updates our estimate by computing the leading right singular space of the imputed data matrix. We prove that the error of `primePCA` converges to zero at a geometric rate in the noiseless case, and when the signal strength is not too small. An important feature of our theoretical guarantees is that they depend on average, as opposed to worst-case, properties of the missingness mechanism. Our numerical studies on both simulated and real data reveal that `primePCA` exhibits very encouraging performance across a wide range of scenarios, including settings where the data are not Missing Completely At Random.

1. Introduction

One of the ironies of working with Big Data is that missing data play an ever more significant role, and often present serious difficulties for analysis. For instance, a common approach to handling missing data is to perform a so-called *complete-case analysis* (Little and Rubin, 2019), where we restrict attention to individuals in our study with no missing attributes. When relatively few features are recorded for each individual, one can frequently expect a sufficiently large proportion of complete cases that, under an appropriate missing at random hypothesis, a complete-case analysis may result in only a relatively small loss of efficiency. On the other hand, in high-dimensional regimes where there are many features of interest, there is often such a small proportion of complete cases that this approach becomes infeasible. As a very simple illustration of this phenomenon, imagine an $n \times d$ data matrix in which each entry is missing independently with probability 0.01. When $d = 5$, a complete-case analysis would result in around 95% of the individuals (rows) being retained, but even when we reach $d = 300$, only around 5% of rows will have no missing entries.

The inadequacy of the complete-case approach in many applications has motivated numerous methodological developments in the field of missing data over the past 60 years or so, including imputation (Ford, 1983; Rubin, 2004), factored likelihood (Anderson, 1957) and maximum likelihood approaches (Dempster, Laird and Rubin, 1977); see, e.g., Little and Rubin (2019) for an introduction to the area. Recent years have also witnessed increasing emphasis on understanding the performance of methods for dealing with missing data in a variety of high-dimensional problems, including sparse regression (Loh and Wainwright, 2012; Belloni, Rosenbaum and Tsybakov, 2017), classification (Cai and Zhang, 2018b), sparse principal component analysis (Elsener and van de Geer, 2018) and covariance and precision matrix estimation (Lounici, 2014; Loh and Tan, 2018).

In this paper, we study the effects of missing data in one of the canonical problems of high-dimensional data analysis, namely dimension reduction via Principal Component Analysis (PCA).

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/rssb.12550

This is closely related to the topic of *matrix completion*, which has received a great deal of attention in the literature over the last decade or so (e.g. Candès and Recht, 2009; Candès and Plan, 2010; Keshavan, Montanari and Oh, 2010; Mazumder, Hastie and Tibshirani, 2010; Koltchinskii, Lounici and Tsybakov, 2011; Candès et al., 2011; Negahban and Wainwright, 2012). There, the focus is typically on accurate recovery of the missing entries, subject to a low-rank assumption on the signal matrix; by contrast, our focus is on estimation of the principal eigenspaces. Previously proposed methods for low-dimensional PCA with missing data include non-linear iterative partial least squares (Wold and Lyttkens, 1969), iterative PCA (Kiers, 1997; Josse and Husson, 2012) and its regularised variant (Josse et al., 2009); see Dray and Josse (2015) for a nice survey and comparative study. More broadly, the `R-miss-tastic` website `https://rmisstastic.netlify.com/` provides a valuable resource on methods for handling missing data.

The importance of the problem of high-dimensional PCA with missing data derives from its wide variety of applications. For instance, in many commercial settings, one may have a matrix of customers and products, with entries recording the number of purchases. Naturally, there will typically be a high proportion of missing entries. Nevertheless, PCA can be used to identify items that distinguish the preferences of customers particularly effectively, to make recommendations to users of products they might like and to summarise efficiently customers' preferences. Later, we will illustrate such an application, on the Million Song Dataset, where we are able to identify particular songs that have substantial discriminatory power for users' preferences as well as other interesting characteristics of the user database. Other potential application areas include health data, where one may seek features that best capture the variation in a population, and where the corresponding principal component scores may be used to cluster individuals into subgroups (that may, for instance, receive different treatment regimens).

To formalise the problem we consider, suppose that the (partially observed) matrix $n \times d$ matrix $\mathbf{Y}$ is of the form

$$\mathbf{Y} = \mathbf{X} + \mathbf{Z}, \tag{1}$$

for independent random matrices $\mathbf{X}$ and $\mathbf{Z}$, where $\mathbf{X}$ is a low-rank matrix and $\mathbf{Z}$ is a noise matrix with independent and identically distributed entries having zero mean. The low-rank property of $\mathbf{X}$ is encoded through the assumption that it is generated via

$$\mathbf{X} = \mathbf{U}\mathbf{V}_K^\top, \tag{2}$$

where $\mathbf{V}_K \in \mathbb{R}^{d\times K}$ has orthonormal columns and $\mathbf{U}$ is a random $n \times K$ matrix (with $n > K$) having independent and identically distributed rows with mean zero and covariance matrix $\mathbf{\Sigma_u}$. Note that when $\mathbf{X}$ and $\mathbf{Z}$ are independent, the covariance matrix of $\mathbf{Y}$ has a K-spiked structure; such covariance models have been studied extensively in both theory and applications (Paul, 2007; Johnstone and Lu, 2009; Cai, Ma and Wu, 2013; Fan, Liao and Mincheva, 2013).

We are interested in estimating the column space of $\mathbf{V}_K$, denoted by $\mathrm{Col}(\mathbf{V}_K)$, which is also the K-dimensional leading eigenspace of $\mathbf{\Sigma_y} := n^{-1}\mathbb{E}(\mathbf{Y}^\top\mathbf{Y})$. Cho, Kim and Rohe (2017) considered a different but related model where $\mathbf{U}$ in (2) is deterministic, and is not necessarily centred, so that $\mathbf{V}_K$ is the top K right singular space of $\mathbb{E}(\mathbf{Y})$. (By contrast, in our setting, $\mathbb{E}(\mathbf{Y}) = \mathbf{0}$, so the mean structure is uninformative for recovering $\mathbf{V}_K$.) Their model can be viewed as being obtained from the model (1) and (2) by conditioning on $\mathbf{U}$. In the context of a *p-homogeneous Missing Completely At Random (MCAR) observation model*, where each entry of $\mathbf{Y}$ is observed independently with probability $p \in (0, 1)$ (independently of $\mathbf{Y}$), Cho, Kim and Rohe (2017) studied the estimation of $\mathrm{Col}(\mathbf{V}_K)$ by $\mathrm{Col}(\widehat{\mathbf{V}}_K)$, where $\widehat{\mathbf{V}}_K$ is a simple estimator formed as the top K eigenvectors of an observed-proportion weighted (OPW) version of the sample covariance matrix (here, the weighting is designed to achieve approximate unbiasedness). Our first contribution, in Section 2, is to provide a detailed, finite-sample analysis of this estimator in the model given by (1) and (2) together with a p-homogeneous MCAR missingness structure, with a noise level of constant order. The differences between the settings necessitate completely different arguments, and reveal in particular a new phenomenon in the form of a phase transition in the attainable risk bound for the $\sin\Theta$ loss function, i.e. the Frobenius norm of the diagonal matrix of the sines of the principal angles between $\widehat{\mathbf{V}}_K$ and $\mathbf{V}_K$. Moreover, we also

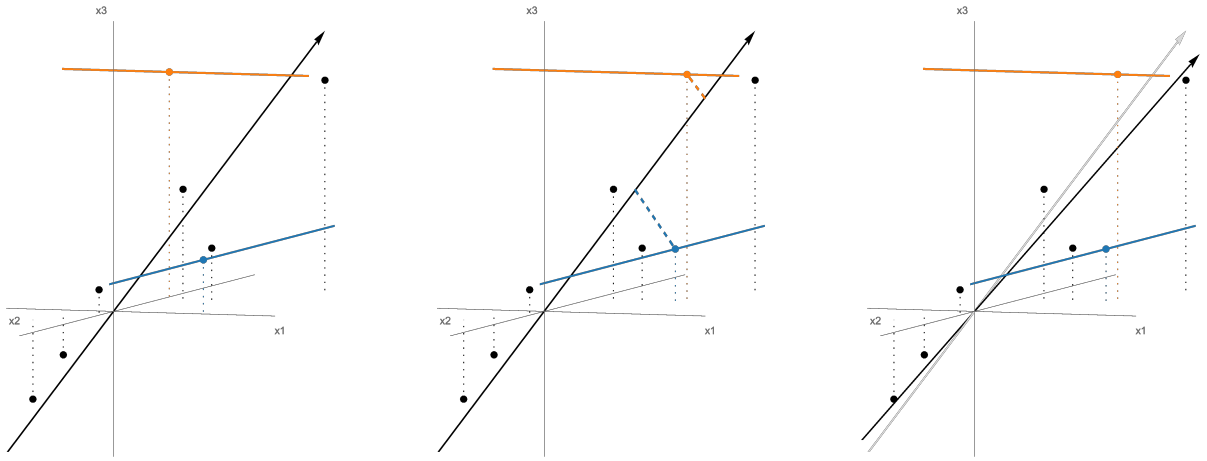

Fig. 1. An illustration of the two steps of a single iteration of the `primePCA` algorithm with $d = 3$ and $K = 1$. Black dots represent fully observed data points, while vertical dotted lines that emanate from them give an indication of their x_3 coordinate values, as well as their projections onto the x_1-x_2 plane. The x_1 coordinate of the orange data point and the x_2 coordinate of the blue data point are unobserved, so the true observations lie on the respective solid lines through those points (which are parallel to the relevant axes). Starting from an input estimate of $\mathbf{V}_K$ (left), given by the black arrow, we impute the missing coordinates as the closest points on the coloured lines to $\mathbf{V}_K$ (middle), and then obtain an updated estimate of $\mathbf{V}_K$ as the leading right singular vector of the imputed data matrix (right, with the old estimate in grey).

provide a minimax lower bound in the case of estimating a single principal component, which reveals that this estimator achieves the minimax optimal rate up to a poly-logarithmic factor.

While this appears to be a very encouraging story for the OPW estimator, it turns out that it is really only the starting point for a more complete understanding of high-dimensional PCA with missing data. For instance, in the noiseless case, the OPW estimator fails to provide exact recovery of the principal components. Moreover, it is the norm rather than the exception in applications that missingness is *heterogeneous*, in the sense that the probability of observing entries of $\mathbf{Y}$ varies (often significantly) across columns. For instance, in recommendation systems, some products will typically be more popular than others, and hence we observe more ratings in those columns. As another example, in meta-analyses of data from several studies, it is frequently the case that some covariates are common across all studies, while others appear only in a reduced proportion of them. In Section 2.2, we present an example to show that, even with an MCAR structure, PCA algorithms can break down entirely for such heterogeneous observation mechanisms when individual rows of $\mathbf{V}_K$ can have large Euclidean norm. Intuitively, if we do not observe the interaction between the jth and kth columns of $\mathbf{Y}$, then we cannot hope to estimate the jth or kth rows of $\mathbf{V}_K$, and this will cause substantial error if these rows of $\mathbf{V}_K$ contain significant signal. This example illustrates that it is only possible to handle heterogeneous missingness in high-dimensional PCA with additional structure, and indicates that it is natural to measure the difficulty of the problem in terms of the *incoherence* among the entries of $\mathbf{V}_K$ — i.e., the maximum Euclidean norm of the rows of $\mathbf{V}_K$.

Our main contribution, then, is to propose a new, iterative algorithm, called `primePCA` (short for projected refinement for imputation of missing entries in Principal Component Analysis), in Section 3, to estimate $\mathbf{V}_K$, even with heterogeneous missingness. The main initialiser that we study for this algorithm is a modified version of the simple estimator discussed above, where the modification accounts for potential heterogeneity. Each iteration of `primePCA` projects the observed entries of $\mathbf{Y}$ onto the column space of the current estimate of $\mathbf{V}_K$ to impute missing entries, and then updates our estimate of $\mathbf{V}_K$ by computing the leading right singular space of the imputed data matrix. An illustration of the two steps of a single iteration of the `primePCA` algorithm in the case $d = 3$, $K = 1$ is given in Figure 1.

Our theoretical results reveal that in the noiseless setting, i.e., $\mathbf{Z} = \mathbf{0}$, `primePCA` achieves exact recovery of the principal eigenspaces (with a geometric convergence rate) when the initial estimator

is close to the truth and a sufficiently large proportion of the data are observed. Moreover, we also provide a performance guarantee for the initial estimator, showing that under appropriate conditions it satisfies the desired requirement with high probability, conditional on the observed missingness pattern. Code for our algorithm is available in the `R` package `primePCA` (Zhu, Wang and Samworth, 2019).

To the best of our knowledge, `primePCA` is the first method for high-dimensional PCA that is designed to cope with settings where missingness is heterogeneous. Indeed, the previously mentioned works on high-dimensional PCA and other high-dimensional statistical problems with missing data have either focused on a uniform missingness setting or have imposed a lower bound on entrywise observation probabilities, which reduces to this uniform case. In particular, such results fail to distinguish in terms of the performance of their algorithms between a setting where one variable is observed with a very low probability p and all other variables are fully observed, and a setting where all variables are observed with probability p. A key contribution of our work is to account explicitly for the effect of a heterogeneous missingness mechanism, where the estimation error depends on average entrywise missingness rather than worst-case missingness; see the discussions after Theorem 4 and Proposition 2 below. In Section 4, the empirical performance of `primePCA` is compared with both that of the initialiser, and a popular method for matrix completion called `softImpute` (Mazumder, Hastie and Tibshirani, 2010; Hastie et al., 2015); we also discuss maximum likelihood approaches implemented via the Expectation–Maximisation (EM) algorithm, which can be used when the dimension is not too high. Our settings include a wide range of signal-to-noise ratios, as well as Missing Completely At Random, Missing At Random and Missing Not At Random examples (Little and Rubin, 2019; Seaman et al., 2013). These comparisons reveal that `primePCA` provides highly accurate and robust estimation of principal components, for instance outperforming the `softImpute` algorithm, even when the latter is allowed access to the oracle choice of regularisation parameter for each dataset. Our analysis of the Million Song Dataset is given in Section 5. In Section 6, we illustrate how some of the ideas in this work may be applied to other high-dimensional statistical problems involving missing data. Proofs of our main results are deferred to Section A in the supplmentary material (Zhu, Wang and Samworth, 2021); auxiliary results and their proofs are given in Section B of the supplementary material.

1.1. Notation

For a positive integer T, we write $[T] := \{1, \ldots, T\}$. For $\mathbf{v} = (v_1, \ldots, v_d)^\top \in \mathbb{R}^d$ and $p \in [1, \infty)$, we define $\|\mathbf{v}\|_p := \big(\sum_{j=1}^d |v_j|^p\big)^{1/p}$ and $\|\mathbf{v}\|_\infty := \max_{j\in[d]} |v_j|$. We let $\mathcal{S}^{d-1} := \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\|_2 = 1\}$ denote the unit Euclidean sphere in $\mathbb{R}^d$.

Given $\mathbf{u} = (u_1, \ldots, u_d)^\top \in \mathbb{R}^d$, we write $\mathrm{diag}(\mathbf{u}) \in \mathbb{R}^{d\times d}$ for the diagonal matrix whose jth diagonal entry is u_j. We let $\mathbb{O}^{d_1\times d_2}$ denote the set of matrices in $\mathbb{R}^{d_1\times d_2}$ with orthonormal columns. For a matrix $\mathbf{A} = (A_{ij}) \in \mathbb{R}^{d_1\times d_2}$, and $p, q \in [1, \infty]$, we write $\|\mathbf{A}\|_p := \big(\sum_{i,j} |A_{ij}|^p\big)^{1/p}$ if $1 \le p < \infty$ and $\|\mathbf{A}\|_\infty := \max_{i,j} |A_{ij}|$ for its entrywise ℓ_p norm, as well as $\|\mathbf{A}\|_{p\to q} := \sup_{\|\mathbf{v}\|_p=1} \|\mathbf{A}\mathbf{v}\|_q$ for its p-to-q operator norm. We provide special notation for the (Euclidean) operator norm and the Frobenius norm by writing $\|\mathbf{A}\|_{\mathrm{op}} := \|\mathbf{A}\|_{2\to 2}$ and $\|\mathbf{A}\|_{\mathrm{F}} := \|\mathbf{A}\|_2$ respectively. We also write $\sigma_j(\mathbf{A})$ for the jth largest singular value of $\mathbf{A}$, and define its nuclear norm by $\|\mathbf{A}\|_* := \sum_{j=1}^{\min(d_1,d_2)} \sigma_j(\mathbf{A})$. If $S \subseteq [n]$, we write $\mathbf{A}_S \in \mathbb{R}^{|S|\times d}$ for the matrix obtained by extracting the rows of $\mathbf{A}$ that are in S. For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d_1\times d_2}$, the Hadamard product of $\mathbf{A}$ and $\mathbf{B}$, denoted $\mathbf{A} \circ \mathbf{B}$, is defined such that $(\mathbf{A} \circ \mathbf{B})_{ij} = A_{ij}B_{ij}$ for any $i \in [d_1]$ and $j \in [d_2]$.

2. The observed-proportion weighted estimator

In this section, we study a simple observed-proportion weighted (OPW) estimator of the matrix of principal components. To define the estimator, let $\mathcal{A}_{ij}$ denote the event that the (i, j)th entry y_{ij} of $\mathbf{Y}$ is observed. We define the revelation matrix $\mathbf{\Omega} = (\omega_{ij}) \in \mathbb{R}^{n\times d}$ by $\omega_{ij} := \mathbb{1}_{\mathcal{A}_{ij}}$, and the partially observed data matrix

$$\mathbf{Y}_{\mathbf{\Omega}} := \mathbf{Y} \circ \mathbf{\Omega}. \tag{3}$$

Our observed data are the pair $(\mathbf{Y}_{\mathbf{\Omega}}, \mathbf{\Omega})$. Importantly, the fact that we observe $\mathbf{\Omega}$ allows us to distinguish between observed zeros and missing entries (even though these also appear as zeros in $\mathbf{Y}_{\mathbf{\Omega}}$). We first consider the simplest possible case, which we refer to as the p-homogeneous observation model, where entries of the data matrix $\mathbf{Y}$ are observed independently and completely at random (i.e., independent of $(\mathbf{U}, \mathbf{Z})$), each with probability p. Thus, $\mathbb{P}(\mathcal{A}_{ij}) = p \in (0,1)$ for all $i \in [n], j \in [d]$, and $\mathcal{A}_{ij}$ and $\mathcal{A}_{i'j'}$ are independent for $(i,j) \neq (i',j')$.

For $i \in [n]$, let $\mathbf{y}_i^\top$ and $\boldsymbol{\omega}_i^\top$ denote the ith rows of $\mathbf{Y}$ and $\mathbf{\Omega}$ respectively, and define $\widetilde{\mathbf{y}}_i := \mathbf{y}_i \circ \boldsymbol{\omega}_i$. Writing $\mathbf{P} := \mathbb{E}\boldsymbol{\omega}_1\boldsymbol{\omega}_1^\top$ and $\mathbf{W}$ for its entrywise inverse, we have that under the p-homogeneous observation model, $\mathbf{P} = p^2\{\mathbf{1}_d\mathbf{1}_d^\top - (1-p^{-1})\mathbf{I}_d\}$ and $\mathbf{W} = p^{-2}\{\mathbf{1}_d\mathbf{1}_d^\top - (1-p)\mathbf{I}_d\}$. Following Lounici (2013, 2014) and Cho, Kim and Rohe (2017), we consider the following weighted sample covariance matrix:

$$\mathbf{G} := \left(\frac{1}{n}\mathbf{Y}_{\mathbf{\Omega}}^\top\mathbf{Y}_{\mathbf{\Omega}}\right) \circ \mathbf{W} = \left(\frac{1}{n}\sum_{i=1}^n \widetilde{\mathbf{y}}_i\widetilde{\mathbf{y}}_i^\top\right) \circ \mathbf{W}.$$

The reason for including the weight $\mathbf{W}$ is to ensure that $\mathbb{E}(\mathbf{G}|\mathbf{Y}) = n^{-1}\mathbf{Y}^\top\mathbf{Y}$, so that $\mathbf{G}$ is an unbiased estimator of $\mathbf{\Sigma}_{\mathbf{y}}$. Related ideas appear in the work of Cai and Zhang (2016) on high-dimensional covariance matrix estimation with missing data; see also Little and Rubin (2019, Section 3.4). In practice, p is typically unknown and needs to be estimated. It is therefore natural to consider the following plug-in estimator $\widehat{\mathbf{G}}$:

$$\widehat{\mathbf{G}} = \left(\frac{1}{n}\mathbf{Y}_{\mathbf{\Omega}}^\top\mathbf{Y}_{\mathbf{\Omega}}\right) \circ \widehat{\mathbf{W}}, \tag{4}$$

where $\widehat{\mathbf{W}} = \widehat{p}^{-2}\{\mathbf{1}_d\mathbf{1}_d^\top - (1-\widehat{p})\mathbf{I}_d\}$ and $\widehat{p} := (nd)^{-1}\|\mathbf{\Omega}\|_1$ denotes the proportion of observed entries in $\mathbf{Y}$. The observed-proportion weighted estimator of $\mathbf{V}_K$, denoted $\widehat{\mathbf{V}}_K^{\mathrm{OPW}}$, is the $d \times K$ matrix formed from the top K eigenvectors of $\widehat{\mathbf{G}}$.

2.1. Theory for homogeneous missingness

We begin by studying the theoretical performance of $\widehat{\mathbf{V}}_K^{\mathrm{OPW}}$ in a simple model that will allow us to reveal an interesting phase transition for the problem. For a random vector $\mathbf{x}$ taking values in $\mathbb{R}^d$ and for $r \geq 1$, we define its (Orlicz) ψ_r-norm and a version that is invariant to invertible affine transformations by

$$\|\mathbf{x}\|_{\psi_r} := \sup_{\mathbf{u}\in\mathcal{S}^{d-1}} \sup_{q\in\mathbb{N}} \frac{\left(\mathbb{E}|\mathbf{u}^\top\mathbf{x}|^q\right)^{1/q}}{q^{1/r}} \quad \text{and} \quad \|\mathbf{x}\|_{\psi_r^*} := \sup_{\mathbf{u}\in\mathcal{S}^{d-1}} \frac{\|\mathbf{u}^\top(\mathbf{x}-\mathbb{E}\mathbf{x})\|_{\psi_r}}{\mathrm{Var}^{1/2}(\mathbf{u}^\top\mathbf{x})}$$

respectively. Recall that we say $\mathbf{x}$ is *sub-Gaussian* if $\|\mathbf{x}\|_{\psi_2^*} < \infty$.

In this preliminary section, we assume that $(\mathbf{Y}_{\mathbf{\Omega}}, \mathbf{\Omega})$ is generated according to (1), (2) and (3), where:

(A1) $\mathbf{U}$, $\mathbf{Z}$ and $\mathbf{\Omega}$ are independent;
(A2) $\mathbf{U}$ has independent and identically distributed rows $(\mathbf{u}_i : i \in [n])$ with $\mathbb{E}\mathbf{u}_1 = 0$ and $\|\mathbf{u}_1\|_{\psi_2^*} \leq \tau$;
(A3) $\mathbf{Z} = (z_{ij})_{i\in[n],j\in[d]}$ has independent and identically distributed entries with $\mathbb{E}z_{11} = 0$, $\mathrm{Var}\, z_{11} = 1$ and $\|z_{11}\|_{\psi_2^*} \leq \tau$;
(A4) $\|y_{1j}^2\|_{\psi_1} \leq M$ for all $j \in [d]$;
(A5) $\mathbf{\Omega}$ has independent Bern(p) entries.

Thus, (A1) ensures that the complete data matrix $\mathbf{Y}$ and the revelation matrix $\mathbf{\Omega}$ are independent; in other words, for now we work in a Missing Completely At Random (MCAR) setting. In a homoscedastic noise model, there is no loss of generality (by a scaling argument) in assuming that each entry of $\mathbf{Z}$ has unit variance, as in (A3). In many places in this work, it will be convenient to think intuitively of τ and M in (A2)–(A4) as constants. In particular, if $\mathbf{U}$ has multivariate normal rows and $\mathbf{Z}$ has normal entries, then we can simply take $\tau = 1$. For M, under the same normality assumptions, we have $\|y_{1j}^2\|_{\psi_1} = \mathrm{Var}(y_{1j})$, so this intuition amounts to thinking of the variance of each component of our data as being of constant order.

A natural measure of the performance of an estimator $\widehat{\mathbf{V}}_K$ of $\mathbf{V}_K$ is given by the Davis–Kahan $\sin\Theta$ loss

$$L(\widehat{\mathbf{V}}_K, \mathbf{V}_K) := \frac{1}{\sqrt{2}}\|\widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top - \mathbf{V}_K\mathbf{V}_K^\top\|_{\mathrm{F}}$$

(Davis and Kahan, 1970)†. Our first theorem controls the risk of the OPW estimator; here and below, we write λ_k for the kth largest eigenvalue of $\mathbf{\Sigma_u}$.

THEOREM 1. *Assume (A1)–(A5) and that $n, d \geq 2$, $dp \geq 1$. Write $R := \lambda_1 + 1$. Then there exists a universal constant $C > 0$ such that*

$$\mathbb{E}L(\widehat{\mathbf{V}}_K^{\mathrm{OPW}}, \mathbf{V}_K) \leq \frac{CK^{1/2}}{\lambda_K p}\left\{\left(\frac{Md(R\tau^2 p + M\log d)\log^2 d}{n}\right)^{1/2} + \frac{Md\log^2 d\log n}{n}\right\}. \tag{5}$$

In particular, if $n \geq d\log^2 d\log^2 n/(\lambda_1 p + \log d)$, then there exists $C_{M,\tau} > 0$, depending only on M and τ, such that

$$\mathbb{E}L(\widehat{\mathbf{V}}_K^{\mathrm{OPW}}, \mathbf{V}_K) \leq \frac{C_{M,\tau}}{\lambda_K p}\left(\frac{Kd(\lambda_1 p + \log d)\log^2 d}{n}\right)^{1/2}. \tag{6}$$

Theorem 1 reveals an interesting phase transition phenomenon. Specifically, if the signal strength is large enough that $\lambda_1 \geq p^{-1}\log d$, then we should regard np as the effective sample size, as might intuitively be expected. On the other hand, if $\lambda_1 < p^{-1}\log d$, then the estimation problem is considerably more difficult and the effective sample size is of order np^2. In fact, by inspecting the proof of Theorem 1, we see that in the high signal case, it is the difficulty of estimating the diagonal entries of $\mathbf{\Sigma_y}$ that drives the rate, while when the signal strength is low, the bottleneck is the challenge of estimating the off-diagonal entries. By comparing (6) with the minimax lower bound result in Theorem 2 below, we see that this phase transition phenomenon is an inherent feature of this estimation problem, rather than an artefact of the proof techniques we used to derived the upper bound.

The condition $n \geq d\log^2 d\log^2 n/(\lambda_1 p + \log d)$ in Theorem 1 is reasonable given the scaling requirement for consistency of the empirical eigenvectors (Shen et al., 2016; Wang and Fan, 2017; Johnstone and Lu, 2009). Indeed, Shen et al. (2016, Theorem 5.1) show that when $\lambda_1 \gg 1$, the top eigenvector of the sample covariance matrix estimator is consistent if and only if $d/(n\lambda_1) \to 0$. If we regard np as the effective sample size in our missing data PCA problem, then it is a sensible analogy to assume that $d/(np\lambda_1) \to 0$ here, which implies that the condition $n \geq d\log^2 d\log^2 n/(\lambda_1 p + \log d)$ holds for large n, up to poly-logarithmic factors.

As mentioned in the introduction, Cho, Kim and Rohe (2017) considered the different but related problem of singular space estimation in a model in which $\mathbf{Y} = \mathbf{\Theta} + \mathbf{Z}$, where $\mathbf{\Theta}$ is a matrix of the form $\mathbf{U}\mathbf{V}_K^\top$ for a *deterministic* matrix $\mathbf{U}$, whose rows are not necessarily centred. In this setting, $\mathbf{V}_K$ is the matrix of top K right singular vectors of $\mathbf{\Theta}$, and the same estimator $\widehat{\mathbf{V}}_K$ can be applied. An important distinction is that, when the rows of $\mathbf{U}$ are not centred and the entries of $\mathbf{\Theta}$ are of comparable magnitude, $\|\mathbf{\Theta}\|_{\mathrm{F}}$ is of order $\sqrt{nd}$, so when K is regarded as a constant, it is natural to think of the singular values of $\mathbf{\Theta}$ as also being of order $\sqrt{nd}$. Indeed, this is assumed in Cho, Kim and Rohe (2017). On the other hand, in our model, where the rows of $\mathbf{U}$ have mean zero, assuming that the eigenvalues are of order $\sqrt{nd}$ would amount to an extremely strong requirement, essentially restricting attention to very highly spiked covariance matrices. Removing this condition in Theorem 1 requires completely different arguments.

In order to state our minimax lower bound, we let $\mathcal{P}_{n,d}(\lambda_1, p)$ denote the class of distributions of pairs $(\mathbf{Y_\Omega}, \mathbf{\Omega})$ satisfying (A1), (A2), (A3) and (A5) with $K = 1$. Since we are now working with vectors instead of matrices, we write $\mathbf{v}$ in place of $\mathbf{V}_1$.

THEOREM 2. *There exists a universal constant $c > 0$ such that*

$$\inf_{\widehat{\mathbf{v}}}\sup_{P\in\mathcal{P}_{n,d}(\lambda_1,p)}\mathbb{E}_P L(\widehat{\mathbf{v}}, \mathbf{v}) \geq c\min\left\{\frac{1}{\lambda_1 p}\left(\frac{d(\lambda_1 p + 1)}{n}\right)^{1/2}, 1\right\},$$

†When $K = 1$, we have that $L(\widehat{\mathbf{V}}_1, \mathbf{V}_1)$ is the sine of the acute angle between $\widehat{\mathbf{V}}_1$ and $\mathbf{V}_1$. More generally, $L^2(\widehat{\mathbf{V}}_K, \mathbf{V}_K)$ is the sum of the squares of the sines of the principal angles between the subspaces spanned by $\widehat{\mathbf{V}}_K$ and $\mathbf{V}_K$.

where the infimum is taken over all estimators $\widehat{\mathbf{v}} = \widehat{\mathbf{v}}(\mathbf{Y}_{\boldsymbol{\Omega}}, \boldsymbol{\Omega})$ *of* $\mathbf{v}$.

Theorem 2 reveals that $\widehat{\mathbf{V}}_1^{\mathrm{OPW}}$ in Theorem 1 achieves the minimax optimal rate of estimation up to a poly-logarithmic factor when M and τ are regarded as constants.

2.2. Heterogeneous observation mechanism

A key assumption of the theory in Section 2.1, which allowed even a very simple estimator to perform well, was that the missingness probability was homogeneous across the different entries of the matrix. On the other hand, the aim of this subsection is to show that the situation changes dramatically once the data can be missing heterogeneously.

To this end, consider the following example. Suppose that $\boldsymbol{\omega}$ is equal to $(1, 0, 1, \ldots, 1)^\top$ or $(0, 1, 1, \ldots, 1)^\top$ with equal probability, so that

$$\mathbf{P} = \mathbb{E}\boldsymbol{\omega}\boldsymbol{\omega}^\top = \begin{pmatrix} 1/2 & 0 & 1/2 & \ldots & 1/2 \\ 0 & 1/2 & 1/2 & \ldots & 1/2 \\ 1/2 & 1/2 & 1 & \ldots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1/2 & 1/2 & 1 & \ldots & 1 \end{pmatrix} \in \mathbb{R}^{d \times d}.$$

In other words, for each $i \in [n]$, we observe precisely one of the first two entries of $\mathbf{y}_i$, together with all of the remaining $(d-2)$ entries. Let $\boldsymbol{\Sigma} = \mathbf{I}_d + \boldsymbol{\alpha}\boldsymbol{\alpha}^\top$, where $\boldsymbol{\alpha} = (2^{-1/2}, 2^{-1/2}, 0, \ldots, 0)^\top \in \mathbb{R}^d$, and $\boldsymbol{\Sigma}' = \mathbf{I}_d + \boldsymbol{\alpha}'(\boldsymbol{\alpha}')^\top$, where $\boldsymbol{\alpha}' = (2^{-1/2}, -2^{-1/2}, 0, \ldots, 0)^\top \in \mathbb{R}^d$. Suppose that $\mathbf{y} \sim N_d(\mathbf{0}, \boldsymbol{\Sigma})$ and let $\widetilde{\mathbf{y}} := \mathbf{y} \circ \boldsymbol{\omega}$, and similarly assume that $\mathbf{y}' \sim N_d(\mathbf{0}, \boldsymbol{\Sigma}')$ and set $\widetilde{\mathbf{y}}' := \mathbf{y}' \circ \boldsymbol{\omega}$. Then $(\widetilde{\mathbf{y}}, \boldsymbol{\omega})$ and $(\widetilde{\mathbf{y}}', \boldsymbol{\omega})$ are identically distributed. However, the leading eigenvectors of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$ are respectively $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$, which are orthogonal!

Thus, it is impossible to simultaneously estimate consistently the leading eigenvector of both $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$ from our observations. We note that it is the disproportionate weight of the first two coordinates in the leading eigenvector, combined with the failure to observe simultaneously the first two entries in the data, that makes the estimation problem intractable in this example. The understanding derived from this example motivates us to seek bounds on the error in high-dimensional PCA that depend on an incoherence parameter $\mu := (d/K)^{1/2}\|\mathbf{V}_K\|_{2\to\infty} \in \big[1, (d/K)^{1/2}\big]$. The intuition here is that the maximally incoherent case is where each column of $\mathbf{V}_K$ is a unit vector proportional to a vector whose entries are either 1 or -1, in which case $\|\mathbf{V}_K\|_{2\to\infty} = (K/d)^{1/2}$ and $\mu = 1$. On the other hand, in the worst case, when the columns of $\mathbf{V}_K$ are the first K standard basis vectors in $\mathbb{R}^d$, we have $\mu = (d/K)^{1/2}$. Bounds involving incoherence have appeared previously in the literature on matrix completion (e.g., Candès and Plan, 2010; Keshavan, Montanari and Oh, 2010), but for a different reason. There, the purpose is to control the principal angles between the true right singular space and the standard basis, which yields bounds on the number of observations required to infer the missing entries of the matrix. In our case, the incoherence condition controls the extent to which the loadings of the principal components of interest are concentrated in any single coordinate, and therefore the extent to which significant estimation error in a few components of the leading eigenvectors can affect the overall statistical performance. In the intractable example above, $\mu = (d/2)^{1/2}$, and with such a large value of μ, heavy corruption from missingness in only a few entries spoils any chance of consistent estimation.

3. Our new algorithm for PCA with missing entries

We are now in a position to introduce and analyse our iterative algorithm `primePCA` to estimate $\mathrm{Col}(\mathbf{V}_K)$, the principal eigenspace of the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{y}}$. The basic idea is to iterate between imputing the missing entries of the data matrix $\mathbf{Y}_{\boldsymbol{\Omega}}$ using a current (input) iterate $\widehat{\mathbf{V}}_K^{(\mathrm{in})}$, and then applying a singular value decomposition (SVD) to the completed data matrix. More precisely, for $i \in [n]$, we let $\mathcal{J}_i$ denote the indices for which the corresponding entry of $\mathbf{y}_i$ is observed, and regress the observed data $\widetilde{\mathbf{y}}_{i,\mathcal{J}_i} = \mathbf{y}_{i,\mathcal{J}_i}$ on $(\widehat{\mathbf{V}}_K^{(\mathrm{in})})_{\mathcal{J}_i}$ to obtain an estimate $\widehat{\mathbf{u}}_i$ of the ith row of $\mathbf{U}$. This is natural

in view of the data generating mechanism $\mathbf{y}_i = \mathbf{V}_K\mathbf{u}_i + \mathbf{z}_i$. We then use $\widehat{\mathbf{y}}_{i,\mathcal{J}_i^c} := (\widehat{\mathbf{V}}_K^{(\mathrm{in})})_{\mathcal{J}_i^c}\widehat{\mathbf{u}}_i$ to impute the missing values $\mathbf{y}_{i,\mathcal{J}_i^c}$, retain the original observed entries as $\widehat{\mathbf{y}}_{i,\mathcal{J}_i} := \widetilde{\mathbf{y}}_{i,\mathcal{J}_i}$, and set our next (output) iterate $\widehat{\mathbf{V}}_K^{(\mathrm{out})}$ to be the top K right singular vectors of the imputed matrix $\widehat{\mathbf{Y}} := (\widehat{\mathbf{y}}_1, \ldots, \widehat{\mathbf{y}}_n)^\top$. To motivate this final choice, observe that when $\mathbf{Z} = \mathbf{0}$, we have $\mathrm{rank}(\mathbf{Y}) = K$; we therefore have the SVD $\mathbf{Y} = \mathbf{L}\mathbf{\Gamma}\mathbf{R}^\top$, where $\mathbf{L} \in \mathbb{O}^{n\times K}, \mathbf{R} \in \mathbb{O}^{d\times K}$ and $\mathbf{\Gamma} \in \mathbb{R}^{K\times K}$ is diagonal with positive diagonal entries. This means that $\mathbf{R} = \mathbf{V}_K\mathbf{U}^\top\mathbf{L}\mathbf{\Gamma}^{-1}$, so the column spaces of $\mathbf{R}$ and $\mathbf{V}_K$ coincide. For convenience, pseudocode of a single iteration of refinement in this algorithm is given in Algorithm 1.

Algorithm 1 `refine`$(K, \widehat{\mathbf{V}}_K^{(\mathrm{in})}, \mathbf{\Omega}, \mathbf{Y}_{\mathbf{\Omega}})$, a single step of refinement of current iterate $\widehat{\mathbf{V}}_K^{(\mathrm{in})}$

Input: $K \in [d]$, $\widehat{\mathbf{V}}_K^{(\mathrm{in})} \in \mathbb{O}^{d\times K}$, $\mathbf{\Omega} \in \{0,1\}^{n\times d}$ with $\min_i \|\boldsymbol{\omega}_i\|_1 \geq 1$, $\mathbf{Y}_{\mathbf{\Omega}} \in \mathbb{R}^{n\times d}$
Output: $\widehat{\mathbf{V}}_K^{(\mathrm{out})} \in \mathbb{O}^{d\times K}$

1: **for** i in $[n]$ **do**
2: $\quad\mathcal{J}_i \leftarrow \{j \in [d] : \omega_{ij} = 1\}$
3: $\quad\widehat{\mathbf{u}}_i \leftarrow (\widehat{\mathbf{V}}_K^{(\mathrm{in})})^\dagger_{\mathcal{J}_i}\widetilde{\mathbf{y}}_{i,\mathcal{J}_i}$, where $(\widehat{\mathbf{V}}_K^{(\mathrm{in})})^\dagger_{\mathcal{J}_i}$ denotes the Moore–Penrose pseudoinverse of $(\widehat{\mathbf{V}}_K^{(\mathrm{in})})_{\mathcal{J}_i}$.
4: $\quad\widehat{\mathbf{y}}_{i,\mathcal{J}_i^c} \leftarrow (\widehat{\mathbf{V}}_K^{(\mathrm{in})})_{\mathcal{J}_i^c}\widehat{\mathbf{u}}_i$
5: $\quad\widehat{\mathbf{y}}_{i,\mathcal{J}_i} \leftarrow \widetilde{\mathbf{y}}_{i,\mathcal{J}_i}$
6: **end for**
7: $\widehat{\mathbf{Y}} \leftarrow (\widehat{\mathbf{y}}_1, \ldots, \widehat{\mathbf{y}}_n)^\top$
8: $\widehat{\mathbf{V}}_K^{(\mathrm{out})} \leftarrow$ top K right singular vectors of $\widehat{\mathbf{Y}}$

We now seek to provide formal justification for Algorithm 1. The recursive nature of the `primePCA` algorithm induces complex relationships between successive iterates, so to facilitate theoretical analysis, we will impose some conditions on the underlying data generating mechanism that may not hold in situations where we would like to apply to algorithm. Nevertheless, we believe that the analysis provides considerable insight into the performance of the `primePCA` algorithm, and these are discussed extensively below; moreover, our simulations in Section 4 consider settings both within and outside the scope of our theory, and confirm its attractive and robust numerical performance.

In addition to the loss function L, it will be convenient to define a slightly different notion of distance between subspaces. For any $\mathbf{V}, \widetilde{\mathbf{V}} \in \mathbb{O}^{d\times K}$, we let $\mathbf{W}_1\mathbf{D}\mathbf{W}_2^\top$ be an SVD of $\widetilde{\mathbf{V}}^\top\mathbf{V}$. The *two-to-infinity distance* between $\widetilde{\mathbf{V}}$ and $\mathbf{V}$ is then defined to be

$$\mathcal{T}(\widetilde{\mathbf{V}}, \mathbf{V}) := \|\widetilde{\mathbf{V}} - \mathbf{V}\mathbf{W}_2\mathbf{W}_1^\top\|_{2\to\infty}.$$

We remark that the definition of $\mathcal{T}(\widetilde{\mathbf{V}}, \mathbf{V})$ does not depend on our choice of SVD and that $\mathcal{T}(\widetilde{\mathbf{V}}, \mathbf{V}) = \mathcal{T}(\widetilde{\mathbf{V}}\mathbf{O}_1, \mathbf{V}\mathbf{O}_2)$ for any $\mathbf{O}_1, \mathbf{O}_2 \in \mathbb{O}^{K\times K}$, so that $\mathcal{T}$ really represents a distance between the subspaces spanned by $\widetilde{\mathbf{V}}$ and $\mathbf{V}$. In fact, there is a sense in which the change-of-basis matrix $\mathbf{W}_2\mathbf{W}_1^\top$ tries to align the columns of $\mathbf{V}$ as closely as possible with those of $\widetilde{\mathbf{V}}$; more precisely, if we change the norm from the two-to-infinity operator norm to the Frobenius norm, then $\mathbf{W}_2\mathbf{W}_1^\top$ uniquely solves the so-called *Procrustes problem* (Schönemann, 1966):

$$\mathbf{W}_2\mathbf{W}_1^\top = \operatorname*{argmin}_{\mathbf{W}\in\mathbb{O}^{K\times K}} \|\widetilde{\mathbf{V}} - \mathbf{V}\mathbf{W}\|_{\mathrm{F}}. \tag{7}$$

The following proposition considers the noiseless setting $\mathbf{Z} = \mathbf{0}$, and shows that, for any estimator $\widehat{\mathbf{V}}_K^{(\mathrm{in})}$ that is close to $\mathbf{V}_K$, a single iteration of refinement in Algorithm 1 contracts the two-to-infinity distance between their column spaces, under appropriate conditions. We define $\mathbf{\Omega}^c := \mathbf{1}_d\mathbf{1}_d^\top - \mathbf{\Omega}$.

Proposition 1. *Let* $\widehat{\mathbf{V}}_K^{(\mathrm{out})} := \mathtt{refine}\big(K, \widehat{\mathbf{V}}_K^{(\mathrm{in})}, \mathbf{\Omega}, \mathbf{Y}_{\mathbf{\Omega}}\big)$ *as in Algorithm 1 and further let* $\Delta := \mathcal{T}(\widehat{\mathbf{V}}_K^{(\mathrm{in})}, \mathbf{V}_K)$. *We assume that* $\min_{i\in[n]}\|\boldsymbol{\omega}_i\|_1 > K$ *and that* $\min_{i\in[n]} \frac{d^{1/2}\sigma_K((\widehat{\mathbf{V}}_K^{(\mathrm{in})})_{\mathcal{J}_i})}{|\mathcal{J}_i|^{1/2}} \geq 1/\sigma_* > 0$. *Suppose that* $\mathbf{Z} = \mathbf{0}$ *and that the SVD of* $\mathbf{Y}$ *is of the form* $\mathbf{L}\mathbf{\Gamma}\mathbf{R}^\top$, *where* $\|\mathbf{L}\|_{2\to\infty} \leq \mu(K/n)^{1/2}$ *and*

$\|\mathbf{R}\|_{2\to\infty} \le \mu(K/d)^{1/2}$ *for some* $\mu \ge 1$. *Then there exist* $c_1, C > 0$, *depending only on* σ_*, *such that whenever*

(i) $\Delta \le \frac{c_1 \sigma_K(\mathbf{\Gamma})}{K^2\mu^4\sigma_1(\mathbf{\Gamma})\sqrt{d}}$,
(ii) $\rho := \frac{CK^2\mu^4\sigma_1(\mathbf{\Gamma})\|\mathbf{\Omega}^c\|_{1\to1}}{\sigma_K(\mathbf{\Gamma})n} < 1$,

we have that

$$\mathcal{T}(\widehat{\mathbf{V}}_K^{(\mathrm{out})}, \mathbf{V}_K) \le \rho\Delta.$$

In order to understand the main conditions of Proposition 1, it is instructive to consider the case $K = 1$, as was illustrated in Figure 1, and initially to think of μ as a constant. In that case, condition (i) asks that the absolute value of every component of the difference between the vectors $\widehat{\mathbf{V}}_1^{(\mathrm{in})}$ and $\mathbf{V}_1$ is $O(d^{-1/2})$; for intuition, if two vectors are uniformly distributed on $\mathcal{S}^{d-1}$, then each of their ℓ_∞ norms is $O_p(d^{-1/2}\log^{1/2} d)$; in other words, we only ask that the initialiser is very slightly better than a random guess. In Condition (ii), ρ being less than 1 is equivalent to the proportion of missing data in each column being less than $1/(C'\mu^4)$ (where C' again depends only on σ_*), and the conclusion is that the `refine` step contracts the initial two-to-infinity distance from $\mathbf{V}_K$ by at least a factor of ρ. In the noiseless setting of Proposition 1, the matrix $\mathbf{R}$ of right singular vectors of $\mathbf{Y}$ has the same column span (and hence the same two-to-infinity norm) as $\mathbf{V}_K$. We can therefore gain some intuition about the scale of μ by considering the situation where $\mathbf{V}_K$ is uniformly distributed on $\mathbb{O}^{d\times K}$, so in particular, the columns of $\mathbf{V}_K$ are uniformly distributed on $\mathcal{S}^{d-1}$. By Vershynin (2018, Theorem 5.1.4), we deduce that $\|\mathbf{V}_K\|_{2\to\infty} = O_p\big(\sqrt{\frac{K\log d}{d}}\big)$. On the other hand, when the distribution of $\mathbf{U}$ is invariant under left multiplication by an orthogonal matrix (e.g. if $\mathbf{U}$ has independent and identically distributed Gaussian rows), then $\mathbf{L}$ is distributed uniformly on $\mathbb{O}^{n\times K}$. Arguing as above, we see that, with high probability, we may take $\mu \lesssim \max\big(\sqrt{\log n}, \sqrt{\log d}\big)$. This calculation suggests that we do not lose too much by thinking of μ as a constant (or at most, growing very slowly with n and d).

To apply Proposition 1, we also require conditions on $\min_{i\in[n]}\|\boldsymbol{\omega}_i\|_1$ and σ_*. In practice, if either of these conditions is not satisfied, we first perform a screening step that restricts attention to a set of row indices for which the data contain sufficient information to estimate the K principal components. This screening step is explicitly accounted for in Algorithm 2 below, as well as in the theory that justifies it. An alternative would be to seek to weight rows according to their utility for principal component estimation, but it seems difficult to implement this in a principled way that facilitates formal justification.

Algorithm 2 `primePCA`, an iterative algorithm for estimating $\mathbf{V}_K$ given initialiser $\widehat{\mathbf{V}}_K^{(0)}$

Input: $K \in [d], \widehat{\mathbf{V}}_K^{(0)} \in \mathbb{O}^{d\times K}, \mathbf{\Omega} \in \{0,1\}^{n\times d}, \mathbf{Y}_{\mathbf{\Omega}} \in \mathbb{R}^{n\times d}$, $n_{\mathrm{iter}} \in \mathbb{N}$, $\sigma_* \in (0,\infty)$, $\kappa^* \in [0,\infty)$
Output: $\widehat{\mathbf{V}}_K \in \mathbb{R}^{d\times K}$

for i in $[n]$ **do**
  $\mathcal{J}_i \leftarrow \{j \in [d] : \omega_{ij} = 1\}$
end for
for t in $[n_{\mathrm{iter}}]$ **do**
  $\mathcal{I}^{(t-1)} \leftarrow \big\{i : \|\boldsymbol{\omega}_i\|_1 > K, \sigma_K\big((\widehat{\mathbf{V}}_K^{(t-1)})_{\mathcal{J}_i}\big) \ge \frac{|\mathcal{J}_i|^{1/2}}{d^{1/2}\sigma_*}\big\}$
  $\widehat{\mathbf{V}}_K^{(t)} \leftarrow \mathtt{refine}(K, \widehat{\mathbf{V}}_K^{(t-1)}, \mathbf{\Omega}_{\mathcal{I}^{(t-1)}}, (\mathbf{Y}_{\mathbf{\Omega}})_{\mathcal{I}^{(t-1)}})$, where `refine` is defined in Algorithm 1.
  if $L(\widehat{\mathbf{V}}_K^{(t)}, \widehat{\mathbf{V}}_K^{(t-1)}) < \kappa^*$ **then break**
  end if
end for
return $\widehat{\mathbf{V}}_K = \widehat{\mathbf{V}}_K^{(t)}$

Algorithm 2 provides pseudocode for the iterative `primePCA` algorithm, given an initial estimator $\widehat{\mathbf{V}}_K^{(0)}$. The iterations continue until either we hit the convergence threshold κ^* or the maximum iteration number n_{iter}. Theorem 3 below guarantees that, in the noiseless setting of Proposition 1, the `primePCA` estimator converges to $\mathbf{V}_K$ at a geometric rate.

THEOREM 3. *For $t \in [n_{\mathrm{iter}}]$, let $\widehat{\mathbf{V}}_K^{(t)}$ be the t^{th} iterate of Algorithm 2 with input K, $\widehat{\mathbf{V}}_K^{(0)}$, $\boldsymbol{\Omega} \in \{0,1\}^{n\times d}$, $\mathbf{Y}_{\boldsymbol{\Omega}} \in \mathbb{R}^{n\times d}$, $n_{\mathrm{iter}} \in \mathbb{N}$, $\sigma_* \in (0,\infty)$ and $\kappa^* = 0$. Write $\Delta := \mathcal{T}(\widehat{\mathbf{V}}_K^{(0)}, \mathbf{V}_K)$ and let*

$$\mathcal{I} := \Big\{ i : \|\boldsymbol{\omega}_i\|_1 > K, \sigma_K\big((\mathbf{V}_K)_{\mathcal{J}_i}\big) \geq \frac{|\mathcal{J}_i|^{1/2}}{d^{1/2}\sigma_*} \Big\},$$

where $\mathcal{J}_i := \{j : \omega_{ij} = 1\}$. Suppose that $\mathbf{Z} = \mathbf{0}$ and that the SVD of $\mathbf{Y}_{\mathcal{I}}$ is of the form $\mathbf{L}\boldsymbol{\Gamma}\mathbf{R}^\top$, where $\|\mathbf{L}\|_{2\to\infty} \leq \mu(K/n)^{1/2}$ and $\|\mathbf{R}\|_{2\to\infty} \leq \mu(K/d)^{1/2}$ for some $\mu \geq 1$. Assume that

$$\epsilon := \min\Big\{ \Big| \frac{\sigma_K\big((\mathbf{V}_K)_{\mathcal{J}_i}\big) d^{1/2}}{|\mathcal{J}_i|^{1/2}} - \frac{1}{\sigma_*} \Big| : i \in [n], \|\boldsymbol{\omega}_i\|_1 > K \Big\} > 0.$$

Then there exist $c_1, C > 0$, depending only σ_* and ϵ, such that whenever

(i) $\Delta \leq \frac{c_1 \sigma_K(\boldsymbol{\Gamma})}{K^2\mu^4\sigma_1(\boldsymbol{\Gamma})\sqrt{d}}$,

(ii) $\rho := \frac{CK^2\mu^4\sigma_1(\boldsymbol{\Gamma})\|\boldsymbol{\Omega}^{\mathrm{c}}_{\mathcal{I}}\|_{1\to 1}}{\sigma_K(\boldsymbol{\Gamma})|\mathcal{I}|} < 1$,

we have that for every $t \in [n_{\mathrm{iter}}]$,

$$\mathcal{T}(\widehat{\mathbf{V}}_K^{(t)}, \mathbf{V}_K) \leq \rho^t \Delta.$$

The condition that $\epsilon > 0$ amounts to the very mild assumption that the algorithmic input σ_* is not exactly equal to any element of the set $\big\{ \frac{|\mathcal{J}_i|^{1/2}}{\sigma_K((\mathbf{V}_K)_{\mathcal{J}_i})d^{1/2}} : i \in [n], \|\boldsymbol{\omega}_i\|_1 > K \big\}$, though the conditions on c_1 and C become milder as ϵ increases.

3.1. Initialisation

Theorem 3 provides a general guarantee on the performance of `primePCA`, but relies on finding an initial estimator $\widehat{\mathbf{V}}_K^{(0)}$ that is sufficiently close to the truth $\mathbf{V}_K$. The aim of this subsection, then, is to propose a simple initialiser and show that it satisfies the requirement of Theorem 3 with high probability, conditional on the missingness pattern.

Consider the following modified weighted sample covariance matrix

$$\widetilde{\mathbf{G}} := \frac{1}{n}\sum_{i=1}^n \widetilde{\mathbf{y}}_i\widetilde{\mathbf{y}}_i^\top \circ \widetilde{\mathbf{W}}, \tag{8}$$

where for any $j,k \in [d]$,

$$\widetilde{\mathbf{W}}_{jk} := \begin{cases} \frac{n}{\sum_{i=1}^n \omega_{ij}\omega_{ik}} & \text{if } \sum_{i=1}^n \omega_{ij}\omega_{ik} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Here, the matrix $\widetilde{\mathbf{W}}$ replaces $\widehat{\mathbf{W}}$ in (4) because, unlike in Section 2.1, we no longer wish to assume homogeneous missingness. We take as our initial estimator of $\mathbf{V}_K$ the matrix of top K eigenvectors of $\widetilde{\mathbf{G}}$, denoted $\widetilde{\mathbf{V}}_K$. Theorem 4 below studies the performance of this initialiser, in terms of its two-to-infinity norm error, and provides sufficient conditions for us to be able to apply Theorem 3. In particular, it ensures that the initialiser is reasonably well-aligned with the target $\mathbf{V}_K$. We write $\mathbb{P}^{\boldsymbol{\Omega}}$ and $\mathbb{E}^{\boldsymbol{\Omega}}$ for probabilities and expectations conditional on $\boldsymbol{\Omega}$.

THEOREM 4. *Assume (A1)–(A4) and that $n, d \geq 2$. Suppose further that $\|\mathbf{V}_K\|_{2\to\infty} \leq \mu(K/d)^{1/2}$, that $\sum_{i=1}^n \omega_{ij}\omega_{ik} > 0$ for all j,k and let $R := \lambda_1 + 1$. Then there exist $c_{M,\tau}, C_{M,\tau} > 0$, depending only on M and τ, such that for every $\xi > 2$, if*

$$\lambda_K > c_{M,\tau}\Big\{ \Big(\frac{\max\big(\|\widetilde{\mathbf{W}}\|_1, R\|\widetilde{\mathbf{W}}\|_{1\to 1}\big)\xi\log d}{n} \Big)^{1/2} + \frac{\xi\|\widetilde{\mathbf{W}}\|_{\mathrm{F}}\log^2 d}{n} \Big\}, \tag{10}$$

then

$$\mathbb{P}^{\mathbf{\Omega}}\biggl\{\mathcal{T}(\widetilde{\mathbf{V}}_K,\mathbf{V}_K)\geq\frac{C_{M,\tau}K\mu^2R^{1/2}}{\lambda_K}\biggl(\frac{K}{d^{1/2}}+\frac{1}{\lambda_K}\biggr)\biggl(\frac{\xi^{1/2}\|\widetilde{\mathbf{W}}\|_{\infty\to\infty}^{1/2}\log^{1/2}d}{n^{1/2}}+\frac{\xi\|\widetilde{\mathbf{W}}\|_{2\to\infty}\log d}{n}\biggr)\biggr\}\\ \leq 2(e^{K\log 5}+K+4)d^{-(\xi-1)}+2d^{-(\xi-2)}.$$

As a consequence, writing

$$\mathcal{A}:=\biggl\{\frac{\sigma_K(\mathbf{Y}_\mathcal{I})}{\sigma_1(\mathbf{Y}_\mathcal{I})}>\frac{C_{M,\tau}K^3\mu^6R^{1/2}}{c_1\lambda_K}\biggl(1+\frac{d^{1/2}}{K\lambda_K}\biggr)\biggl(\frac{\xi^{1/2}\|\widetilde{\mathbf{W}}\|_{\infty\to\infty}^{1/2}\log^{1/2}d}{n^{1/2}}+\frac{\xi\|\widetilde{\mathbf{W}}\|_{2\to\infty}\log d}{n}\biggr)\biggr\},$$

where $\mathcal{I}$ and c_1 are as in Theorem 3, we have that

$$\mathbb{P}^{\mathbf{\Omega}}\biggl(\mathcal{T}(\widetilde{\mathbf{V}}_K,\mathbf{V}_K)>\frac{c_1\sigma_K(\mathbf{Y}_\mathcal{I})}{K^2\sigma_1(\mathbf{Y}_\mathcal{I})d^{1/2}}\biggr)\leq 2(e^{K\log 5}+K+4)d^{-(\xi-1)}+2d^{-(\xi-2)}+\mathbb{P}^{\mathbf{\Omega}}(\mathcal{A}^{\mathrm{c}}).$$

The first part of Theorem 4 provides a general probabilistic upper bound for $\mathcal{T}(\widetilde{\mathbf{V}}_K,\mathbf{V}_K)$, after conditioning on the missingness pattern. This allows us, in the second part, to provide a guarantee on the probability with which $\widetilde{\mathbf{V}}_K$ is a good enough initialiser for Theorem 3 to apply. For intuition regarding $\mathbb{P}^{\mathbf{\Omega}}(\mathcal{A}^{\mathrm{c}})$, consider the MCAR setting with $p_{jk}:=\mathbb{E}(\omega_{1j}\omega_{1k})$ for $j,k\in[d]$. In that case, by Lemma 6, typical realisations of $\widetilde{\mathbf{W}}$ have $\|\widetilde{\mathbf{W}}\|_{\infty\to\infty}\leq 2\max_{j\in[d]}\sum_{k\in[d]}p_{jk}^{-1}$ and $\|\widetilde{\mathbf{W}}\|_{2\to\infty}\leq 2\max_{j\in[d]}\bigl(\sum_{k\in[d]}p_{jk}^{-2}\bigr)^{1/2}$ when $\sum_{j,k\in[d]}e^{-np_{jk}/8}$ is small. In particular, when $n\min_{j,k\in[d]}p_{jk}\geq\log d$, we expect $\mathbb{P}^{\mathbf{\Omega}}(\mathcal{A}^{\mathrm{c}})$ to be small when λ_1 and λ_K are both of the same order, and grow faster than

$$\max\biggl\{\biggl(\frac{d\log d}{n}\max_{j\in[d]}\sum_{k=1}^d\frac{1}{p_{jk}}\biggr)^{1/3},\frac{\log d}{n}\max_{j\in[d]}\sum_{k=1}^d\frac{1}{p_{jk}}\biggr\}.$$

As a special case, in the p-homogeneous model where $p_{jk}=p^2\mathbb{1}_{\{j\neq k\}}+p\mathbb{1}_{\{j=k\}}$ for $j,k\in[d]$, the requirement on λ_K above is that it should grow faster than $\max\bigl\{\bigl(\frac{d^2\log d}{np^2}\bigr)^{1/3},\frac{d\log d}{np^2}\bigr\}$.

One of the attractions of our analysis is the fact that we are able to provide bounds that only depend on entrywise missingness probabilities in an average sense, as opposed to worst-case missingness probabilities. The refinements conferred by such bounds are particularly important when the missingness mechanism is heterogeneous, as typically encountered in practice. The averaging of missingness probabilities can be partially seen in Theorem 4, since $\|\widetilde{\mathbf{W}}\|_{\infty\to\infty}$ and $\|\widetilde{\mathbf{W}}\|_{2\to\infty}$ depend only on the ℓ_1 and ℓ_2 norms of each row of $\widetilde{\mathbf{W}}$, but is even more evident in the proposition below, which gives a probabilistic bound on the original $\sin\Theta$ distance between $\widetilde{\mathbf{V}}_K$ and $\mathbf{V}_K$.

PROPOSITION 2. *Assume the same conditions as in Theorem 4. Then there exists a universal constant $C>0$ such that for any $\xi>1$, if*

$$\lambda_K>C\biggl\{\biggl(\frac{M\tau^2R\|\widetilde{\mathbf{W}}\|_{1\to1}\xi\log d}{n}\biggr)^{1/2}+\frac{M\|\widetilde{\mathbf{W}}\|_{\mathrm{op}}\xi\log^2 d}{n}\biggr\},\tag{11}$$

then

$$\mathbb{P}^{\mathbf{\Omega}}\biggl\{L(\widetilde{\mathbf{V}}_K,\mathbf{V}_K)\geq\frac{2^{9/2}eK\tau\mu}{\lambda_K}\biggl(\frac{MR}{d}\biggr)^{1/2}\biggl(\frac{\xi^{1/2}\|\widetilde{\mathbf{W}}\|_1^{1/2}\log^{1/2}d}{n^{1/2}}+\frac{\xi\|\widetilde{\mathbf{W}}\|_{\mathrm{F}}\log d}{n}\biggr)\biggr\}\leq(2K+4)d^{-(\xi-1)}.$$

In this bound, we see that $L(\widetilde{\mathbf{V}}_K,\mathbf{V}_K)$ only depends on $\widetilde{\mathbf{W}}$ through the entrywise ℓ_1 and ℓ_2 norms of the whole matrix. Lemma 6 again provides probabilistic control of these norms under the p-homogeneous missingness mechanism. In general, if the rows of $\mathbf{\Omega}$ are independent and identically distributed, but different covariates are missing with different probabilities, then off-diagonal entries of $\widetilde{\mathbf{W}}$ will concentrate around the reciprocals of the simultaneous observation probabilities of pairs of covariates. As such, for a typical realisation of $\mathbf{\Omega}$, our bound in Proposition 2 depends only on the harmonic averages of these simultaneous observation probabilities and their squares. Such an averaging effect ensures that our method is effective in a much wider range of heterogeneous settings than previously allowed in the literature.

3.2. Weakening the missingness proposition condition for contraction

Theorem 3 provides a geometric contraction guarantee for the `primePCA` algorithm in the noiseless case. The price we pay for this strong conclusion, however, is a strong condition on the proportion of missingness that enters the contraction rate parameter ρ through $\|\mathbf{\Omega}_{\mathcal{I}}^{\mathrm{c}}\|_{1\to 1}$; indeed in an asymptotic framework where the incoherence parameter μ grows with the sample size and/or dimension, the proportion of missingness would need to vanish asymptotically. Therefore, to complement our earlier theory, we present below Proposition 3 and Corollary 1. Proposition 3 is an analogue of the deterministic Proposition 1 in that it demonstrates that a single iteration of the `primePCA` algorithm yields a contraction provided that the input $\widehat{\mathbf{V}}_K^{(\mathrm{in})}$ is sufficiently close to $\mathbf{V}_K$. The two main differences are first that the contraction is in terms of a Procrustes-type loss (see the discussion around (7)), which turns out to be convenient for Corollary 1; and second, the bound depends only on the incoherence of the matrix $\mathbf{V}_K$, and not on the corresponding quantity for $\mathbf{U}$.

PROPOSITION 3. *Let* $\widehat{\mathbf{V}}_K^{(\mathrm{in})} \in \mathbb{O}^{d\times K}$, *let* $\mathbf{O} := \mathrm{argmin}_{\widetilde{\mathbf{O}}\in\mathbb{O}^{K\times K}} \|\widehat{\mathbf{V}}_K^{(\mathrm{in})} - \mathbf{V}_K\widetilde{\mathbf{O}}\|_{\mathrm{F}}$ *and let* $\mathbf{\Xi} := \widehat{\mathbf{V}}_K^{(\mathrm{in})} - \mathbf{V}_K\mathbf{O}$. *Fix* $\mathbf{U} \in \mathbb{R}^{n\times K}$ *and* $\mathbf{V}_K \in \mathbb{O}^{d\times K}$ *with* $\|\mathbf{V}_K\|_{2\to\infty} \le \mu(K/d)^{1/2}$, *and let* $\mathbf{Y} := \mathbf{U}\mathbf{V}_K^\top$, *with* $c := \sigma_K(\mathbf{Y})/\|\mathbf{Y}\|_{\mathrm{F}}$. *Suppose that* $\kappa_1, \kappa_2, \kappa_3 > 0$ *are such that for every* $i \in [n]$,

$$\frac{\|\mathbf{V}_{\mathcal{J}_i,K}^\top \mathbf{\Xi}_{\mathcal{J}_i}\|_{\mathrm{op}}}{|\mathcal{J}_i|} \le \frac{\|\mathbf{\Xi}\|_{\mathrm{op}}^2}{d} + \kappa_1\frac{\mu\|\mathbf{\Xi}\|_{\mathrm{op}}}{d^{3/2}}, \quad \frac{\|\mathbf{\Xi}_{\mathcal{J}_i}\|_{\mathrm{op}}^2}{|\mathcal{J}_i|} \le \kappa_2\frac{\|\mathbf{\Xi}\|_{\mathrm{op}}^2}{d}, \quad \|\mathbf{\Xi}_{\mathcal{J}_i^c}\|_{\mathrm{op}}^2 \le \kappa_3\|\mathbf{\Xi}\|_{\mathrm{op}}^2. \tag{12}$$

Assume further that

$$\|\mathbf{\Xi}\|_{\mathrm{op}} \le \min\left\{\left(\frac{c}{4\sigma_*^2(\kappa_1+\kappa_2)}\right)^{1/2}, \frac{c}{4\mu\kappa_1\sigma_*^2 K}\left(\frac{d}{\log K}\right)^{1/2}\right\}. \tag{13}$$

Then the output $\widehat{\mathbf{V}}_K^{(\mathrm{out})} := \mathtt{refine}\big(K, \widehat{\mathbf{V}}_K^{(\mathrm{in})}, \mathbf{\Omega}, \mathbf{Y}_{\mathbf{\Omega}}\big)$ *of Algorithm 1 satisfies*

$$\|\widehat{\mathbf{V}}_K^{(\mathrm{out})} - \mathbf{V}_K\widehat{\mathbf{O}}\|_{\mathrm{op}} \le \frac{16\|\mathbf{\Xi}\|_{\mathrm{op}}}{c}\left\{\sigma_*^2(\kappa_1+\kappa_2)\|\mathbf{\Xi}\|_{\mathrm{op}} + \sigma_*^2\kappa_1\mu K\left(\frac{\log K}{d}\right)^{1/2} + \kappa_3^{1/2}\left(1+\frac{c}{2}\right)\right\},$$

where $\widehat{\mathbf{O}} := \mathrm{argmin}_{\widetilde{\mathbf{O}}\in\mathbb{O}^{K\times K}} \|\widehat{\mathbf{V}}_K^{(\mathrm{out})} - \mathbf{V}_K\widetilde{\mathbf{O}}\|_{\mathrm{F}} \in \mathbb{O}^{K\times K}$.

Interestingly, the proof of Proposition 3 proceeds in a very different fashion from that of Proposition 1. The key step is to bound the discrepancy between the principal components of the imputed data matrix $\widehat{\mathbf{Y}}$ in Algorithm 1 and $\mathbf{V}_K$ using a modified version of Wedin's theorem (Wang, 2016). To achieve the desired contraction rate, instead of viewing the true data matrix $\mathbf{Y}$ as the reference matrix when calculating the perturbation, we choose a different reference matrix $\widetilde{\mathbf{Y}}$ with the same top K right singular space as $\mathbf{Y}$ but which is closer to $\widehat{\mathbf{Y}}$ in terms of the Frobenius norm. Such a reference shift sharpens the eigenspace perturbation bound.

The contraction rate in Proposition 3 is a sum of three terms, the first two of which are small provided that $\|\mathbf{\Xi}\|_{\mathrm{op}}$ is small and d is large respectively. On the other hand, the final term is small provided that no small subset of the rows of $\mathbf{\Xi}$ contributes excessively to its operator norm. For different missingness mechanisms, such a guarantee would need to be established probabilistically on a case-by-case basis; in Corollary 1 we illustrate how this can be done to achieve a high probability contraction in the simplest missingness model. Importantly, the proportion of missingness allowed, and hence the contraction rate parameter, no longer depend on the incoherence of $\mathbf{V}_K$, and can be of constant order.

COROLLARY 1. *Consider the p-homogeneous MCAR setting. Fix* $\mathbf{U} \in \mathbb{R}^{n\times K}$ *and* $\mathbf{V}_K \in \mathbb{O}^{d\times K}$ *with* $\|\mathbf{V}_K\|_{2\to\infty} \le \mu(K/d)^{1/2}$, *and let* $\mathbf{Y} := \mathbf{U}\mathbf{V}_K^\top$, *with* $c := \sigma_K(\mathbf{Y})/\|\mathbf{Y}\|_{\mathrm{F}}$. *Suppose that* $\widehat{\mathbf{V}}_K^{(\mathrm{in})}, \widehat{\mathbf{V}}_K^{(\mathrm{out})} \in \mathbb{O}^{d\times K}$, $\mathbf{O}, \widehat{\mathbf{O}} \in \mathbb{O}^{K\times K}$ *and* $\mathbf{\Xi} \in \mathbb{R}^{d\times K}$ *are as in Proposition 3, let* $C_* := \|\mathbf{\Xi}\|_{\mathrm{op}}/\|\mathbf{\Xi}\|_{2\to\infty}$, *and suppose that*

$$\|\mathbf{\Xi}\|_{\mathrm{op}} \le \min\left\{\frac{p(1-p)^{1/2}}{44\mu K^{3/2}(\sigma_*\vee 1)\log(24nK/\delta)}, \left(\frac{c}{8\sigma_*^2}\right)^{1/2}, \frac{c}{4\mu\sigma_*^2 K}\left(\frac{d}{\log K}\right)^{1/2}\right\}.$$

Fix $\delta \in (0,1]$ and suppose that $dp \geq 8\log(3/\delta)$. Then with probability at least $1-\delta$, the output $\widehat{\mathbf{V}}_K^{(\mathrm{out})} := \mathtt{refine}\big(K, \widehat{\mathbf{V}}_K^{(\mathrm{in})}, \boldsymbol{\Omega}, \mathbf{Y}_{\boldsymbol{\Omega}}\big)$ of Algorithm 1 satisfies

$$\|\widehat{\mathbf{V}}_K^{(\mathrm{out})} - \mathbf{V}_K\widehat{\mathbf{O}}\|_{\mathrm{op}} \leq \frac{125}{c}\biggl\{K^{1/2}(1-p)^{1/2} + \frac{\log^{1/2}(3/\delta)}{C_*}\biggr\}\|\widehat{\mathbf{V}}_K^{(\mathrm{in})} - \mathbf{V}_K\mathbf{O}\|_{\mathrm{op}}.$$

To understand the conclusion of Corollary 1, it is instructive to consider the special case $K = 1$. Here, $c = 1$ and C_* is the ratio of the ℓ_2 and ℓ_∞ norms of the vector $\widehat{\mathbf{V}}_1^{(\mathrm{in})} - \mathrm{sgn}(\mathbf{V}_1^\top\widehat{\mathbf{V}}_1^{(\mathrm{in})})\mathbf{V}_1$. When the entries of this vector are of comparable magnitude, C_* is therefore of order $d^{1/2}$, so the contraction rate is of order $(1-p)^{1/2} + d^{-1/2}$.

3.3. Other missingness mechanisms

Another interesting aspect of our theory is that the guarantees provided in Theorem 3 are deterministic. Provided we start with a sufficiently good initialiser, Theorem 3 describes the way in which the performance of `primePCA` improves over iterations. An attraction of this approach is that it offers the potential to study the performance of `primePCA` under more general missingness mechanisms. For instance, one setting of considerable practical interest is the Missing At Random (MAR) model, which postulates that our data vector $\mathbf{y} = (y_1, \ldots, y_d)$ and observation indicator vector $\boldsymbol{\omega}$ satisfy

$$\mathbb{P}(\boldsymbol{\omega} = \boldsymbol{\epsilon} \mid \mathbf{y} = \mathbf{a}) = \mathbb{P}\biggl(\boldsymbol{\omega} = \boldsymbol{\epsilon} \;\Big|\; \bigcap_{j:\epsilon_j=1}\{y_j = a_j\}\biggr) \tag{14}$$

for all $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_d)^\top \in \{0,1\}^d$ and $\mathbf{a} = (a_1, \ldots, a_d)^\top \in \mathbb{R}^d$. In other words, the probability of seeing a particular missingness pattern only depends on the data vector through components of this vector that are observed. Thus, if we want to understand the performance of `primePCA` under different missingness mechanisms, such as specific MAR (or even Missing Not At Random (MNAR)) models, all we require is an analogue of Theorem 4 on the performance of the initialiser in these new missingness settings. Such results, however, are likely to be rather problem-specific in nature, and it can be that choosing an initialiser based on available information on the dependence between the observations and the missingness mechanism makes it easier to prove the desired performance guarantees.

We now provide an example to illustrate how such initialisers can be constructed and analysed. Consider an MAR setting where the missingness pattern depends on the data matrix only through a fully observed categorical variable. In this case, we can construct a variant of the OPW estimator, denoted $\widehat{\mathbf{V}}_K^{\mathrm{OPWv}}$, by modifying the weighted sample covariance matrix in (8) to condition on the fully observed covariate, and then take the leading eigenvectors of an appropriate average of these conditional weighted sample covariance matrices. Specifically, suppose that our data consist of independent and identically distributed copies $(\mathbf{y}_1, \boldsymbol{\omega}_1), \ldots, (\mathbf{y}_n, \boldsymbol{\omega}_n)$ of $(\mathbf{y}, \boldsymbol{\omega}) = (y_0, y_1, \ldots, y_d, \omega_0, \omega_1, \ldots, \omega_d)$, where $\omega_0 = 1$, where y_0 is a categorical random variable taking values in $\{1, \ldots, L\}$ and where $(y_1, \ldots, y_d) \mid y_0 \sim N_d(\mathbf{0}, \boldsymbol{\Sigma}_{y_0})$ is independent of $\omega_j \mid y_0 \overset{\mathrm{iid}}{\sim} \mathrm{Bern}(p_{y_0})$ for all $j \in [d]$. Writing $\mathbf{y}_{-0} := (y_1, \ldots, y_d)^\top$, $\boldsymbol{\omega}_{-0} := (\omega_1, \ldots, \omega_d)^\top$ and $\widetilde{\mathbf{y}}_{-0} := \mathbf{y}_{-0} \circ \boldsymbol{\omega}_{-0}$, we have that $\mathrm{Cov}(y_0, \mathbf{y}_{-0}) = \mathbf{0}$, i.e. $\mathrm{Cov}(\mathbf{y})$ is block diagonal. Thus, introducing the subscript i for our ith observation, as a starting point to construct $\widehat{\mathbf{V}}_K^{\mathrm{OPWv}}$, it is natural to consider an oracle estimator of $\mathrm{Cov}(\mathbf{y}_{-0})$, given by

$$\mathbf{G} := \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i:y_{i0}=\ell}\widetilde{\mathbf{y}}_{i,-0}\widetilde{\mathbf{y}}_{i,-0}^\top \circ \mathbf{W}_\ell,$$

where $\mathbf{W}_\ell := p_\ell^{-2}\bigl\{\mathbf{1}_d\mathbf{1}_d^\top - (1-p_\ell)\mathbf{I}_d\bigr\}$. Observe that we can write

$$\mathbf{G} = \sum_{\ell=1}^{L}\frac{n_\ell}{n}\mathbf{G}^{(\ell)},$$

where $n_\ell := |\{i : y_{i0} = \ell\}|$ and where $\mathbf{G}^{(\ell)} := n_\ell^{-1} \sum_{i:y_{i0}=\ell} \widetilde{\mathbf{y}}_{i,-0} \widetilde{\mathbf{y}}_{i,-0}^\top \circ \mathbf{W}_\ell$ is the OPW estimator of $\mathbf{\Sigma}_\ell$ based on the observations with $y_{i0} = \ell$. Hence, $\mathbf{G}$ is unbiased for $\mathrm{Cov}(\mathbf{y}_{-0})$, because

$$\mathbb{E}(\mathbf{G}) = \sum_{\ell=1}^{L} \mathbb{E}\biggl(\frac{n_\ell}{n} \mathbb{E}(\mathbf{G}^{(\ell)} \mid y_{10}, \ldots, y_{n0})\biggr) = \sum_{\ell=1}^{L} \frac{\mathbb{E}(n_\ell)}{n} \mathbf{\Sigma}_\ell = \mathrm{Cov}(\mathbf{y}_{-0}).$$

In practice, when p_ℓ is unknown, we can estimate it by

$$\widehat{p}_\ell := \frac{1}{d n_\ell} \sum_{i:y_{i0}=\ell} \sum_{j=1}^{d} \omega_{ij},$$

and substitute this estimate into $\mathbf{W}_\ell$ to obtain empirical estimators $\widetilde{\mathbf{G}}^{(\ell)}$ and $\widetilde{\mathbf{G}}$ of $\mathbf{G}^{(\ell)}$ and $\mathbf{G}$ respectively. Finally, $\widehat{\mathbf{V}}_K^{\mathrm{OPWv}}$ can be obtained as the matrix of top K eigenvectors of the $(d+1)\times(d+1)$ matrix

$$\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n y_{i0}^2 - (\frac{1}{n}\sum_{i=1}^n y_{i0})^2 & \mathbf{0}^\top \\ \mathbf{0} & \widetilde{\mathbf{G}} \end{pmatrix}.$$

To sketch the way to bound the $\sin\Theta$ loss of such an initialiser, we can condition on $y_{10}, \ldots, y_{n0}$ and apply matrix Bernstein concentration inequalities similarly to those in the proof of Theorem 1 to show that $\widetilde{\mathbf{G}}^{(\ell)}$ is close to $\mathbf{\Sigma}_\ell$ for each ℓ. Simple binomial concentration bounds then allow us to combine these to control $\|\widetilde{\mathbf{G}} - \mathrm{Cov}(\mathbf{y}_{-0})\|_{\mathrm{op}}$, and then apply a variant the Davis–Kahan theorem to obtain a final result.

While different initialisers can be designed and analysed theoretically in specific missingness settings, as shown in the example above, our empirical experience, nevertheless, is that regardless of the missingness mechanism, `primePCA` is extremely robust to the choice of initialiser. This is evident from the discussion of the performance of `primePCA` in MAR and MNAR settings given in Section 4.4.

4. Simulation studies

In this section, we assess the empirical performance of `primePCA`, as proposed in Algorithm 2, with initialiser $\widetilde{\mathbf{V}}_K$ from Section 3.1, and denote the output of this algorithm by $\widehat{\mathbf{V}}_K^{\mathrm{prime}}$. In Sections 4.1, 4.2 and 4.3, we generate observations according to the model described in (1), (2) and (3) where the rows of the matrix $\mathbf{U}$ are independent $N_d(\mathbf{0}, \mathbf{\Sigma_u})$ random vectors, for some positive semi-definite $\mathbf{\Sigma_u} \in \mathbb{R}^{d\times d}$. We further generate the observation indicator matrix $\mathbf{\Omega}$, independently of $\mathbf{U}$ and $\mathbf{Z}$, and investigate the following four missingness mechanisms that represent different levels of heterogeneity:

(H1) Homogeneous: $\mathbb{P}(\omega_{ij} = 1) = 0.05$ for all $i \in [n], j \in [d]$;

(H2) Mildly heterogeneous: $\mathbb{P}(\omega_{ij} = 1) = P_i Q_j$ for $i \in [n], j \in [d]$, where $P_1, \ldots, P_n \overset{\mathrm{iid}}{\sim} U[0, 0.2]$ and $Q_1, \ldots, Q_d \overset{\mathrm{iid}}{\sim} U[0.05, 0.95]$ independently;

(H3) Highly heterogeneous columns: $\mathbb{P}(\omega_{ij} = 1) = 0.19$ for $i \in [n]$ and all odd $j \in [d]$ and $\mathbb{P}(\omega_{ij} = 1) = 0.01$ for $i \in [n]$ and all even $j \in [d]$.

(H4) Highly heterogeneous rows: $\mathbb{P}(\omega_{ij} = 1) = 0.18$ for $j \in [d]$ and all odd $i \in [n]$ and $\mathbb{P}(\omega_{ij} = 1) = 0.02$ for $j \in [d]$ and all even $i \in [n]$.

In Sections 4.1, 4.2 and 4.3 below, we investigate `primePCA` in noiseless, noisy and misspecified settings respectively. Section 4.4 is devoted to MAR and MNAR settings. In all cases, the average statistical error was estimated from 100 Monte Carlo repetitions of the experiment. For comparison, we also studied the `softImpute` algorithm (Mazumder, Hastie and Tibshirani, 2010; Hastie et al., 2015), which is considered to be state-of-the-art for matrix completion (Chi, Lu and Chen, 2018). This algorithm imputes the missing entries of $\mathbf{Y}$ by solving the following nuclear-norm-regularised optimisation problem:

$$\widehat{\mathbf{Y}}^{\mathrm{soft}} := \operatorname*{argmin}_{\mathbf{X}\in\mathbb{R}^{n\times d}} \biggl\{ \frac{1}{2} \|\mathbf{Y}_{\mathbf{\Omega}} - \mathbf{X}_{\mathbf{\Omega}}\|_{\mathrm{F}}^2 + \lambda \|\mathbf{X}\|_* \biggr\},$$

where $\lambda > 0$ is to be chosen by the practitioner. The `softImpute` estimator of $\mathbf{V}_K$ is then given by the matrix of top K right singular vectors $\widehat{\mathbf{V}}_K^{\mathrm{soft}}$ of $\widehat{\mathbf{Y}}^{\mathrm{soft}}$. In practice, the optimisation is carried out by representing $\mathbf{X}$ as $\mathbf{A}\mathbf{B}^\top$, and performing alternating projections to update $\mathbf{A} \in \mathbb{R}^{n\times K}$ and $\mathbf{B} \in \mathbb{R}^{d\times K}$ iteratively. The fact that the `softImpute` algorithm was originally intended for matrix completion means that it treats the left and right singular vectors symmetrically, whereas the `primePCA` algorithm, which has the advantage of a clear geometric interpretation as exemplified in Figure 1, focuses on the target of inference in PCA, namely the leading right singular vectors.

Figure 2 presents Monte Carlo estimates of $\mathbb{E}L(\widehat{\mathbf{V}}_K^{\mathrm{prime}}, \mathbf{V}_K)$ for different choices of σ_* in two different settings. The first uses the noiseless set-up of Section 4.1, together with missingness mechanism (H1); the second uses the noisy setting of Section 4.2 with parameter $\nu = 20$ and missingness mechanism (H2). We see that the error barely changes when σ_* varies within $[2, 10]$; very similar plots were obtained for different data generation and missingness mechanisms, though we omit these for brevity. For definiteness, we therefore fixed $\sigma_* = 3$ throughout our simulation study.

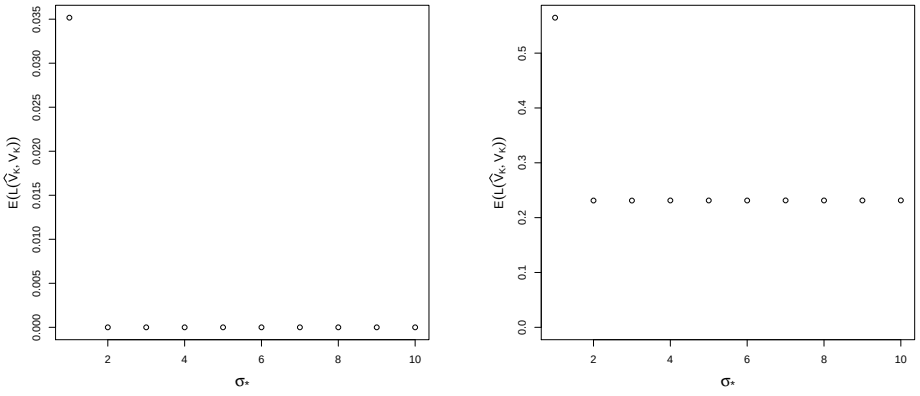

Fig. 2. Estimates of $\mathbb{E}L(\widehat{\mathbf{V}}_K^{\mathrm{prime}}, \mathbf{V}_K)$ for various choices of σ_* under (H1) in the noiseless setting of Section 4.1 (left) and (H2) in the noisy setting of Section 4.2 with $\nu = 20$ (right).

4.1. Noiseless case

In the noiseless setting, we let $\mathbf{Z} = \mathbf{0}$, and also fix $n = 2000$, $d = 500$, $K = 2$ and $\mathbf{\Sigma_u} = 100\mathbf{I}_2$. We set

$$\mathbf{V}_K = \sqrt{\frac{1}{500}} \begin{pmatrix} \mathbf{1}_{250} & \mathbf{1}_{250} \\ \mathbf{1}_{250} & -\mathbf{1}_{250} \end{pmatrix} \in \mathbb{R}^{500\times 2}.$$

In Figure 3, we present the (natural) logarithm of the estimated average loss of `primePCA` and `softImpute` under (H1), (H2), (H3) and (H4). We set the range of y-axis to be the same for each method to facilitate straightforward comparison. We see that the statistical error of `primePCA` decreases geometrically as the number of iterations increases, which confirms the conclusion of Theorem 3 in this noiseless setting. Moreover, after a moderate number of iterations, its performance is a substantial improvement on that of the `softImpute` algorithm, even if this latter algorithm is given access to an oracle choice of the regularisation parameter λ. The high statistical error of `softImpute` in these settings can be partly explained by the default value of the tuning parameter `thresh` in the `softImpute` package in `R`, namely 10^{-5}, which corresponds to the red curve in the right-hand panels of Figure 3. By reducing the values of `thresh` to 10^{-7} and 10^{-9}, corresponding to the green and blue curves in Figure 3 respectively, we were able to improve the performance of `softImpute` to some extent, though the statistical error is sensitive to the choice of the regularisation parameter λ. Moreover, even with the optimal choice of λ, it is not competitive with `primePCA`. Finally, we mention that for the 2000 iterations of setting (H2), `primePCA` took on average just under 10 minutes per repetition to compute, whereas the solution path of `softImpute` with `thresh` $= 10^{-9}$ took around 36 minutes per repetition.

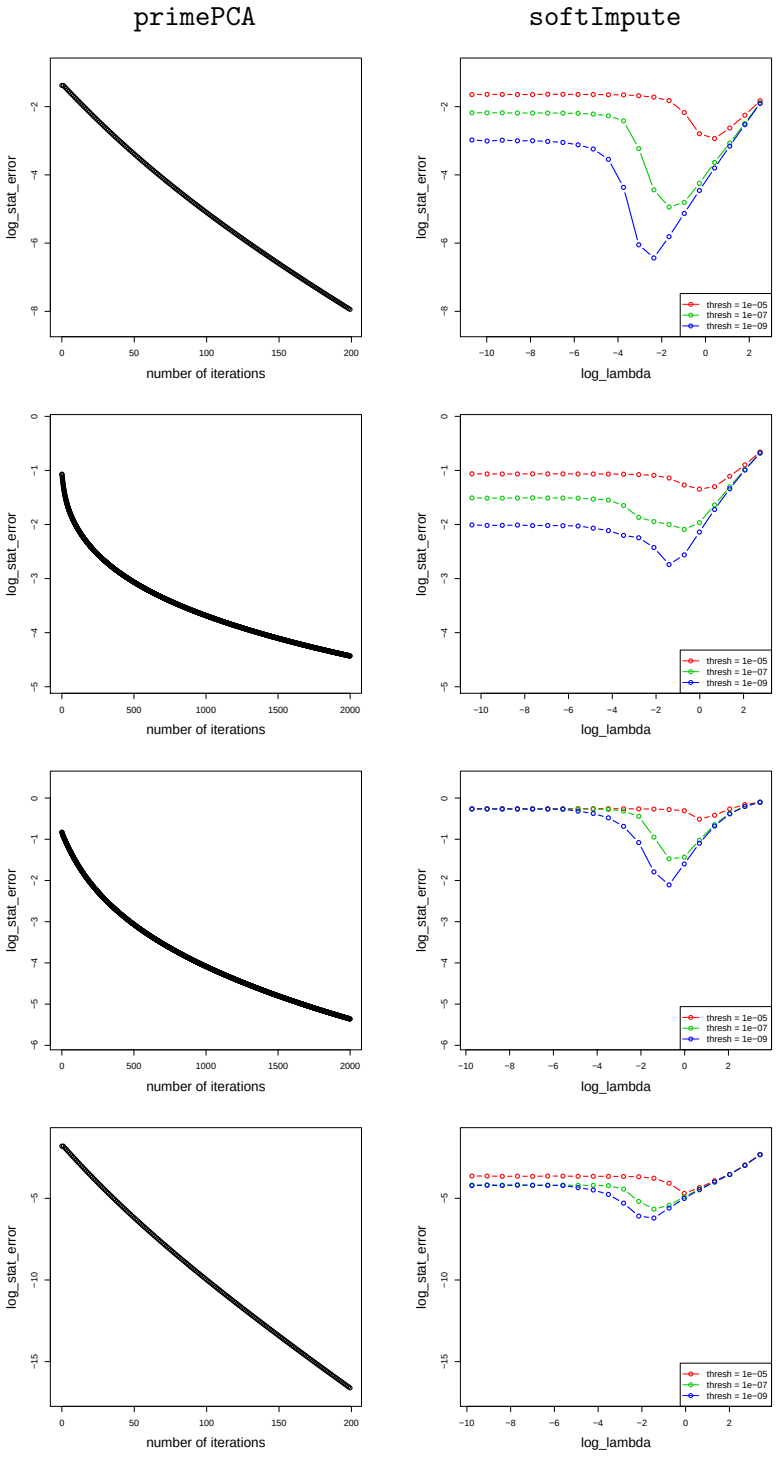

Fig. 3. Logarithms of the average Frobenius norm $\sin\Theta$ error of `primePCA` and `softImpute` under various heterogeneity levels of missingness in absence of noise. The four rows of plots above, from the top to bottom, correspond to (H1), (H2), (H3) and (H4).

Table 1. Average losses (with standard errors in brackets) under (H1), (H2), (H3) and (H4).

		$\nu = 10$	$\nu = 20$	$\nu = 40$	$\nu = 60$
(H1)	`hardImpute`	$0.891_{(0.005)}$	$0.444_{(0.001)}$	$0.251_{(0.001)}$	$0.186_{(0.0005)}$
	`softImpute(oracle)`	$0.377_{(0.0009)}$	$0.186_{(0.0004)}$	$0.095_{(0.0002)}$	$0.064_{(0.0002)}$
	`primePCA_init`	$0.449_{(0.001)}$	$0.306_{(0.001)}$	$0.266_{(0.001)}$	$0.259_{(0.001)}$
	`primePCA`	$0.368_{(0.001)}$	$0.171_{(0.0004)}$	$0.084_{(0.0002)}$	$0.056_{(0.0001)}$
(H2)	`hardImpute`	$0.920_{(0.006)}$	$0.473_{(0.001)}$	$0.291_{(0.001)}$	$0.236_{(0.001)}$
	`softImpute(oracle)`	$0.519_{(0.001)}$	$0.308_{(0.001)}$	$0.185_{(0.001)}$	$0.141_{(0.001)}$
	`primePCA_init`	$0.549_{(0.002)}$	$0.399_{(0.002)}$	$0.357_{(0.001)}$	$0.349_{(0.001)}$
	`primePCA`	$0.475_{(0.002)}$	$0.232_{(0.001)}$	$0.115_{(0.001)}$	$0.077_{(0.0005)}$
(H3)	`hardImpute`	$0.792_{(0.003)}$	$0.479_{(0.001)}$	$0.385_{(0.001)}$	$0.427_{(0.001)}$
	`softImpute(oracle)`	$0.622_{(0.002)}$	$0.374_{(0.001)}$	$0.222_{(0.001)}$	$0.170_{(0.001)}$
	`primePCA_init`	$0.624_{(0.002)}$	$0.486_{(0.001)}$	$0.449_{(0.001)}$	$0.442_{(0.001)}$
	`primePCA`	$0.581_{(0.002)}$	$0.290_{(0.001)}$	$0.145_{(0.001)}$	$0.097_{(0.0004)}$
(H4)	`hardImpute`	$0.368_{(0.001)}$	$0.174_{(0.0005)}$	$0.089_{(0.0003)}$	$0.062_{(0.0003)}$
	`softImpute(oracle)`	$0.243_{(0.0006)}$	$0.121_{(0.0002)}$	$0.062_{(0.0001)}$	$0.042_{(0.0001)}$
	`primePCA_init`	$0.290_{(0.0007)}$	$0.203_{(0.001)}$	$0.175_{(0.0005)}$	$0.169_{(0.0004)}$
	`primePCA`	$0.238_{(0.0006)}$	$0.116_{(0.0003)}$	$0.058_{(0.0002)}$	$0.038_{(0.0001)}$

4.2. Noisy case

Here, we generate the rows of $\mathbf{Z}$ as independent $N_d(\mathbf{0}, \mathbf{I}_d)$ random vectors, independent of all other data. We maintain the same choices of n, d, K and $\mathbf{V}_K$ as in Section 4.1, set $\mathbf{\Sigma_u} = \nu^2\mathbf{I}_2$ and vary $\nu > 0$ to achieve different signal-to-noise ratios. In particular, defining $\mathrm{SNR} := \operatorname{tr}\operatorname{Cov}(\mathbf{x}_1)/\operatorname{tr}\operatorname{Cov}(\mathbf{z}_1)$, the choices $\nu = 10, 20, 40, 60$ correspond to the very low, low, medium and high signal-to-noise ratios $\mathrm{SNR} = 0.4, 1.6, 6.4, 14.4$, respectively. For an additional comparision, we consider a variant of the `softImpute` algorithm called `hardImpute` (Mazumder, Hastie and Tibshirani, 2010), which retains only a fixed number of top singular values in each iteration of matrix imputation; this can be achieved by setting the argument λ in the `softImpute` function to be 0.

To avoid confounding our study of the statistical performance of the `softImpute` algorithm with the choice of regularisation parameter λ, we gave the `softImpute` algorithm a particularly strong form of oracle choice of λ, namely where λ was chosen for each individual repetition of the experiment, so as to minimise the loss function. Naturally, such a choice is not available to the practitioner. Moreover, in order to ensure the range of λ was wide enough to include the best `softImpute` solution, we set the argument `rank.max` in that algorithm to be 20.

In Table 1, we report the statistical error of `primePCA` after 2000 iterations of refinement, together with the corresponding statistical errors of our initial estimator `primePCA_init` and those of `softImpute(oracle)` and `hardImpute`. Remarkably, `primePCA` exhibits stronger performance than these other methods across each of the signal-to-noise ratio regimes and different missingness mechanisms. We also remark that `hardImpute` is inaccurate and unstable, because it might converge to a local optimum that is far from the truth.

4.3. Near low-rank case

In this subsection, we set $n = 2000$, $d = 500$, $K = 10$, $\mathbf{\Sigma_u} = \operatorname{diag}(2^{10}, 2^9, \ldots, 2)$, and fixed $\mathbf{V}_K$ once for all experiments to be the top K eigenvectors of one realisation‡ of the sample covariance matrix of n independent $N_d(\mathbf{0}, \mathbf{I}_d)$ random vectors. Here $d^{1/2}\|\mathbf{V}_K\|_{2\to\infty} < 1.72$, and we again generated the rows of $\mathbf{Z}$ as independent $N_d(\mathbf{0}, \mathbf{I}_d)$ random vectors. Table 2 reports the average loss of estimating the top $\widehat{K}$ eigenvectors of $\mathbf{\Sigma_y}$, where $\widehat{K}$ varies from 1 to 5. Interestingly, even in this misspecified setting, `primePCA` is competitive with the oracle version of `softImpute`.

‡In `R`, we set the random seed to be 2019 before generating $\mathbf{V}_K$.

Table 2. Average losses (with standard errors in brackets) in the setting of Section 4.3 under (H1), (H2), (H3) and (H4).

		$\widehat{K}=1$	$\widehat{K}=2$	$\widehat{K}=3$	$\widehat{K}=4$	$\widehat{K}=5$
(H1)	`hardImpute`	$0.308_{(0.002)}$	$0.507_{(0.002)}$	$0.764_{(0.004)}$	$1.199_{(0.006)}$	$1.524_{(0.004)}$
	`softImpute(oracle)`	$0.107_{(0.001)}$	$0.182_{(0.001)}$	$0.275_{(0.001)}$	$0.401_{(0.001)}$	$0.596_{(0.001)}$
	`primePCA_init`	$0.203_{(0.001)}$	$0.345_{(0.001)}$	$0.554_{(0.003)}$	$1.074_{(0.007)}$	$1.427_{(0.006)}$
	`primePCA`	$0.141_{(0.001)}$	$0.200_{(0.001)}$	$0.269_{(0.001)}$	$0.374_{(0.001)}$	$0.580_{(0.001)}$
(H2)	`hardImpute`	$0.298_{(0.002)}$	$0.466_{(0.002)}$	$0.696_{(0.003)}$	$1.124_{(0.006)}$	$1.452_{(0.004)}$
	`softImpute(oracle)`	$0.188_{(0.001)}$	$0.283_{(0.001)}$	$0.410_{(0.001)}$	$0.562_{(0.001)}$	$0.751_{(0.001)}$
	`primePCA_init`	$0.285_{(0.001)}$	$0.443_{(0.004)}$	$0.757_{(0.013)}$	$1.201_{(0.004)}$	$1.533_{(0.003)}$
	`primePCA`	$0.190_{(0.002)}$	$0.267_{(0.002)}$	$0.368_{(0.003)}$	$0.543_{(0.008)}$	$0.797_{(0.009)}$
(H3)	`hardImpute`	$0.302_{(0.001)}$	$0.482_{(0.002)}$	$0.695_{(0.002)}$	$1.004_{(0.006)}$	$1.373_{(0.004)}$
	`softImpute(oracle)`	$0.206_{(0.001)}$	$0.338_{(0.001)}$	$0.492_{(0.001)}$	$0.664_{(0.002)}$	$0.878_{(0.002)}$
	`primePCA_init`	$0.341_{(0.001)}$	$0.528_{(0.019)}$	$1.097_{(0.008)}$	$1.306_{(0.008)}$	$1.597_{(0.004)}$
	`primePCA`	$0.222_{(0.001)}$	$0.330_{(0.002)}$	$0.452_{(0.003)}$	$0.641_{(0.008)}$	$0.919_{(0.007)}$
(H4)	`hardImpute`	$0.090_{(0.001)}$	$0.148_{(0.001)}$	$0.226_{(0.001)}$	$0.346_{0.002}$	$0.589_{(0.007)}$
	`softImpute(oracle)`	$0.071_{(0.001)}$	$0.112_{(0.001)}$	$0.164_{(0.001)}$	$0.233_{(0.001)}$	$0.332_{(0.001)}$
	`primePCA_init`	$0.139_{(0.001)}$	$0.220_{(0.001)}$	$0.325_{(0.001)}$	$0.475_{(0.002)}$	$0.805_{(0.012)}$
	`primePCA`	$0.098_{(0.001)}$	$0.135_{(0.001)}$	$0.176_{(0.001)}$	$0.236_{(0.001)}$	$0.328_{(0.001)}$

4.4. Other missingness mechanisms

Finally in this section, we investigate the performance of `primePCA`, as well as other alternative algorithms, in settings where the MCAR hypothesis is not satisfied. We consider two simulation frameworks to explore both MAR (see (14)) and MNAR mechanisms. In the first, we assume that missingness depends on the data matrix $\mathbf{Y}$ only through a fully observed covariate, as in the example in Section 3.3. Specifically, for some $\alpha \geq 0$, for $K = 2$, and for two matrices§ $\mathbf{V}_+, \mathbf{V}_- \in \mathbb{O}^{d\times 2}$, the pair $(\mathbf{y}_1, \boldsymbol{\omega}_1) = (y_{10}, y_{11} \ldots, y_{1d}, \omega_{10}, \omega_{11}, \ldots, \omega_{1d})$ is generated as follows:

$$\begin{aligned} &\omega_{10} = 1, \quad y_{10} \sim \mathrm{Unif}\{-1, 1\}, \\ &(y_{11}, \ldots, y_{1d}), \omega_{11}, \ldots, \omega_{1d} \text{ are conditionally independent given } y_{10}, \\ &(y_{11}, \ldots, y_{1d})^\top \mid y_{10} \sim \begin{cases} N_d(\mathbf{0}, \mathbf{V}_+ \mathrm{diag}(40, 10)\mathbf{V}_+^\top + \mathbf{I}_d) & \text{if } y_{10} = 1 \\ N_d(\mathbf{0}, \mathbf{V}_- \mathrm{diag}(40, 10)\mathbf{V}_-^\top + \mathbf{I}_d) & \text{if } y_{10} = -1, \end{cases} \\ &\mathbb{P}(\omega_{1j} = 1 \mid y_{10}) = \left\{1 + \exp\left(\frac{j}{d} + y_{10}\alpha\right)\right\}^{-1}, \quad \text{for } j \in [d]. \end{aligned} \tag{15}$$

The other rows of $(\mathbf{Y}, \boldsymbol{\Omega})$ are taken to be as independent copies of $(\mathbf{y}_1, \boldsymbol{\omega}_1)$. Thus, when $\alpha = 0$, the matrices $\mathbf{Y}$ and $\boldsymbol{\Omega}$ are independent, and we are in an MCAR setting; when $\alpha \neq 0$, the data are MAR but not MCAR, and α measures the extent of departure from the MCAR setting. The covariance matrix of $\mathbf{y}_1$ is

$$\boldsymbol{\Sigma}_{\mathbf{y}} = \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \frac{1}{2}\mathbf{V}_+ \mathrm{diag}(40, 10)\mathbf{V}_+^\top + \frac{1}{2}\mathbf{V}_- \mathrm{diag}(40, 10)\mathbf{V}_-^\top + \mathbf{I}_d \end{pmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}.$$

In this example, we can construct a variant of the OPW estimator, which we call the OPWv estimator, by exploiting the fact that, conditional on the fully observed first column of $\mathbf{Y}$, the data are MCAR. To do this, let

$$\widehat{\boldsymbol{\Sigma}}^{\mathrm{OPWv}} := \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \frac{1}{2}\widetilde{\mathbf{G}}_+ + \frac{1}{2}\widetilde{\mathbf{G}}_- \end{pmatrix},$$

where $\widetilde{\mathbf{G}}_+$ and $\widetilde{\mathbf{G}}_-$ are the weighted sample covariance matrices computed as in (8), based on data $(y_{ij}, \omega_{ij})_{i:y_{i0}=1, j\in[d]}$ and $(y_{ij}, \omega_{ij})_{i:y_{i0}=-1, j\in[d]}$ respectively. The OPWv estimator is the matrix of the

§To be completely precise, in our simulations, $\mathbf{V}_+$ and $\mathbf{V}_-$ were generated independently (and independently of all other randomness) and were drawn from Haar measure on $\mathbb{O}^{d\times 2}$; however, these matrices were then fixed for every replication, so it is convenient to regard them as deterministic for the purposes of this description.

Table 3. Root mean squared errors of the $\sin\Theta$ loss function (with standard errors in brackets) over 100 repetitions from the data generating mechanism in (15) for OPW estimator and its class-weighted variant (OPWv), EM and `primePCA` with both the OPW or OPWv initialisers.

d	α	OPW	OPWv	EM	EMv	primePCA	primePCAv
25	0.1	$0.266_{(0.005)}$	$0.247_{(0.004)}$	$0.414_{(0.045)}$	$0.464_{(0.053)}$	$0.206_{(0.004)}$	$0.206_{(0.004)}$
25	0.5	$0.346_{(0.009)}$	$0.248_{(0.005)}$	$0.445_{(0.047)}$	$0.378_{(0.056)}$	$0.248_{(0.014)}$	$0.248_{(0.008)}$
50	0.1	$0.287_{(0.003)}$	$0.265_{(0.003)}$	$0.350_{(0.032)}$	$0.346_{(0.032)}$	$0.220_{(0.002)}$	$0.220_{(0.002)}$
50	0.5	$0.591_{(0.025)}$	$0.290_{(0.003)}$	$0.588_{(0.033)}$	$0.369_{(0.03)}$	$0.255_{(0.005)}$	$0.255_{(0.005)}$

first two eigenvectors of $\widehat{\boldsymbol{\Sigma}}^{\mathrm{OPWv}}$. Both the OPW and OPWv estimators are plausible initialisers for the `primePCA` algorithm.

In low-dimensional settings, likelihood-based approaches, often implemented via an EM algorithm, are popular for handling MAR data (14) (Rubin, 1976). In Table 3, we compare the performance of `primePCA` in this setting with that of an EM algorithm derived from the suggestion in Little and Rubin (2019, Section 11.3), and considered both the OPW and OPWv estimators as initialisers. We set $n = 500$, $d \in \{25, 50\}$, $\alpha \in \{0.1, 0.5\}$ and took $\widehat{K} = 2$ for both the `primePCA` and the EM algorithms. From the table, we see that the OPWv estimator is able to exploit the group structure of the data to improve upon the OPW estimator, especially for the larger value of α. It is reassuring to find that the performance of `primePCA` is completely unaffected by the choice of initialiser, and, remarkably, it outperforms the OPWv estimator, even though the latter has access to additional model structure information. The worse root mean squared error of the EM algorithm is mainly due its numerical instability when performing Schur complement computations¶.

The second simulation framework is as follows. Let $\boldsymbol{\Sigma} := (\min\{j,k\})_{j,k\in[d]} \in \mathbb{R}^{d\times d}$ and let $\boldsymbol{\xi} = (\xi_{ij})_{i\in[n],j\in[d]}$ be a latent Bernoulli thinning matrix. The data matrix $\mathbf{Y} = (y_{ij})_{i\in[n],j\in[d]}$ and revelation matrix $\boldsymbol{\Omega} = (\omega_{ij})_{i\in[n],j\in[d]}$ are generated in such a way that $\mathbf{Y}$ and $\boldsymbol{\xi}$ are independent,

$$\begin{aligned} (y_{i1}, \ldots, y_{id})^\top &\overset{\text{iid}}{\sim} N_d(\mathbf{0}, \boldsymbol{\Sigma}), \text{ for } i \in [n], \\ \xi_{ij} &\overset{\text{iid}}{\sim} \mathrm{Bern}(p), \\ \omega_{ij} &= \xi_{ij} \mathbb{1}_{\{\max_{1\le t<j} |y_{it}| < \tau\}}, \text{ for some } \tau > 0, \end{aligned} \tag{16}$$

(where the maximum of the empty set is $-\infty$ by convention). As usual, we observe $(\mathbf{Y} \circ \boldsymbol{\Omega}, \boldsymbol{\Omega})$. In other words, viewing each $(y_{i1}, \ldots, y_{id})$ as a d-step standard Gaussian random walk, we observe Bernoulli-thinned paths of the process up to (and including) the hitting time of the threshold $\pm\tau$. We note that the observations satisfy the MAR hypothesis if and only if $p = 1$, and as p decreases from 1, the mechanism becomes increasingly distant from MAR, as we become increasingly likely to fail to observe the threshold hitting time. We take $K = 1$.

In Table 4, we compare the performance of `primePCA` with that of the EM algorithm, and in both cases, we can initialise with either the OPW estimator or a mean-imputation estimator, obtained by imputing all missing entries by their respective population column means. We set $n = 500$, $d = 100$, $\tau = d^{1/2}$, took $\widehat{K} = 1$ for both `primePCA` and the EM algorithm, and took $p \in \{0.25, 0.5, 0.75, 1\}$. From the table, we see that `primePCA` outperforms the EM algorithm except in the MAR case where $p = 1$, which is tailor-made for the likelihood-based EM approach. In fact, `primePCA` is highly robust statistically and stable computationally, performing well consistently across different missingness settings and initialisers. On the other hand, the EM algorithm exhibits a much heavier dependence on the initialiser: its statistical performance suffers when initialised with the poorer mean-imputation estimator and runs into numerical instability issues when initialised with the OPW estimator in the MNAR settings. We found that these instability issues are exacerbated in higher dimensions, and

¶In fact, to try to improve the numerical stability of the EM procedure, we prevented the sample covariance estimators from exiting the cone of positive semi-definite matrices during iterations and took Moore–Penrose pseudoinverses with eigenvalues below 10^{-10} regarded as 0. Both of these modifications did indeed improve the algorithm, but some instability persists. Moreover, use of the SWEEP operator (Beaton, 1964), which is designed to compute the Schur complement in a numerically stable way, failed to remedy the situation, yielding identical (increasing) log-likelihood trajectories as the vanilla algorithm.

Table 4. Root mean squared errors of the $\sin\Theta$ loss function (with standard errors in brackets) over 100 repetitions from the data generating mechanism in (16) for mean-imputation estimator (MI), OPW estimator, EM and `primePCA` with both MI and OPW initialisers (distinguished by subscripts in table header).

p	MI	OPW	EM_{MI}	EM_{OPW}	$\text{primePCA}_{\text{MI}}$	$\text{primePCA}_{\text{OPW}}$
1	$0.548_{(0.004)}$	$0.282_{(0.004)}$	$0.086_{(0.003)}$	$0.056_{(0.002)}$	$0.096_{(0.002)}$	$0.096_{(0.002)}$
0.75	$0.551_{(0.004)}$	$0.285_{(0.004)}$	$0.117_{(0.004)}$	$0.353_{(0.041)}$	$0.097_{(0.002)}$	$0.097_{(0.002)}$
0.5	$0.557_{(0.005)}$	$0.29_{(0.004)}$	$0.186_{(0.025)}$	$0.944_{(0.013)}$	$0.1_{(0.002)}$	$0.1_{(0.002)}$
0.25	$0.575_{(0.005)}$	$0.309_{(0.005)}$	$0.228_{(0.005)}$	$0.989_{(0.001)}$	$0.112_{(0.002)}$	$0.123_{(0.006)}$

moreover, that the EM algorithm quickly becomes computationally infeasible‖. This explains why we did not run the EM algorithm on the larger-scale problems in Sections 4.1, 4.2 and 4.3, as well as the real data example in Section 5 below.

5. Real data analysis: Million Song Dataset

We apply `primePCA` to a subset of the Million Song Dataset** to analyse music preferences. The original data can be expressed as a matrix with 110,000 users (rows) and 163,206 songs (columns), with entries representing the number of times a song was played by a particular user. The proportion of non-missing entries in the matrix is 0.008%. Since the matrix is very sparse, and since most songs have very few listeners, we enhance the signal-to-noise ratio by restricting our attention to songs that have at least 100 listeners (1,777 songs in total). This improves the proportion of non-missing entries to 0.23%. Further summary information about the filtered data is provided below:

(a) Quantiles of non-missing matrix entry values:

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
1	1	1	1	1	1	2	3	5	8	500

90%	91%	92%	93%	94%	95%	96%	97%	98%	99%	100%
8	9	9	10	11	13	15	18	23	33	500

(b) Quantiles of the number of listeners for each song:

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
100	108	117	126	139	154	178	214	272.8	455.6	5043

(c) Quantiles of the total play counts of each user:

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0	0	1	3	4	6	9	14	21	38	1114

90%	91%	92%	93%	94%	95%	96%	97%	98%	99%	100%
38	41	44	48	54	60	68	79	97	132	1114

We mention here a respect in which the data set does not conform exactly to the framework studied in the paper, namely that we treat zero entries as missing data (this is very common for analyses of user-preference data sets). In practice, while it may indeed be the case that a zero play count for

‖Each iteration of the EM algorithm involves the inversion of n matrices, where the dimension of the ith such matrix is $\sum_{j=1}^{d}\omega_{ij} \times \sum_{j=1}^{d}\omega_{ij}$ (i.e. $O(d) \times O(d)$). Using standard matrix inversion algorithms, then, each iteration has computational complexity of order nd^3, and moreover the number of iterations required for numerical convergence can be very large in higher dimensions. This meant that even when $d = 100$, `primePCA` was nearly 50 times faster than the EM algorithm.

**`https://www.kaggle.com/c/msdchallenge/data`

song j by user i provides no indication of their level of preference for that song, it may also be the case that it reflects a dislike of that song. To address this issue, following our main analysis we will present a study of the robustness of our conclusions to different levels of true zeros in the data.

From point (a) above, we see that the distribution of play counts has an extremely heavy tail, and in particular the sample variances of the counts will be highly heterogeneous across songs. To guard against excessive influence from the outliers, we discretise the play counts into five interest levels as follows:

Play count	1	2 – 3	4 – 6	7 – 10	≥ 11
Level of interest	1	2	3	4	5

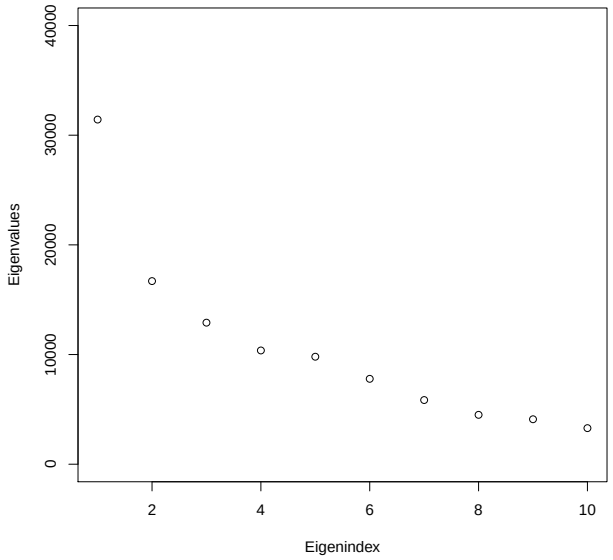

Fig. 4. Leading eigenvalues of $\widehat{\mathbf{\Sigma}}_{\mathbf{y}}$.

We are now in a position to analyse the data using `primePCA`, noting that one of the attractions of estimating the principal eigenspaces in this setting (as opposed to matrix completion, for instance), is that it becomes straightforward to make recommendations to new users, instead of having to run the algorithm again from scratch. For $i = 1, \ldots, n = 110{,}000$ and $j = 1, \ldots, d = 1{,}777$, let $y_{ij} \in \{1, \ldots, 5\}$ denote the level of interest of user i in song j, let $\widehat{K} = 10$ and let $\mathcal{I} = \{i : \|\omega_i\|_1 > \widehat{K}\}$. Our initial goal is to assess the top $\widehat{K}$ eigenvalues of $\mathbf{\Sigma}_{\mathbf{y}}$ to see if there is low-rank signal in $\mathbf{Y} = (y_{ij})$. To this end, we first apply Algorithm 2 to obtain $\widehat{\mathbf{V}}^{\texttt{prime}}_{\widehat{K}}$; next, for each $i \in \mathcal{I}$, we run Steps 2–5 of Algorithm 1 to obtain the estimated principal score $\widehat{\mathbf{u}}_i$, so that we can approximate $\mathbf{y}_i$ by $\widehat{\mathbf{y}}_i = \widehat{\mathbf{V}}^{\texttt{prime}}_{\widehat{K}}\widehat{\mathbf{u}}_i$. This allows us to estimate $\mathbf{\Sigma}_{\mathbf{y}}$ by $\widehat{\mathbf{\Sigma}}_{\mathbf{y}} = n^{-1}\sum_{i\in\mathcal{I}}\widehat{\mathbf{y}}_i\widehat{\mathbf{y}}_i^\top$. Figure 4 displays the top $\widehat{K}$ eigenvalues of $\widehat{\mathbf{\Sigma}}_{\mathbf{y}}$, which exhibit a fairly rapid decay, thereby providing evidence for the existence of low-rank signal in $\mathbf{Y}$.

In the left panel of Figure 5, we present the estimate $\widehat{\mathbf{V}}^{\text{prime}}_2$ of the top two eigenvectors of the covariance matrix $\mathbf{\Sigma}_{\mathbf{y}}$, with colours indicating the genre of the song. The outliers in the x-axis of this plot are particularly interesting: they reveal songs that polarise opinion among users (see Table 5) and that best capture variation in individuals' preferences for types of music measured by the first principal component. It is notable that Rock songs are overrepresented among the outliers (see Table 6), relative to, say, Country songs. Users who express a preference for particular songs are also more likely to enjoy songs that are nearby in the plot. Such information is therefore potentially commercially valuable, both as an efficient means of gauging users' preferences, and for providing recommendations.

The right panel of Figure 5 presents the principal scores $\{\widehat{\mathbf{u}}_i\}_{i=1}^n$ of the users, with frequent users (whose total song plays are in the top 10% of all users) in red and occasional users in blue. This plot reveals, for instance, that the second principal component is well aligned with general interest in the website. Returning to the left plot, we can now interpret a positive y-coordinate for a particular song

Table 5. Titles, artists and genres of the 22 outlier songs in Figure 5.

ID	Title	Artist	Genre
1	Your Hand In Mine	Explosions In The Sky	Rock
2	All These Things That I've Done	The Killers	Rock
3	Lady Marmalade	Christina Aguilera / Lil' Kim/ Mya / Pink	Pop
4	Here It Goes Again	Ok Go	Rock
5	I Hate Pretending (Album Version)	Secret Machines	Rock
6	No Rain	Blind Melon	Rock
7	Comatose (Comes Alive Version)	Skillet	Rock
8	Life In Technicolor	Coldplay	Rock
9	New Soul	Yael Naïm	Pop
10	Blurry	Puddle Of Mudd	Rock
11	Give It Back	Polly Paulusma	Pop
12	Walking On The Moon	The Police	Rock
13	Face Down (Album Version)	The Red Jumpsuit Apparatus	Rock
14	Savior	Rise Against	Rock
15	Swing Swing	The All-American Rejects	Rock
16	Without Me	Eminem	Rap
17	Almaz	Randy Crawford	Pop
18	Hotel California	Eagles	Rock
19	Hey There Delilah	Plain White T's	Rock
20	Revelry	Kings Of Leon	Rock
21	Undo	Björk	Rock
22	You're The One	Dwight Yoakam	Country

(which is the case for the large majority of songs) as being associated with an overall interest in the music provided by the site.

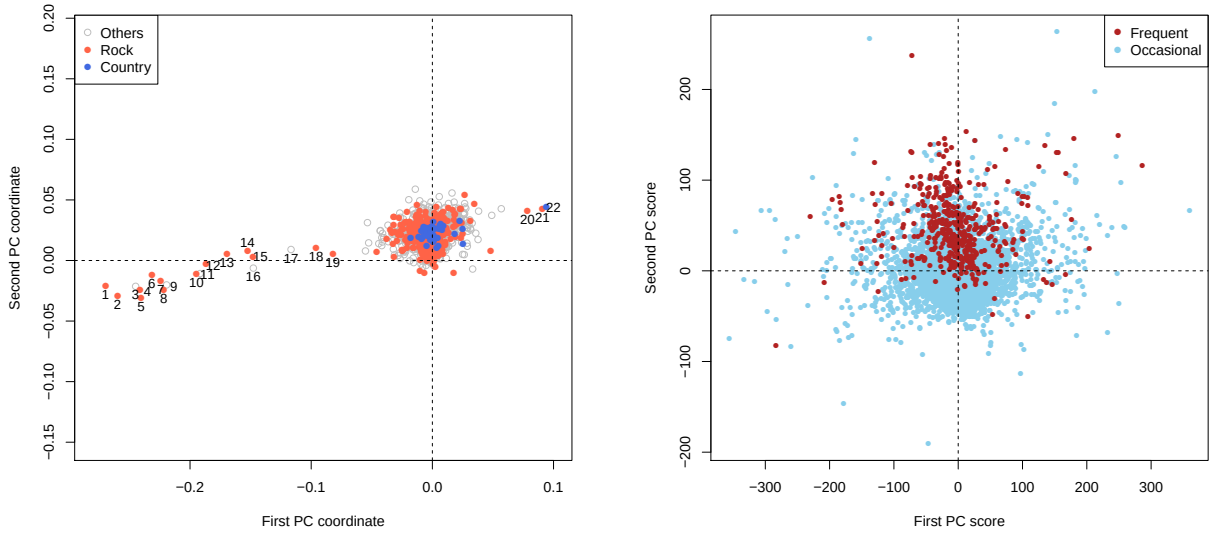

Fig. 5. Plots of the first two principal components $\widehat{\mathbf{V}}_2^{\text{prime}}$ (left) and the associated scores $\{\widehat{\mathbf{u}}_i\}_{i=1}^n$ (right).

As discussed above, it may be the case that some of the entries that we have treated as missing in fact represent a user's aversion to a particular song. We therefore studied the robustness of our conclusions by replacing some of the missing entries with an interest level of 1 (i.e. the lowest level available). More precisely, for some $\alpha \in \{0.05, 0.1, 0.2\}$, and independently for each user $i \in [n]$, we generated $R_i \sim \text{Poisson}(\alpha\|\boldsymbol{\omega}_i\|_1)$, and assigned an interest level of 1 to R_i uniformly-random chosen

Table 6. Genre distribution of the outliers (songs whose corresponding coordinate in the estimated leading principal component is of magnitude larger than 0.07).

	Rock	Pop	Electronic	Rap	Country	RnB	Latin	Others
Population (Total = 1,777)	48.92%	18.53%	9.12%	7.15%	4.33%	2.35%	2.26%	7.34%
Outliers (Total = 22)	72.73%	18.18%	0%	4.54%	4.54%	0%	0%	0%

Table 7. Robustness assessment: average inner products (over 100 repetitions) between top two eigenvectors obtained by running `primePCA` on the original data and with some of the missing entries imputed with an interest level of 1. Standard errors are given in brackets.

α	$\langle \widehat{\mathbf{v}}_1, \widehat{\mathbf{v}}_1' \rangle$	$\langle \widehat{\mathbf{v}}_1, \widehat{\mathbf{v}}_2' \rangle$	$\langle \widehat{\mathbf{v}}_2, \widehat{\mathbf{v}}_1' \rangle$	$\langle \widehat{\mathbf{v}}_2, \widehat{\mathbf{v}}_2' \rangle$
0.05	$0.816_{(0.018)}$	$-0.042_{(0.007)}$	$-0.012_{(0.007)}$	$0.910_{(0.002)}$
0.1	$0.756_{(0.018)}$	$-0.027_{(0.007)}$	$-0.070_{(0.008)}$	$0.893_{(0.002)}$
0.2	$0.546_{(0.025)}$	$-0.067_{(0.010)}$	$-0.085_{(0.010)}$	$0.859_{(0.002)}$

songs that this user had not previously heard through the site. We then ran `primePCA` on this imputed dataset, obtaining estimators $\widehat{\mathbf{v}}_1'$ and $\widehat{\mathbf{v}}_2'$ of the two leading principal components. Denoting the original `primePCA` estimators for the two columns of $\mathbf{V}_2$ by $\widehat{\mathbf{v}}_1$ and $\widehat{\mathbf{v}}_2$ respectively, Table 7 reports the average of the inner product $\langle \widehat{\mathbf{v}}_j, \widehat{\mathbf{v}}_k' \rangle$, where $j, k \in \{1, 2\}$, based on 100 independent Monte Carlo experiments. Bearing in mind that the average absolute inner product between two independent random vectors chosen uniformly on $\mathcal{S}^{1776}$ is around 0.020, this table is reassuring that the conclusions are robust to the treatment of missing entries.

6. Discussion

Heterogeneous missingness is ubiquitous in contemporary, large-scale data sets, yet we currently understand very little about how existing procedures perform or should be adapted to cope with the challenges this presents. Here we attempt to extract the lessons learned from this study of high-dimensional PCA, in order to see how related ideas may be relevant in other statistical problems where one wishes to recover low-dimensional structure with data corrupted in a heterogeneous manner.

A key insight, as gleaned from Section 2.2, is that the way in which the heterogeneity interacts with the underlying structure of interest is crucial. In the worst case, the missingness may be constructed to conceal precisely the structure one seeks to uncover, thereby rendering the problem infeasible by any method. The only hope, then, in terms of providing theoretical guarantees, is to rule out such an adversarial interaction. This was achieved via our incoherence condition in Section 3, and we look forward to seeing how the relevant interactions between structure and heterogeneity can be controlled in other statistical problems such as those mentioned in the introduction. For instance, in sparse linear regression, one would anticipate that missingness of covariates with strong signal would be much more harmful than corresponding missingness for noise variables.

Our study also contributes to the broader understanding of the uses and limitations of spectral methods for estimating hidden low-dimensional structures in high-dimensional problems. We have seen that the OPW estimator is both methodologically simple and, in the homogeneous missingness setting, achieves near-minimax optimality when the noise level is of constant order. Similar results have been obtained for spectral clustering for network community detection in stochastic block models (Rohe et al., 2011) and in low-rank-plus-sparse matrix estimation problems (Fan, Liao and Mincheva, 2013). On the other hand, the OPW estimator fails to provide exact recovery of the principal components in the noiseless setting. In these other aforementioned problems, it has also been observed that refinement of an initial spectral estimator can enhance performance, particularly in high signal-to-noise ratio regimes (Gao et al., 2016; Zhang, Cai and Wu, 2018), as we were able to show for our `primePCA` algorithm. This suggests that such a refinement has the potential to confer a sharper dependence of the statistical error rate on the signal-to-noise ratio compared with a vanilla spectral algorithm, and understanding this phenomenon in greater detail provides another interesting avenue for future

research.

Acknowledgements: The authors thank the anonymous reviewers for helpful feedback, which helped to improve the paper. Z.Z. was supported by NSF grant DMS-2015366, T.W. was supported by EPSRC grant EP/T02772X/1 and R.J.S. was supported by EPSRC grants EP/P031447/1 and EP/N031938/1, as well as ERC grant Advanced Grant 101019498.

References

Anderson, T. W. (1957) Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *J. Amer. Statist. Assoc.*, **52**, 200–203.

Beaton, A. E. (1964) The use of special matrix operators in statistical calculus. *ETS Research Bulletin Series*, **2**, i–222.

Belloni, A., Rosenbaum, M. and Tsybakov, A. B. (2017) Linear and conic programming estimators in high dimensional errors-in-variables models. *J. Roy. Statist. Soc., Ser. B*, **79**, 939–956.

Cai, T. T., Ma, Z. and Wu, Y. (2013) Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.*, **41**, 3074–3110.

Cai, T. T. and Zhang, A. (2016) Minimax rate-optimal estimation of high-dimensional covariance matrices with incomplete data. *J. Multivar. Anal.*, **150**, 55–74.

Cai, T. T. and Zhang, L. (2018b) High-dimensional linear discriminant analysis: optimality, adaptive algorithm, and missing data. `arXiv:1804.03018`.

Candès, E. J., Li, X., Ma, Y. and Wright, J. (2011) Robust principal component analysis? *J. ACM*, **58**, 11:1–11:37.

Candès, E. J. and Plan, Y. (2010) Matrix completion with noise. *Proc. IEEE*, **98**, 925–936.

Candès, E. J. and Recht, B. (2009) Exact matrix completion via convex optimization. *Found. Comput. Math.*, **9**, 717–772.

Chi Y., Lu Y. and Chen Y. (2018) Nonconvex optimization meets low-rank matrix factorization: An overview. arXiv preprint arXiv:1809.09573.

Cho, J., Kim, D. and Rohe, K. (2017) Asymptotic theory for estimating the singular vectors and values of a partially-observed low rank matrix with noise. *Statist. Sinica*, **27**, 1921–1948.

Davis, C. and Kahan, W. M. (1970) The rotation of eigenvectors by a perturbation III. *SIAM J. Numer. Anal.*, **7**, 1–46.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc., Ser. B*, **39**, 1–38.

Dray, S. and Josse J. (2015) Principal component analysis with missing values: a comparative survey of methods. *Plant Ecol.*, **216**, 657–667.

Elsener, A and van de Geer, S. (2018) Sparse spectral estimation with missing and corrupted measurements. `arXiv:1811.10443`.

Fan, J., Liao, Y. and Micheva, M. (2013) Large covariance estimation by thresholding principal orthogonal complements. *J. Roy. Statist. Soc., Ser. B*, **75**, 603–680.

Ford, B. L. (1983) An overview of hot-deck procedures. In W. G. Madow, I. Olkin and D. B. Rubin (Eds.) *Incomplete Data in Sample Surveys, Vol. 2: Theory and Bibliographies*, 185–207. Academic Press, New York.

Gao, C., Ma, Z., Zhang, A. Y. and Zhou, H. H. (2016) Achieving optimal misclassification proportion in stochastic block models. *J. Mach. Learn. Res.*, **18**, 1–45.

Hastie T., Mazumder R., Lee J. D. and Zadeh R. (2015) Matrix completion and low-rank SVD via fast alternating least squares. *J. Mach. Learn. Res.*, **16**, 3367–3402.

Johnstone, I. M. and Lu, A. Y. (2009) On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.*, **104**, 682–693.

Josse J. and Husson F. (2012) Handling missing values in exploratory multivariate data analysis methods. *J. de la Société Française de Statistique*, **153**, 1–21.

Josse J., Pagès J. and Husson F. (2009) Gestion des donneés manquantes en analyse en composantes principales. *J. de la Société Française de Statistique*, **150**, 28–51.

Keshavan, R. H., Montanari, A. and Oh, S. (2010) Matrix completion from a few entries. *IEEE Trans. Inform. Theory*, **56**, 2980–2998.

Kiers H. A. L. (1997) Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika*, **62**, 251–266.

Koltchinskii, V., Lounici, K. and Tsybakov, A. B. (2011) Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, **39**, 2302–2329.

Little, R. J. and Rubin, D. B. (2019) *Statistical Analysis with Missing Data*, John Wiley & Sons, Hoboken.

Loh, P.-L. and Tan, X. L. (2018) High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electron. J. Statist.*, **12**, 1429–1467.

Loh, P.-L. and Wainwright, M. J. (2012) High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. *Ann. Statist.*, **40**, 1637–1664.

Lounici, K. (2013) Sparse principal component analysis with missing observations. In C. Houdré et al. (Eds.) *High Dimensional Probability VI*, 327–356. Birkhäuser, Basel.

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

Mazumder, R., Hastie, T. and Tibshirani R. (2010) Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.*, **11**, 2287–2322.

Negahban, S. and Wainwright, M. J. (2012) Restricted strong convexity and weighted matrix completion: optimal bounds with noise. *J. Mach. Learn. Res.*, **13**, 1665–1697.

Paul, D. (2007) Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica.* **17**, 1617–1642.

Rohe, K., Chatterjee, S. and Yu, B. (2011) Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.*, **39**, 1878–1915.

Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.

Rubin, D. B. (2004) *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Hoboken.

Schönemann, P. (1966) A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, **31**, 1–10.

Seaman, S., Galati, J., Jackson, D. and Carlin, J. (2013) What is meant by "missing at random"?. *Statist. Sci.*, **28**, 257–268.

Shen, D., Shen, H., Zhu, H. and Marron, J. (2016) The statistics and mathematics of high dimension low sample size asymptotics. *Statist. Sinica*, **26**, 1747–1770.

Vershynin, R. (2018) High-Dimensional Probability: An Introduction with Applications in Data Science. Cambridge University Press, Cambridge.

Wainwright, M. J. (2019) *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, Cambridge.

Wang, W. and Fan, J. (2017) Asymptotics of empirical eigen-structure for ultra-high dimensional spiked covariance model. *Ann. Statist.*, **45**, 1342–1374.

Wold, H. and Lyttkens, E. (1969) Nonlinear iterative partial least squares (NIPALS) estimation procedures. *Bull. Int. Stat. Inst.*, **43**, 29–51.

Zhang, A., Cai, T. T. and Wu, Y. (2018) Heteroskedastic PCA: Algorithm, optimality, and applications. `arXiv:1810.08316`.

Zhu, Z., Wang, T. and Samworth, R. J. (2019) `primePCA`: projected refinement for imputation of missing entries in principal component analysis. `R` package, version 1.2. https://cran.r-project.org/web/packages/primePCA/.