LOL Extra

Joshua T. Vogelstein, Mauro Maggioni

As scientific measurements become increasing easy and inexpensive to obtain, the dimensionality of scientific data is exploding. Common examples include genomics, wherein an individual is encoded by a 6 billion dimensional sequence, and connectomics, wherein an individual is encoded by up to 1 million nodes, and therefore up to 10^{12} potential interaction terms. Such rich representations provide hope for mining these data to find, for example, prognostic biomarkers or personalized therapies. That said, utilizing these data at their native resolution and dimensionality has been hindered thus far.

The most pernicious impediment is perhaps the low sample size relative to the number of observed dimensions: while the observed dimensionality has ballooned, the sample sizes have *not* witnessed a concomitant increase. It is common for sample sizes to be in the tens; hundreds is considered a very large experiment, and thousands is all but unheard of (see, for example, Table 2 of [**?**] in connectomics, and XXX in genomics). This means that we often do not have a large enough sample size to obtain high power (or accuracy) without assuming some additional structure in the data. In other words, even though the information is likely present in the high dimensional data, our ability to discover it is impeded. To combat this problem, we can assume some structure, and then use that assumption to significantly reduce the number of parameters to estimate. Although the set of such possible assumptions is vast, we characterize all such options into one of three philosophical methodologies.

One recently very popular approach is to assume sparsity, that is, that a small subset of the observed dimensions contain sufficient information to make quality predictions [**?** **?**]. Sparse methods include screening methods such as higher criticism screening [**?**], L1 penalized methods such as LASSO [**?**], and more recent sparse discriminant analysis methods [**?** **?** **?**]. While very useful because they admit readily interpretable results, they are guaranteed to work only in very restricted situations [**?**]. Specifically, they require certain orthogonality conditions on the individual dimensions (or features), and require that a small number of those features have large predictive accuracy.

Another alternative is to "pre-process" the data via applying some unsupervised manifold learning technique, which search for a low dimensional representation. The most well-known such example of such a methodology is principle components analysis (PCA) [**?** **?**]. In the $>$110 years since its original development, many extensions and generalizations of PCA have been proposed, including random projections [**?**], NMF [**?** **?**], and kernel variants of PCA such as isomap [**?**], LLE [**?**], and Laplacian Eigenmaps [**?**]. These methods can find more general representations than their sparse counterparts by discovering linear or nonlinear combinations of the observed dimensions to represent the data. However, these approaches do not utilized the supervised information to discover the best representation for accurate predictions or tests. Thus, even though they sometimes provide strong theoretical support for discovering the true underlying manifold under rather general settings [1**?**], the goal in this work is *not* to find the *true* manifold of the data, but rather, to find the manifold that maximizes accuracy.

We therefore consider a third philosophy, which we call supervised manifold learning. This approach takes the best of both worlds from the previously mentioned approaches: it can in theory find very complex linear or nonlinear combinations of the original observed dimensions, and it does so in a fully supervised manner. In fact, the two previous approaches can accurately be thought of as special cases of this more general approach (see Appendix for rigorous definitions). A number of investigators have previously developed supervised manifold learning methods. A set of methods from the statistics community is collectively referred to as "sufficient dimensionality reduction" methods [**?** **?** **?** **?** **?**]. These methods are theoretically elegant, but typically require the sample size to be larger than the number of observed dimensions (although see [**?**] for some promising work). Other approaches formulate an optimization problem, such as projection pursuit [**?**] or supervised dictionary learning [**?**]. These methods are limited because they are prone to fall into local minima, they require costly iterative algorithms, and lack any theoretical guarantees [**?**]. Thus, their remains a gap in the literature: a supervised learning method with theoretical guarantees under general conditions with an efficient algorithm.

In this work, we present such a method, which we term LOL, for Low-rank Optimal Linear projection. The key idea emerges from considering perhaps the simplest supervised learning task: a two-class classification problem. It is immediate that even for multivariate Gaussian data, the discriminant boundary (that is, the boundary that discriminates between the two classes) is a function of both: (i) the difference of the means, and (ii) the covariance of the two classes. Therefore, we find a low dimensional representation of the data utilizing both of these objects.

The simplicity of this idea translates into a methodology that readily admits theoretical investigation, efficient implementations, and generalizations. We provide all of them in this work. Moreover, we demonstrate improved performance over competing ideas in terms of both accuracy and speed, for a number of simulated and real data examples. Moreover, straightforward extensions of this idea lead to similarly simple and useful other tasks, such as multi-class classification, regression and testing. We provide open source MATLAB and R code to enable further extensions and applications.

Results

The Generality of LOL

To understand the generality of applicability of LOL, we consider possibly the simplest supervised manifold learning setting. Specifically, we consider a two-class classification problem, where both class have multivariate Gaussian distributions. To further simplify, we assume that the two classes have the same covariance, so that the only difference between the two classes are the means. Formally, we write $\boldsymbol{X}|Y = y \overset{iid}{\sim} \mathcal{N}_D(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$, where $\boldsymbol{X}$ is a D-dimensional vector of observed features, y is a binary class label, $\overset{iid}{\sim}$ indicates that each $\boldsymbol{X}$ is sampled identically and independently according to some distribution, and $\mathcal{N}_D(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$ indicates a D-dimensional Gaussian distribution with mean $\boldsymbol{\mu}_y$ (different for the two classes) and covariance $\boldsymbol{\Sigma}$ (the same for the two classes). In such a scenario, the Bayes optimal classifier is Fisher's Linear Discriminant Analysis (LDA). Specifically, the Bayes classifier projects a new sample, $\boldsymbol{x}$, onto the subspace spanned by the dot product of the inverse covariance matrix with the difference of the mean vector:

$$g_*(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{x}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} > t, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $\boldsymbol{\delta} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$, and t is a simple linear function of the class priors and the mean of the class means. Eq. (1) shows that the optimal manifold for discriminating between two multivariate Gaussian distributions with the same covariance is a linear manifold spanned by $\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$. From this, it would seem that any manifold learning methodology that we employ would benefit from including both an estimate of the mean difference, $\hat{\boldsymbol{\delta}}$, as well as the pooled covariance, $\hat{\boldsymbol{\Sigma}}$. Unsupervised manifold learning techniques ignore $\hat{\boldsymbol{\delta}}$, only utilizing $\hat{\boldsymbol{\Sigma}}$. Sparse methods struggle with correlated data, that is, data for which the covariance matrix is dense in the off-diagonal elements. This motivates the use of supervised manifold learning techniques that overcome these two problems.

To illustrate the differences between these three methods, we consider three different parameter settings for the above specified model. In each, we let D, the number of observed dimensions, be 1000, whereas we only sampled $n = 100$ points total, so that $D \gg n$. Fisher's Linear Discriminant Analysis (LDA) would be optimal in these settings if the parameters were known or easily estimable, but because sample size equals the ambient dimension, the estimated covariance matrix is singular, and so LDA does not even run. Instead, therefore, we consider an example of three different methodological principles. For each method, we embed the data into five dimensions.

First, we consider the simplest parameter setting: a single observation dimension contains nearly all the discriminant signal, and that dimension is the dimension of maximal variance (see Figure **??**(A)). This situation is ideal for using PCA to pre-process the data, as PCA finds the directions of maximal variance. However, PCA in high-dimensions is inconsistent, and has relatively high variance [**?** **?**]. Therefore, even in this idealized scenario, running LDA on the data projected onto the first 5 principal components yields subpar performance (second row of (A)). On the other hand, because the signal is so sparse, a method designed specifically for sparse multivariate Gaussian classification (Regularized Optimal Affine

Discriminant; ROAD [**?**]), performs very well (third row of (A)). Similarly, because LOL uses both the discriminant dimension and the principle components, LOL also does essentially optimally (fourth row of (A)). Finally, we also show the Bayes optimal performance (bottom row of (A)) for comparison purposes.

Second, we consider a related parameter setting, but now, the dimension with most of the discrimination information is aligned with the dimension of maximal variance. This is the worse possible scenario for using PCA to pre-process the data. Unfortuantely for PCA, we can readily prove that this scenario is much more common than the prior scenario, under relatively general settings (see Appendix).

Column (A) depicts the simplest setting: the optimal one-dimensional discriminant boundary is aligned with the direction of maximal variance. Thus, PCA finds this discriminant boundary, sparse methods do as well, as does LOL. Column (B) depicts a slightly more difficult setting, the optimal one-dimensional discriminant boundary is aligned with the direction of minimal variance. This is the most difficult setting for PCA, because PCA finds the directions of maximal variance, and thus discards directions of minimal variance (which, in this case, includes the optimal discriminant direction). Sparse methods could recover the optimal discriminant direction, but typically fail to find it. Rather, they find a random collection of other dimensions, each of which is equally and minimally informative. LOL performs significantly better, as it is able to fuse information across multiple dimensions. Finally, column (C) depicts the same setting as column (B), but all data have been randomly rotated. Both LDA $\circ$ PCA and LOL are invariant to such rotations, so there expected performance does not change. Sparse methods now randomly select a few dimensions, as each ambient dimension is now equally informative.

Finding the Best (not the True) Manifold

LOL for Wide Regression and Testing

Discussion

Big data are abundant. The applications of big data range from government, to industry, to science, with serious potential societal implications. Most big data analytics thus far, however, deal with huge sample size (e.g., $n = 10^7$) and large dimensionality (e.g., $p = 10^4$). Big scientific data, however, has the opposite characteristic: the dimensionality is huge (e.g., number of base pairs in a genome, or potential edges in a connectome), and the sample size is relatively small (e.g., hundreds typically). Optimally extracting information from such big scientific data therefore requires radically different approaches than typical industry or government big data problems. We are particularly interested in "supervised" learning problems here, such as building a classifier that operates on a whole genome or connectome, to make a diagnosis. In such cases, the genome or connectome comprise the *predictor* variables and the diagnosis represents the *target* variable.

To gain insight into the challenges associated with big scientific data, consider the MNIST dataset, consisting of images of the digits, 0-9 [2]. The top row of Figure **??** depicts four random samples of digits 3, 7, and 8 each. Our task is to learn a classifier $\hat{g}_n(\cdot)$ from a set of n such samples, such that given a new image, $\hat{g}_n(\cdot)$ would accurately determine the true class of the image. Each of these images is 28×28 pixels, meaning it has dimensionality $p = 784$. To make this scenario similar to the big scientific data scenario of interest, we have limited the number of training samples to $n = 100$, so that $n \ll p$. In such scenarios, it is well known that classic tools, such as linear discriminant analysis fail [**?**], because we lack sufficient information to obtain reliable estimates of the parameters. The statistics and machine learning communities have developed a number of mitigating strategies for such problems.

One recently very popular approach is to assume sparsity, that is, that a small subset of the given dimensions contain sufficient information to make quality predictions [**?** **?**]. Sparse methods include screening methods such as higher criticism screening [**?**], L1 penalized methods such as LASSO [**?**], and more recent sparse discriminant analysis methods [**?** **?** **?**]. These methods are inappropriate and inadequate for such applications, because there does not exist a small number of ambient dimensions that adequately capture the discriminant information. This is evidenced by panel (D), which shows the most discriminating

pixels, and panel (G) which shows 100 test images embedded into the top three dimensions discovered by a sparse approach. The three different classes are clearly not well distinguished, as quantified by the cyan line in panel (J), which shows the error rate (1-accuracy) for a variety of methods as a function of the number of embedded dimensions to keep.

Another alternative is to "pre-process" the data via applying some unsupervised manifold learning technique, which search for a low dimensional representation (potentially a manifold or collection thereof) of the predictor variables. Examples of such approaches includes PCA [**?**], random projections [**?**], NMF [**? ?**], and LLE [**?**]. Panel (E) shows the top four eigenimages of these data, that is, the four images that collectively capture the most variance of all possible four images. Clearly, each image has aspects of 3, 7, and 8 in it, which is intuitively pleasing. However, panel (H) shows 100 test images embedded into the top three such dimensions. Clearly, these data are again not very well separated, although much better separated than the sparse methods, as evidenced by the magenta line in panel (J).

Finally, the philosophy that we adopt in our work can be characterized as a supervised manifold learning technique. Such techniques seek to utilize both the predictor dimensions as well as the target variables to discover a low dimensional representation of the discriminant boundary. Sparse methods, are in fact, a special case of this approach, in which the discriminant boundary is assumed to lie along a subset of the predictor variables. More general supervised manifold learning techniques allow the low dimensional representation to be a (potentially nonlinear) combination of predictor variables. "Sufficient dimensionality reduction" methods [**? ? ? ? ?**], while theoretically appealing, require the sample size to be larger than the ambient dimensionality (although see [**?**] for some promising work). Other methods formulate an optimization problem, such as projection pursuit [**?**] or supervised dictionary learning [**?**]. These methods are limited because they are prone to fall into local minima, they require costly iterative algorithms, and lack any theoretical guarantees.

Our approach is termed LOL, for Low-rank Optimal Linear projection. On the MNIST dataset, the first few projection vectors that it learns look much like the unsupervised ones (see panel (F)); indeed, they are also linear combinations of the training samples. However, LOL results in embeddings that clearly separate the three classes (panel (I)). The improvement in classification performance over the other methods is dramatic, as quantified by the green line in panel (J). Moreover, computational time is just slightly longer than the unsupervised approach (panel (K)), and faster than the sparse method.

Our construction follows from a simple geometric argument given perhaps the simplest big scientific data supervised learning problem. This intuition leads to a formulation that admits an approach that improves over the current state of the art along many factors, including clarity, ease of implementation, computational efficiency, scalability, and theoretical justification. Moreover, its simplicity immediately leads to a number of extensions and generalizations. We demonstrate its properties on a range of simulations and real data experiments. Moreover, we provide open source MATLAB and R code to enable further extensions and applications.

Motivating Intuition

To gain insight into the challenges associated with big scientific data, we considered the simplest such problem we could think of: a two-class classification problem, where each class conditional distribution is a multivariate gaussian, and the classes have a shared covariance (see Figure **??**, top left panel).

Figure **??** shows a couple examples with diagonal covariance matrices. Because the sample size ($n = 50$) is significantly smaller than the ambient dimension ($p = 100$), classical methods will fail. Thus, a technique, such as the above, could potentially alleviate the issue. However, this problem is difficult because the principle direction of variance is *orthogonal* to the principle discriminant direction. Thus, unsupervised methods will discard the discriminant signal in favor of the variance dimensions. Moreover, because the discriminant dimension is not a canonical dimension, but rather a linear combination thereof, sparse methods will also fail. We therefore are left with supervised dictionary learning techniques. When the dimensionality is large, F2M methods will fail, and iterative optimization methods will be too computationally burdensome.

In such a scenario, a simple linear algebraic calculation reveals that the optimal projection matrix is the dot

product of the difference between the means with the shared inverse covariance matrix, $\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}$. However, when the number of dimensions p is much greater than the number of samples n, we obtain estimates of $\boldsymbol{\Sigma}$ that are badly scaled, such that inverting the matrix is numerically unstable.

I Results

I.A Main Approach

I.B Numerical Results

I.B(0).1 Illustrative example on real data

I.B(0).2 Generalizations

The simplicity of LOL enables a wide range of applications and extensions (Figure **??**).

1. In the "Trunk" example, each of the 500 ambient dimensions contains signal, although the amount of signal is decreasing geometrically, while the variance is increasing geometrically. With only 50 training samples, LDA$\circ$PCA never performs as well as LOL with only a single dimension. This example was originally proposed to demonstrate that the optimal number of dimensions to keep when estimating a classifier is a function of the sample size.

2. It is increasingly common to assume sparsity in the data. Recent advances is sparse linear discriminant analysis demonstrate impressive theoretical and empirical results. Four recently proposed methods include the ℓ_1-Fisher's discriminant analysis (FSDA) [**?**], sparse optimal scoring (SOS) [**?**], direct sparse discriminant analysis (DSDA) [**?**], and regularized optimal affine discriminant (ROAD) [**?**]. [**?**] proved that the first three are equivalent, and [**?**] numerically compares them with ROAD, as well as nearest shrunken centroids (PAM) [**?**] and features annealed independence rules (FAIR) [**?**]. While the first four use the correlation structure, the last two discard it, much like the naïve Bayes classifier as discussed in [**?**]. [**?**] showed that including covariance is useful. All of these approaches, however, are rotationally sensitive. That is, they all benefit from the data living in a sparse subset of the ambient bases, rather than a learned dictionary in that subspace. In other words, all these methods suffer if the data are arbitrarily rotated. To demonstrate this, we chose a simulation setting provided by [**?**], with $D = 1000$ but $n = 300$, and $s = 10$ (number of signal dimensions) in which ROAD performs nearly optimally, and rotated the data by a random rotation matrix. The performance of ROAD, as well as all the other algorithms, significantly deteriorates in this context. On the other hand LOL is rotation invariant, and moreover, significantly faster than these other covariance dependent algorithms, as LOL utilizes extremely optimized numerical linear algebra routines such as `svd` and `qr`, whereas the other methods require relatively computationally intensive iterative optimization algorithms. Because of this, LOL also scales up to millions or billions of dimensions, whereas none of the other methods have this level of scalability.

3. The third example we constructed to exhibit slower spectral decay than the previous ones, to reflect real data scenarios. More specifically, the covariance matrix is Toeplitz, and the difference between the means is a square wave, which lives in the subspace spanned by the eigenvectors corresponding to the lowest eigenvalues. Thus, this example corresponds to a difficult, but realistic, scenario. For simplicity, we let $D = n = 50$ in this example, although we can easily tune the parameters to yield identical performance for D and n of any size. As (C) shows, LOL again significantly outperforms LDA$\circ$PCA in terms of misclassification rate for almost all embedding dimensions. Moreover, both algorithms exhibit a noticeable unimodal shape, with misclassification rate improving as it keeps including more informative dimensions, and then begins to degrade as subsequent dimensions are more noise than signal.

4. The flexibility of LOL enables us to consider a wide range of extensions. Fast via randomized svd, or random projections

5. Multiple subspaces
6. Robust
7. In semi-supervised settings, many training samples may lack labels. In such a scenario, we can still leverage the unlabeled samples to contribute to the pooled covariance estimate. Thus, under certain settings [3] we can improve performance via utilizing those additional data. Panel (D) is the same setting as panel (C), but now we have 50 labeled samples, and an additional 450 unlabeled samples, thus only 10% of the samples are labeled.

I.B(0).3 Scalability

I.B(0).4 Novel Applications

I.C Theoretical Results

In particular, let ρ denote the correlation between the mean difference vector and the diagonal of the diagonal covariance matrix. Let $\rho_{PCA}(i)$ be the correlation between the Bayes optimal projection $\boldsymbol{A}_*$ and the ith eigenvector, and let $\rho_{LOL}(i)$ be defined similarly, for the i^{th} projection vector of LOL. Similarly, let θ_i be the correlation between the mean difference vector δ and the i^{th} eigenvector. Let the unsubscripted version of ρ and θ denote their maximum value, for example, $\rho_Z = \max_i \rho_Z(i)$. Under suitable regularity conditions (see Methods for details).

Theorem 1. *LOL has finite sample bounded error*

$$\mathbb{P}[R_{d_n}^{LOL} - R^* < \varepsilon(d, \boldsymbol{\theta})] \geq \eta(n, d_n, \boldsymbol{\theta}) \tag{2}$$

Theorem 2. *LOL has better finite sample error than LDA o PCA or RP o PCA*

$$\mathbb{P}[R_{d_n}^{PCA} - R_n^{LOL} < \varepsilon(d, \boldsymbol{\theta})] \geq \eta(n, d_n, \boldsymbol{\theta}) \tag{3}$$
$$\mathbb{P}[R_{d_n}^{RP} - R_n^{LOL} < \varepsilon(d, \boldsymbol{\theta})] \geq \eta(n, d_n, \boldsymbol{\theta}) \tag{4}$$
$$\tag{5}$$

Theorem 3. *when subspace is a subset of the element of the Stiefel manifold, we recover it just as quickly.*

II Discussion

II.(0).5 Extensions

Regression (vs. GMRA-regression) and two-sample testing (vs. Wainwright [**?**]).

Figure 1: top: regression, bottom: two sample testing

II.(0).6 Related work

[4**? ? ? ? ? ? ? ? ? ? ? ? ?** –6]

A Approaches for dealing with Wide Data

I.A Unsupervised Manifold Learning

I.B Sparse Methods

I.C Supervised Manifold Learning

B Theory

C Supplement

III.A Notation

Let $\mathbb{X}\colon \Omega \to \mathcal{X} \subseteq \mathbb{R}^D$ be a high dimensional predictor-valued random variable and let $\mathbb{Y}\colon \Omega \to \mathcal{Y}$ is be a target-valued random variable. Assume that $\mathbb{X}$ and $\mathbb{Y}$ are jointly distributed according to some distribution amongst a parametric model, $\mathbb{X}, \mathbb{Y} \overset{iid}{\sim} P_{XY} \equiv P \in \mathcal{P}$. Throughout, we will assume that the density (measure) P_{XY} has a corresponding probability distribution function $p \equiv p_{XY}$, and moreover, that p decomposes $p_{X|Y}p_Y$. Let $\Phi = \{\phi\colon \mathcal{X} \to \mathcal{Y}\}$ so that each $\phi \in \mathcal{Y}^{\mathcal{X}}$ is a set of measurable functions from $\mathcal{X}$ to $\mathcal{Y}$. Let $\boldsymbol{\mu}_y = \mathbb{E}[\mathbb{X}|\mathbb{Y} = y]$ be the conditional mean and $\boldsymbol{\Sigma} = \int_{\mathcal{Y}} \mathbb{E}[\mathbb{X} \cdot \mathbb{X}|\mathbb{Y}]dY$ be the joint covariance matrix.

When $\mathcal{Y}$ is a categorical measurable metric space, that is, $\mathcal{Y}$ is shorthand for the metric space with a $0-1$ metric, we let $\mathcal{Y} = \{\widetilde{y}_1, \ldots, \widetilde{y}_k\}$, and $k = \#(\mathcal{Y})$ is the total number of classes. In such cases, we let $\mathcal{P}_{Dk}$ denote the set of distributions for which $\mathbb{X}$ is a d-dimensional random variable and $\mathbb{Y}$ is categorical with two categories. When $\mathcal{Y}$ is a finite or countable set of classes, each ϕ is called a *classifier function*. Let $\boldsymbol{\Sigma}_y = \mathbb{E}[\mathbb{X} \cdot \mathbb{X}|\mathbb{Y} = y] = \int_{\mathcal{Y}} \mathbb{E}[\mathbb{X}|\mathbb{Y}]F_Y dY$ be the class conditional covariance matrix.

A *loss function* is a measurable map $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+ \equiv [0, +\infty)$ that we use to assess the performance of a classifier, where $\ell \in \mathcal{L}$. Specifically, we are interested in minimizing risk. Let the *risk*, $R\colon \mathcal{L} \times \mathcal{P} \times \Phi \to \mathbb{R}_+$, of a classifier be the expected loss under the true distribution of the data:

$$R_{\ell,P}(\phi) \equiv \mathbb{E}_P[\ell(\phi(x), y)] \equiv \int_{\mathcal{X}\times\mathcal{Y}} \ell(\phi(x), y)dP. \tag{6}$$

For the remainder of this work, let ℓ be $0-1$ loss, that is, $\ell = \mathbb{I}\{x = y\}$.

The Bayes optimal classifier for a given distribution is that classifier that minimizes risk:

$$\phi^*(\cdot) = \underset{\phi\in\Phi}{\operatorname{argmin}}\, R_{\ell,P}(\phi) = \underset{\phi\in\Phi}{\operatorname{argmin}}\, \mathbb{E}_P[\ell(\phi(\cdot), y)], \tag{7}$$

and $R^*_{\ell,P}$ is the minimum.

Let $[\boldsymbol{S}, \boldsymbol{V}] = \texttt{eig}(\boldsymbol{\Sigma})$, such that $\boldsymbol{V}_j$ is the j^{th} eigenvector of $\boldsymbol{\Sigma}$, $\boldsymbol{S}$ is a diagonal matrix whose $(j,j)^{th}$ entry is is j^{th} eigenvalue, such that $s_1 \geq s_2 \geq \cdots s_D$. Let $[\boldsymbol{Q}, \boldsymbol{R}] = \texttt{qr}(\boldsymbol{Y})$ yield the QR decomposition of $\boldsymbol{Y} \in \mathbb{R}^{D\times D}$ where $\boldsymbol{Q}$ is an orthogonal matrix, $\boldsymbol{Q}\boldsymbol{Q}^\mathsf{T} = \boldsymbol{I}$, and $\boldsymbol{R}$ is an upper triangular matrix, both $D \times D$ in general. Let $[\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{V}] = (\boldsymbol{X})$ be the singular value decomposition of $\boldsymbol{X}$. Let $\Phi_D\colon \mathbb{R}^D \to \mathcal{Y}$ be classifiers that operate on $\boldsymbol{X}$'s that live in $\mathbb{R}^D$.

III.B P is Given, $k = 2$

III.B(1) LDA

Linear disciminant analysis (LDA) function is defined as follows:

$$\phi^L_D(\boldsymbol{x}; \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) = \underset{y\in\mathcal{Y}}{\operatorname{argmax}} \frac{\pi_y}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_y)^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_y)\right\} \tag{8}$$

Note that LDA operates on the joint mean and covariance, as well as the class-conditional means. Further note that it is not restricted to two-class classification problems, rather, the LDA classifier finds the most likely class *a posteriori*. As it turns out, LDA is the Bayes optimal classifiers under Gaussian data with either the same covariance. In particular, LDA is Bayes optimal under the following model

$$\mathcal{P}_{D2}^{L} = \{P_{X|Y}P_Y : p_{x|y} = \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}), p_y = \mathcal{B}ern(\pi_{\widetilde{y}}), \text{ for } y \in \{0,1\}\} \tag{9}$$

where $(\boldsymbol{\mu}_y, \pi_y) \in \mathbb{R}^D \times (0,1)$ for $y \in \{0,1\}$ and $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^\mathsf{T}\boldsymbol{\Lambda}$, for $\boldsymbol{\Lambda} \in \mathbb{R}^{D\times D}\}$.

Also note that LDA can be written explicitly by the following:

$$\phi_D^L(\boldsymbol{x}; \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) = \mathbb{I}\{(\boldsymbol{x} - \bar{\boldsymbol{\mu}})^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} + \log(\pi_1/\pi_0) > 0\} \tag{10}$$

where $\boldsymbol{\delta} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$ and $\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}_0\pi_0 + \boldsymbol{\mu}_1\pi_1$. Thus, clearly LDA is a *linear* function of the moments.

Let $\Psi_{D,d} = \{\psi_{D,d} \colon \mathbb{R}^D \to \mathbb{R}^d\}$ be linear projection matrices. Let and $\widetilde{\Phi} = \{\widetilde{\phi}_{D,d} := \phi_d \circ \psi_{D,d}\}$ be classifiers that take the following form:

$$\widetilde{\phi}_{D,d}(\boldsymbol{x}; \psi_{D,d}, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) = \mathbb{I}\{(\widetilde{\boldsymbol{x}} - \bar{\widetilde{\boldsymbol{\mu}}})^\mathsf{T}\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{\delta}} + \log(\pi_1/\pi_0) > 0\} \tag{11}$$

where $\widetilde{\boldsymbol{x}} = \psi_{D,d}\boldsymbol{x} \in \mathbb{R}^d$, and similarly for $\bar{\boldsymbol{\mu}}$, $\boldsymbol{\delta}$, and $\widetilde{\boldsymbol{\Sigma}} = \psi_{D,d}\boldsymbol{\Sigma}\psi_{D,d} \in \mathbb{R}^{d\times d}_{\succeq 0}$, where $\mathbb{R}^{d\times d}_{\succeq 0}$ is the set of $d \times d$ positive definite matrices.

Let $\psi_{D,d}^L$ be the matrix defined by the QR'ed concatenated $\boldsymbol{\delta}$ with the first $d-1$ eigenvectors of $\boldsymbol{\Sigma}$:

$$\psi_{D,d}^L = \mathtt{qr}([\boldsymbol{\delta}; (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_{d-1})]) \tag{12}$$

Let $\widetilde{\phi}_{D,d}^L := \phi_d \circ \psi_{D,d}^L$. Let $R_{D,d}^L := R_{0-1,P}(\widetilde{\phi}_{D,d}^L)$ and $R_{D,d}^* = R_{0-1,P}^*$.

Let

$$R_{D,d^*}^L = \min_{d\in[D]} R_{D,d}^L \tag{13}$$

Remark 1. *When $P \in \mathcal{P}_{D2}^{LDA}$, $d^* = 1$ and $\psi_{D,1} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$ is optimal. This is true regardless of π_y's and $\bar{\boldsymbol{\mu}}$, that is, even for highly imbalanced classes.*

III.B(2) QDA

The QDA model is given by the following

Model 1.

$$\mathcal{P}_{D2}^{Q} = \{P_{X|Y}P_Y : p_{x|y} = \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y), p_y = \mathcal{B}ern(\pi_{\widetilde{y}}), \text{ for } y \in \{0,1\}\} \tag{14}$$

that is, the two classes have different class-conditional covariances, but is otherwise the same as the LDA model.

Quadratic discriminant analysis (QDA) functions are defined as follows:

$$\phi_D^Q(\boldsymbol{x}; \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1) = \operatorname*{argmax}_{y\in\mathcal{Y}} \frac{\pi_y}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_y|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_y)^\mathsf{T}\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_y)\right\} \tag{15}$$

Let $\psi_{D,d}^Q$ be the matrix defined by the QR'ed concatenated $\boldsymbol{\delta}$ with the largest $d-1$ eigenvectors of $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$. That is, let $\boldsymbol{S}_y$, $\boldsymbol{V}_y$ be the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}_y$ for $y = \{\widetilde{y}_0, \widetilde{y}_1\}$. Let $\boldsymbol{s} = [\boldsymbol{s}_0, \boldsymbol{s}_1]$ be the concatenation of $\boldsymbol{s}_0 = diag(\boldsymbol{S}_0)$ and $\boldsymbol{s}_1 = diag(\boldsymbol{S}_1)$, and let $\boldsymbol{s}^{(1)} \geq \boldsymbol{s}^{(2)} \cdots \boldsymbol{s}^{(2D)}$ be the sorted eigenvalues. We therefore define

$$\psi_{D,d}^Q = \mathtt{qr}([\boldsymbol{\delta}; (\boldsymbol{V}_{(1)}, \ldots, \boldsymbol{V}_{(d-1)})]) \tag{16}$$

Let $\widetilde{\phi}_{D,d}^Q := \phi_d \circ \psi_{D,d}^Q$. Let $R_{D,d}^Q := R_{0-1,P}(\widetilde{\phi}_{D,d}^L)$. Let $R_{D,d^*}^Q = \min_{d\in[D]} R_{D,d}^Q$.

In this case, it seems like a right thing to do is fit a union of hyperplanes, and then to classifier a new point, first find its cluster, and then project onto that cluster's hyperplane, and use LDA there. This has got to work. The tree depth will set the approximation error (along with some other properties of the distribution).

III.C P is not given

Assume that we obtain n independent and identical realizations from this joint distribution, $(\mathbb{X}_i, \mathbb{Y}_i) \overset{iid}{\sim} P$, for $i \in [n] = \{1, \ldots, n\}$ and $n \ll D$. Let $\mathbb{D}_n \equiv \{(\mathbb{X}_i, \mathbb{Y}_i)\} \in \mathcal{D}_n$ be the data corpus. Let n_y denote the number of samples in class y. Let $\boldsymbol{x}_i \in \mathbb{R}^D$ be a realization of $\mathbb{X}_i$ and let $\boldsymbol{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. Let $\hat{\boldsymbol{x}}$ denote an estimate of $\boldsymbol{x}$. We use the following estimators:

$$\hat{\boldsymbol{\mu}}_y^n = \frac{1}{n_y} \sum_{i:y_i=y} \boldsymbol{x}_i \tag{17}$$

$$\hat{\pi}_y^n = n_y/n \tag{18}$$

$$\hat{\boldsymbol{\delta}}^n = \hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1 \tag{19}$$

$$\widehat{\boldsymbol{\Sigma}}^n = \frac{1}{n-1} \sum_{i \in [n]} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{y_i})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{y_i})^\mathsf{T} \tag{20}$$

$$\hat{\phi}_D^n = \mathbb{I}\{(\boldsymbol{x} - (\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1)/2)^\mathsf{T} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\delta}} + \log(\hat{\pi}_1/\hat{\pi}_0) > 0\} \tag{21}$$

$$[U^n, S^n, V^n] = (\boldsymbol{X}) \tag{22}$$

$$\hat{\psi}_{D,d}^{L,n} = \mathtt{qr}([\hat{\boldsymbol{\delta}}^n; \hat{\boldsymbol{V}}_1^n, \ldots, \hat{\boldsymbol{V}}_{d-1}^n]) \tag{23}$$

$$\widehat{\widetilde{\boldsymbol{X}}}^{L,n} = \hat{\psi}_{D,d}^L \boldsymbol{X}_n \tag{24}$$

$$\hat{\boldsymbol{\Sigma}}_d = \hat{\psi}_{D,d}^L \hat{\boldsymbol{\Sigma}} \hat{\psi}_{D,d}^L \tag{25}$$

$$\hat{\Phi}_{D,d}^n = \{\hat{\phi}_{D,d}^n := \phi_d^n \circ \hat{\psi}_{D,d}^n\} \tag{26}$$

$$\hat{\phi}_{D,d}^{L,n} = \hat{\phi}_d^n \circ \hat{\psi}_{D,d}^{L,n} \tag{27}$$

And let

$$\boldsymbol{X}_i \in \mathbb{R}^D, \quad \boldsymbol{X} \in \mathbb{R}^{D \times n}, \quad \boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}, \quad \boldsymbol{\delta} \in \mathbb{R}^D, \quad \boldsymbol{V} \in \mathbb{R}^{d_n \times D}. \tag{28}$$

Theorem 4. *$R_{D,d_n}^L \to R^*$ as $n \to \infty$ for any D and $d_n \to D$ as $n \to \infty$ and any $P \in \mathcal{P}$.*

Proof. WLOG, let $\pi_0 = \pi_1$ and let $\bar{\boldsymbol{\mu}} = \boldsymbol{0}$, so that risk is determined entirely by $\boldsymbol{X}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta}$. Thus, to show that our classifier is asymptotically consistent, it is sufficient to show that

$$(V\boldsymbol{X})^\mathsf{T} (V \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{V}}^\mathsf{T})^{-1} \hat{\boldsymbol{V}} \hat{\boldsymbol{\delta}} \to \boldsymbol{X}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta} \tag{29}$$

as $n \to \infty$. We know that for $\boldsymbol{V} \in \mathbb{R}^{D \times D}$ orthonormal matrix, and $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^\mathsf{T}$ and $\boldsymbol{\Sigma}^{-1} = \boldsymbol{U}\boldsymbol{S}^{-1}\boldsymbol{U}$, we have

$$(\boldsymbol{V}\boldsymbol{X})^\mathsf{T} (\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\mathsf{T})^{-1} \boldsymbol{V}\boldsymbol{\delta} = \tag{30}$$

$$\boldsymbol{X}^\mathsf{T} \boldsymbol{V}^\mathsf{T} (\boldsymbol{V}\boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^\mathsf{T}\boldsymbol{V}^\mathsf{T})^{-1} \boldsymbol{V}\boldsymbol{\delta} = \tag{31}$$

$$\boldsymbol{X}^\mathsf{T} \boldsymbol{V}^\mathsf{T} (\boldsymbol{V}\boldsymbol{U}\boldsymbol{S}^{-1}\boldsymbol{U}^\mathsf{T}\boldsymbol{V}^\mathsf{T}) \boldsymbol{V}\boldsymbol{\delta} = \tag{32}$$

$$\boldsymbol{X}^\mathsf{T} \boldsymbol{U}\boldsymbol{S}^{-1}\boldsymbol{U}^\mathsf{T} \boldsymbol{\delta} = \tag{33}$$

$$\boldsymbol{X}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta} \tag{34}$$

Therefore, the $R_{D,d_n}^L \to R^*$ as $n \to \infty$ since $\hat{\boldsymbol{\delta}} \to \boldsymbol{\delta}$, V converges to a $D \times D$ orthonormal matrix, and $\hat{\boldsymbol{\Sigma}} \to \boldsymbol{\Sigma}$.

□

Remark 2. *This theorem means that LDA ∘ PCA and LDA ∘ δPCA and, in fact, any $D \times D$ projection matrix is asymptotically consistent. We would like a statement demonstrating that the latter performs better than the former for $n \ll \infty$ or $d_n < D$.*

Theorem 5. *As $n \to \infty$, if $spect(\boldsymbol{\Sigma}) = k\mathbf{1}$, then δ contains all the information, and $R^L < R^P = R^R$ with high probability.*

Proof. $\boldsymbol{\Sigma}^{-1} = k^{-1}\boldsymbol{I}\boldsymbol{\delta} = k^{-1}\boldsymbol{\delta}$. Our classifier uses δ. However, PCA chooses a random vector. □

Theorem 6. *For all $P \in \mathcal{P}_{D2}^L$,*

$$\mathbb{P}[R(\hat{\phi}_{D,d}^{L,n}) \geq R^* + \varepsilon(d, \boldsymbol{\delta}, \boldsymbol{\Sigma})] < \eta(n, d_n, \boldsymbol{\delta}, \boldsymbol{\Sigma}) \tag{35}$$

We know that for a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times D}$ with sub-Gaussian rows (for example, rows sampled iid according to a multivariate normal distribution) that

$$\left\| \hat{\boldsymbol{\Sigma}}^n - \boldsymbol{\Sigma} \right\| \leq \max(\eta, \eta^2) := \varepsilon \quad \text{where } \eta = C_K \sqrt{\frac{n}{D}} + \frac{t}{\sqrt{n}}. \tag{36}$$

with at least probability $1 - 2\exp\{-t^2/2\}$ where C_K depends only on the sub-Gaussian norm $K = \max_i \|\boldsymbol{X}_i\|_{\psi_2}$. (Thm 5.39 of Vershynin 2011).

Theorem 7. *For all $P \in \mathcal{P}_{D2}^L$,*

$$\mathbb{P}[R(\hat{\phi}_{D,d_n}^{L,n}) \geq R^* + \varepsilon(d_n, \boldsymbol{\delta}, \boldsymbol{\Sigma})] < \eta(n, d_n, \boldsymbol{\delta}, \boldsymbol{\Sigma}) \tag{37}$$

Corollary 1. $\varepsilon(d_n, \boldsymbol{\delta}, \boldsymbol{\Sigma}) \to 0$ *as* $n \to \infty$.

III.C(1) don't forget this stuff

Consider the case in which δ lives in the first k dimensional subspace of $\boldsymbol{\Sigma}$

$$\frac{\left\| \hat{\psi}_k \hat{\delta}_n \right\|}{\hat{\delta}_n} \approx \frac{\|\psi_k \delta\|}{\delta} \tag{38}$$

In this case, the above ratio should have a "spectral gap" at k. We then want to show

$$\mathbb{P}[\left\| \hat{\boldsymbol{V}}_{d_n} - \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} \right\| \geq \varepsilon_n] > \eta(P) \tag{39}$$

III.D subsection name

Theorem 8. *(from Tropp) Let $Z_1, \ldots, Z_n \in \mathbb{R}^{D \times D}$ independent symmetric. Define*

$$\sigma^2 := \left\| \sum \mathbb{E} Z_i^2 \right\|, \qquad \bar{D} = 4\frac{\operatorname{tr}\left(\sum \mathbb{E} Z_i^2\right)}{\|\sum \mathbb{E} Z_i^2\|} \leq 4D. \tag{40}$$

If $\mathbb{E}[Z_i] = \mathbf{0}$ and $\|Z_i\| \leq R$ for all $i \in [n]$ almost surely, then, $\forall t \geq 1$, with $\mathbb{P} \geq 1 - e^{-t}$,

$$\left\| \sum Z_i \right\| \leq 2\max\{\sigma\sqrt{t + \log \bar{D}}, R(t + \log \bar{D})\}. \tag{41}$$

Define the *self-adjoint dilation* of a rectangular matrix $\boldsymbol{B}$

$$\mathfrak{D}(\boldsymbol{B}) := \begin{bmatrix} \mathbf{0} & \boldsymbol{B} \\ \boldsymbol{B}^* & \mathbf{0} \end{bmatrix}. \tag{42}$$

$\mathfrak{D}(\boldsymbol{B})$ is, of course, always self-adjoint, and therefore the eigenvalues of $\mathfrak{D}(\boldsymbol{B})$ are equal to those of $\mathfrak{D}(\boldsymbol{B})^\mathsf{T}$, including, in particular, the minimum and maximum. A short calculation yields the important identity

$$\mathfrak{D}(\boldsymbol{B})^2 = \begin{bmatrix} \boldsymbol{B}\boldsymbol{B}^* & \mathbf{0} \\ \mathbf{0} & \boldsymbol{B}^*\boldsymbol{B} \end{bmatrix}. \tag{43}$$

It can also be verified that the self-adjoint dilation preserves spectral information:

$$\lambda_{\max}(\mathfrak{D}(\boldsymbol{B})) = \|\mathfrak{D}(\boldsymbol{B})\| = \|\boldsymbol{B}\|. \tag{44}$$

III.D(1) Applications

Empirical mean: Let x_i be bounded: $\|x_i\| \le \frac{R}{2}$ almost surely. Define $\hat{m} = \frac{1}{n}\sum x_i$.

$$\|m - \hat{m}\| = \left\| \frac{1}{n} \sum_{i\in[n]} \begin{bmatrix} \mathbf{0} & (x_i - m)^\mathsf{T} \\ x_i - m & \mathbf{0} \end{bmatrix} \right\|. \tag{45}$$

Then,

$$\sigma^2 \le \frac{1}{n^2}\sum \mathbb{E}\|x_i - m\|^2 \le \frac{\mathrm{tr}(\Sigma)}{n}, \qquad \bar{D} \le 4\frac{\mathrm{tr}\left(\sum \mathbb{E}Z_i^2\right)}{\|\sum \mathbb{E}Z_i^2\|} \le 4\frac{\mathrm{tr}\left(\mathbb{E}Z_i^2\right)}{\|\mathbb{E}Z_i^2\|} = 8. \tag{46}$$

We obtain

$$\|m - \hat{m}\| \le 2\max\left\{\frac{\mathrm{tr}(\Sigma)}{\sqrt{n}}\sqrt{t + \log 8}, R(t + \log 8)\right\} \tag{47}$$

$$\le 2\frac{\mathrm{tr}(\Sigma)}{\sqrt{n}}\sqrt{t + \log 8}. \tag{48}$$

Empirical covariance:

$$\left\|\hat{\Sigma} - \Sigma\right\| \le \left\|\frac{1}{n}\sum (x_i - m)(x_i - m)^\mathsf{T} - \Sigma\right\| + \left\|(m - \hat{m})(m - \hat{m})^\mathsf{T}\right\|, \tag{49}$$

$$\sigma^2 := \left\|\frac{1}{n}\sum \mathbb{E}Z_i^2\right\| \le \left\|\mathbb{E}\|x_i - m\|^2 (x_i - m)(x_i - m)^\mathsf{T}\right\| \le R^2\|\Sigma\|, \tag{50}$$

$$\bar{D} = 4\frac{\mathrm{tr}\left(\mathbb{E}Z_i^2\right)}{\|\mathbb{E}Z_i^2\|} \le 4D. \tag{51}$$

So,

$$\left\|\hat{\Sigma} - \Sigma\right\| \le 2\max\left\{\sqrt{\frac{R^2}{n}\|\Sigma\|\left(t + \log\frac{4\mathrm{tr}(\mathbb{E}Z_i^2)}{\|\mathbb{E}Z_i^2\|}\right)}, \frac{R}{n}\left(t + \log\frac{4\mathrm{tr}\left(\mathbb{E}Z_i^2\right)}{\|\mathbb{E}Z_i^2\|}\right)\right\}, \tag{52}$$

$$= \sqrt{\frac{R^2}{n}\|\Sigma\|\left(t + \log\frac{4\mathrm{tr}(\mathbb{E}Z_i^2)}{\|\mathbb{E}Z_i^2\|}\right)}. \tag{53}$$

III.E Numerical Results

III.E(1) Simulations

III.E(2) Cancer

III.E(3) Decision Theoretic Task 1: Classification

Let $\Phi = \{\phi\colon \mathcal{X} \to \mathcal{Y}\}$ so that each $\phi \in \mathcal{Y}^{\mathcal{X}}$ is a set of measurable functions from $\mathcal{X}$ to $\mathcal{Y}$. When $\mathcal{Y}$ is a finite or countable set of classes, each ϕ is called a *classifier function*. A *loss function* is a measurable map $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+ \equiv [0, +\infty)$ that we use to assess the performance of a classifier, where $\ell \in \mathcal{L}$. Specifically, we are interested in minimizing risk. Let the *risk*, $R\colon \mathcal{L} \times \mathcal{P} \times \Phi \to \mathbb{R}_+$, of a classifier be the expected loss under the true distribution of the data:

$$R_{\ell,P}(\phi) \equiv \mathbb{E}_P[\ell(\phi(x), y)] \equiv \int_{\mathcal{X}\times\mathcal{Y}} \ell(\phi(x), y) dP. \tag{54}$$

The Bayes optimal classifier for a given distribution is that classifier that minimizes risk:

$$\phi^*(\cdot) = \operatorname*{argmin}_{\phi\in\Phi} R_{\ell,P}(\phi) = \operatorname*{argmin}_{\phi\in\Phi} \mathbb{E}_P[\ell(\phi(\cdot), y)]. \tag{55}$$

Here, we consider $0-1$ loss, $\ell_\phi(x, y) = \mathbb{I}\{\phi(x) \neq y\}$, where $\mathbb{I}\{\cdot\}$ is the identity function taking value unity when its argument is true, and zero otherwise. It is easy to show that when the true distribution of the data P is given, that the Bayes "plug-in" classifier is optimal under $0-1$ loss:

$$\phi^*(\cdot) = \operatorname*{argmax}_{y\in\mathcal{Y}} p_{X|Y=y}(\cdot) p_Y(y). \tag{56}$$

In our setting, we do not know P, however, we can obtain an estimate $\widehat{P}_n$ from the data to plug in. We can think of $\mathcal{P}$ as a collection of densities, each indexed by a parameter $\boldsymbol{\theta}$, $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, where $\boldsymbol{\Theta} \subseteq \mathbb{R}^q$ for some $q \in \mathbb{N}_1 \equiv \{1, 2, \ldots\}$. Thus, obtaining an estimate of P (or p) is equivalent to finding an estimate of the parameter, $\boldsymbol{\theta}$. In particular, we can define the *induced Bayes Classifier*:

$$\widehat{\phi}_n(\cdot) = \operatorname*{argmax}_{y\in\mathcal{Y}} \widehat{p}^n_{X|Y=y}(\cdot) \widehat{p}^n_Y(y), \tag{57}$$

where $\widehat{p}^n$ is the distribution indexed by $\widehat{\boldsymbol{\theta}}_n$. A learning function (algorithm) f is a map $f\colon \cup_{n\in\mathbb{N}_1} \mathcal{D}_n \to \Gamma_F \equiv \Gamma$, such that $\mathbb{D}_n \mapsto \widehat{\gamma}_n \in \Gamma$, where Γ is some functional of the density of F. Two important examples of Γ's for us are the (i) Bayes optimal classifier (which is a functional of F) and (ii) the parameter of the distribution.

Let f_ϕ be a classifier learning algorithm whose output is a function $\widehat{\phi}_n\colon \mathcal{X} \times \mathcal{D}_n \to \mathcal{Y}$, and $\widehat{\phi}_n$ is called the *induced classifier*. Thus, to use an estimate of the Bayes plug-in classifier, $\widehat{\phi}_n$ is determined by an estimate of the parameters, $\widehat{\boldsymbol{\theta}}_n$. A parameter estimate is said to be *consistent* if $\widehat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}$ for some suitable notion of convergence [**?**]. Whenever there exists a consistent estimator for $\boldsymbol{\theta}$, then the induced Bayes plugin classifier is a *consistent classifier*, in the sense that $\widehat{\phi}_n \to \phi$, again for a suitable notion of consistency.[1] Thus, f_ϕ is *uniformly consistent* and is universally uniformly consistent whenever $\mathcal{P}$ is parametric, meaning that $\boldsymbol{\Theta} \subseteq \mathbb{R}^q$ for some $q < \infty$ [**? ?**].

III.E(4) A Simple Example with a Big Data Problem

Let $P_{X|Y}$ be a class-conditional D-dimensional Gaussian distribution and let P_Y be a categorical distribution with only K classes (Bernoulli). Thus, we write $X|Y = \widetilde{y}_k \sim \mathcal{N}_D(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and let $Y \sim \mathcal{C}at(\boldsymbol{w})$, where each $\boldsymbol{\mu}_k \in \mathbb{R}^D$ and $\boldsymbol{\Sigma}_k$ is a $D \times D$ dimensional positive definite matrix denoted $\boldsymbol{\Sigma}_k \succ \mathbf{0}_D$, and $\boldsymbol{w} \in \triangle_K$ which is the K-dimensional simplex, such that $\sum_{k\in[K]} w_k = 1$ and $w_k \geq 0$. Thus, $\boldsymbol{\theta} = \{\boldsymbol{w}, \{\boldsymbol{\mu}_k\}, \{\boldsymbol{\Sigma}_k\}\} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^q$,

[1] I should really learn what I mean here.

and $q = K - 1 + K \times D + K \times D^2/2$. The Bayes plugin classifier for this very simple scenario is simply the classifier that assigns a new observation to its most likely class, k^*

$$k^*(\cdot) = \underset{k \in [K]}{\operatorname{argmax}} \frac{1}{(2\pi)^D |\boldsymbol{\Sigma}_k|} \exp\{-\frac{1}{2}(\cdot - \boldsymbol{\mu}_k)^\mathsf{T} \boldsymbol{\Sigma}_k^{-1} (\cdot - \boldsymbol{\mu}_k)\} w_k. \tag{58}$$

An induced Bayes plugin classifier simply defines $\widehat{k}_n$ like k^* but replaces all the parameters with their corresponding estimates.

All is well and good when we have $n \to \infty$ and $q < \infty$. However, when $n \approx q$ or even $n \ll q$, our parameter estimates, and therefore our classifier, can misbehave badly. For example, when $n \approx q$, our estimates of $\boldsymbol{\Sigma}_k$'s might not even be positive definite. We therefore cannot rely on the above consistency results, and must consider something else.

D Bibliography

[1] V. de Silva and J. B. Tenenbaum, "Global Versus Local Methods in Nonlinear Dimensionality Reduction," in *Neural Information Processing Systems*, 2003, pp. 721–728. 1

[2] Y. LeCun, C. Cortes, and C. Burges, "MNIST handwritten digit database." [Online]. Available: http://yann.lecun.com/exdb/mnist/ 3

[3] T. Yang and C. E. Priebe, "The Effect of Model Misspecification on Semi-supervised Classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 2093–2103, 2011. 6

[4] C. Bouveyron, S. Girard, and C. Schmid, "High-dimensional data clustering," *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 502–519, sep 2007. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0167947307000692 7

[5] T. T. Ngo, M. Bellalij, and Y. Saad, "The Trace Ratio Optimization Problem," pp. 545–569, 2012.

[6] Q. Mai, H. Zou, and M. Yuan, "A direct approach to sparse discriminant analysis in ultra-high dimensions," *Biometrika*, vol. 99, no. 1, pp. 29–42, jan 2012. [Online]. Available: http://biomet.oxfordjournals.org/content/99/1/29.abstracthttp://biomet.oxfordjournals.org/content/early/2012/01/05/biomet.asr066 7