An Empirical Bayes Approach to Partially Labeled and Shuffled Data Sets

Alex Dytso* and H. Vincent Poor†

Department of Electrical Engineering, Princeton University
Email: adytso@princeton.edu*, poor@princeton.edu†

***Abstract*—This work outlines a method for an application of empirical Bayes in the setting of semi-supervised learning. That is, we consider a scenario in which the training set is partially or entirely unlabeled. In addition to the missing labels, we also consider a scenario where the available training data might be shuffled (i.e., the features and labels are not matched).**

Specifically, we propose to train model-based empirical Bayes separately on the set of features and the set of labels and combine/mix the two models based on the proportion of unlabeled pairs. The method then can be used to recover the missing labels (i.e., create pseudo-labels) of the data set and, in addition, if the data is shuffled, recover the correct permutation of the data. The technique is evaluated for a multivariate Gaussian model and is shown to consistently outperform a maximum likelihood approach. Moreover, the procedure is shown to be a consistent estimator for a multivariate Gaussian model with an arbitrary (non-degenerate) covariance matrix.

I. Introduction

Consider a learning task where $\mathcal{U}$ denotes the underlying space of features, and $\mathcal{V}$ denotes the underlying space of labels. Let $\mathcal{S}_{\mathcal{U}}$ denote the available training set of features, and let $\mathcal{S}_{\mathcal{V}}$ denote the available training set of labels. Moreover, let $\mathcal{S}_{\mathcal{U}\times\mathcal{V}}$ denote the set of matched pairs between $\mathcal{S}_{\mathcal{U}}$ and $\mathcal{S}_{\mathcal{V}}$. Throughout the paper, we assume that the data generating process follows some joint probability distribution P_{UV}. More specifically, the pairs $(u, v) \in \mathcal{S}_{\mathcal{U}\times\mathcal{V}}$ are assumed to have been generated independently and identically (i.i.d.) according to P_{UV}. In machine learning the design of learning models based on the joint probability distribution P_{UV} (or the family of joint distributions) goes under the name of generative probabilistic model, and the design of the models based only on the conditionals $P_{V|U}$ or $P_{U|V}$ goes under the name of discriminative probabilistic models [1].

Due to various data collecting methods, we might encounter either one or both of the following scenarios:

1) One of the training sets is partially or entirely unlabeled. The former is referred to as semi-supervised learning and the latter as unsupervised learning [2]; and
2) The labels and features in the training set might not be properly matched. In other words, we have access to $\mathcal{S}_{\mathcal{V}}$ and $\mathcal{S}_{\mathcal{U}}$, but not to $\mathcal{S}_{\mathcal{U}\times\mathcal{V}}$. This is often referred to as a permuted or shuffled data scenario. The interested reader is referred to [3] and [4] for examples of such scenarios.

This work was supported in part by the U. S. National Science Foundation under Grant CCF-1908308.

In this work, we propose a learning method that can operate in either or both of the aforementioned scenarios. After using $\mathcal{S}_{\mathcal{V}}$ and $\mathcal{S}_{\mathcal{U}}$ to train the model the proposed method can be used 1) as a stand-alone regression method; 2) to recover the correct permutation of the data and match the pairs; and 3) to create pseudo-labels to repopulate the missing labels in the training set.

A. Contributions and Paper Outline

In what follows:

1) Section II discusses the details of the proposed method;
2) Section III applies the proposed method to a multivariate Gaussian model. In particular, it is shown that the proposed method is a consistent estimator. Moreover, several simulation results are presented, which evaluate the performance of the proposed method. For example, the proposed method is shown to consistently outperform a maximum likelihood approach; and
3) Section IV is dedicated to the implementation details of the Gaussian model.

II. Proposed Method

Consider a conditional expectation of V given U

$$\mathbb{E}[V|U=u] = \int V \mathrm{d}P_{V|U=u}, u \in \mathcal{U}. \tag{1}$$

The conditional expectation in (1) is an optimal Bayesian estimator under a very large family of loss functions, namely Bregman divergences [5]. For example, the family of Bregman divergences includes the ubiquitous squared error loss.

At the heart of our method is the idea of empirical Bayes [6]. To apply our procedure the conditional expectation must satisfy the following property: *for a fixed $P_{U|V=v}$, the conditional expectation in* (1) *is said to be empirical Bayes compatible (EBC) if for all $u \in \mathcal{U}$ and all admissible marginal distributions P_U there exists an operator $\mathrm{F}_{\mathcal{U}}$ such that*

$$\mathbb{E}[V|U=u] = \mathrm{F}_{\mathcal{U}}(u; P_U). \tag{2}$$

In other words, the conditional expectation is EBC if it depends on the joint distribution P_{VU} only through the marginal P_U. EBC implies, that from a knowledge of P_U alone, we can compute $\mathbb{E}[V|U]$ without using any information about P_V.

Note that the conditional expectation can alway be written as a functional of P_V, that

$$\mathbb{E}[V|U=u] = \mathrm{F}_{\mathcal{V}}(u; P_V) = \frac{\int V \mathrm{d}P_{V|U=u} \mathrm{d}P_V}{\mathrm{d}P_U(u)} \tag{3}$$

where $P_U = \int P_{U|V} \mathrm{d}P_V$.

Examples of probabilistic models that are EBC are given in Table I.

TABLE I
EXAMPLES OF PROBABILISTIC MODELS THAT ARE EMPIRICAL BAYES COMPATIBLE.

$P_{U\|V}$	$\mathbb{E}[V\|U=u]$
Gaussian Model I $\mathcal{U}=\mathcal{V}=\mathbb{R}^n$ $P_{U\|V=v}=\mathcal{N}(v,\mathsf{K})$, $0 \preceq \mathsf{K} \in \mathbb{R}^{n\times n}$ (covariance matrix)	$u+\mathsf{K}\frac{\nabla_u f_U(u)}{f_U(u)}$
Gaussian Model II $\mathcal{U}=\mathbb{R}^n$, $\mathcal{V}=\mathbb{R}^m$ $P_{U\|V=v}=\mathcal{N}(\mathsf{H}v,\mathsf{I}_n)$, $\mathsf{H}\in\mathbb{R}^{n\times m}$ (parameter matrix)	$(\mathsf{H}^\mathsf{T}\mathsf{H})^{-1}\mathsf{H}^\mathsf{T}\left(u+\frac{\nabla_u f_U(u)}{f_U(u)}\right)$
Poisson Model $\mathcal{U}=\mathbb{N}\cup\{0\}$, $\mathcal{V}=\mathbb{R}^+\cup\{0\}$ $P_{U\|V=v}=\frac{(av+\lambda)^u}{u!}\mathrm{e}^{-(av+\lambda)}$ $a>0, \lambda\geq 0$ (parameters)	$\frac{1}{a}\frac{(u+1)P_U(u+1)}{P_U(u)}-\frac{\lambda}{a}$
Geometric Model $\mathcal{U}=\mathbb{N}\cup 0$, $\mathcal{V}=(0,1]$ $P_{U\|V=v}=(1-v)v^u$,	$\frac{P_U(u+1)}{P_U(u)}$

Now under the EBC assumption, we propose a method for reconstructing the missing labels and recovering the correct permutation. The output of the procedure is a set $\widehat{\mathcal{S}}_{\mathcal{U}\times\mathcal{V}}$, which is an estimate of $\mathcal{S}_{\mathcal{U}\times\mathcal{V}}$. The method consists of the following steps and assumptions:

1) *(Assumption I):* Assume that a transition probability $P_{U|V}$ captures the mapping from V to U. The distribution $P_{U|V}$ is assumed to be known;
2) *(Assumption II):* Assume EBC assumption in (2) holds;
3) *(Estimator I):* Construct an empirical estimate $\widehat{\mathrm{F}}_{\mathcal{U}}(u;\mathcal{S}_{\mathcal{U}})$ of $\mathrm{F}_{\mathcal{U}}(u;P_U)$ based only on the set of features $\mathcal{S}_{\mathcal{U}}$. Note this is possible in view of the EBC assumption;
4) *(Estimator II):* Construct an empirical estimate $\widehat{\mathrm{F}}_{\mathcal{V}}(u;\mathcal{S}_{\mathcal{V}})$ of $\mathrm{F}_{\mathcal{V}}(u;P_V)$ based only on the set of labels $\mathcal{S}_{\mathcal{V}}$;
5) *(Mixture Estimator):* Choose some $\gamma\in[0,1]$ and construct

$$\widehat{\mathrm{M}}(u;\mathcal{S}_{\mathcal{U}},\mathcal{S}_{\mathcal{V}})=(1-\gamma)\widehat{\mathrm{F}}_{\mathcal{V}}(u;\mathcal{S}_{\mathcal{V}})+\gamma\widehat{\mathrm{F}}_{\mathcal{U}}(u;\mathcal{S}_{\mathcal{U}}), \quad (4)$$

as an estimator of $\mathbb{E}[V|U]$. The mixture parameter γ reflects our confidence about which of the estimators is a better approximation of $\mathbb{E}[V|U]$;
6) *(Match the Pairs or Recover the Correct Permutation):* Choose some appropriate distance/metric d on $\mathcal{V}$ and match the pairs as follows:

$$\widehat{\mathcal{S}}^{\pi}_{\mathcal{U}\times\mathcal{V}}=\Big\{(u,v):(u,v)=\arg\min_{v\in\mathcal{S}_{\mathcal{V}},u\in\mathcal{S}_{\mathcal{U}}}\mathsf{d}(v,\hat{v}(u)),\ \hat{v}(u)=\widehat{\mathrm{M}}(u;\mathcal{S}_{\mathcal{U}},\mathcal{S}_{\mathcal{V}})\Big\}; \quad (5)$$

7) *(Generate Pseudo-Labels):* Let $\mathcal{S}^m_{\mathcal{U}}\subseteq\mathcal{S}_{\mathcal{U}}$ denote the set of features that have missing labels and have not been paired in step 6). Using $\widehat{\mathrm{M}}(u;\mathcal{S}_{\mathcal{U}},\mathcal{S}_{\mathcal{V}})$ recover the missing labels as follows:

$$\widehat{\mathcal{S}}^m_{\mathcal{U}\times\mathcal{V}}=\{(u,\hat{v}):\hat{v}=\widehat{\mathrm{M}}(u;\mathcal{S}_{\mathcal{U}},\mathcal{S}_{\mathcal{V}}),u\in\mathcal{S}^m_{\mathcal{U}}\};\ \text{and} \quad (6)$$

8) *(Construct an Estimate of $\mathcal{S}_{\mathcal{U}\times\mathcal{V}}$):* Using sets in step 6) and step 7) let

$$\widehat{\mathcal{S}}_{\mathcal{U}\times\mathcal{V}}=\widehat{\mathcal{S}}^{\pi}_{\mathcal{U}\times\mathcal{V}}\cup\widehat{\mathcal{S}}^m_{\mathcal{U}\times\mathcal{V}}. \quad (7)$$

A few comments are now in order. First, besides Assumption I and Assumption II, the procedure makes no other assumptions about the joint distribution P_{UV}. Moreover, in what follows, we also offer solutions on how to avoid the exact knowledge of $P_{U|V}$ and show that the procedure can also be implemented if only the family of $P_{U|V}$ is known but not $P_{U|V}$ exactly.

Second, the mixing parameter γ should be a function of the cardinalities of $\mathcal{S}_{\mathcal{V}}$ and $\mathcal{S}_{\mathcal{U}}$. For example, a possible, but not optimal, choice is $\gamma=\frac{|\mathcal{S}_{\mathcal{U}}|}{|\mathcal{S}_{\mathcal{U}}|+|\mathcal{S}_{\mathcal{V}}|}$. Ideally, γ should also depend on the rates of convergence of $\widehat{\mathrm{F}}_{\mathcal{U}}(u;\mathcal{S}_{\mathcal{U}})$ and $\widehat{\mathrm{F}}_{\mathcal{V}}(u;\mathcal{S}_{\mathcal{V}})$ to $\mathbb{E}[V|U]$. As we will see, the convergence rates can be very different.

The empirical Bayes method has been proposed by Robbins in [6]. For a historical account of the impact of the empirical Bayes, the interested reader is referred to [7] and [8]. In the context of a Gaussian model, the authors of [9] and [10] combined the empirical Bayes formula with kernel density estimation methods. In this work, we will also use the theory of kernel density estimation; however, unlike previous works, we allow a Gaussian model to have an arbitrary covariance matrix. In [11], in the case of a scalar Gaussian model, the authors considered an empirical Bayes procedure that assumed the variance of the model is unknown. In this work, we will also allow the covariance matrix of the model to be unknown.

To the best of our knowledge, the empirical Bayes has not been used to recover permuted data.

III. GAUSSIAN MODEL I

In this section, we implement the empirical Bayes procedure from Section II for the Gaussian I model (see Table I). The implementation of the Gaussian II model (see Table I) that can handle the case when $\mathcal{V}$ and $\mathcal{U}$ are of different dimensions is deferred to the extended version of the paper. That is, we assume that $\mathcal{V}=\mathcal{U}=\mathbb{R}^n$ and $P_{U|V}$ takes the following Gaussian form:

$$P_{U|V=v}=\mathcal{N}(v,\mathsf{K}) \quad (8)$$

where K is some covariance matrix. In other words, we assume that given a label $V=v$, the feature U is according to a Gaussian distribution with the covariance matrix K.

We begin by verifying that Assumption I and Assumption II hold. For this model, the empirical Bayes formula takes the following form [12]:

$$\mathbb{E}[V|U=u]=u+\mathsf{K}\rho(u),\ \rho(u)=\frac{\nabla_u f_U(v)}{f_U(v)}, \quad (9)$$

where $f_U(v)$ is the probability density function (pdf) of U. In the statistical literature, $\rho(u)$ is known as the score function. As desired, (9) only depends on the marginal distribution of U, and, hence, the EBC assumption holds. As stated before, we make no assumptions about the distribution of P_V, and the

only assumption we make is that $P_{U|V}$ belongs to a Gaussian family.

Having verified that Assumption I and Assumption II hold, it remains to describe the implementation of steps 3), 4) and 5). A brief outline of this is given next with the exact implementation details postponed to Section IV.

- *Noise Covariance Construction:* If K is unknown, construct an estimate $\widehat{\mathsf{K}}$ of the covariance matrix K. A method for generating $\widehat{\mathsf{K}}$ is discussed in Section IV-B;
- *Score Function Estimation:* Construct two estimates of the score function $\widehat{\rho}(u;\mathcal{S}_{\mathcal{U}})$ and $\widehat{\rho}(u;\mathcal{S}_{\mathcal{V}})$ based on $\mathcal{S}_{\mathcal{U}}$ and $\mathcal{S}_{\mathcal{V}}$, respectively. The estimation of the score function is discussed in Section IV-C;
- *Empirical Bayes:* For some $\gamma \in [0,1]$, combine the two estimates of the score function

$$\widehat{\mathrm{M}}(u;\mathcal{S}_{\mathcal{U}},\mathcal{S}_{\mathcal{V}}) = u + (1-\gamma)\widehat{\mathsf{K}}\widehat{\rho}(u;\mathcal{S}_{\mathcal{V}}) + \gamma\widehat{\mathsf{K}}\widehat{\rho}(u;\mathcal{S}_{\mathcal{U}}); \text{ and} \tag{10}$$

- *Permutation Recovery:* Let d in (5) be the ℓ_2 distance.

Next, we show that the proposed empirical Bayes method is a consistent estimator of the conditional expectation.

Theorem 1. *Assume that $P_{U|V=v} = \mathcal{N}(v,\mathsf{K})$. Consider the following cases:*

- *Suppose that K is known. Then, with probability one*

$$\lim_{|\mathcal{S}_{\mathcal{U}}|\to\infty} \widehat{\mathrm{M}}(\cdot;\mathcal{S}_{\mathcal{U}},\mathcal{S}_{\mathcal{V}}) = \mathbb{E}[V|U], \text{ for all } \gamma \in (0,1],$$
$$\lim_{|\mathcal{S}_{\mathcal{V}}|\to\infty} \widehat{\mathrm{M}}(\cdot;\mathcal{S}_{\mathcal{U}},\mathcal{S}_{\mathcal{V}}) = \mathbb{E}[V|U], \text{ for all } \gamma \in [0,1).$$

- *Suppose that K is unknown. Then, with probability one*

$$\lim_{\min(|\mathcal{S}_{\mathcal{U}}|,|\mathcal{S}_{\mathcal{V}}|)\to\infty} \widehat{\mathrm{M}}(\cdot;\mathcal{S}_{\mathcal{U}},\mathcal{S}_{\mathcal{V}}) = \mathbb{E}[V|U], \text{ for all } \gamma \in [0,1].$$

The proof of Theorem 1, along with the convergence rates, is deferred to the extended version of this paper.

A. Simulation Results for a Gaussian Model

In this section, we evaluate the performance of our method for a Gaussian model I. We first test the ability of the proposed method to recover the correct permutation and match the pairs (i.e., we examine the performance of step 6). With this purpose in mind, we assume that there are no missing labels (i.e., $|\mathcal{S}_{\mathcal{V}}| = |\mathcal{S}_{\mathcal{U}}|$), however, the pairs are not properly matched.

Fig. 1 demonstrates simulation results for the recovery of correct permutation of the data and where we show:

- an error when the exact matching needs to be recovered (black curve); and
- an error when the matching needs to be recovered to within a certain prescribed tolerance level (orange curve). Specifically, for the approximate matching, we choose some tolerance parameter ϵ and declare an error only if $|1 - \frac{\mathrm{d}(v,\hat{v}(u))}{\mathrm{d}(v,\hat{v}(u'))}| > \epsilon$ where u is the correct matching and u' is the estimated matching. In Fig. 1 we set $\epsilon = 0.1$.

We now test the performance of the empirical Bayes method in the case of partially labeled data. Specifically, assume that

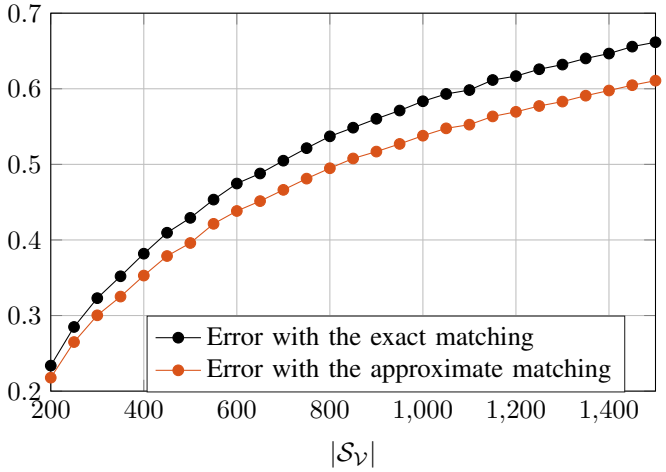

Fig. 1. The percentage of incorrectly recovered labels vs. the size of data sets $|\mathcal{S}_{\mathcal{V}}| = |\mathcal{S}_{\mathcal{U}}|$. In this example, $\mathsf{K} = 0.04\mathsf{I}_n$ and where V follows a mixed Gaussian distribution with the equal mixture, the means are set to $\mu_1 = 8\cdot\mathbf{1}_n$ and $\mu_2 = -\mu_1$, with variances equal to I_n and $n = 3$. Each point in the figure is obtained by averaging over 200 simulations.

the labels are matched, but $\mathcal{S}_{\mathcal{V}}$ is only partially filled with respect to $\mathcal{S}_{\mathcal{U}}$ and where the proportion is denoted by

$$\lambda = \frac{|\mathcal{S}_{\mathcal{V}}|}{|\mathcal{S}_{\mathcal{U}}|} \in [0,1]. \tag{11}$$

The case of $\lambda = 0$ correspond to the case when all of the labels are missing and $\lambda = 1$ corresponds to the case when all labels are available. Fig. 2 demonstrates a simulation of the mean squared error (MSE) for the recovery of the missing labels vs. the proportion λ and where we show:

- (black dashed curve) the normalized MSE of $\widehat{\mathrm{M}}(\cdot;\mathcal{S}_{\mathcal{U}},\mathcal{S}_{\mathcal{V}})$ with $\gamma = 1$. That is, the empirical Bayes is only trained from the set of features $\mathcal{S}_{\mathcal{U}}$. The choice of $\gamma = 1$ can be interpreted as using an unsupervised learning method since no labels are used to train the model. As expected, from Fig. 2, we see that this method performs best when the proportion of the labeled data is low;
- (black dotted curve) the MSE of $\widehat{\mathrm{M}}(\cdot;\mathcal{S}_{\mathcal{U}},\mathcal{S}_{\mathcal{V}})$ with $\gamma = 0$. That is, the empirical Bayes is only trained from the set of labels $\mathcal{S}_{\mathcal{V}}$. The case of $\gamma = 0$ can be interpreted as using a supervised learning method. As expected, from Fig. 2, the performance of the method improves with better labeling of the data. Interestingly, the method begins to perform well when only ten percent of the data is labeled; and
- (black solid curve) the MSE of the maximum likelihood estimator (MLE) (i.e., $\hat{V} = U$). Since we are assuming that $P_{U|V=v}$ is available, we can also use the MLE as a baseline for the performance of the empirical Bayes.

IV. Implementation Details for the Gaussian Model I

In this section, we discuss the details of implementation for a Gaussian model I.

A. Estimation of the Marginal Density and it's Gradient

In order to implement the estimation procedure outlined in Section III we must find an estimator of the density f_U and

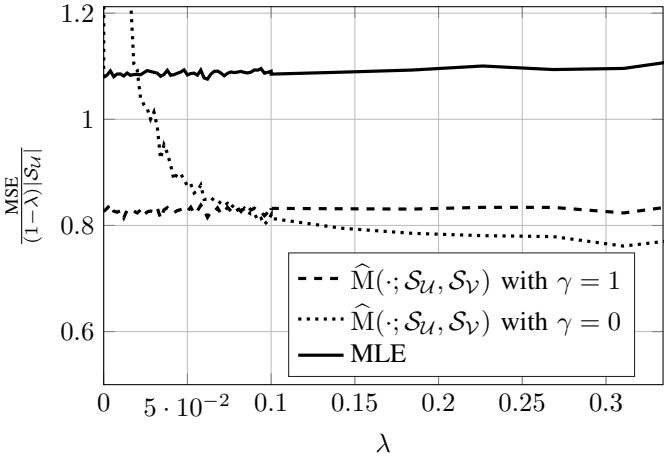

Fig. 2. Plot of the MSE normalized by $(1-\lambda)|\mathcal{S}_{\mathcal{U}}|$ vs. proportion λ. In this example, $\mathsf{K} = 0.04\mathsf{I}_n$ and where V follows a mixed Gaussian distribution with the equal mixture, the means are set to $\mu_1 = 8\cdot\mathbf{1}_n$ and $\mu_2 = -\mu_1$, with variances equal to I_n and $n = 3$. The training set is taken to be $|\mathcal{S}_{\mathcal{U}}| = 600$. Each point in the figure is obtained by averaging over 200 simulations.

an estimator of the gradient of f_U based only on the set of labels $\mathcal{S}_{\mathcal{U}}$. To do this, we employ the method of kernel density estimation and let

$$\widehat{f}(t;\mathcal{S}_{\mathcal{U}}) = \frac{1}{|\mathcal{S}_{\mathcal{U}}|}\sum_{u\in\mathcal{S}_{\mathcal{U}}}\frac{1}{\sqrt{\det(\mathsf{B})}}k\left(\mathsf{B}^{-\frac{1}{2}}(t-u)\right), \quad (12)$$

where we take the kernel $k(t) = \frac{1}{(2\pi)^{\frac{n}{2}}}\mathrm{e}^{-\frac{\|t\|^2}{2}}$, $t\in\mathbb{R}^n$ and where $\mathsf{B}\in\mathbb{R}^{n\times n}$ is a positive-definite bandwidth matrix. Moreover, the estimator of the gradient can be constructed by simply differentiating (12), that is

$$\widehat{f}'(t;\mathcal{S}_{\mathcal{U}}) = \frac{1}{|\mathcal{S}_{\mathcal{U}}|}\sum_{u\in\mathcal{S}_{\mathcal{U}}}\frac{\mathsf{B}^{-\frac{1}{2}}}{\sqrt{\det(\mathsf{B})}}\nabla k\left(\mathsf{B}^{-\frac{1}{2}}(t-u)\right). \quad (13)$$

In the paper, we set the bandwidth matrix to be $a\mathsf{I}_n$. Moreover, we set $a = |\mathcal{S}_{\mathcal{U}}|^{-\frac{1}{n+6}}$ to guarantee convergence [13].

B. Estimation of the Model Covariance

In this section, we discuss methods for constructing an estimator of K. Specifically, we propose to use

$$\widehat{\mathsf{K}} = (1-\alpha)\widehat{\mathsf{K}}_\ell(\mathcal{S}_{\mathcal{U}},\mathcal{S}_{\mathcal{V}}) + \alpha\widehat{\mathsf{K}}_m(\mathcal{S}_{\mathcal{U}}) \quad (14)$$

where $\widehat{\mathsf{K}}_\ell(\mathcal{S}_{\mathcal{U}},\mathcal{S}_{\mathcal{V}})$ is the estimator of K that uses all of the training set (i.e., supervised learning), and $\widehat{\mathsf{K}}_m(\mathcal{S}_{\mathcal{U}})$ is the estimator of K that only uses the set of features $\mathcal{S}_{\mathcal{U}}$ (i.e., unsupervised learning). The weight parameter α can, for example, be set to $\alpha = \frac{|\mathcal{S}_{\mathcal{U}}|}{|\mathcal{S}_{\mathcal{U}}|+|\mathcal{S}_{\mathcal{V}}|}$.

To estimate $\widehat{\mathsf{K}}_\ell(\mathcal{S}_{\mathcal{U}},\mathcal{S}_{\mathcal{V}})$ observe that (8) implies that

$$U = V + N, \quad (15)$$

where N is a zero mean Gaussian random vector independent of V with a covariance matrix K. Based on this observation, we adopt the following estimator:

$$\widehat{\mathsf{K}}_\ell(\mathcal{S}_{\mathcal{U}},\mathcal{S}_{\mathcal{V}}) = \widehat{\mathsf{K}}_U - \widehat{\mathsf{K}}_V, \quad (16)$$

where $\widehat{\mathsf{K}}_U$ and $\widehat{\mathsf{K}}_V$ are empirical covariances of U and V, respectively. For a recent survey on the estimation of covariance matrices, the interested reader is referred to [14].

We now turn to the estimation of K only from the set $\mathcal{S}_{\mathcal{U}}$. In general, such a problem is ill-defined. For example, suppose that V has a Gaussian component say $V = V_0 + V_G$ where V_G is Gaussian with covariance K_G. Then,

$$U = V + N = V_0 + V_G + N = V_0 + \tilde{N}, \quad (17)$$

where $\tilde{N}$ is a Gaussian vector with covariance $\mathsf{K}+\mathsf{K}_G$. Since we do not know the variance of N, we cannot decide between K and $\mathsf{K}+\mathsf{K}_G$.

We, however, can identify the 'largest' Gaussian component. To this end, observe that by using Jensen's inequality

$$f_U(t) = \mathbb{E}\left[\frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det(\mathsf{K})}}\mathrm{e}^{-\frac{(t-V)^T\mathsf{K}^{-1}(t-V)}{2}}\right] \quad (18)$$

$$\geq \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det(\mathsf{K})}}\mathrm{e}^{-\frac{\mathbb{E}\left[(t-V)^T\mathsf{K}^{-1}(t-V)\right]}{2}}. \quad (19)$$

From (19), we see that the tail of f_U is controlled from below by a Gaussian tail with the covariance K. Using this observation, we propose the following estimator:

$$\widehat{\mathsf{K}}_m(\mathcal{S}_{\mathcal{U}}) = \arg\min_{\mathsf{A}:0\preceq\mathsf{A}}\sum_{u\in\mathcal{S}_{\mathcal{U}}}(\widehat{f}(u;\mathcal{S}_{\mathcal{U}}) - \phi_{\mathsf{A}}(u))^2, \quad (20)$$

where $\widehat{f}(u;\mathcal{S}_{\mathcal{U}})$ is the estimator of f_U in (12).

C. Estimation of the Score Function

We now discuss an approach for constructing $\widehat{\rho}(u;\mathcal{S}_{\mathcal{V}})$ and $\widehat{\rho}(u;\mathcal{S}_{\mathcal{U}})$. To estimate $\widehat{\rho}(u;\mathcal{S}_{\mathcal{V}})$ observe that the true score function, given that we know the covariance K, can be written as an expectation over V as follows:

$$\rho(u) = \frac{\nabla_u f_U(u)}{f_U(u)} = \frac{\mathbb{E}\left[\mathsf{K}^{-1}(u-V)\phi_{\mathsf{K}}(u-V)\right]}{\mathbb{E}\left[\phi_{\mathsf{K}}(u-V)\right]}, \quad (21)$$

where $\phi_{\mathsf{K}}(t) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det(\mathsf{K})}}\mathrm{e}^{-\frac{t^T\mathsf{K}^{-1}t}{2}}$ is a Gaussian pdf. Using (21) and the fact that we operate over the set $\mathcal{S}_{\mathcal{V}}$ we adopt the empirical score function as the estimator

$$\widehat{\rho}(u;\mathcal{S}_{\mathcal{V}}) = \frac{\sum_{v\in\mathcal{S}_{\mathcal{V}}}\hat{\mathsf{K}}^{-1}(u-v)\phi_{\hat{\mathsf{K}}}(u-v)}{\sum_{v\in\mathcal{S}_{\mathcal{V}}}\phi_{\hat{\mathsf{K}}}(u-v)}, \quad (22)$$

where $\hat{\mathsf{K}}$ is the estimate of K.

To estimate $\widehat{\rho}(u;\mathcal{S}_{\mathcal{U}})$ we use the kernel density method discussed in Section IV-A. That is, we use the following estimator:

$$\widehat{\rho}(u;\mathcal{S}_{\mathcal{U}}) = \frac{\widehat{f}'(t;\mathcal{S}_{\mathcal{U}})}{\widehat{f}(t;\mathcal{S}_{\mathcal{U}})}, \quad (23)$$

where $\widehat{f}(u;\mathcal{S}_{\mathcal{U}})$ and $\widehat{f}'(u;\mathcal{S}_{\mathcal{U}})$ are defined in (12) and (13), respectively.

It can be shown that both $\widehat{\rho}(u;\mathcal{S}_{\mathcal{V}})$ and $\widehat{\rho}(u;\mathcal{S}_{\mathcal{U}})$ are consistent estimators of the score function. Details of the convergence, as well as the rates of the convergence, can be found in the extended version of the paper.

REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] X. J. Zhu, "Semi-supervised learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2005.

[3] A. Pananjady, M. J. Wainwright, and T. A. Courtade, "Denoising linear models with permuted data," in *Proc. IEEE Int. Symp. Inf. Theory*. IEEE, 2017, pp. 446–450.

[4] A. Dytso, M. Cardone, M. S. Veedu, and H. V. Poor, "On estimation under noisy order statistics," *arXiv preprint arXiv:1901.06294*, 2019.

[5] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, no. Oct, pp. 1705–1749, 2005.

[6] H. Robbins, "An empirical Bayes approach to statistics," in *Herbert Robbins Selected Papers*. Springer, 1985, pp. 41–47.

[7] B. Efron and T. Hastie, *Computer Age Statistical Inference*. Cambridge University Press, 2016, vol. 5.

[8] B. Efron *et al.*, "Bayes, oracle Bayes and empirical Bayes," *Statistical Science*, vol. 34, no. 2, pp. 177–201, 2019.

[9] C.-H. Zhang, "Compound decision theory and empirical Bayes methods," *The Annals of Statistics*, vol. 31, no. 2, pp. 379–390, 2003.

[10] L. D. Brown and E. Greenshtein, "Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means," *arXiv preprint arXiv:0908.1712*, 2009.

[11] D. Donoho and G. Reeves, "Achieving Bayes MMSE performance in the sparse signal+ Gaussian white noise model when the noise level is unknown," in *Proc. IEEE Int. Symp. Inf. Theory*. IEEE, 2013, pp. 101–105.

[12] R. Esposito, "On a relation between detection and estimation in decision theory," *Inf. Control*, vol. 12, no. 2, pp. 116–120, February 1968.

[13] J. E. Chacón, T. Duong, and M. Wand, "Asymptotics for general multivariate kernel density derivative estimators," *Statistica Sinica*, pp. 807–840, 2011.

[14] J. Fan, Y. Liao, and H. Liu, "An overview of the estimation of large covariance and precision matrices," *The Econometrics Journal*, vol. 19, no. 1, pp. C1–C32, 2016.