Empirical Bayes PCA in high dimensions

Xinyi Zhong | Chang Su | Zhou Fan

1Department of Statistics and Data Science, Yale University, New Haven, USA
2Department of Biostatistics, Yale University, New Haven, USA

Correspondence
Zhou Fan, Yale University, Department of Statistics and Data Science, New Haven, USA.
Email: zhou.fan@yale.edu

Funding information
NSF, Grant/Award Number: DMS-1916198

Abstract

When the dimension of data is comparable to or larger than the number of data samples, principal components analysis (PCA) may exhibit problematic high-dimensional noise. In this work, we propose an empirical Bayes PCA method that reduces this noise by estimating a joint prior distribution for the principal components. EB-PCA is based on the classical Kiefer–Wolfowitz non-parametric maximum likelihood estimator for empirical Bayes estimation, distributional results derived from random matrix theory for the sample PCs and iterative refinement using an approximate message passing (AMP) algorithm. In theoretical ‘spiked’ models, EB-PCA achieves Bayes-optimal estimation accuracy in the same settings as an oracle Bayes AMP procedure that knows the true priors. Empirically, EB-PCA significantly improves over PCA when there is strong prior structure, both in simulation and on quantitative benchmarks constructed from the 1000 Genomes Project and the International HapMap Project. An illustration is presented for analysis of gene expression data obtained by single-cell RNA-seq.

KEYWORDS
Principal components analysis, empirical Bayes, random matrix theory, AMP algorithms

1 | INTRODUCTION

Principal components analysis (PCA) is a widely used technique for dimensionality reduction. However, when the dimension of the data may be comparable to or larger than the number of...
available data samples, it is known that the sample principal components (PCs) may exhibit phenomena of high-dimensional noise (Johnstone & Lu, 2009; Lu, 2002). We propose a method called EB-PCA for reducing this noise, using the classical statistical idea of empirical Bayes (Efron, 2012; Robbins, 1956).

Figure 1 illustrates EB-PCA on a genetics example. Panel (a) displays the top four PCs of a genotype matrix from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015), containing genotypes of 2504 individuals at 100,000 common single nucleotide polymorphisms (SNPs). The PCs depict the stratification of these individuals according to five broad ethnic populations. Here, the number of SNPs far exceeds the dimension 2504 of each PC, and the estimation noise is small. This allows us to interpret the PCs in panel (a) as an approximate ‘ground truth’.

The phenomenon of high-dimensional noise is illustrated in panel (b), which displays the top four PCs for genotypes of the same 2504 individuals subsampled at only 1000 randomly selected SNPs. Applying EB-PCA to this reduced data of 1000 SNPs yields the PC estimates displayed in panel (c). These are remarkably close to the PCs in panel (a) computed on all 100,000 SNPs, even though EB-PCA has only access to the 1000 subsampled SNPs. In Section 4, we use this subsampling approach to demonstrate a sizeable quantitative improvement of EB-PCA over PCA. We also illustrate an application to single-cell RNA-seq gene expression data where a ground truth is unknown.

A central component of the method is a Bayes approximate message passing (AMP) procedure (Montanari & Venkataramanan, 2021; Rangan & Fletcher, 2012) that implements approximate Bayesian inference for low-rank matrix estimation in high dimensions. EB-PCA adapts Bayes AMP, which requires knowledge of the true prior distributions, to more typical settings in

![Figure 1](https://example.com/figure1.png)

**FIGURE 1** Illustration of Empirical Bayes principal components analysis (EB-PCA) on genotype data from the 1000 Genomes Project. (a) 1st versus 2nd PC and 3rd versus 4th PC, for the genotypes of 2504 individuals across 100,000 common single nucleotide polymorphisms (SNPs). Each scatterplot has 2504 data points, representing the embedding of these individuals into a 4-dimensional space, with points colored by the individuals’ ethnicity. We take these PCs as the ground truth. (b) PCs computed from a random subsample of 1000 SNPs. Substantial high-dimensional noise is observed in these PCs. (c) The EB-PCA estimates of the top four PCs, computed from the same subsampled data as in panel (b). These estimates are much closer to the ground-truth PCs in panel (a) and have quantitatively lower estimation error [Colour figure can be viewed at wileyonlinelibrary.com]
practice where such information is unavailable, by non-parametrically estimating the priors from the sample PCs and the AMP iterates. Similar strategies can be applied to Bayes AMP algorithms for other applications.

To describe the main ideas behind EB-PCA, consider a rank-one signal-plus-noise model for the observed data,

\[ Y = \frac{S}{n} \cdot uv^T + W \in \mathbb{R}^{nxd} \]  

(1)

where \( u \in \mathbb{R}^n \) and \( v \in \mathbb{R}^d \) are the left and right true PCs of interest, with associated signal strength \( s > 0 \), and \( W \in \mathbb{R}^{nxd} \) is i.i.d. Gaussian observational noise. We discuss possible extensions to more general noise in Section 6. We will refer to the leading left- and right-singular vectors \( f \in \mathbb{R}^n \) and \( g \in \mathbb{R}^d \) of \( Y \) as the sample PCs.

The EB-PCA approach consists of three main ideas, each of which is individually well studied:

(a) **Kiefer–Wolfowitz NPMLE.** Consider the classical compound decision problem of estimating \( \theta \in \mathbb{R}^n \) from a Gaussian observation vector \( x \sim \mathcal{N}(\mu \cdot \theta, \sigma^2 \cdot \text{Id}_{nxn}) \), for two known scalar parameters \( \mu, \sigma^2 > 0 \). The empirical Bayes paradigm first posits a prior distribution \( \pi_* \) for the coordinates of \( \theta \), then estimates \( \pi_* \) by an estimator \( \pi \) based on the marginal density of the observed coordinates of \( x \), and finally applies Bayes’s rule defined by \( \pi \) to ‘denoise’ \( x \) and obtain the estimate of \( \theta \).

A non-parametric implementation of this paradigm was described in Robbins (1950); Kiefer and Wolfowitz (1956), which suggested estimating \( \pi_* \) by the non-parametric maximum likelihood estimator (NPMLE) that maximizes the likelihood of \( x \) over all prior probability distributions \( \pi \) on the real line. It was shown in Kiefer and Wolfowitz (1956); Laird (1978) and Lindsay (1983a) that such a maximizer \( \pi \) exists with discrete and finite support. We denote by

\[ \theta(x | \mu, \sigma^2, \pi) = \mathbb{E}_\pi[\theta | x] \]

the empirical Bayes posterior mean estimate of \( \theta \) using this estimated prior \( \pi \).

(b) **Random matrix asymptotics for sample PCs.** In the model of (1), an influential line of work (Baik et al., 2005; Benaych-Georges & Nadakuditi, 2012; Nadler, 2008; Paul, 2007) has quantified the asymptotic error of the sample PCs \( (f, g) \) for the true PCs \( (u, v) \) when \( n, d \to \infty \) simultaneously such that \( d/n \to \gamma \in (0, \infty) \). This work showed that in this high-dimensional limit,

\[ \langle f, u \rangle \to \mu_* \equiv \mu_*(s, \gamma), \quad \langle g, v \rangle \to \mu_* \equiv \mu_*(s, \gamma) \]

for two inner products \( \mu_*, \mu_* \in [0, 1) \) that depend only on the signal strength \( s \) and the dimension ratio \( \gamma \). For \( s \) larger than a certain phase transition threshold \( s_*(\gamma) \), the leading singular value of \( Y \) emerges as an outlier from the bulk distribution of its remaining singular values, the inner products \( \mu_*, \mu_* \) are strictly positive, and \( g \) has an approximate entrywise Gaussian law

\[ g \approx \mathcal{N}(\mu_* \cdot v, \sigma_*^2 \cdot \text{Id}_{dxd}), \quad \sigma_*^2 = 1 - \mu_*^2. \]

An analogous approximation holds for \( f \) and \( u \). This provides a connection to the compound decision problem above. EB-PCA estimates \( (\mu_*, \sigma_*^2) \) by estimating \( s \), and applies the Kiefer–Wolfowitz NPMLE to obtain an empirical Bayes estimate \( \hat{v} \) for \( v \).
Iterative refinement via AMP. If this estimate $\hat{v}$ is more accurate than the original sample PC $g$ for $v$, then we expect $Y\hat{v}$ to be more accurate than $Yg \propto f$ for $u$. This suggests that empirical Bayes denoising should be applied to $Y\hat{v}$ instead of $f$ to estimate $u$, and leads to an iterative idea (Wang & Stephens, 2021) of initializing $g^0 = g$ and computing

$$
\begin{align*}
\nu_t &= \theta(g_t | \mu_t, \sigma^2_t, \pi_t), \\
\nu_t &= \theta(f_t | \mu_t, \sigma^2_t, \pi_t), \\
g^{t+1} &= Y^\top u_t.
\end{align*}
$$

(2)

Here, $\pi_t, \pi_t$ are non-parametrically estimated priors and $\mu_t, \sigma^2_t, \mu_t, \sigma^2_t$ are scalar parameters in each iteration. In the first iteration, $\nu^0 = \hat{v}$ is the above empirical Bayes estimate of $v$.

Unfortunately, this procedure does not ensure that $(f_t, g_t)$ have approximate entrywise Gaussian laws after this first iteration, breaking the connection to the compound decision problem in subsequent iterations. EB-PCA applies instead an AMP algorithm as developed in Rangan and Fletcher (2012); Montanari and Venkataramanan (2021),

$$
\begin{align*}
\nu_t &= \theta(g_t | \mu_t, \sigma^2_t, \pi_t), \\
\nu_t &= \theta(f_t | \mu_t, \sigma^2_t, \pi_t), \\
g^{t+1} &= Y^\top u_t - b_t v_t.
\end{align*}
$$

The Onsager corrections $b_t u_t^{-1}$ and $b_t v_t$ are defined so as to remove a bias of $(f_t, g_t)$ in the directions of $(u_t^{-1}, v_t)$ and restore the entrywise Gaussian approximations.

EB-PCA is most effective when there is strong prior structure for the true PCs. We described the rank-one model of (1) for clarity, but in many examples including Figure 1, there is stronger structure jointly over several PCs. In these examples, we learn a joint prior in $k > 1$ dimensions, where $k$ is the number of PCs to be simultaneously estimated. The result of Figure 1c is obtained by joint empirical Bayes estimation for all $k = 4$ depicted PCs, rather than estimating each PC individually. We describe the method in more detail in Section 2.4 and present theoretical guarantees in Section 5.

This application of empirical Bayes methodology to PCA via an iterative algorithm is closely related to earlier and inspirational work by Wang and Stephens (2021), who proposed an empirical Bayes matrix factorization (EBMF) method that yields the iterations of Equation (2). EBMF is derived from a ‘naive mean-field’ variational approximation to the posterior distribution of $(u, v)$, and we discuss further in Section 2.5 the relation between EB-PCA and this naive mean-field approach.

1.1 Related literature

The possible inconsistency of PCA in high dimensions has been discussed in Lu (2002); Johnstone and Lu (2009); Johnstone and Paul (2018), and improving PCA using prior structure has been a long-standing goal. A large body of literature has notably studied sparse PCA methods, which improve over PCA under sparsity assumptions (d’Aspremont et al., 2005; Amini & Wainwright, 2008; Birnbaum et al., 2013; Cadima & Jolliffe, 1995; Cai et al., 2013; Fan et al., 2013; Jolliffe et al., 2003; Ma, 2013; Vu et al., 2013; Zou et al., 2006). Figure 1 illustrates an example where the PCs indeed have strong prior structure, but this structure is not well-characterized by entrywise sparsity. We believe that such examples may be common across scientific applications, and this forms the primary motivation for our work.
EB-PCA is complementary to spectral shrinkage methods that preserve the sample PCs but shrink or truncate the singular values (Cai et al., 2010; Chatterjee, 2015; Gavish & Donoho, 2017; Ledoit & Wolf, 2012; Nadakuditi, 2014; Shabalin & Nobel, 2013). These methods have been motivated in part by a perspective that, in the absence of prior structural knowledge about the PCs, ‘...it is reasonable to require that covariance matrix estimators be rotation-equivariant [and have] the same eigenvectors as the sample covariance matrix’ (Ledoit & Wolf, 2012). Our work stands contrary to this perspective, illustrating that empirical Bayes ideas can substantially improve over such equivariant procedures even without knowledge of prior structure, as long as some structure is present.

EB-PCA is an empirical Bayes implementation of the multivariate Bayes AMP algorithm described by Montanari and Venkataramanan (2021). AMP algorithms were first developed for CDMA, compressed sensing and generalized linear model applications by Kabashima (2003); Donoho et al. (2009) and Rangan (2011). Empirical Bayes versions of AMP for compressed sensing and GLMs were studied by Vila and Schniter (2011); Krzakala et al. (2012); Vila and Schniter (2013) and Kamilov et al. (2014), in univariate and parametric contexts that are different from the non-parametric perspective of our work.

AMP algorithms for PCA have been studied in a line of work including Rangan and Fletcher (2012); Matsushita and Tanaka (2013); Deshpande and Montanari (2014); Montanari and Richard (2015); Lesieur et al. (2015b); Kabashima et al. (2016) and Deshpande et al. (2017). These algorithms originally required an informative initialization independent of \( Y \), and Montanari and Venkataramanan (2021) provided the practical extension of initializing at the sample PCs. A related line of work (Alaoui & Krzakala, 2018; Barbier & Macris, 2019; Barbier et al., 2016; Lelarge &MIolane, 2019; Lesieur et al., 2015a and MIolane, 2017) has explored more generally the limits of low-rank matrix estimation with Bayesian priors. In particular, Deshpande and Montanari (2014); Barbier et al. (2016) and Deshpande et al. (2017) showed in various Bayesian rank-one models that AMP algorithms can achieve the asymptotically optimal squared-error Bayes risk, which has been characterized in Lesieur et al. (2015a); Barbier et al. (2016); MIolane (2017) and Lelarge and MIolane (2019). A second motivation for our work is to bring this important body of statistical theory a step closer to statistical practice. Our results imply that AMP algorithms can achieve Bayes-optimal estimation even without knowledge of the true priors.

The initial step of EB-PCA relies on quantitative understanding of spectral behaviour in spiked random matrix models (Bai & Yao, 2008; Bai & Silverstein, 2006; Bai et al., 2005 and Johnstone, 2001). We assume in this work a Gaussian model, where the error of the sample singular vectors was first studied in Paul (2007) and Nadler (2008). Such results have been extended to non-Gaussian settings in Capitaine et al. (2011); Knowles and Yin (2013, 2014); Bloemendal et al. (2016); Capitaine (2018) and Ding (2020), models with non-white noise in Mestre (2008); Benaych-Georges and Nadakuditi (2011, 2012) and Bai and Yao (2012), and more general asymptotic regimes in Jung and Marron (2009); Shen et al. (2013) and Wang and Fan (2017). Related distributional properties of singular vectors were recently studied in Capitaine and Donati-Martin (2018) and Bao et al. (2020, 2021).

The Kiefer–Wolfowitz NPMLE was proposed in Robbins (1950) and Kiefer and Wolfowitz (1956). Identifiability, existence and uniqueness, asymptotic consistency and discreteness of the support were studied in Kiefer and Wolfowitz (1956); Simar (1976); Laird (1978); Jewell (1982); Lindsay (1983a, b) and Lindsay and Roeder (1993), and a detailed treatment of these topics is provided in Lindsay (1995). Ghosal and van der Vaart (2001); Zhang (2009); Jiang and Zhang (2009) and Saha and Gun t suboyina (2020) studied the rate of convergence of the NPMLE and associated empirical Bayes estimator, and our analyses draw on their techniques. Recently, Polyanskiy and...
Wu (2020) showed that these results on estimation rates are connected also to the size of the discrete NPMLE support. Computing and approximating the NPMLE has been discussed in Bohning et al. (1992); Böhning (1999); Lashkari and Golland (2008); Koenker and Mizera (2014) and Feng and Dicker (2016).

2 | THE EB-PCA METHOD

2.1 | Model

The EB-PCA algorithm is derived in the following rank-$k$ version of the model in (1),

\[
Y = \frac{1}{n} \cdot USV^T + W = \sum_{i=1}^{k} \frac{S_i}{n} \cdot u_i v_i^T + W \in \mathbb{R}^{n \times d}.
\] (3)

The columns of \(U = (u_1, \ldots, u_k) \in \mathbb{R}^{n \times k}\) and \(V = (v_1, \ldots, v_k) \in \mathbb{R}^{d \times k}\) are $k$ left and right principal components of interest, and \(S = \text{diag}(s_1, \ldots, s_k) \in \mathbb{R}^{k \times k}\) contains the signal strengths of these PCs. \(W \in \mathbb{R}^{n \times d}\) is observational noise, which we assume has entries \(w_{ij} \sim \mathcal{N}(0, 1/n)\).

Remark 1 We write the noise variance of \(w_{ij}\) for convenience as \(1/n\), rather than a more general \(\tau^2/n\), to avoid carrying \(\tau^2\) throughout our formulas. This is without loss of generality, as \(Y, S, W\) may be rescaled by a common factor \(\tau\). In practice, we estimate this noise variance and rescale the data to match this scaling. Given \(Y_{\text{obs}} \in \mathbb{R}^{n \times d}\), we may estimate its entrywise residual variance upon regressing out its top \(k\) PCs,

\[
\hat{\tau}^2 = \frac{d-1}{n} \|R\|_F^2, \quad R = Y_{\text{obs}} - \text{top } k \text{ PCs of } Y_{\text{obs}}.
\] (4)

We then set \(Y = Y_{\text{obs}}/\hat{\tau}\). Consistency of \(\hat{\tau}^2\) is discussed in Appendix A.2.

We study this model in the high-dimensional limit \(n, d \to \infty\) such that \(k\) and \(\gamma \equiv d/n\) are both fixed. It is helpful to keep in mind a Bayesian setting where the rows of \(U\) and \(V\) are generated according to two fixed prior probability distributions \(\pi_*\) and \(\pi_*\) on \(\mathbb{R}^k\) (although we will only require empirical convergence to these priors in the later theory). The goal of EB-PCA is then to estimate these two priors from the data \(Y\), and to use these estimated priors to perform Bayesian estimation of \(U\) and \(V\).

To fix the scaling of the PCs, we normalize \(\pi_*\) and \(\pi_*\) to satisfy \(\mathbb{E}_{U \sim \pi_*}[U_i^2] = 1\) and \(\mathbb{E}_{V \sim \pi_*}[V_i^2] = 1\) for all \(i = 1, \ldots, k\). This ensures

\[
n^{-1}\|u_i\|_2^2 \to 1, \quad d^{-1}\|v_i\|_2^2 \to 1.
\] (5)

We will also assume \(\mathbb{E}_{U \sim \pi_*}[U_iU_j] = 0\) and \(\mathbb{E}_{V \sim \pi_*}[V_iV_j] = 0\) for all \(1 \leq i \neq j \leq k\), so that

\[
n^{-1}u_i^T u_j \to 0, \quad d^{-1}v_i^T v_j \to 0.
\] (6)

lending to the interpretations of \(u_i\) and \(v_i\) as the (orthogonal) principal components. Under these scalings, the \(k\) singular values of \(n^{-1}USV^T\) converge to the limits \(\sqrt{\gamma s_1} > \cdots > \sqrt{\gamma s_k}\) and we make the simplifying assumption that these limit values are distinct. Note that we will not enforce the orthogonality conditions \(\mathbb{E}_{U \sim \pi_*}[U_iU_j] = 0\) and \(\mathbb{E}_{V \sim \pi_*}[V_iV_j] = 0\) for the estimated priors in the later
algorithm, but approximate orthogonality will automatically hold from initializing the algorithm at the sample PCs.

Turning to the sample PCs, let us write the best rank-*k* approximation for *Y* as

$$
\frac{1}{n} \cdot F \Lambda G^T = \sum_{i=1}^{k} \frac{\lambda_i}{n} \cdot f_i g_i^T.
$$

Here, the columns of $F = (f_1, \ldots, f_k) \in \mathbb{R}^{n \times k}$ and $G = (g_1, \ldots, g_k) \in \mathbb{R}^{d \times k}$ are the top *k* left and right singular vectors of *Y*, normalized analogously with a sign convention so that for all $1 \leq i \neq j \leq k$,

$$
d^{-1}||g_i||^2 = n^{-1}||f_i||^2 = 1, \quad u_i^T f_i \geq 0, \quad v_i^T g_i \geq 0, \quad d^{-1} g_i^T g_j = n^{-1} f_i^T f_j = 0. \quad (7)
$$

We set $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_k)$. Then the *k* largest singular values of *Y* are given by $\sqrt{\gamma} \lambda_1 \geq \cdots \geq \sqrt{\gamma} \lambda_k$.

Under this model, the following phase transition occurs for the leading *k* sample singular values and singular vectors of *Y* (Baik et al., 2005; Benaych-Georges & Nadakuditi, 2012; Paul, 2007): setting $s_*(\gamma) = \gamma^{-1/4}$, for supercritical PCs such that $s_i > s_*(\gamma)$, we have

$$
\lim_{n,d \to \infty} \sqrt{\gamma} \cdot \lambda_i > \lambda_+, \quad \lim_{n,d \to \infty} n^{-1} f_i^T u_i > 0, \quad \lim_{n,d \to \infty} d^{-1} g_i^T v_i > 0
$$

where $\lambda_+ = 1 + \sqrt{\gamma}$ is the upper edge of the ‘bulk distribution’ of the noise singular values. Conversely, for subcritical PCs such that $s_i \leq s_*(\gamma)$,

$$
\lim_{n,d \to \infty} \sqrt{\gamma} \cdot \lambda_i = \lambda_+, \quad \lim_{n,d \to \infty} n^{-1} f_i^T u_i = 0, \quad \lim_{n,d \to \infty} d^{-1} g_i^T v_i = 0.
$$

Thus the *i*th sample singular value is absorbed into the bulk, and the sample PCs are nearly orthogonal to the true PCs. For notational and expositional clarity, we will assume

$$
s_i > s_*(\gamma) \quad \text{for all} \quad i = 1, \ldots, k,
$$

i.e. all *k* of the leading PCs are supercritical. Our theoretical results may be extended to more general settings having both supercritical and subcritical PCs, where EB-PCA is applied only to the supercritical PCs that have positive alignment with the truth.

**Remark 2** If the rows of *U* are drawn from $\pi_* = \mathcal{N}(0, \text{Id}_{k \times k})$, then the rows of $\sqrt{n} \cdot Y$ marginalized over *U* are distributed as $\mathcal{N}(0, \Sigma)$ where

$$
\Sigma = \sum_{i=1}^{k} s_i^2 \cdot \frac{v_i v_i^T}{n} + \text{Id}_{d \times d}.
$$

Thus $Y^T Y$ follows the spiked covariance model introduced in Johnstone (2001), and our results pertain also to estimating the spike eigenvectors of *Σ*. In this model, it would be reasonable to consider a version of EB-PCA that fixes $\pi_* = \mathcal{N}(0, \text{Id})$, only estimates $\pi_*$, and performs Bayesian estimation of *V* but not of *U*. We will focus instead on the more general scenario where both *U* and *V* may have non-Gaussian structure, and describe EB-PCA for estimating both matrices.
2.2 | Empirical Bayes for the multivariate compound decision problem

Let $\pi_*$ be a probability distribution on $\mathbb{R}^k$. For two given matrices $M, \Sigma \in \mathbb{R}^{k \times k}$ where $\Sigma$ is symmetric positive-definite, consider the compound decision model

$$\Theta \sim \pi_*, \quad X | \Theta \sim \mathcal{N}(M \cdot \Theta, \Sigma)$$

(8)

for $\Theta, X \in \mathbb{R}^k$. We will denote the Bayes posterior mean estimate of $\Theta$ based on $X$ as

$$\theta(X | M, \Sigma, \pi_*) = \mathbb{E}_{\pi_*}[\Theta | X].$$

(9)

Suppose now that $\pi_*$ is unknown, but belongs to a known class of probability distributions $\mathcal{P}$ over $\mathbb{R}^k$. In a model of $n$ i.i.d. samples $x_1, \ldots, x_n$ distributed according to Equation (8), stacked as the rows of a matrix $X \in \mathbb{R}^{n \times k}$, consider the maximum likelihood estimator

$$\pi = \text{MLE}(X | M, \Sigma, \mathcal{P})$$

$$\equiv \arg \max_{\pi \in \mathcal{P}} \prod_{i=1}^{n} \int \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \cdot \exp \left( -\frac{(x_i - M \cdot \theta_i)^\top \Sigma^{-1} (x_i - M \cdot \theta_i)}{2} \right) d\pi(\theta_i).$$

(10)

This integral is the marginal Gaussian mixture density of $x_i$ in the model of Equation (8), and the notation makes explicit the dependence of $\pi$ on the prior class $\mathcal{P}$. We will be interested primarily in non-parametric classes $\mathcal{P}$, and $\pi$ is a non-parametric maximum likelihood estimate (NPMLE) for $\pi_*$. In our implementation, we take $\mathcal{P}$ as the class of all probability distributions on $\mathbb{R}^k$, and approximate this class $\mathcal{P}$ using a discrete support by applying the ‘exemplar method’ of Lashkari and Golland (2008). See Appendix E for details.

Stacking $\theta_1, \ldots, \theta_n$ as the rows of $\Theta \in \mathbb{R}^{n \times k}$, the model for $X$ may be written as

$$X = \Theta M^\top + Z \Sigma^{1/2}, \quad Z \in \mathbb{R}^{n \times k} \text{ has i.i.d. } \mathcal{N}(0, 1) \text{ entries.}$$

(11)

The NPMLE $\pi$ defines an empirical Bayes estimate of $\Theta$, which applies the posterior mean function $\theta(\cdot)$ for the estimated prior $\pi$ row-wise to $X$. We denote this by

$$\theta(X | M, \Sigma, \pi) = \mathbb{E}_{\pi}[\Theta | X].$$

2.3 | Initial denoising of the sample PCs

In the model of (3), as $n, d \to \infty$, the precise forms of the limits of the supercritical singular values $\sqrt{\gamma} \cdot \lambda_i$ and corresponding PCs $f_i, g_i$ of $Y$ are given by

$$\sqrt{\gamma} \cdot \lambda_i \to \sqrt{(\gamma s_i^2 + 1)(s_i^2 + 1)/s_i^2},$$

$$n^{-1} f_i^\top u_i \to \mu_{s,i} \equiv \sqrt{1 - \sigma_{s,i}^2}, \quad d^{-1} g_i^\top v_i \to \mu_{s,i} \equiv \sqrt{1 - \sigma_{s,i}^2},$$

$$\sigma_{s,i}^2 \equiv \frac{1 + s_i^2}{s_i^2(\gamma s_i^2 + 1)}, \quad \sigma_{s,i}^2 \equiv \frac{1 + \gamma s_i^2}{\gamma s_i^2(s_i^2 + 1)}.$$
Iterative refinement using AMP

Let $\mu_s, \ldots, \mu_{s,k}$ be the true matrices $M_s, \ldots, M_{s,k}$ for $s = 1, \ldots, k$. Initialize $G_j$ of Montanari and Venkataramanan (2021). This may begin with either the estimate for $\Sigma_s$, which may be consistently estimated by $\hat{\Sigma}_s = (\gamma \lambda_i^2 - (1 + \gamma) + \sqrt{(\gamma \lambda_i^2 - (1 + \gamma))^2 - 4\gamma})/(2\gamma)$. (See Lemma A.3.)

\begin{align}
\overline{M}_s &= \text{diag}(\overline{\mu}_{s,1}, \ldots, \overline{\mu}_{s,k}), \quad M_s = \text{diag}(\mu_{s,1}, \ldots, \mu_{s,k}), \\
\overline{\Sigma}_s &= \text{diag}(\overline{\sigma}_{s,1}^2, \ldots, \overline{\sigma}_{s,k}^2), \quad \Sigma_s = \text{diag}(\sigma_{s,1}^2, \ldots, \sigma_{s,k}^2),
\end{align}

a consequence is that $F \in \mathbb{R}^{n \times k}$ and $G \in \mathbb{R}^{d \times k}$ have the Gaussian approximations

\begin{align}
F \approx U M_s^\top + \overline{Z} \Sigma_s^{1/2}, \quad G \approx V M_s^\top + Z \Sigma_s^{1/2}
\end{align}

for large $n$ and $d$, where $\overline{Z}, Z \in \mathbb{R}^{n \times k}$ have i.i.d. $\mathcal{N}(0, 1)$ entries. This relates the behaviour of the sample PCs $F$ and $G$ to the multivariate compound decision model in (11).

As the true matrices $\overline{M}_s, M_s, \overline{\Sigma}_s, \Sigma_s$ are unknown, we replace them by consistent estimates to derive empirical Bayes estimators for $U$ and $V$: observe that Equation (12) implies each value $\lambda_i^2$ may be consistently estimated by

\begin{align}
\hat{\lambda}_i^2 = (\gamma \lambda_i^2 - (1 + \gamma) + \sqrt{(\gamma \lambda_i^2 - (1 + \gamma))^2 - 4\gamma})/(2\gamma).
\end{align}

These may be used to obtain plug-in estimators $\overline{M}_s, M_s, \overline{\Sigma}_s, \Sigma_s$ for $\overline{M}_s, M_s, \overline{\Sigma}_s, \Sigma_s$, which substitute $\hat{s}_i$ for $s_i$ in Equation (12). This yields the initial empirical Bayes estimates of $U$ and $V$ given by

\begin{align}
\bar{\pi} &= \text{MLE}(F | \overline{M}, \overline{\Sigma}, \mathcal{P}), \quad \hat{U} = \theta(F | \overline{M}, \overline{\Sigma}, \overline{\pi}), \\
\pi &= \text{MLE}(G | M, \Sigma, \mathcal{P}), \quad \hat{V} = \theta(G | M, \Sigma, \pi).
\end{align}

2.4 Iterative refinement using AMP

We now describe iterative refinement using an AMP algorithm, as discussed also in Appendix J of Montanari and Venkataramanan (2021). This may begin with either the estimate for $U$ or $V$—here, we begin with $V$.

The algorithm takes the following form: let $u_1, u_2, \ldots : \mathbb{R}^k \to \mathbb{R}^k$ and $v_1, v_2, \ldots : \mathbb{R}^k \to \mathbb{R}^k$ be two arbitrary sequences of Lipschitz functions. Initialize $G^0 = G$ as the right sample PCs, and compute for $t = 0, 1, 2, \ldots$

\begin{align}
V^t &= v_t(G^t), \quad F^t = YV^t - U^{t-1} \cdot \gamma \langle d v_t(G^t) \rangle^\top, \\
U^t &= u_t(F^t), \quad G^{t+1} = Y^\top U^t - V^t \cdot \langle d u_t(F^t) \rangle^\top.
\end{align}

Here $u_t(F^t) \in \mathbb{R}^{n \times k}$ and $v_t(G^t) \in \mathbb{R}^{d \times k}$ denote the applications of $u_t$ and $v_t$ row-wise to $F^t$ and $G^t$, $du_t : \mathbb{R}^k \to \mathbb{R}^{k \times k}$ and $dv_t : \mathbb{R}^k \to \mathbb{R}^{k \times k}$ denote Jacobian matrices of these functions, and $\langle du_t(F^t) \rangle$ and $\langle dv_t(G^t) \rangle$ are the averages of $du_t$ and $dv_t$ across the rows of $F^t$ and $G^t$.

Under the model of (3), Gaussian approximations analogous to Equation (15) continue to hold for $F^t$ and $G^t$ across iterations, where

\begin{align}
F^t \approx U M_{s,t}^\top + \overline{Z} \Sigma_{s,t}^{1/2}, \quad G^t \approx V M_{s,t}^\top + Z \Sigma_{s,t}^{1/2}.
\end{align}
Here $\overline{M}_{s,t}, M_{s,t}, \overline{\Sigma}_{s,t}, \Sigma_{s,t}$ are deterministic matrices that prescribe the parameters of the compound decision model associated to each iteration. In contrast to the initial state of Equations (13)--(14), these matrices are no longer diagonal in later iterations, if the prior is a general multivariate distribution on $\mathbb{R}^k$. They evolve over iterations according to a state evolution

$$(M_{s,0}, \Sigma_{s,0}) \leftrightarrow (\overline{M}_{s,0}, \overline{\Sigma}_{s,0}) \leftrightarrow (M_{s,1}, \Sigma_{s,1}) \leftrightarrow (\overline{M}_{s,1}, \overline{\Sigma}_{s,1}) \leftrightarrow \cdots$$

given by the initializations $(M_{s,0}, \Sigma_{s,0}) \equiv (M_s, \Sigma_s)$ describing the sample PCs $G^0 \equiv G$ in Equations (13)--(14), and by the updates

$$\overline{M}_{s,t} = \gamma \mathbb{E}[v_t(G_t) V_t^\top] S, \quad \overline{\Sigma}_{s,t} = \gamma \mathbb{E}[v_t(G_t)V_t(G_t)^\top],$$

$$M_{s,t+1} = \mathbb{E}[u_t(F_t) U_t^\top] S, \quad \Sigma_{s,t+1} = \mathbb{E}[u_t(F_t)u_t(F_t)^\top].$$

(21)

$S = \text{diag}(s_1, \ldots, s_k)$ is the diagonal matrix of signal strengths in Equation (3), and the expectations are over the random vectors

$$U \sim \overline{\pi}_s, \quad F_t \sim \mathcal{N}(\overline{M}_{s,t} \cdot U, \overline{\Sigma}_{s,t}), \quad V \sim \pi_s, \quad G_t \sim \mathcal{N}(M_{s,t} \cdot V, \Sigma_{s,t}).$$

These laws of $F_t$ and $G_t$ approximate the row-wise distributions of $F^t$ and $G^t$.

If $\overline{\pi}_s, \pi_s, \overline{M}_{s,t}, M_{s,t}, \overline{\Sigma}_{s,t}, \Sigma_{s,t}$ are all known, then applying this algorithm with the Bayes posterior mean functions

$$u_t(X) = \theta(X | \overline{M}_{s,t}, \overline{\Sigma}_{s,t}, \overline{\pi}_s), \quad v_t(X) = \theta(X | M_{s,t}, \Sigma_{s,t}, \pi_s)$$

(22)

implements an iterative variational Bayesian inference scheme (Montanari & Venkataramanan, 2021). We will call this the ‘oracle’ Bayes AMP algorithm. For these $u_t, v_t$, we have the identities $\mathbb{E}[u_t(F_t)U^\top] = \mathbb{E}[u_t(F_t)u_t(F_t)^\top]$ and $\mathbb{E}[v_t(G_t)V^\top] = \mathbb{E}[v_t(G_t)v_t(G_t)^\top]$, so Equation (21) yields

$$\overline{M}_{s,t} = \overline{\Sigma}_{s,t} \cdot S, \quad \overline{M}_{s,t+1} = \Sigma_{s,t+1} \cdot S.$$

(23)

EB-PCA uses the posterior mean functions defined instead by NPMLEs of $\overline{\pi}_s$ and $\pi_s$, together with the empirical estimates $\overline{\Sigma} = n^{-1}(V^\top) V$, $\overline{M} = \overline{S} \hat{S}$, $\Sigma_{t+1} = n^{-1}(U^\top) U$, and $M_{t+1} = \Sigma_{t+1} \hat{S}$ where $\hat{S} = \text{diag}(\delta_1, \ldots, \delta_k)$ is the estimate of $S$ from Equation (16). (For $\overline{\Sigma}_t$, we have applied $\gamma/d = 1/n$.) These empirical estimates avoid the need to perform Gaussian integrations to analytically evaluate the expectations that define the true matrices $\overline{M}_{s,t}, M_{s,t}, \overline{\Sigma}_{s,t}, \Sigma_{s,t}$.

EB-PCA is initialized at the right sample PCs $G^0 \equiv G$ and the plug-in estimates $(M_0, \Sigma_0) \equiv (M, \Sigma)$ from the preceding section. In particular, $V^0$ is the initial empirical Bayes estimate for $V$ based on $G$ as previously described.

We summarize the full EB-PCA method as Algorithm 1.

**Remark 3** In Lines 6 and 10 of Algorithm 1, we form new NPMLEs for $\overline{\pi}_s$ and $\pi_s$ in each iteration. This allows for the possibility of improving these estimates as the signal-to-noise ratios reflected by the state parameters $(M_t, \Sigma_t)$ and $(\overline{M}_t, \overline{\Sigma}_t)$ improve across iterations. The theory does not require this re-estimation of $\overline{\pi}_s$ and $\pi_s$, as it will show that both $\overline{\pi}_s$ and $\pi_s$ may be estimated consistently in the first iteration. In practice, for data examples with strong signal, this re-estimation of $\overline{\pi}_s$ and $\pi_s$ may also be removed to improve computational efficiency. See Appendix E.2 for further discussion.
Remark 4 For the state evolution to correctly describe the AMP iterates under the PCA initialization $G^0 = G$, the first AMP iteration for $F^0$ in Line 8 should use $U^{-1} = F \cdot \Sigma_0^{1/2}$ as initialized in Line 4, rather than $U^{-1} = 0$ as described in Montanari and Venkataramanan (2021). We elaborate on this in Appendix C.

Algorithm 1. EB-PCA

Input: Data matrix $Y \in \mathbb{R}^{n \times d}$, normalized as in Remark 1 to have average entrywise noise variance $1/n$. Number of PCs $k$, number of AMP iterations $T$, prior class $\mathcal{P}$.

// Initialization
1: Let $\gamma = d/n$. Let $(\sqrt{\gamma} \lambda_1, \ldots, \sqrt{\gamma} \lambda_k)$, $F = (f_1, \ldots, f_k)$, and $G = (g_1, \ldots, g_k)$ be the top $k$ singular values and singular vectors of Y, with $\|f_i\|_2 = \sqrt{n}$ and $\|g_i\|_2 = \sqrt{d}$.
2: Define $s_i^2$ by (2.14). Set $\sigma_i^2 = (1 + \gamma s_i^2)/(\gamma s_i^2 + \gamma s_t^2)$, $\mu_i = 1 - \sigma_i^2$, $M_0 = \text{diag}(\mu_1, \ldots, \mu_k)$, $\Sigma_0 = \text{diag}(\sigma_1^2, \ldots, \sigma_k^2)$, $\tilde{S} = \text{diag}(s_1, \ldots, s_k)$.
3: $G^0 \leftarrow G$ and $U^{-1} \leftarrow F \cdot \Sigma_0^{1/2}$

// Iterative refinement
4: for $t = 0, 1, 2, \ldots, T$ do
    // Denoise left PCs
5: $\pi_t \leftarrow \text{MLE}(G^t \mid M_t, \Sigma_t, \mathcal{P})$
6: $V^t \leftarrow \theta(G^t \mid M_t, \Sigma_t, \pi_t)$
7: $F^t \leftarrow YV^T - U^{-1} \cdot \gamma (\theta(G^t \mid M_t, \Sigma_t, \pi_t))^T$
8: $\Sigma_t \leftarrow (V^T)^T V^n/n$ and $M_t \leftarrow \hat{\Sigma} \cdot \tilde{S}$
    // Denoise right PCs
9: $\hat{\pi}_t \leftarrow \text{MLE}(F^t \mid \hat{M}_t, \hat{\Sigma}_t, \mathcal{P})$
10: $U^t \leftarrow \theta(F^t \mid \hat{M}_t, \hat{\Sigma}_t, \hat{\pi}_t)$
11: $G^{t+1} \leftarrow Y^T U^t - V^t \cdot (\theta(F^t \mid \hat{M}_t, \hat{\Sigma}_t, \hat{\pi}_t))^T$
12: $\Sigma_{t+1} \leftarrow (U^T)^T U^n/n$ and $M_{t+1} \leftarrow \Sigma_{t+1} \cdot \tilde{S}$
4: end for

Output: Final estimates $(\hat{U}, \hat{\Sigma}, \hat{V}) = (U^T, \hat{\Sigma}, \hat{V}^T)$

2.5 Relation to naive mean field variational Bayes

Wang and Stephens (2021) propose an empirical Bayes matrix factorization (EBMF) algorithm similar to EB-PCA, based instead on naive mean-field variational Bayes: in the rank-one model of (1), this approximates the posterior law $p(u, v \mid Y)$ by a factorized form $\bar{q}(u)q(v) = \prod_{i=1}^n \bar{q}_i(u_i) \prod_{j=1}^d q_j(v_j)$. The distributions $\bar{q}_i, q_j$ are chosen to minimize the Kullback–Leibler divergence $D_{KL}(\bar{q}(u)q(v)||p(u,v \mid Y))$ or, equivalently, to maximize the evidence lower bound

$$F(\bar{q}, q) = \mathbb{E}_{u \sim \bar{q}, v \sim q}[\log p(Y, u, v) - \log \bar{q}(u)q(v)].$$

Here, the joint density $p(Y, u, v)$ depends on the priors $(\bar{\pi}, \pi)$ for $(u, v)$, and EBMF estimates these by maximizing $F$ jointly over $(\bar{q}, \bar{\pi}, q, \pi)$. This maximization is performed via the iterative coordinate ascent variational inference (CAVI) updates

$$(q_t, \pi_t) = \arg \max_{q, \pi} F(q_{t-1}, \bar{\pi}_{t-1}, q, \pi), \quad (\bar{q}_t, \bar{\pi}_t) = \arg \max_{\bar{q}, \bar{\pi}} F(\bar{q}, \bar{\pi}, q_t, \pi_t).$$
As shown in Wang and Stephens (2021), the CAVI updates admit a simple form in terms of the quantities

\[ \mathbf{v}^t \equiv \mathbb{E}_{\mathbf{v} \sim q_{t}}[\mathbf{v}], \quad \sigma_v^2 \equiv n^{-1} \mathbb{E}_{\mathbf{v} \sim q_{t}}[\|\mathbf{v}\|^2], \quad \mathbf{u}^t \equiv \mathbb{E}_{\mathbf{u} \sim q_{t}}[\mathbf{u}], \quad \sigma_u^2 \equiv n^{-1} \mathbb{E}_{\mathbf{u} \sim q_{t}}[\|\mathbf{u}\|^2], \]

which is very similar to the iterations of Algorithm 1 but does not incorporate the AMP Onsager correction terms. The Onsager terms correct for weak dependences in the true posterior distributions of \( u_1, \ldots, u_n \) and of \( v_1, \ldots, v_d \) in high dimensions. These dependences are omitted in the naive mean field approximation, and this is discussed further in Ghorbani et al. (2019) and Fan et al. (2021). The differences between these approaches vanish in the limit of infinite signal-to-noise ratio \( s \to \infty \), but are non-negligible in any bounded signal-to-noise setting, and become more pronounced for weak signals near the phase transition threshold \( s_{\ast}(\gamma) = \gamma^{-1/4} \).

We believe there are two particular appeals of applying AMP over the naive mean field approximation in this specific application: first, in the limit \( s \to \infty \), the sample PCs become increasingly accurate, and there is less to gain from an empirical Bayes approach. It is precisely in settings of weaker signals that empirical Bayes may yield the largest improvements over PCA. Second, even if the model of (3) is correctly specified, the connection between the CAVI iterates \( f^t, g^t \) and the Gaussian compound decision model is inexact—see Figures 3 and S3—whereas the AMP iterates \( f^t, g^t \) are exactly described by the Gaussian models as \( n, d \to \infty \). This provides a stronger justification for applying empirical Bayes procedures based on these Gaussian models to the iterates of AMP. However, we note that EBMF is developed in a more general model of heteroscedastic noise \( w_{ij} \sim \mathcal{N}(0, 1/\tau_{ij}) \), whereas our current derivation and analysis of EB-PCA are limited to a setting of uniform noise variance.

There are a few other distinctions in perspective between EBMF and our work: for the rank-\( k \) model, Wang and Stephens (2021) propose a CAVI scheme that iteratively updates each rank-one component separately, and hence does not learn a joint prior for the multivariate distribution of several PCs. For each rank-one component, there is a stronger emphasis in (Wang & Stephens, 2021) on sparsity-inducing priors that are unimodal at 0, connecting the approach more to sparse PCA. Depending on the characteristics of the data at hand, we believe that a multivariate approach of learning a fully non-parametric joint prior for several PC has the potential of yielding improved accuracy.

### 3 SIMULATED EXAMPLES

#### 3.1 Univariate priors

We compare EB-PCA with several other methods on simulated data from the rank-one model of (1), using four different univariate priors: (a) standard Gaussian \( \mathcal{N}(0, 1) \), (b) Uniform\([-\sqrt{3}, \sqrt{3}]\),  (c) Bernoulli\{+1, -1\} and (d) sparse point-normal \( 0.9\delta_0 + 0.1\mathcal{N}(0, 10) \). For simplicity, we use the same prior for both \( u \) and \( v \). The two-point and point-normal priors (c–d) represent simple clustering and sparse-PCA applications.

We compare EB-PCA with standard PCA, the oracle Bayes AMP procedure of Montanari and Venkataramanan (2021) that knows the true priors, and naive mean-field variational Bayes with priors estimated by NPMLE, corresponding to a version of EBMF in Wang and Stephens (2021) with known homoscedastic noise variance. For the sparse point-normal prior (d), we compare also to the \texttt{spca} method of Zou et al. (2006). Simulation details can be found in Appendix E.
FIGURE 2  Estimation accuracy of principal components analysis (PCA), oracle Bayes approximate message passing (AMP), empirical Bayes-PCA and naive mean-field VB, across 50 simulations of a rank-one model with four priors. For (d), comparison with spca is also shown. In all cases, EB-PCA nearly matches the accuracy of the oracle Bayes AMP procedure and improves over mean-field VB. EB-PCA also improves over standard PCA for the non-Gaussian priors (b–d), and over spca for the sparse prior (d) [Colour figure can be viewed at wileyonlinelibrary.com]

Figure 2 displays the accuracy of these procedures, in terms of the alignments $\langle \hat{u}, u \rangle / ||u|| ||\hat{u}||$ and $\langle \hat{v}, v \rangle / ||v|| ||\hat{v}||$. These alignments measure the accuracy of the estimated PC directions, and do not account for further improvements of EB-PCA/Bayes-AMP resulting from the shrinkage of $(u, v)$. The dimensions tested are $(n, d) = (2000, 4000)$, so that the estimate of $u \in \mathbb{R}^n$ is more accurate than that of $v \in \mathbb{R}^d$. The phase-transition point for supercritical signal strength is $s^*_\gamma = 0.84$, and we tested the range of signal strengths $s \in \{1.1, 1.3, 1.5, 2.0\}$.

Under the standard Gaussian prior (a), the posterior mean function in Equation (9) for oracle Bayes AMP is linear in each iteration. Thus its iterates remain proportional to the sample PCs, and the asymptotic accuracy of both EB-PCA and Bayes-AMP in the above alignment metric coincide with standard PCA. Thus the Gaussian prior (a) provides a ‘control’ setting in which we hope to match the performance of PCA.

The following observations summarize these comparisons:

- **Oracle Bayes AMP.** EB-PCA is nearly as accurate as oracle Bayes AMP in all cases, without knowing the true priors. There is a small decrease in accuracy of EB-PCA for very weak signals, due to estimation variance for $s$.
- **Standard PCA.** In the control setting of the $\mathcal{N}(0, 1)$ prior, EB-PCA yields accuracy comparable to PCA. EB-PCA yields improved accuracy in all remaining settings, with this improvement being more substantial for weaker signals and for the two-point and sparse point-normal priors that reflect stronger prior structure.
- **SPCA.** EB-PCA and mean-field VB both improve significantly over spca for the sparse point-normal prior, with the added advantage of being tuning-parameter free.
- **Mean-field VB.** EB-PCA improves over mean-field VB in all cases. The improvement is small for the point-normal prior and for larger signal strengths, but is larger in the remaining settings. Mean-field VB seems to yield worse estimation accuracy than standard PCA for the $\mathcal{N}(0, 1)$ and continuous uniform priors.
FIGURE 3 Iterates of empirical Bayes principal components analysis versus mean-field VB with Uniform$[-\sqrt{3}, \sqrt{3}]$ priors for $u$. (a) Distribution of entries of the EB-PCA iterates $f_0, f_1, f_4$. A close agreement is observed with the overlaid convolution densities Uniform$[-\sqrt{3}, \sqrt{3}] \ast \mathcal{N}(0, \sigma_t^2 / \mu_t^2)$ for $t = 0, 1, 4$, where $\mu_t^2 / \sigma_t^2$ is the estimated signal-to-noise ratio (SNR) used to perform empirical Bayes denoising. (b) Analogous picture for mean-field VB, with $\mu_t^2 / \sigma_t^2$ as estimated in mean-field VB. A discrepancy is observed between the iterates and the convolution densities. (c) Accuracy of $u$, across iterations, for EB-PCA versus mean-field VB. Results for $(v, g)$ are similar and omitted for brevity [Colour figure can be viewed at wileyonlinelibrary.com]

To provide a more detailed comparison of EB-PCA with mean-field VB at a weak signal strength, panels (a) and (b) of Figures 3 and S3 display the entrywise distributions of several iterates for the uniform and two-point priors with $s = 1.3$. Overlaid are the convolution densities of the true prior with the estimated levels of Gaussian noise. Panel (c) displays the accuracies across iterations. The overlaid convolution densities in EB-PCA closely match the empirical distributions of the iterates, whereas discrepancies accumulate for mean-field VB. These discrepancies can cause mean-field VB to estimate an increasingly incorrect prior, and to have decreasing accuracy across iterations.

3.2 Bivariate priors

We now demonstrate that multivariate EB-PCA, which estimates a joint prior over several PCs, can improve substantially over EB-PCA applied marginally to each PC when there is strong multivariate structure. We consider dimensions $(n, d) = (1000, 1000)$ and signal strengths $(s_1, s_2) = (4, 2)$, for the two bivariate priors presented in Figure 4a: a discrete three-point prior and a uniform prior on a circle.

Titles in Figure 4 display the estimation error as defined by the subspace distance (see section 2.5 of Van Loan & Golub, 2013) between the column spans of $\hat{U}$ and $U$ and also between the spans of the individual PCs. Table S1 reports the average of such estimation errors across 50 simulations. These results indicate that by leveraging the underlying joint structure, multivariate EB-PCA learns a more accurate prior and has lower error both for the estimated two-dimensional subspace and for the individual PCs. The simultaneous estimation of $k$ PCs has the
FIGURE 4 Comparison of standard principal components analysis (PCA), empirical Bayes PCA (EB-PCA) applied marginally and EB-PCA applied jointly for two bivariate priors: a discrete three-point mixture prior (top row) and a uniform prior on a circle (bottom row). Displayed are scatter-plots of the columns of $U$ for (a) the true PCs, (b) the sample PCs, (c) EB-PCA applied marginally for each PC and (d) EB-PCA applied jointly to learn the bivariate prior. Estimation errors are displayed in the figure titles. Marginal EB-PCA improves over standard PCA, and joint EB-PCA further improves over marginal EB-PCA [Colour figure can be viewed at wileyonlinelibrary.com]

additional benefit of reducing the computation time by a factor of roughly $k$ over the univariate approach.

4 | APPLICATIONS

We illustrate EB-PCA on three high-dimensional genetics datasets: genotype data from the 1000 Genomes Project and the third phase of the International HapMap Project (HapMap3) (International HapMap 3 Consortium, 2010), and single cell RNA-seq (scRNA-seq) gene expression data on Peripheral Blood Mononuclear Cells (PBMC) from 10X Genomics. Preprocessing procedures and implementation details are provided in Appendix E.

4.1 | 1000 Genomes Project genotypes

PCA is commonly used to correct for population stratification in genome-wide association studies. As the number of SNPs $n$ often far exceeds the number of individuals $d$, the estimated PCs in $\mathbb{R}^d$ suffer minimally from high-dimensional noise for the leading PCs. This provides a ground truth by which we may quantitatively compare estimation accuracy on subsampled data.

Implementing this experiment for genotype data from the 1000 Genomes Project, we extracted 100,000 common SNPs for 2504 individuals, from which we computed the ground truth PCs. We then estimated PCs on subsamples of 100, 1000 or 10,000 randomly selected SNPs using both EB-PCA and PCA. Figure S2 plots the singular values for a typical subsample of 1000 SNPs, in which 4 clear outlier values are apparent. Thus we chose to estimate the leading four PCs using EB-PCA.

Table 1 compares errors across 50 random subsamples of each size, where the error is the subspace distance (as previously used in Section 3.2) against the ground truth PCs. The results
show a clear improvement of EB-PCA over PCA, with slightly larger improvement for the lower PCs having smaller singular values.

A visual comparison for 1000 subsampled SNPs was presented previously in Figure 1. The EB-PCA estimates are closer to the ground truth than the sample PCs, and better separate the subjects by ethnicity. For example, the Caucasian, African and East Asian populations are mixed in the two-dimensional plot of the third versus fourth sample PC, whereas they are separated in the EB-PCA estimates and also in the ground truth. The information used by EB-PCA for performing this separation in PCs 3 and 4 comes from learning a joint prior with PCs 1 and 2, where these populations have clear separation.

### 4.2 HapMap3 genotypes

We performed a similar experiment on genotype data from HapMap3. For the 1397 individuals in HapMap3, we computed ground truth PCs from 142,185 common SNPs. Based on the singular value distribution in Figure S2 on a subsample of 5000 SNPs, we chose to estimate the leading 4 PCs using EB-PCA.

Table S2 compares the errors of EB-PCA and PCA across 50 subsampled data sets of 1000, 5000 and 10,000 SNPs, and Figure 5 depicts results for 5000 SNPs. We again observe a consistent decrease in estimation error for EB-PCA, which is larger for the weaker PCs.

### 4.3 10x Genomics PBMC single-cell RNA-seq

PCA is often the first step in single-cell gene expression data analysis pipelines such as Seurat and scanpy, to capture the signatures of cell identity. Non-linear dimensionality reduction methods are often then applied with the estimated PCs as input, to perform cell clustering and to infer cell types.

We illustrate an application of EB-PCA on scRNA-seq data of $d = 2626$ peripheral blood mononuclear cells (PBMCs) with $n = 13,711$ gene expressions, from 10X Genomics. This data is more representative of typical applications of EB-PCA, in which there are insufficient samples to provide a known ground truth. Therefore, we applied EB-PCA without subsampling. The singular values are shown in Figures S2, and 3 large outlier values are apparent. The corresponding leading three PCs are depicted as two scatterplots in Figure 6a. Qualitatively, these scatterplots exhibit estimation noise that resembles the noise previously observed in the subsampled 1000 matrices.
FIGURE 5  Illustration of Empirical Bayes principal components analysis (EB-PCA) on genotype data from HapMap3, similar to Figure 1. (a) Ground truth PCs, with points colored by the individuals' ethnicity. (b) Sample PCs computed on a subsample of 5000 SNPs. (c) EB-PCA estimates of the four PCs, using the same subsample as in panel (b) [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 6  Estimated principal components (PCs) using (a) standard PC analysis and (b) Empirical Bayes-PCA, for peripheral blood mononuclear cells single-cell RNA-seq gene expression data, with points coloured by cell type. The EB-PCA estimates exhibit reduced estimation noise and clearer separation of cells by cell type [Colour figure can be viewed at wileyonlinelibrary.com]

Genomes and HapMap3 genotype matrices, suggesting that the noise may be a consequence of high dimensionality.

Figure 6b displays the estimated PCs using EB-PCA. Although there is not a basis for quantitative comparison with PCA in this example, we believe that the EB-PCA results may be more accurate for an underlying ground truth. Qualitatively, the estimation noise exhibited in Figure 6a has been reduced. There is a clearer separation between different cell types. For example, Naïve CD4 T cells are separated from CD14+ monocytes in the plot of the second versus third PCs estimated by EB-PCA, whereas these overlap in the plot of the sample PCs.
5 | THEORETICAL GUARANTEES

In this section, we summarize theoretical guarantees for EB-PCA. In the context of the signal-plus-noise model (3), our results show that EB-PCA asymptotically achieves the same (first-order) estimation accuracy as the oracle Bayes AMP procedure. Consequently, the asymptotic estimation error provably decreases across iterations, and can approach the Bayes-optimal error as the number of iterations \( t \to \infty \) under general conditions for the true priors.

The strategy of proof is to first establish asymptotic consistency of the NPMLEs for the prior distributions, and then, through an inductive comparison argument, show that the iterates of EB-PCA are characterized by the same state evolution as oracle Bayes AMP. We note that prior work on asymptotic consistency for the NPMLE assumes that a compound decision model (8) with i.i.d. Gaussian errors holds exactly, whereas such an error model holds only approximately for the sample PCs and AMP iterates. Our arguments show that this approximation is sufficient for asymptotic consistency.

5.1 | Assumptions

Consider the rank-\( k \) model of (3). We assume the empirical distributions of rows of \( U \in \mathbb{R}^{n \times k} \) and \( V \in \mathbb{R}^{d \times k} \) converge in Wasserstein-2 distance to fixed distributions \( \pi_s, \pi_s \) on \( \mathbb{R}^k \), as \( n, d \to \infty \). This means that for any continuous function \( \psi : \mathbb{R}^k \to \mathbb{R} \) with \( E_{U \sim \pi_s}[\psi(U)^2] < \infty \), we have \( \lim_{n \to \infty} n^{-1} \sum_{i=1}^n \psi(u_{i1}, \ldots, u_{ik}) = E_{U \sim \pi_s}[\psi(U)] \), and similarly for \( V \). We denote this convergence both as \( U \overset{W^2}{\rightharpoonup} \pi_s \) and as \( U \overset{w_2}{\rightharpoonup} U \) for a random vector \( U \sim \pi_s \). Our model assumptions are then summarized as follows.

**Assumption 1** \( \mathcal{P} \) is a family of probability distributions on \( \mathbb{R}^k \) having finite second moment, and \( n, d \to \infty \) such that

(a) \( W \in \mathbb{R}^{n \times d} \) has entries \( w_{ij} \) iid \( \mathcal{N}(0, 1/n) \).
(b) \( k, s_1, \ldots, s_k, \) and \( \gamma \equiv d/n \) remain constant, where \( s_1 > \cdots > s_k > s_s(\gamma) \equiv \gamma^{-1/4} \).
(c) \( U \overset{W}{\rightharpoonup} \pi_s \) and \( V \overset{W}{\rightharpoonup} \pi_s \) for two distributions \( \pi_s, \pi_s \in \mathcal{P} \) that satisfy the normalizations, for all \( 1 \leq i \neq j \leq k \),

\[ E_{U \sim \pi_s}[U_i^2] = 1, \quad E_{U \sim \pi_s}[U_iU_j] = 0, \quad E_{V \sim \pi_s}[V_i^2] = 1, \quad E_{V \sim \pi_s}[V_iV_j] = 0 \]

(d) For any non-singular \( M_s \in \mathbb{R}^{k \times k} \), symmetric positive-definite \( \Sigma_s \in \mathbb{R}^{k \times k} \), and \( \pi_s \in \mathcal{P} \), there is a weakly open neighbourhood \( O \) of \( \pi_s \) such that \( \theta(x|M_s, \Sigma_s, \pi) \) is Lipschitz in \( x \) uniformly over \( \pi \in O \).

Part (a) makes a Gaussian assumption for the noise, and part (b) assumes for simplicity that the signal values \( s_1, \ldots, s_k \) are distinct and supercritical.

Part (c) ensures the normalizations in Equations (5) and (6). This can hold both when \( U \in \mathbb{R}^{n \times k} \) and \( V \in \mathbb{R}^{d \times k} \) are deterministic matrices whose columns are the (exactly) orthogonal true PCs, as well as almost surely in a Bayesian setting when \( U \) and \( V \) are random with independent
rows generated from $\overline{\pi}_s$ and $\pi_s$. We will assume $U$ and $V$ are deterministic that is our results apply conditional on $(U, V)$ in the Bayesian setting.

Part (d) assumes a Lipschitz property for the posterior mean functions, as is common in analyses of AMP. This places a small restriction on the prior class $P$; for example, the assumption holds if $P$ is the class of all priors supported on a compact domain of $\mathbb{R}^k$.

We analyse EB-PCA in a slightly idealized setting where the noise variance $1/n$ in part (a) is known, rather than estimated as in Remark 1, and where the NPMLE in Lines 5 and 9 of Algorithm 1 are computed exactly in each iteration. Our results may be extended to incorporate a consistent estimate of the noise variance and an approximate NPMLE computed on a sufficiently fine discretization of the support, by a standard comparison argument with this idealized setting—we omit the details of such an argument for brevity.

5.2 Limiting risk for the initial empirical Bayes estimates

In the compound decision model $\Theta \sim \pi$ and $X|\Theta \sim \mathcal{N}(M \cdot \Theta, \Sigma)$, we denote the squared error Bayes risk for estimating $\Theta$ based on $X$ as

$$\text{mmse}(\pi | M, \Sigma) = \mathbb{E}[||\Theta - \mathbb{E}[\Theta | X]||_2^2].$$

The entrywise Gaussian approximation in Equation (15) for the sample PCs is formalized as the following proposition. Lemma C.1 of Montanari and Venkataramanan (2021) proves a similar result for the symmetric spiked model.

**Proposition 1** Let $\overline{M}_s, M_s, \overline{\Sigma}_s, \Sigma_s \in \mathbb{R}^{k \times k}$ be defined by Equations (13) and (14). Under Assumption 1, almost surely $(U, F) \xrightarrow{w_2} (U, F)$ and $(V, G) \xrightarrow{w_2} (V, G)$ where $U \sim \overline{\pi}_s$, $F|U \sim \mathcal{N}(\overline{M}_s \cdot U, \overline{\Sigma}_s)$, $V \sim \pi_s$, and $G|V \sim \mathcal{N}(M_s \cdot V, \Sigma_s)$.

Combined with an asymptotic consistency result for the NPMLE in approximate compound decision models, shown in Lemma B.2 and Corollary B.3, this yields the following asymptotic squared-error risks for the initial empirical Bayes estimates of $U$ and $V$.

**Corollary 1** Let $\hat{U}$ and $\hat{V}$ be the initial empirical Bayes estimators in Equations (17) and (18). Let $\overline{M}_s, M_s, \overline{\Sigma}_s, \Sigma_s \in \mathbb{R}^{k \times k}$ be defined by Equations (13) and (14). Under Assumption 1, almost surely

$$n^{-1}||\hat{U} - U||_F^2 \rightarrow \text{mmse}(\overline{\pi}_s | \overline{M}_s, \overline{\Sigma}_s), \quad d^{-1}||\hat{V} - V||_F^2 \rightarrow \text{mmse}(\pi_s | M_s, \Sigma_s).$$

Thus the asymptotic squared-error risk for $\hat{U}$ is the Bayes risk for estimating $\Theta \sim \overline{\pi}_s$ based on $X|\Theta \sim \mathcal{N}(\overline{M}_s \cdot \Theta, \overline{\Sigma}_s)$, and similarly for $\hat{V}$. In contrast, the analogous risks for the naive sample PCs $F$ and $G$, or more generally of the best shrinkage estimators obtained by rescaling their columns, correspond to the Bayes risks of the best linear estimators of $\Theta$ in these compound decision problems. These linear risks may be substantially larger if the priors $\overline{\pi}_s$ and $\pi_s$ are far from the standard Gaussian law.
5.3 Limiting risk and Bayes optimality of EB-PCA

For the full EB-PCA method, the following verifies that the NPMLEs \( \bar{\pi}_t, \pi_t \) remain consistent for \( \pi_s, \pi_e \) across iterations. Consequently, the EB-PCA iterates satisfy the same Gaussian approximations Equation (20) and are tracked by the same state evolution (21) as the oracle Bayes AMP algorithm.

**Theorem 1** Suppose Assumption 1 holds. Then almost surely for each fixed iteration \( t \in \{0, \ldots, T\} \) of Algorithm 1, \( \pi_t \) converges weakly to \( \pi_e \) and \( \bar{\pi}_t \) converges weakly to \( \bar{\pi}_e \). Furthermore,

\[
(U, F^t) \xrightarrow{W^t} (U, F_t) \quad \text{and} \quad (V, G^t) \xrightarrow{W^t} (V, G_t)
\]

where \( U \sim \bar{\pi}_e, F_t | U \sim \mathcal{N}(\bar{M}_{s,t} \cdot U, \bar{\Sigma}_{s,t}), V \sim \pi_e, \) and \( G_t | V \sim \mathcal{N}(M_{s,t} \cdot V, \Sigma_{s,t}) \). The matrices \( \bar{M}_{s,t}, \bar{\Sigma}_{s,t}, M_{s,t}, \Sigma_{s,t} \) are defined iteratively by the state evolution in Equation (21).

As a corollary, the asymptotic squared-error risk for each EB-PCA iterate is the same as that achieved by the oracle Bayes AMP algorithm with known priors.

**Corollary 2** Let \( \bar{M}_{s,t}, \bar{\Sigma}_{s,t}, M_{s,t}, \Sigma_{s,t} \) be defined iteratively by the state evolution in Equation (21). Under Assumption 1, almost surely for each fixed iteration \( t \in \{0, \ldots, T\}, \)

\[
\frac{1}{n} \|U^t - U\|_F^2 \to \text{mmse}(\bar{\pi}_s | \bar{M}_{s,t}, \bar{\Sigma}_{s,t}), \quad \frac{1}{d} \|V^t - V\|_F^2 \to \text{mmse}(\pi_e | M_{s,t}, \Sigma_{s,t}).
\]

To study the decrease of these errors across iterations, we follow Miolane (2017) and introduce the positive-definite state matrices

\[
\bar{Q}_{s,t} = \frac{\gamma}{S^{1/2}} M_{s,t}^\top \Sigma_{s,t}^{-1} M_{s,t} S^{-1/2}, \quad Q_{s,t} = S^{1/2} M_{s,t}^\top \Sigma_{s,t}^{-1} M_{s,t} S^{-1/2}.
\]

These are matrix-valued measures of the signal-to-noise ratios in the compound decision models associated to \( F^t \) and \( G^t \) in each iteration. For the standardized compound decision model

\[
\Theta \sim \pi, \quad X|\Theta \sim \mathcal{N}(\Theta, Q^{-1})
\]

(with \( M = \text{Id} \)) parametrized by \( \pi \) and \( Q \), define the map

\[
F_{\pi}(Q) = \mathbb{E}[\mathbb{E}[\Theta|X] \cdot \mathbb{E}[\Theta|X]^	op].
\]

We verify in Appendix D that the state evolution in Equation (21) is equivalently expressed as

\[
\bar{Q}_{s,t} = F^{S^{1/2}\pi_e}(Q_{s,t}), \quad Q_{s,t+1} = F^{S^{1/2}\bar{\pi}_e}(\gamma \cdot \bar{Q}_{s,t}),
\]

where \( S^{1/2}/\pi \) is the distribution of \( S^{1/2} \cdot \Theta \) when \( \Theta \sim \pi \). The following shows that the state evolution converges to a fixed point of these maps, the squared-error risks of the EB-PCA iterates improve over the initial empirical Bayes estimates, and these risks decrease monotonically over iterations.

**Proposition 2** Under Assumption 1, for each \( t = 0, 1, 2, \ldots \)
(a) \( \bar{Q}_{s,t+1} \geq \bar{Q}_{s,t} \) and \( Q_{s,t+1} \geq Q_{s,t} \). Furthermore, as \( t \to \infty \), these matrices converge to a fixed point of

\[
\bar{Q} = F_{S \mid \pi_s}(Q), \quad Q = F_{S \mid \pi_s}(\gamma \cdot Q). \tag{26}
\]

(b) Let \( \text{mmse}(\pi_s | M_s, \Sigma_s) \) and \( \text{mmse}(\pi_s | M_s, \Sigma_s) \) be the risks of the initial empirical Bayes estimates in Corollary 1. Then the asymptotic risks in Equation (24) satisfy

\[
\text{mmse}(\pi_s | M_{s,t+1}, \Sigma_{s,t+1}) \leq \text{mmse}(\pi_s | M_{s,t}, \Sigma_{s,t}) \leq \text{mmse}(\pi_s | M_s, \Sigma_s),
\]

\[
\text{mmse}(\pi_s | M_{s,t+1}, \Sigma_{s,t+1}) \leq \text{mmse}(\pi_s | M_{s,t}, \Sigma_{s,t}) \leq \text{mmse}(\pi_s | M_s, \Sigma_s).
\]

Finally, suppose \( \bar{\pi}_s, \pi_s \) are such that Equation (26) has a unique fixed point \((\bar{Q}, Q)\). In the Bayesian setting where rows of \( U \) and \( V \) are random and i.i.d., Proposition 3 of Molane (2017) shows that this fixed point characterizes the Bayes-optimal squared-error risk for estimating \( USV^T \):

\[
\lim_{n, d \to \infty} \left( \inf_{\hat{X}} \frac{1}{n d} E[\|\hat{X}(Y) - USV^T\|_F^2] \right) = \text{Tr}E_{(Y)}[UU^T SVV^T S] - \text{Tr}\bar{Q}Q,
\]

where the infimum is over all (measurable) estimators \( \hat{X}(Y) \) and achieved at the Bayes estimator \( E[USV^T | Y] \). Exact computation of this Bayes estimator may be intractable. The below verifies that the asymptotic error of EB-PCA (and hence also of the oracle Bayes AMP algorithm) approaches this Bayes-optimal error as \( t \to \infty \). This type of result has been stated for rank \( k = 1 \) in Barbier et al. (2016) and Montanari and Venkataaramanan (2021).

**Proposition 3** Under Assumption 1, suppose the fixed point \((\bar{Q}, Q)\) of Equation (26) is unique. For any fixed \( t \geq 1 \), let \((U^t, V^t)\) be the EB-PCA estimates of \((U, V)\) in iteration \( t \) of Algorithm 1, and let \( \hat{S} = \text{diag}(\hat{s}_1, \ldots, \hat{s}_k) \) be the estimate of \( S \) from Equation (6). Then almost surely as \( n, d \to \infty \),

\[
\frac{1}{n d} \|U^t \hat{S}(V^t)^T - USV^T\|_F^2 \to \text{Tr}E_{(U, V)}[UU^T SVV^T S] - \text{Tr}\bar{Q}Q - o_t(1)
\]

where \( o_t(1) \) is a deterministic quantity satisfying \( o_t(1) \to 0 \) as \( t \to \infty \).

For rank \( k = 1 \), we refer readers to Barbier et al. (2016); Montanari and Venkataaramanan (2021) and Lelarge and Molane (2019) for examples of priors \( \bar{\pi}_s, \pi_s \) for which uniqueness of the fixed point to Equation (26) does and does not hold. It has been conjectured that in examples where this fixed point is not unique, the asymptotic risk corresponding to the fixed point that is reached by EB-PCA/oracle-Bayes AMP is the smallest Bayes risk that is attainable by any estimator in polynomial time (Antenucci et al., 2019; Barbier et al., 2016; Lelarge & Molane, 2019; Lesieur et al., 2015b).

6 | CONCLUSION

We have described an EB-PCA procedure for performing PCA in high dimensions, which couples classical empirical Bayes ideas with high-dimensional asymptotic theory. In applications where the joint distribution of PCs has non-Gaussian structure, EB-PCA can improve estimation accuracy by obtaining a non-parametric estimate of this structure.
EB-PCA is an example of a more general paradigm of carrying out ‘TAP-corrected’ variational Bayesian inference in high dimensions using an empirical Bayes approach. The high dimensionality of the latent variable space becomes a blessing in such problems, enabling the estimation of complex and non-parametric prior distributions for these latent variables using empirical Bayes ideas. This general paradigm may potentially be implemented with other inference algorithms and extended to other inference problems.

We conclude with a discussion of a direction for future work. The quantitative form of EB-PCA—and of AMP-based procedures more generally—is derived assuming that the noise matrix $W$ has independent entries with common variance. (To keep the proofs simple, we have also assumed that these entries are Gaussian, but it is known that the Gaussian assumption can be substantially weakened: universality results of this type have been shown for the sample PCs in Bao et al. (2020, 2021), and for the AMP state evolution in Bayati et al. (2015) and Chen and Lam (2021).)

It is an open question to extend the procedure to settings where $W$ has correlation structure, which is commonly reflected in data by an overdispersed singular value distribution. One such setting that is partially understood is that of bi-rotationally invariant matrices $W$ that satisfy the equality in law $W^L = O^\top WQ$ for any orthogonal matrices $O \in \mathbb{R}^{n \times n}$ and $Q \in \mathbb{R}^{d \times d}$. This reflects an assumption that $W$ may have an arbitrary distribution of singular values but ‘generic’ singular vectors. The asymptotic behaviour of sample PCs in this setting was studied by Benaych-Georges and Nadakuditi (2012), who provided the quantitative forms for the limits Equation (12). In the rank-one case, it was shown also in (Fan, 2020) that the Onsager correction in Equation (19) and AMP state evolution (21) are to be replaced by certain series expressions defined by the free cumulants of $W$, which may be estimated from the singular value distribution of $Y$.

Other models for $W$—for example general covariance models $W = ZB^{1/2}$ (Bai & Yao, 2012; Bao et al., 2020) or separable covariance models $W = A^{1/2}ZB^{1/2}$ (Ding & Yang, 2021; Yang, 2019) where $Z \in \mathbb{R}^{n \times d}$ has i.i.d. entries—may also be studied. Developing variational Bayesian procedures that are both asymptotically exact and computationally efficient in these and related models is an interesting direction for future work, and we believe it is likely that such developments may improve and robustify the EB-PCA procedure, when they become available.

ACKNOWLEDGEMENTS

We thank Yixuan Ye, Chao Zhou and Hongyu Zhao for their help in collecting and interpreting data, and Yihong Wu for helpful discussions about the Kiefer–Wolfowitz NPMLE and ideas in the early stages of the theoretical analyses. This research was supported in part by NSF Grant DMS-1916198.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at https://urldefense.com/v3/__https://www.internationalgenome.org/__;!!N11eV2iwtf5IoE_IBPCpoiNQjRdFhMo6CzKZC8Xz4A7WbJaCE45mEVagvD3226wcILB42rimC1GGyh352PP50yWF4dZ8A$, https://urldefense.com/v3/__https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html__;!!N11eV2iwtf5IoE_IBPCpoiNQjRdFhMo6CzKZC8Xz4A7WbJaCE45mEVagvD3226wcILB42rimC1GGyh352PP50zT3XiLLw$, and https://urldefense.com/v3/__https://support.10xgenomics.com/single-cell-gene-expression/datasets__;!!N11eV2iwtf5IoE_IBPCpoiNQjRdFhMo6CzKZC8Xz4A7WbJaCE45mEVagvD3226wcILB42rimC1GGyh352PP50zIdoUWBA$. 
REFERENCES

d’Aspremont, A., Ghaoui, L.E., Jordan, M.I. & Lanckriet, G.R. (2005) A direct formulation for sparse PCA using semidefinite programming. In: Advances in neural information processing systems, 41–48.

Alaoui, A.E. & Krzakala, F. (2018) Estimation in the spiked Wigner model: A short proof of the replica formula. In IEEE International Symposium on Information Theory, 1874–1878.

Amini, A.A. & Wainwright, M.J. (2008) High-dimensional analysis of semidefinite relaxations for sparse principal components. In IEEE International Symposium on Information Theory, 2454–2458.

Antenucci, F., Franz, S., Urbani, P. & Zdeborová, L. (2019) Glassy nature of the hard phase in inference problems. Physical Review X, 9, 011020.

Bai, Z. & Yao, J. (2008) Central limit theorems for eigenvalues in a spiked population model. Annales de l'IHP Probabilités et statistiques, 44, 447–474.

Bai, Z. & Yao, J. (2012) On sample eigenvalues in a generalized spiked population model. Journal of Multivariate Analysis, 106, 167–177.

Baik, J. & Silverstein, J.W. (2006) Eigenvalues of large sample covariance matrices of spiked population models. Journal of Multivariate Analysis, 97, 1382–1408.

Baik, J., Ben Arous, G. & Péché, S. (2005) Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. The Annals of Probability, 33, 1643–1697.

Bao, Z., Ding, X., Wang, J. & Wang, K. (2020) Statistical inference for principal components of spiked covariance matrix. arXiv preprint arXiv:2008.11903.

Bao, Z., Ding, X. & Wang, K. (2021) Singular vector and singular subspace distribution for the matrix denoising model. The Annals of Statistics, 49, 370–392.

Barbier, J. & Macris, N. (2019) The adaptive interpolation method: a simple scheme to prove replica formulas in Bayesian inference. Probability Theory and Related Fields, 174, 1133–1185.

Barbier, J., Dia, M., Macris, N., Krzakala, F., Lesieur, T. & Zdeborová, L. (2016) Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. In: Advances in Neural Information Processing Systems, 424–432.

Bayati, M., Lelarge, M. & Montanari, A. (2015) Universality in polytope phase transitions and message passing algorithms. The Annals of Applied Probability, 25, 753–822.

Benaych-Georges, F. & Nadakuditi, R.R. (2011) The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. Advances in Mathematics, 227, 494–521.

Benaych-Georges, F. & Nadakuditi, R.R. (2012) The singular values and vectors of low rank perturbations of large rectangular random matrices. Journal of Multivariate Analysis, 111, 120–135.

Birnbaum, A., Johnstone, I.M., Nadler, B. & Paul, D. (2013) Minimax bounds for sparse PCA with noisy high-dimensional data. The Annals of Statistics, 41, 1055.

Bloemendal, A., Knowles, A., Yau, H.-T. & Yin, J. (2016) On the principal components of sample covariance matrices. Probability Theory and Related Fields, 164, 459–552.

Böhning, D. (1999) Computer-assisted analysis of mixtures and applications: meta-analysis, disease mapping and others, vol. 81. Boca Raton: CRC press.

Böhning, D., Schlattmann, P. & Lindsay, B. (1992) Computer-assisted analysis of mixtures (CA MAN): statistical algorithms. Biometrics, 48, 283–303.

Cadima, J. & Jolliffe, I.T. (1995) Loading and correlations in the interpretation of principal components. Journal of Applied Statistics, 22, 203–214.

Cai, J.-F., Candès, E.J. & Shen, Z. (2010) A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization, 20, 1956–1982.

Cai, T.T., Ma, Z. & Wu, Y. (2013) Sparse PCA: optimal rates and adaptive estimation. The Annals of Statistics, 41, 3074–3110.

Capitaine, M. (2018) Limiting eigenvectors of outliers for spiked information-plus-noise type matrices. In: Séminaire de Probabilités XLIX, 119–164. Springer.
Capitaine, M. & Donati-Martin, C. (2018) Non universality of fluctuations of outlier eigenvectors for block diagonal deformations of Wigner matrices. arXiv preprint arXiv:1807.07773.

Capitaine, M., Donati-Martin, C., Féral, D. & Février, M. (2011) Free convolution with a semicircular distribution and eigenvalues of spiked deformations of Wigner matrices. Electronic Journal of Probability, 16, 1750–1792.

Chatterjee, S. (2015) Matrix estimation by universal singular value thresholding. The Annals of Statistics, 43, 177–214.

Chen, W.-K. & Lam, W.-K. (2021) Universality of approximate message passing algorithms. Electronic Journal of Probability, 26, 1–44.

Deshpande, Y. & Montanari, A. (2014) Information-theoretically optimal sparse PCA. In: IEEE International Symposium on Information Theory, 2197–2201.

Deshpande, Y., Abbe, E. & Montanari, A. (2017) Asymptotic mutual information for the balanced binary stochastic block model. Information and Inference: A Journal of the IMA, 6, 125–170.

Ding, X. (2020) High dimensional deformed rectangular matrices with applications in matrix denoising. Bernoulli, 26, 387–417.

Ding, X. & Yang, F. (2021) Spiked separable covariance matrices and principal components. The Annals of Statistics, 49, 1113–1138.

Donoho, D.L., Maleki, A. & Montanari, A. (2009) Message-passing algorithms for compressed sensing. Proceedings of the National Academy of Sciences, 106, 18914–18919.

Efron, B. (2012) Large-scale inference: empirical Bayes methods for estimation, testing, and prediction, vol. 1. Cambridge: Cambridge University Press.

Fan, Z. (2020) Approximate message passing algorithms for rotationally invariant matrices. arXiv preprint arXiv:2008.11892.

Fan, J., Liao, Y. & Mincheva, M. (2013) Large covariance estimation by thresholding principal orthogonal complements. Journal of the Royal Statistical Society. Series B, Statistical methodology, 75, 603–680.

Fan, Z., Mei, S. & Montanari, A. (2021) TAP free energy, spin glasses and variational inference. The Annals of Probability, 49, 1–45.

Feng, L. & Dicker, L.H. (2016) Approximate nonparametric maximum likelihood inference for mixture models via convex optimization. arXiv preprint arXiv:1606.02011.

Gavish, M. & Donoho, D.L. (2017) Optimal shrinkage of singular values. IEEE Transactions on Information Theory, 63, 2137–2152.

Ghorbani, B., Javadi, H. & Montanari, A. (2019) An instability in variational inference for topic models. In: International Conference on Machine Learning, 2221–2231.

Ghosal, S. & van der Vaart, A.W. (2001) Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. Annals of Statistics, 00, 1233–1263.

International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. Nature, 467, 52.

Jewell, N.P. (1982) Mixtures of exponential distributions. The Annals of Statistics, 10, 479–484.

Jiang, W. & Zhang, C.-H. (2009) General maximum likelihood empirical Bayes estimation of normal means. The Annals of Statistics, 37, 1647–1684.

Johnstone, I.M. (2001) On the distribution of the largest eigenvalue in principal components analysis. The Annals of Statistics, 29, 295–327.

Johnstone, I.M. & Lu, A.Y. (2009) On consistency and sparsity for principal components analysis in high dimensions. Journal of the American Statistical Association, 104, 682–693.

Johnstone, I.M. & Paul, D. (2018) PCA in high dimensions: an orientation. Proceedings of the IEEE, 106, 1277–1292.

Jolliffe, I.T., Trendafilov, N.T. & Uddin, M. (2003) A modified principal component technique based on the LASSO. Journal of Computational and Graphical Statistics, 12, 531–547.

Jung, S. & Marron, J.S. (2009) PCA consistency in high dimension, low sample size context. The Annals of Statistics, 37, 4104–4130.

Kabashima, Y. (2003) A CDMA multiuser detection algorithm on the basis of belief propagation. Journal of Physics A: Mathematical and General, 36, 11111.

Kabashima, Y., Krzakala, F., Mézard, M., Sakata, A. & Zdeborová, L. (2016) Phase transitions and sample complexity in Bayes-optimal matrix factorization. IEEE Transactions on Information Theory, 62, 4228–4265.
Kamilov, U.S., Rangan, S., Fletcher, A.K. & Unser, M. (2014) Approximate message passing with consistent parameter estimation and applications to sparse learning. *IEEE Transactions on Information Theory*, 60, 2969–2985.

Kiefer, J. & Wolfowitz, J. (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 27, 887–906.

Knowles, A. & Yin, J. (2013) The isotropic semicircle law and deformation of Wigner matrices. *Communications on Pure and Applied Mathematics*, 66, 1663–1749.

Knowles, A. & Yin, J. (2014) The outliers of a deformed Wigner matrix. *The Annals of Probability*, 42, 1980–2031.

Koenker, R. & Mizera, I. (2014) Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *Journal of the American Statistical Association*, 109, 674–685.

Krzakala, F., Mézard, M., Sausset, F., Sun, Y. & Zdeborová, L. (2012) Statistical-physics-based reconstruction in compressed sensing. *Physical Review X*, 2, 021005.

Laird, N. (1978) Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73, 805–811.

Lashkari, D. & Golland, P. (2008) Convex clustering with exemplar-based models. In: *Advances in Neural Information Processing Systems*, 825–832.

Ledoit, O. & Wolf, M. (2012) Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40, 1024–1060.

Lelarge, M. & Mila, L. (2019) Fundamental limits of symmetric low-rank matrix estimation. *Probability Theory and Related Fields*, 173, 859–929.

Lesieur, T., Krzakala, F. & Zdeborová, L. (2015a) MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel. In: 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), 680–687. IEEE.

Lesieur, T., Krzakala, F. & Zdeborová, L. (2015b) Phase transitions in sparse PCA. In: 2015 IEEE International Symposium on Information Theory (ISIT), 1635–1639. IEEE.

Lindsay, B.G. (1983a) The geometry of mixture likelihoods: a general theory. *The Annals of Statistics*, 11, 86–94.

Lindsay, B.G. (1983b) The geometry of mixture likelihoods, part II: the exponential family. *The Annals of Statistics*, 11, 783–792.

Lindsay, B.G. (1995) *Mixture models: theory, geometry and applications*. Institute of Mathematical Statistics.

https://www.jstor.org/stable/4153184

Lindsay, B. G. & Roeder, K. (1993) Uniqueness of estimation and identifiability in mixture models. *Canadian Journal of Statistics*, 21, 139–147.

Lu, A.Y. (2002) Sparse principal component analysis for functional data. Ph.D. thesis, Stanford University.

Ma, Z. (2013) Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41, 772–801.

Matsushita, R. & Tanaka, T. (2013) Low-rank matrix reconstruction and clustering via approximate message passing. In: *Advances in Neural Information Processing Systems*, 917–925.

Mestre, X. (2008) On the asymptotic behavior of the sample estimates of eigenvalues and eigenvectors of covariance matrices. *IEEE Transactions on Signal Processing*, 56, 5353–5368.

Miolane, L. (2017) Fundamental limits of low-rank matrix estimation: the non-symmetric case. *arXiv preprint arXiv:1702.00473*.

Montanari, A. & Richard, E. (2015) Non-negative principal component analysis: message passing algorithms and sharp asymptotics. *IEEE Transactions on Information Theory*, 62, 1458–1484.

Montanari, A. & Venkataramanan, R. (2021) Estimation of low-rank matrices via approximate message passing. *The Annals of Statistics*, 49, 321–345.

Nadakuditi, R.R. (2014) Optshrink: an algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage. *IEEE Transactions on Information Theory*, 60, 3002–3018.

Nadler, B. (2008) Finite sample approximation results for principal component analysis: a matrix perturbation approach. *The Annals of Statistics*, 36, 2791–2817.

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

Polyanskiy, Y. & Wu, Y. (2020) Self-regularizing property of nonparametric maximum likelihood estimator in mixture models. *arXiv preprint arXiv:2008.08244*. 
Rangan, S. (2011) Generalized approximate message passing for estimation with random linear mixing. In: *IEEE International Symposium on Information Theory Proceedings*, 2168–2172.

Rangan, S. & Fletcher, A.K. (2012) Iterative estimation of constrained rank-one matrices in noise. In: *IEEE International Symposium on Information Theory Proceedings*, 1246–1250.

Robbins, H. (1950) A generalization of the method of maximum likelihood: estimating a mixing distribution (abstract). *Annals of Mathematical Statistics*, 21, 314–315.

Robbins, H. (1956) An empirical Bayes approach to statistics. In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Saha, S. & Guntuboyina, A. (2020) On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising. *The Annals of Statistics*, 48, 738–762.

Shabalin, A.A. & Nobel, A.B. (2013) Reconstruction of a low-rank matrix in the presence of Gaussian noise. *Journal of Multivariate Analysis*, 118, 67–76.

Shen, D., Shen, H. & Marron, J.S. (2013) Consistency of sparse PCA in high dimension, low sample size contexts. *Journal of Multivariate Analysis*, 115, 317–333.

Simar, L. (1976) Maximum likelihood estimation of a compound Poisson process. *The Annals of Statistics*, 4, 1200–1209.

The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature*, 526, 68–74.

Van Loan, C.F. & Golub, G.H. (2013) *Matrix computations*. Baltimore: Johns Hopkins University Press, 4th edn.

Vila, J. & Schniter, P. (2011) Expectation-maximization Bernoulli-Gaussian approximate message passing. In: 2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), 799–803. IEEE.

Vila, J.P. & Schniter, P. (2013) Expectation-maximization Gaussian-mixture approximate message passing. *IEEE Transactions on Signal Processing*, 61, 4658–4672.

Vu, V.Q. Cho, J., Lei, J. & Rohe, K. (2013) Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In: *Advances in Neural Information Processing Systems*, 2670–2678.

Wang, W. & Fan, J. (2017) Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *The Annals of Statistics*, 45, 1342.

Wang, W. & Stephens, M. (2021) Empirical bayes matrix factorization. *Journal of Machine Learning Research*, 22, 1–40.

Yang, F. (2019) Edge universality of separable covariance matrices. *Electronic Journal of Probability*, 24, 1–57.

Zhang, C.-H. (2009) Generalized maximum likelihood estimation of normal mixture densities. *Statistica Sinica*, 19, 1297–1318.

Zou, H., Hastie, T. & Tibshirani, R. (2006) Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15, 265–286.

**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of the article at the publisher’s website.

**How to cite this article:** Zhong, X., Su, C. & Fan, Z. (2022) Empirical Bayes PCA in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(3), 853–878. Available from: https://doi.org/10.1111/rssb.12490