Topics in Empirical Bayes Method

Heqi Yin*

December 7, 2023

Contents

*This is report for independent study program.

1 EBM of Normal Means

1.1 Introduction

Empirical Bayes for the estimation is a common and important method. We introduce this concept by proposing a general maximum likelihood empirical Bayes (GMLEB) method for the estimation of a mean vector based on observations with *i.i.d.* normal errors [Jiang and Zhang, 2009]. This is important because normal errors have been considered to be the canonical model or motivating example in the development of empirical Bayes, admissibility, adaptive nonparametric regression, variable selection, multiple testing, and other areas in statistics. Since the observed data are often understood, represented or summarized as the sum of a signal vector and white noise, it also has significant practical relevance in statistical applications. And when it comes to estimating normal means by compounding, there are three main approaches. There is a first method called general empirical Bayes (EB)[Robbins, 1951],[Robbins, 1983], which relies on the empirical distribution of the unknowns to produce the same performance as oracle separable estimators.

1.2 Methods & Main Results

In this paper, they proposed a general maximum likelihood EB (GMLEB) in which first estimate the empirical distribution of the unknown means by the generalized maximum likelihood estimator (MLE) [Kiefer and Wolfowitz, 1956] and then plug the estimator into the oracle general EB rule. In other words, we treat the unknown means as *i.i.d.* variable, with a completely unknown common "prior" distribution (for the purpose of deriving the GMLEB, whether the unknowns are actually deterministic or random), estimate the nominal prior with the generalized MLE, and then use the Bayes rule for the estimated prior. The basic idea was discussed in the last paragraph of [Robbins, 1951] in order to derive solutions to compound decision problems, even though MLE was a concept without a parametric model at the time, and there hasn't been much research on using generalized MLE to estimate a nominal prior in compound estimation since then. The results in [Jiang and Zhang, 2009] shows that Greedier general EB reduces risk significantly over linear and threshold approaches for a wide range of unknown signal vectors for moderate and large samples by aiming at the lowest risk of all separable estimators. They also demonstrates that the GMLEB estimator has a risk within an infinitesimal fraction of the general EB benchmark when the risk is of the order $n^{-1}(\log n)^5$ or greater depending on the magnitude of the weak ℓ_p norm of the unknown means, $0 < p \leq \infty$. Finally, for moderately sparse and dense means [Zhang, 2005], this adaptive minimaxity result unifies and improves upon the adaptive minimaxity of threshold estimators for sparse means [Abramovich et al., 2006], [Donoho and Johnstone, 1995], [Johnstone and Silverman, 2004]. Through simulation experiments and algorithms, they demonstrate the excellent risk performance of the GMLEB for moderate samples.

1.3 Set Up

Let X_i be independent statistics with

$$X_i \sim \varphi(x - \theta_i) \sim N(\theta_i, 1), \quad i = 1, \ldots, n, \tag{1}$$

under a probability measure $P_{n,\boldsymbol{\theta}}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is an unknown signal vector. Our problem is to estimate $\boldsymbol{\theta}$ under the compound loss defined as:

$$L_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = n^{-1}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\theta}_i - \theta_i\right)^2 \tag{2}$$

for any given estimator $\widehat{\boldsymbol{\theta}} = \left(\widehat{\theta}_1, \ldots, \widehat{\theta}_n\right)$. Notice that the unknown means θ_i are assumed to be deterministic as in the standard compound decision theory [Robbins, 1951]. We use θ only with boldface as a deterministic mean vector in $\mathbb{R}^n$ or with subscripts as elements of θ. A random mean is denoted by ξ as in (2.3) below. We first describe the general and restricted EB:

1.4 Empirical Bayes

The compound estimation of a vector of deterministic normal means is closely related to the Bayes estimation of a single random mean. Recall that in this Bayes problem, we estimate an uninvariant random parameter ξ based on an uninvariate Y such that

$$Y \mid \xi \sim N(\xi, 1), \quad \xi \sim G, \quad \text{under } P_G \tag{3}$$

where the prior distribution is $G = G_n$. And the empirical distribution is:

$$G_n(u) = G_{n,\theta}(u) = \frac{1}{n} \sum_{i=1}^{n} I\left\{\theta_i \leq u\right\}. \tag{4}$$

where $_{n,\boldsymbol{\theta}}$ indicates dependence of distribution or probability upon n and the unknown deterministic vector $\boldsymbol{\theta}$. In the context of the ℓ_2 loss, the fundamental theorem of compound decisions [Robbins, 1951] states that a separable rule has a compound risk $\widehat{\boldsymbol{\theta}} = t(\mathbf{X})$ under the probability $P_{n,\boldsymbol{\theta}}$ in the multivariate model (1) is identical to the MSE of the same rule $\widehat{\xi} = t(Y)$ under the prior (4) in the univariate model (3):

$$E_{n,\boldsymbol{\theta}} L_n(t(\mathbf{X}), \boldsymbol{\theta}) = E_{G_n}(t(Y) - \xi)^2 \tag{5}$$

For any true or nominal priors G, the Bayes rule is:

$$t_G^* = \arg\min_t E_G(t(Y) - \xi)^2 = \frac{\int u\varphi(Y-u)G(du)}{\int \varphi(Y-u)G(du)} \tag{6}$$

and the minimum Bayes risk as

$$R^*(G) = E_G\left(t_G^*(Y) - \xi\right)^2 \tag{7}$$

where the minimum is taken over all Borel functions. Thus, the general EB benchmark is

$$R^*\left(G_n\right) = E_{n,\boldsymbol{\theta}} L_n\left(t_{G_n}^*(\mathbf{X}), \boldsymbol{\theta}\right) = \min_{t(\cdot)} E_{n,\boldsymbol{\theta}} L_n(t(\mathbf{X}), \boldsymbol{\theta}). \tag{8}$$

The general EB approach seeks procedures which approximate the Bayes rule $t_{G_n}^*(\mathbf{X})$ or approximately achieve the risk benchmark $R^*\left(G_n\right)$ in (8). Recall in [Zhang, 1997], the author proposed a minimax risk for (rectangle) kernel estimator, There is an asymptotically minimized Fourier general EB estimator, which approximates the general EB benchmark of (8) uniformly for sparse and dense signals(by using some constraints of the distribution decay rate), provided that the oracle Bayes risk is $n^{-\frac{1}{2}}(\log n)^{\frac{3}{2}}$ or greater [Zhang, 1997]. Indeed, the Fourier general EB requires selection of certain tuning parameters (like the bandwidth) and its proven theoretical properties are not completely satisfying. Thus, we then introduce another investigation.

1.5 GMLEB

In this subsection, we introduce the general maximum likelihood empirical Bayes for the estimation of a mean vector as [Zhang, 1997]. The GMLEB method replaces the unknown prior G_n of the oracle rule $t^*_{G_n}$ by its generalized MLE [Kiefer and Wolfowitz, 1956]

$$\widehat{G}_n = \widehat{G}_n(\cdot; \mathbf{X}) = \arg\max_{G \in \mathscr{G}} \prod_{i=1}^{n} f_G(X_i), \tag{9}$$

where $\mathscr{G}$ is the family of all distribution functions and f_G is the density

$$f_G(x) = \int \varphi(x-u) G(du) \tag{10}$$

of the normal location mixture by distribution G. The estimator in (9) is called the generalized MLE and can be typically solved by iterative algorithms. Thus, the generalized MLE is any solution of

$$\widehat{G}_n \in \mathscr{G}, \quad \prod_{i=1}^{n} f_{\widehat{G}_n}(X_i) \geq q_n \sup_{G \in \mathscr{G}} \prod_{i=1}^{n} f_G(X_i) \tag{11}$$

with $q_n = \left(e\sqrt{2\pi}/n^2\right) \wedge 1$. Specifically, the GMLEB estimator is defined as

$$\widehat{\boldsymbol{\theta}} = t^*_{\widehat{G}_n}(\mathbf{X}) \quad \text{or equivalently} \quad \widehat{\theta}_i = t^*_{\widehat{G}_n}(X_i), \quad i = 1, \ldots, n, \tag{12}$$

where t^*_G is the Bayes rule in (6) and $\widehat{G}_n$ is any approximate generalized MLE (11) for the nominal prior in (4). Clearly, the GMLEB estimator (12) is nonparametric and does not require any restriction for function class like Sobolev Class, regularization, bandwidth selection or other forms of tuning. As a result, the GMLEB is location equivariant

$$t^*_{\widehat{G}_n(\cdot; \mathbf{X}+c\mathbf{e})}(\mathbf{X} + c\mathbf{e}) = t^*_{\widehat{G}_n(\cdot; \mathbf{X})}(\mathbf{X}) + c\mathbf{e} \tag{13}$$

for all real c, where $\mathbf{e} = (1, \ldots, 1) \in \mathbb{R}^n$. This is due to the location equivariance of the generalized MLE: $\widehat{G}_n(x; \mathbf{X} + c\mathbf{e}) = \widehat{G}_n(x - c; \mathbf{X})$. Compared with the Fourier & Wavelet general EB estimators [Zhang, 1997] and [Zhang, 2005], the GMLEB (12) is more appealing since the function $t^*_{\widehat{G}_n}(x)$ of x enjoys all analytical properties of Bayes rules: monotonicity, infinite differentiability and more. However, the GMLEB is much harder to be computational tractable than the Fourier general EB. We then discuss the computational problems.

1.6 Computation of the GMLEB

It is well-known that MLE might be hard to compute. Here we consider discrete approximation of (9) by Carathéodory's theorem [Carathéodory, 1911]. It has shown that there exists a discrete solution of (9) with no more than $n+1$ support points. The discrete approximate generalized MLE $\widehat{G}_n$ with m support points can be written as

$$\widehat{G}_n = \sum_{j=1}^{m} \widehat{w}_j \delta_{u_j}, \quad \widehat{w}_j \geq 0, \quad \sum_{j=1}^{m} \widehat{w}_j = 1, \tag{14}$$

where δ_u is the probability distribution giving its entire mass to u. Given (14), the GMLEB estimator can be easily computed as

$$\widehat{\theta}_i = t^*_{\widehat{G}_n}(X_i) = \frac{\sum_{j=1}^m u_j \varphi(X_i - u_j)\,\widehat{w}_j}{\sum_{j=1}^m \varphi(X_i - u_j)\,\widehat{w}_j}, \tag{15}$$

since $t^*_C(x)$ is the conditional expectation as in (6). Several algorithms can be used to solve (11), but all rely on iterative approximations. Due to the monotonicity of $\varphi(t)$ in t^2, the generalized MLE (9) puts all its mass in the interval $I_0 = [\min_{1\le i\le n} X_i, \max_{1\le i\le n} X_i]$. Given a fine grid $\{u_j\}$ in I_0, the EM-algorithm can be used [Dempster et al., 1977], [Vardi and Lee, 1993]

$$\widehat{w}_j^{(k)} = \frac{1}{n}\sum_{i=1}^n \frac{\widehat{w}_j^{(k-1)}\varphi(X_i - u_j)}{\sum_{\ell=1}^m \widehat{w}_\ell^{(k-1)}\varphi(X_i - u_\ell)} \tag{16}$$

which optimizes the weights $\{\widehat{w}_j\}$. Notice that they [Jiang and Zhang, 2009] also provide a statistical criterion on $\{u_j\}$ and an EM-stopping rule to guarantee (11).
This work can be see an improvement of [Zhang, 1997]. Because in order to deal with the regime of low density, some restrictions are made in the original paper.

2 EB PCA in High Dimensions

2.1 Introduction

Based on the discussion above and [Saha and Guntuboyina, 2020], where they studied the rate of convergence of the NPMLE (Nonparametric Maximum Likelihood Estimation) and associated empirical Bayes estimator, the following discussion on Empirical Bayes PCA in high dimensions [Zhong et al., 2022] draw on these techniques. In this paper, they described an EB-PCA procedure for performing PCA in high dimensions, which couples classical empirical Bayes ideas with high-dimensional asymptotic theory.
It is well-known that principal components analysis (PCA) is a widely used tool for dimensionality reduction. However, the sample principal components (PCs) may exhibit high-dimensional noise phenomena when the dimension of the data is greater than or comparable to the number of available data samples.[Lu, 2002], [Johnstone and Lu, 2009] Thus, it is worthy to discuss a new regime of PCA, which can reduce this noise by estimating a joint prior distribution for the principal components. To summarize, In their method called EB-PCA, the empirical Bayes estimation method is based on a classical non-parametric maximum likelihood estimator called Kiefer–Wolfowitz, distributions for sample PCs are derived from random matrix theory, and iterative refinement is performed by using approximate message passing (AMP). And they conclude that 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. They finally showed that in both simulations and quantitative benchmarks derived from the 1000 Genomes Project and the International HapMap Project, EB-PCA significantly outperforms PCA when there is a strong prior structure.

2.2 Main Idea of EB-PCA

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

$$\mathbf{Y} = \frac{s}{n}\cdot \mathbf{u}\mathbf{v}^\top + \mathbf{W} \in \mathbb{R}^{n\times d} \tag{17}$$

where $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^d$ are the left and right true PCs of interest, with associated signal strength $s > 0$, and $\mathbf{W} \in \mathbb{R}^{n \times d}$ is *i.i.d.* Gaussian observational noise. There are three main ideas in EB-PCA, each of which has been thoroughly studied before:

Kiefer–Wolfowitz NPMLE The model considered by [Kiefer and Wolfowitz, 1956] is a classical compound decision problem of estimating $\boldsymbol{\theta} \in \mathbb{R}^n$ from a Gaussian observation vector $\mathbf{x} \sim \mathcal{N}(\mu \cdot \boldsymbol{\theta}, \sigma^2 \cdot \mathrm{Id}_{n \times n})$, for two known scalar parameters $\mu, \sigma^2 > 0$. An empirical Bayes paradigm first proposes a prior distribution π_* for the coordinates of θ, then estimates π_* by an estimator π based on the marginal density of the observed coordinates of $\mathbf{x}$, and finally applies Bayes's rule defined by π to denoise $\mathbf{x}$ and obtain the estimate of $\boldsymbol{\theta}$. In [Kiefer and Wolfowitz, 1956], they proposed estimating π_* by the non-parametric maximum likelihood estimator (NPMLE) that maximizes the likelihood of $\mathbf{x}$ over all prior probability distributions π on the real line. It has shown that such a maximizer π exists with discrete and finite support.

$$\theta\left(\mathbf{x} \mid \mu, \sigma^2, \pi\right) = \mathbb{E}_\pi[\boldsymbol{\theta} \mid \mathbf{x}] \tag{18}$$

the empirical Bayes posterior mean estimate of θ using this estimated prior π.

Random matrix asymptotics for sample PCs Recall in model (17), the leading left- and right-singular vectors $\mathbf{f} \in \mathbb{R}^n$ and $\mathbf{g} \in \mathbb{R}^d$ of $\mathbf{Y}$ as the sample PCs. Some previous works [Baik et al., 2005], [Benaych-Georges and Nadakuditi, 2012], [Nadler, 2008], [Paul, 2007] have shown that the asymptotic error of the sample PCs $(\mathbf{f}, \mathbf{g})$ for the true PCs $(\mathbf{u}, \mathbf{v})$ when $n, d \to \infty$ simultaneously such that $d/n \to \gamma \in (0, \infty)$. This work showed that in this high-dimensional limit,

$$\langle \mathbf{f}, \mathbf{u} \rangle \to \bar{\mu}_* \equiv \bar{\mu}_*(s, \gamma), \quad \langle \mathbf{g}, \mathbf{v} \rangle \to \mu_* \equiv \mu_*(s, \gamma) \tag{19}$$

for two inner products $\mu_*, \bar{\mu}_* \in [0, 1)$ that depend only on the signal strength s and the dimension ratio γ. For values of s exceeding a critical phase transition threshold, denoted as $s_*(\gamma)$, a notable phenomenon occurs with the singular values of the matrix $\mathbf{Y}$. Specifically, its largest singular value distinctly separates from the main cluster formed by the other singular values. Concurrently, the inner products $\bar{\mu}_*$ and μ_* acquire non-zero positive values. Moreover, the vector $\mathbf{g}$ demonstrates a behavior that closely resembles an entrywise Gaussian distribution, which can be mathematically expressed as:

$$\mathbf{g} \sim \mathcal{N}\left(\mu_* \cdot \mathbf{v}, \sigma_*^2 \cdot \mathbf{I}_{d \times d}\right), \quad \text{where } \sigma_*^2 = 1 - \mu_*^2 \tag{20}$$

In a similar vein, the vectors $\mathbf{f}$ and $\mathbf{u}$ adhere to a comparable approximation pattern. This observation establishes a link to the previously discussed compound decision scenario. In this context, EB-PCA functions by estimating the parameters (μ_*, σ_*^2) through the calculation of s, followed by employing the Kiefer-Wolfowitz NPMLE method to derive an empirical Bayes estimation, denoted as $\hat{\mathbf{v}}$, for the vector $\mathbf{v}$.

Iterative refinement via AMP If this estimate $\hat{\mathbf{v}}$ is more accurate than the original sample PC $\mathbf{g}$ for $\mathbf{v}$, then we expect $\mathbf{Y}\hat{\mathbf{v}}$ to be more accurate than $\mathbf{Y}\mathbf{g} \propto \mathbf{f}$ for $\mathbf{u}$. This suggests that empirical Bayes denoising should be applied to $\mathbf{Y}\hat{\mathbf{v}}$ instead of $\mathbf{f}$ to estimate $\mathbf{u}$, and leads to an iterative idea (Wang & Stephens, 2021) of initializing $\mathbf{g}^0 = \mathbf{g}$ and computing

$$\mathbf{v}^t = \theta\left(\mathbf{g}^t \mid \mu_t, \sigma_t^2, \pi_t\right), \quad \mathbf{f}^t = \mathbf{Y}\mathbf{v}^t, \tag{21}$$

$$\mathbf{u}^t = \theta\left(\mathbf{f}^t \mid \bar{\mu}_t, \bar{\sigma}_t^2, \bar{\pi}_t\right), \quad \mathbf{g}^{t+1} = \mathbf{Y}^\top \mathbf{u}^t, \tag{22}$$

$$\tag{23}$$

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

this procedure does not ensure that $(\mathbf{f}^t, \mathbf{g}^t)$ 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],

$$\mathbf{v}^t = \theta\left(\mathbf{g}^t \mid \mu_t, \sigma_t^2, \pi_t\right), \mathbf{f}^t = \mathbf{Y}\mathbf{v}^t - b_t\mathbf{u}^{t-1} \tag{24}$$

$$\mathbf{u}^t = \theta\left(\mathbf{f}^t \mid \bar{\mu}_t, \bar{\sigma}_t^2, \bar{\pi}_t\right), \mathbf{g}^{t+1} = \mathbf{Y}^\top\mathbf{u}^t - \bar{b}_t\mathbf{v}^t \tag{25}$$

The Onsager corrections $b_t\mathbf{u}^{t-1}$ and $\bar{b}_t\mathbf{v}^t$ are defined so as to remove a bias of $(\mathbf{f}^t, \mathbf{g}^t)$ in the directions of $(\mathbf{u}^{t-1}, \mathbf{v}^t)$ and restore the entrywise Gaussian approximations.

2.3 EB-PCA Method

The following assumptions are necessary for theoretical guarantees for EB-PCA method. The details will be omitted here, however.

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 $\mathbf{W} \in \mathbb{R}^{n\times d}$ has entries $w_{ij} \overset{iid}{\sim} \mathcal{N}(0, 1/n)$.*

Assumption 2. *$k, s_1, \ldots, s_k$, and $\gamma \equiv d/n$ remain constant, where $s_1 > \cdots > s_k > s_*(\gamma) \equiv \gamma^{-1/4}$.*

Assumption 3. *$\mathbf{U} \overset{W_2}{\to} \bar{\pi}_*$ and $\mathbf{V} \overset{W_2}{\to} \pi_*$ for two distributions $\bar{\pi}_*, \pi_* \in \mathcal{P}$ that satisfy the normalizations, for all $1 \le i \ne j \le k$,*

$$\mathbb{E}_{U\sim\bar{\pi}_*}\left[U_i^2\right] = 1, \quad \mathbb{E}_{U\sim\bar{\pi}_*}\left[U_iU_j\right] = 0, \quad \mathbb{E}_{V\sim\pi_*}\left[V_i^2\right] = 1, \quad \mathbb{E}_{V\sim\pi_*}\left[V_iV_j\right] = 0 \tag{26}$$

Assumption 4. *For any non-singular $M_* \in \mathbb{R}^{k\times k}$, symmetric positive-definite $\Sigma_* \in \mathbb{R}^{k\times k}$, and $\pi_* \in \mathcal{P}$, there is a weakly open neighbourhood O of π_* such that $\theta(x \mid M_*, \Sigma_*, \pi)$ is Lipschitz in x uniformly over $\pi \in O$.*

2.4 Set Up

The EB-PCA algorithm is derived in generalized rank- k version of the model in (17),

$$\mathbf{Y} = \frac{1}{n}\cdot\mathbf{U}\mathbf{S}\mathbf{V}^\top + \mathbf{W} = \sum_{i=1}^k \frac{S_i}{n}\cdot\mathbf{u}_i\mathbf{v}_i^\top + \mathbf{W} \in \mathbb{R}^{n\times d} \tag{27}$$

The columns of $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_k) \in \mathbb{R}^{n\times k}$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_k) \in \mathbb{R}^{d\times k}$ are k left and right principal components of interest, and $S = \mathrm{diag}(s_1, \ldots, s_k) \in \mathbb{R}^{k\times k}$ contains the signal strengths of these PCs. $\mathbf{W} \in \mathbb{R}^{n\times d}$ is observational noise, which we assume has entries $w_{ij} \overset{iid}{\sim} \mathcal{N}(0, 1/n)$. As for the sample PCs, we can write the best rank- k approximation for $\mathbf{Y}$ as following

$$\frac{1}{n}\cdot\mathbf{F}\Lambda\mathbf{G}^\top = \sum_{i=1}^k \frac{\lambda_i}{n}\cdot\mathbf{f}_i\mathbf{g}_i^\top. \tag{28}$$

And the columns of $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_k) \in \mathbb{R}^{n\times k}$ and $\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_k) \in \mathbb{R}^{d\times k}$ are the top k left and right singular vectors of $\mathbf{Y}$, normalized with a sign convention so that for all $1 \le i \ne j \le k$,

$$d^{-1}\left\|\mathbf{g}_i\right\|^2 = n^{-1}\left\|\mathbf{f}_i\right\|^2 = 1, \quad \mathbf{u}_i^\top\mathbf{f}_i \ge 0, \quad \mathbf{v}_i^\top\mathbf{g}_i \ge 0, \quad d^{-1}\mathbf{g}_i^\top\mathbf{g}_j = n^{-1}\mathbf{f}_i^\top\mathbf{f}_j = 0, \tag{29}$$

where we set $\Lambda = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_k)$. Then the k largest singular values of $\mathbf{Y}$ are given by $\sqrt{\gamma}\lambda_1 \geq \cdots \geq \sqrt{\gamma}\lambda_k$. By the above definitions of the model, the following phase transition occurs for the leading k sample singular values and singular vectors of Y [Baik et al., 2005],
[Benaych-Georges and Nadakuditi, 2012], [Paul, 2007]: setting $s_*(\gamma) = \gamma^{-1/4}$, for supercritical PCs such that $s_i > s_*(\gamma)$, we have

$$\lim_{n,d\to\infty} \sqrt{\gamma} \cdot \lambda_i > \lambda_+, \quad \lim_{n,d\to\infty} n^{-1}\mathbf{f}_i^\top \mathbf{u}_i > 0, \quad \lim_{n,d\to\infty} d^{-1}\mathbf{g}_i^\top \mathbf{v}_i > 0, \tag{30}$$

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

$$\lim_{n,d\to\infty} \sqrt{\gamma} \cdot \lambda_i = \lambda_+, \quad \lim_{n,d\to\infty} n^{-1}\mathbf{f}_i^\top \mathbf{u}_i = 0, \quad \lim_{n,d\to\infty} d^{-1}\mathbf{g}_i^\top \mathbf{v}_i = 0. \tag{31}$$

Therefore, 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

$$s_i > s_*(\gamma) \quad \text{for all} \quad i = 1, \ldots, k, \tag{32}$$

i.e. all k of the leading PCs are supercritical. In EB-PCA, only the supercritical PCs that align with the truth are considered when there are both supercritical and subcritical PCs.

2.5 Empirical Bayes

Recall that the classical compound decision problem, let π_* be a probability distribution on $\mathbb{R}^k$. For two given matrices $M, \Sigma \in \mathbb{R}^{k\times k}$ where Σ is symmetric positive-definite, we still consider the compound decision model

$$\Theta \sim \pi_*, \quad X \mid \Theta \sim \mathcal{N}(M \cdot \Theta, \Sigma), \tag{33}$$

for $\Theta, X \in \mathbb{R}^k$. Using X as our input, we calculate the Bayes posterior mean of Θ

$$\theta\left(X \mid M, \Sigma, \pi_*\right) = \mathbb{E}_{\pi_*}[\Theta \mid X] \tag{34}$$

Suppose now that π_* 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 (33), arranged into a matrix $\mathbf{X} \in \mathbb{R}^{n\times k}$, consider the maximum likelihood estimator

$$\pi = \mathrm{MLE}(\mathbf{X} \mid M, \Sigma, \mathcal{P}) \tag{35}$$

$$\equiv \arg\max_{\pi\in\mathcal{P}} \prod_{i=1}^{n} \int \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \cdot \exp\left(-\frac{\left(x_i - M\cdot\theta_i\right)^\top \Sigma^{-1}\left(x_i - M\cdot\theta_i\right)}{2}\right) \mathrm{d}\pi\left(\theta_i\right). \tag{36}$$

This integral represents the marginal Gaussian mixture density for x_i as described in the model of (33). The notation here explicitly indicates the dependence of π on the chosen prior class, denoted as $\mathcal{P}$. Our focus is primarily on non-parametric classes $\mathcal{P}$, where π is derived as a non-parametric maximum likelihood estimate (NPMLE) of π_*. We assemble $\theta_1, \ldots, \theta_n$ into the rows of a matrix $\mathbf{\Theta} \in \mathbb{R}^{n\times k}$. Accordingly, the model for $\mathbf{X}$ is expressed as follows:

$$\mathbf{X} = \mathbf{\Theta} M^\top + \mathbf{Z}\Sigma^{1/2}, \text{ where } \mathbf{Z} \in \mathbb{R}^{n\times k} \text{ comprises i.i.d. } \mathcal{N}(0,1) \text{ entries.} \tag{37}$$

The NPMLE π facilitates an empirical Bayes estimate of $\mathbf{\Theta}$, achieved by applying the posterior mean function $\theta(\cdot)$ to each row of $\mathbf{X}$, based on the estimated prior π. This is denoted as:

$$\theta(\mathbf{X} \mid M, \Sigma, \pi) = \mathbb{E}_\pi[\mathbf{\Theta} \mid \mathbf{X}] \tag{38}$$

2.6 Iterative refinement using AMP

Recall the initial denoising method of the sample PCs.

Lemma 1. *Under Assumptions, for each $i = 1, \ldots, k$ and some choices of signs for $\mathbf{f}_i$ and $\mathbf{g}_i$, a.s. as $n, d \to \infty$,*

$$\sqrt{\gamma} \cdot \lambda_i \to \sqrt{\left(\gamma s_i^2 + 1\right)\left(s_i^2 + 1\right)/s_i^2}, \tag{39}$$

$$n^{-1}\mathbf{f}_i^\top \mathbf{u}_i \to \bar{\mu}_{*,i}, \quad d^{-1}\mathbf{g}_i^\top \mathbf{v}_i \to \mu_{*,i} \tag{40}$$

for the values $\bar{\mu}_{*,i}$ and $\mu_{*,i}$ defined in (2.10). Furthermore, $n^{-1}\mathbf{f}_i^\top \mathbf{u}_j \to 0$ and $d^{-1}\mathbf{g}_i^\top \mathbf{v}_j \to 0$ a.s. for all $j \in \{1, \ldots, k\} \backslash \{i\}$.

Proof. Recall our model is

$$\mathbf{Y} = \frac{1}{n}\mathbf{U}S\mathbf{V}^\top + \mathbf{W}, \tag{41}$$

where $\tilde{\mathbf{U}} = (\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_k) \in \mathbb{R}^{n \times k}$ stands for the Gram-Schmidt orthogonalization of $\mathbf{U}$, with columns scaled such that $n^{-1}\left\|\tilde{\mathbf{u}}_i\right\|^2 = 1$. And our thrid assumption implies that $n^{-1}\mathbf{U}^\top\mathbf{U} \to \mathrm{Id}_{k\times k}$, enabling Gram-Schmidt's method to be verified

$$n^{-1}\|\mathbf{U} - \tilde{\mathbf{U}}\|_F^2 \to 0 \tag{42}$$

Similarly, letting $\tilde{\mathbf{V}} = (\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_k) \in \mathbb{R}^{d \times k}$ be the Gram-Schmidt orthogonalization of $\mathbf{V}$ scaled such that $d^{-1}\left\|\tilde{\mathbf{v}}_i\right\|^2 = 1$, we then have

$$d^{-1}\|\mathbf{V} - \tilde{\mathbf{V}}\|_F^2 \to 0. \tag{43}$$

Define $\tilde{\mathbf{Y}} = n^{-1}\tilde{\mathbf{U}}S\tilde{\mathbf{V}}^\top + \mathbf{W}$, and denote its leading rank- k singular component as

$$n^{-1}\tilde{\mathbf{F}}\tilde{\Lambda}\tilde{\mathbf{G}}^\top = \sum_{i=1}^{n} \frac{\tilde{\lambda}_i}{n} \cdot \tilde{\mathbf{f}}_i\tilde{\mathbf{g}}_i^\top \tag{44}$$

Then (42) and (43) together imply $\|\tilde{\mathbf{Y}} - \mathbf{Y}\| \to 0$ in operator norm, so that by the condition of distinct singular values in second assumption and the Weyl and Davis-Kahan inequalities,

$$\|S - \tilde{S}\| \to 0, \quad n^{-1}\|\mathbf{F} - \tilde{\mathbf{F}}\|_F^2 \to 0, \quad d^{-1}\|\mathbf{G} - \tilde{\mathbf{G}}\|_F^2 \to 0. \tag{45}$$

□

In reference to the above lemma, we define the matrices as follows:

$$\begin{array}{ll} \bar{M}_* = \mathrm{diag}(\bar{\mu}_{*,1}, \ldots, \bar{\mu}_{*,k}), & M_* = \mathrm{diag}(\mu_{*,1}, \ldots, \mu_{*,k}), \\ \bar{\Sigma}_* = \mathrm{diag}(\bar{\sigma}_{*,1}^2, \ldots, \bar{\sigma}_{*,k}^2), & \Sigma_* = \mathrm{diag}(\sigma_{*,1}^2, \ldots, \sigma_{*,k}^2), \end{array} \tag{46}$$

As a result, for large values of n and d, the matrices $\mathbf{F} \in \mathbb{R}^{n\times k}$ and $\mathbf{G} \in \mathbb{R}^{d\times k}$ can be approximated by Gaussian distributions:

$$\mathbf{F} \approx \mathbf{U}\bar{M}_*^\top + \overline{\mathbf{Z}}\bar{\Sigma}_*^{1/2}, \quad \mathbf{G} \approx \mathbf{V}M_*^\top + \mathbf{Z}\Sigma_*^{1/2}, \tag{47}$$

where $\overline{\mathbf{Z}}, \mathbf{Z} \in \mathbb{R}^{n\times k}$ consist of i.i.d. $\mathcal{N}(0,1)$ entries. This relationship connects the behavior of the sample PCs $\mathbf{F}$ and $\mathbf{G}$ to our multivariate compound decision model. Since the true matrices $\bar{M}_*, M_*, \bar{\Sigma}_*, \Sigma_*$ are not known, we use consistent estimations of them to derive empirical Bayes

estimators for $\mathbf{U}$ and $\mathbf{V}$. Notably, Equation (12) suggests that each s_i^2 can be consistently estimated by:

$$\hat{s}_i^2 = \frac{\gamma\lambda_i^2 - (1+\gamma) + \sqrt{(\gamma\lambda_i^2 - (1+\gamma))^2 - 4\gamma}}{2\gamma} \tag{48}$$

These estimations facilitate the creation of plug-in estimators $\bar{M}, M, \bar{\Sigma}, \Sigma$ for $\bar{M}_*, M_*, \bar{\Sigma}_*, \Sigma_*$, substituting $\hat{s}_i$ for s_i as indicated in Equation (12). Consequently, the initial empirical Bayes estimates of $\mathbf{U}$ and $\mathbf{V}$ are given by:

$$\begin{aligned} \bar{\pi} &= \mathrm{MLE}(\mathbf{F} \mid \bar{M}, \bar{\Sigma}, \mathcal{P}), \quad \hat{\mathbf{U}} = \theta(\mathbf{F} \mid \bar{M}, \bar{\Sigma}, \bar{\pi}), \\ \pi &= \mathrm{MLE}(\mathbf{G} \mid M, \Sigma, \mathcal{P}), \quad \hat{\mathbf{V}} = \theta(\mathbf{G} \mid M, \Sigma, \pi). \end{aligned} \tag{49}$$

We then outline the process of iterative refinement using an Approximate Message Passing (AMP) algorithm, as detailed in [Montanari and Venkataramanan, 2021]. The refinement process can start with either the estimate of $\mathbf{U}$ or $\mathbf{V}$; in this description, we initiate with $\mathbf{V}$. The algorithm is structured as follows: consider two sequences of Lipschitz functions, $u_1, u_2, \ldots : \mathbb{R}^k \to \mathbb{R}^k$ and $v_1, v_2, \ldots : \mathbb{R}^k \to \mathbb{R}^k$. Begin by setting $\mathbf{G}^0$ equal to $\mathbf{G}$, representing the right sample Principal Components (PCs). Then, for iterations $t = 0, 1, 2, \ldots$, compute:

$$\mathbf{V}^t = v_t\left(\mathbf{G}^t\right), \quad \mathbf{F}^t = \mathbf{Y}\mathbf{V}^t - \mathbf{U}^{t-1}\gamma\left\langle \mathrm{d}v_t\left(\mathbf{G}^t\right)\right\rangle^\top \quad \mathbf{U}^t = u_t\left(\mathbf{F}^t\right), \quad \mathbf{G}^{t+1} = \mathbf{Y}^\top\mathbf{U}^t - \mathbf{V}^t\left\langle \mathrm{d}u_t\left(\mathbf{F}^t\right)\right\rangle^\top \tag{50}$$

In this context, $u_t\left(\mathbf{F}^t\right) \in \mathbb{R}^{n\times k}$ and $v_t\left(\mathbf{G}^t\right) \in \mathbb{R}^{d\times k}$ represent the row-wise application of u_t and v_t to $\mathbf{F}^t$ and $\mathbf{G}^t$, respectively. The notations $\mathrm{d}u_t : \mathbb{R}^k \to \mathbb{R}^{k\times k}$ and $\mathrm{d}v_t : \mathbb{R}^k \to \mathbb{R}^{k\times k}$ refer to the Jacobian matrices of these functions. The terms $\langle \mathrm{d}u_t\left(\mathbf{F}^t\right)\rangle \in \mathbb{R}^{k\times k}$ and $\langle \mathrm{d}v_t\left(\mathbf{G}^t\right)\rangle \in \mathbb{R}^{k\times k}$ denote the average of $\mathrm{d}u_t$ and $\mathrm{d}v_t$ over the rows of $\mathbf{F}^t$ and $\mathbf{G}^t$. Under the model of (17), Gaussian approximations continue to hold for $\mathbf{F}^t$ and $\mathbf{G}^t$ across iterations, where

$$\mathbf{F}^t \approx \mathbf{U}\bar{M}_{*,t}^\top + \overline{\mathbf{Z}}\bar{\Sigma}_{*,t}^{1/2}, \quad \mathbf{G}^t \approx \mathbf{V}M_{*,t}^\top + \mathbf{Z}\Sigma_{*,t}^{1/2} \tag{51}$$

The matrices $\bar{M}*, t, M, t, \bar{\Sigma}_{,t}, \Sigma_{,t}$ are deterministic and define the parameters for the compound decision model relevant to each iteration. In divergence from the initial conditions presented in (39), these matrices become non-diagonal in subsequent iterations, particularly when the prior is a general multivariate distribution over $\mathbb{R}^k$. The evolution of these matrices through iterations follows a state evolution process:

$$\begin{aligned} (M_{*,0}, \Sigma_{*,0}) &\mapsto \left(\bar{M}_{*,0}, \bar{\Sigma}_{*,0}\right) \mapsto (M_{*,1}, \Sigma_{*,1}) \\ &\mapsto \left(\bar{M}_{*,1}, \bar{\Sigma}_{*,1}\right) \mapsto \cdots \end{aligned} \tag{52}$$

This progression is determined by the initializations

$$(M_{*,0}, \Sigma_{*,0}) \equiv (M_*, \Sigma_*), \tag{53}$$

which describe the sample PCs $\mathbf{G}^0 \equiv \mathbf{G}$, and by the update equations

$$\bar{M}_{*,t} = \gamma\mathbb{E}\left[v_t\left(G_t\right)V^\top\right]S, \qquad \bar{\Sigma}_{*,t} = \gamma\mathbb{E}\left[v_t\left(G_t\right)v_t\left(G_t\right)^\top\right], \tag{54}$$

$$M_{*,t+1} = \mathbb{E}\left[u_t\left(F_t\right)U^\top\right]S, \qquad \Sigma_{*,t+1} = \mathbb{E}\left[u_t\left(F_t\right)u_t\left(F_t\right)^\top\right]. \tag{55}$$

Here, $S = \mathrm{diag}\left(s_1, \ldots, s_k\right)$ is the diagonal matrix of signal strengths as defined in (17). The expectations are computed over the random vectors

$$U \sim \bar{\pi}*, \quad F_t \left| U \sim \mathcal{N}\left(\bar{M}, t\cdot U, \bar{\Sigma}_{,t}\right), \quad V \sim \pi, \quad G_t\right| V \sim \mathcal{N}\left(M_{,t}\cdot V, \Sigma_{*,t}\right) \tag{56}$$

These distributions for F_t and G_t serve as approximations for the row-wise distributions of $\mathbf{F}^t$ and $\mathbf{G}^t$. If $\bar{\pi}_*, \pi_*$ and $\bar{M}_{*,t}, M_{*,t}, \bar{\Sigma}_{*,t}, \Sigma_{*,t}$ are all known, then applying this algorithm with the Bayes posterior mean functions, then

$$u_t(\mathbf{X}) = \theta\left(\mathbf{X} \mid \bar{M}_{*,t}, \bar{\Sigma}_{*,t}, \bar{\pi}_*\right), \quad v_t(\mathbf{X}) = \theta\left(\mathbf{X} \mid M_{*,t}, \Sigma_{*,t}, \pi_*\right), \tag{57}$$

implements an iterative variational Bayesian inference scheme [Montanari and Venkataramanan, 2021]. Which is called the oracle Bayes AMP algorithm. For these u_t, v_t, we have the identities $\mathbb{E}\left[u_t\left(F_t\right) U^\top\right] = \mathbb{E}\left[u_t\left(F_t\right) u_t\left(F_t\right)^\top\right]$ and $\mathbb{E}\left[v_t\left(G_t\right) V^\top\right] = \mathbb{E}\left[v_t\left(G_t\right) v_t\left(G_t\right)^\top\right]$, so (54) yields

$$\bar{M}_{*,t} = \bar{\Sigma}_{*,t} \cdot S, \quad M_{*,t+1} = \Sigma_{*,t+1} \cdot S \tag{58}$$

EB-PCA uses the posterior mean functions defined instead by NPMLEs of $\bar{\pi}_*$ and π_*, together with the empirical estimates $\bar{\Sigma}_t = n^{-1}\left(\mathbf{V}^t\right)^\top \mathbf{V}^t, \bar{M}_t = \bar{\Sigma}_t \hat{S}, \Sigma_{t+1} = n^{-1}\left(\mathbf{U}^t\right)^\top \mathbf{U}^t$, and $M_{t+1} = \Sigma_{t+1}\hat{S}$ where $\hat{S} = \operatorname{diag}\left(\hat{s}_1, \ldots, \hat{S}_k\right)$ is the estimate of S from (48). (For $\bar{\Sigma}_t$, 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 $\bar{M}_{*,t}, M_{*,t}, \bar{\Sigma}_{*,t}, \Sigma_{*,t}$. EB-PCA is initialized at the right sample $\mathbf{PCsG}^0 = \mathbf{G}$ and the plug-in estimates $(M_0, \Sigma_0) \equiv (M, \Sigma)$ from the preceding section. In particular, $\mathbf{V}^0$ is the initial empirical Bayes estimate for $\mathbf{V}$ based on $\mathbf{G}$ as previously described. We summarize the full EB-PCA method as the following Algorithm.

Algorithm 1 EB-PCA

Input: Data matrix $\mathbf{Y} \in \mathbb{R}^{n \times d}$, after normalized to have average entrywise noise variance $1/n$. Number of PCs k, number of AMP iterations T, prior class $\mathcal{P}$.
Initialization:
Let $\gamma = d/n$.
Let $\left(\sqrt{\gamma}\lambda_1, \ldots, \sqrt{\gamma}\lambda_k\right), \mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_k)$, and $\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_k)$ be the top k singular values and singular vectors of $\mathbf{Y}$, with $\|\mathbf{f}_i\|_2 = \sqrt{n}$ and $\|\mathbf{g}_i\|_2 = \sqrt{d}$.
Define $\hat{s}_i^2$ by (48).
Set $\sigma_i^2 = \left(1 + \gamma \hat{s}_i^2\right) / \left(\gamma \hat{s}_i^4 + \gamma \hat{s}_i^2\right), \mu_i^2 = 1 - \sigma_i^2, M_0 = \operatorname{diag}\left(\mu_1, \ldots, \mu_k\right)$, $\Sigma_0 = \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_k^2\right), \hat{S} = \operatorname{diag}\left(\hat{s}_1, \ldots, \hat{s}_k\right)$.
$\mathbf{G}^0 \leftarrow \mathbf{G}$ and $\mathbf{U}^{-1} \leftarrow \mathbf{F} \cdot \Sigma_0^{1/2}$
Iterative refinement:
for $t = 0, 1, 2, \ldots, T$ **do**
    // Denoise left PCs
    $\pi_t \leftarrow \operatorname{MLE}\left(\mathbf{G}^t \mid M_t, \Sigma_t, \mathcal{P}\right)$
    $\mathbf{V}^t \leftarrow \theta\left(\mathbf{G}^t \mid M_t, \Sigma_t, \pi_t\right)$
    $\mathbf{F}^t \leftarrow \mathbf{Y}\mathbf{V}^t - \mathbf{U}^{t-1} \cdot \gamma \left\langle \mathbf{d}\theta\left(\mathbf{G}^t \mid M_t, \Sigma_t, \pi_t\right)\right\rangle^\top$
    $\bar{\Sigma}_t \leftarrow \left(\mathbf{V}^t\right)^\top \mathbf{V}^t/n$ and $\bar{M}_t \leftarrow \bar{\Sigma}_t \cdot \hat{S}$
    // Denoise right PCs
    $\bar{\pi}_t \leftarrow \operatorname{MLE}\left(\mathbf{F}^t \mid \bar{M}_t, \bar{\Sigma}_t, \mathcal{P}\right)$
    $\mathbf{U}^t \leftarrow \theta\left(\mathbf{F}^t \mid \bar{M}_t, \bar{\Sigma}_t, \bar{\pi}_t\right)$
    $\mathbf{G}^{t+1} \leftarrow \mathbf{Y}^\top \mathbf{U}^t - \mathbf{V}^t \cdot \left\langle \mathrm{d}\theta\left(\mathbf{F}^t \mid \bar{M}_t, \bar{\Sigma}_t, \bar{\pi}_t\right)\right\rangle^\top$
    $\Sigma_{t+1} \leftarrow \left(\mathbf{U}^t\right)^\top \mathbf{U}^t/n$ and $M_{t+1} \leftarrow \Sigma_{t+1} \cdot \hat{S}$
Output: Final estimates $(\hat{\mathbf{U}}, \hat{S}, \hat{\mathbf{V}}) = \left(\mathbf{U}^T, \hat{S}, \mathbf{V}^T\right)$

2.7 Relation to Mean Field Approach to Empirical Bayes Estimation

[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 (17), this approximates the posterior law $p(\mathbf{u}, \mathbf{v} \mid \mathbf{Y})$ by a factorized form $\bar{q}(\mathbf{u})q(\mathbf{v}) = \prod_{i=1}^{n} \bar{q}_i(u_i) \prod_{j=1}^{d} q_j(v_j)$. The distributions $\bar{q}_i, q_j$ are chosen to minimize the Kullback-Leibler divergence $D_{\mathrm{KL}}(\bar{q}(\mathbf{u})q(\mathbf{v}) \| p(\mathbf{u}, \mathbf{v} \mid \mathbf{Y}))$ or, equivalently, to maximize the evidence lower bound (ELBO)

$$\mathcal{F}(\bar{q}, q) = \mathbb{E}_{\mathbf{u}\sim\bar{q},\mathbf{v}\sim q}[\log p(\mathbf{Y}, \mathbf{u}, \mathbf{v}) - \log \bar{q}(\mathbf{u})q(\mathbf{v})]. \tag{59}$$

As shown in [Wang and Stephens, 2021], their updating admits a simple formula

$$\mathbf{v}^t \equiv \mathbb{E}_{\mathbf{v}\sim q_t}[\mathbf{v}], \quad \bar{\sigma}_t^2 \equiv n^{-1}\mathbb{E}_{\mathbf{v}\sim q_t}\left[\|\mathbf{v}\|^2\right], \quad \mathbf{u}^t \equiv \mathbb{E}_{\mathbf{u}\sim\bar{q}_t}[\mathbf{u}], \quad \sigma_{t+1}^2 \equiv n^{-1}\mathbb{E}_{\mathbf{u}\sim\bar{q}_t}\left[\|\mathbf{u}\|^2\right], \tag{60}$$

which is very similar to the iterations of Algorithm 1 but does not incorporate the AMP Onsager correction terms. In high-dimensional settings, the Onsager correction terms are crucial for addressing the subtle dependencies present in the true posterior distributions of $u_1, \ldots, u_n$ and $v_1, \ldots, v_d$. These dependencies are not accounted for in the simplistic mean field approximation. Further exploration of this topic can be found in the works of [Ghorbani et al., 2019] and [Fan et al., 2021]. While the discrepancy between these methodologies diminishes when the signal-to-noise ratio s approaches infinity ($s \to \infty$), the differences are still significant in scenarios with a bounded signal-to-noise ratio. In particular, these differences become more evident for weaker signals, especially when nearing the phase transition threshold defined as $s_*(\gamma) = \gamma^{-1/4}$.
For more information and related to High Dimensional Linear Regression, see [Mukherjee et al., 2023].

References

[Abramovich et al., 2006] Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate.

[Baik et al., 2005] Baik, J., Ben Arous, G., and Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices.

[Benaych-Georges and Nadakuditi, 2012] Benaych-Georges, F. and 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.

[Carathéodory, 1911] Carathéodory, C. (1911). Über den variabilitätsbereich der fourier'schen konstanten von positiven harmonischen funktionen. *Rendiconti Del Circolo Matematico di Palermo (1884-1940)*, 32(1):193–217.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

[Donoho and Johnstone, 1995] Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association*, 90(432):1200–1224.

[Fan et al., 2021] Fan, Z., Mei, S., and Montanari, A. (2021). Tap free energy, spin glasses and variational inference.

[Ghorbani et al., 2019] Ghorbani, B., Javadi, H., and Montanari, A. (2019). An instability in variational inference for topic models. In *International conference on machine learning*, pages 2221–2231. PMLR.

[Jiang and Zhang, 2009] Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical bayes estimation of normal means.

[Johnstone and Lu, 2009] Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693.

[Johnstone and Silverman, 2004] Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences.

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

[Lu, 2002] Lu, A. Y. (2002). *Sparse principal component analysis for functional data*. Stanford University.

[Montanari and Venkataramanan, 2021] Montanari, A. and Venkataramanan, R. (2021). Estimation of low-rank matrices via approximate message passing.

[Mukherjee et al., 2023] Mukherjee, S., Sen, B., and Sen, S. (2023). A mean field approach to empirical bayes estimation in high-dimensional linear regression. *arXiv preprint arXiv:2309.16843*.

[Nadler, 2008] Nadler, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach.

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

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

[Robbins, 1951] Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, volume 2, pages 131–149. University of California Press.

[Robbins, 1983] Robbins, H. (1983). Some thoughts on empirical bayes estimation. *The Annals of Statistics*, pages 713–723.

[Saha and Guntuboyina, 2020] Saha, S. and Guntuboyina, A. (2020). On the nonparametric maximum likelihood estimator for gaussian location mixture densities with application to gaussian denoising.

[Vardi and Lee, 1993] Vardi, Y. and Lee, D. (1993). From image deblurring to optimal investments: Maximum likelihood solutions for positive linear inverse problems. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 55(3):569–598.

[Wang and Stephens, 2021] Wang, W. and Stephens, M. (2021). Empirical bayes matrix factorization. *The Journal of Machine Learning Research*, 22(1):5332–5371.

[Zhang, 1997] Zhang, C.-H. (1997). Empirical bayes and compound estimation of normal means. *Statistica Sinica*, 7(1):181–193.

[Zhang, 2005] Zhang, C.-H. (2005). General empirical bayes wavelet methods and exactly adaptive minimax estimation.

[Zhong et al., 2022] Zhong, X., Su, C., and Fan, Z. (2022). Empirical bayes pca in high dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):853–878.