Supervised Manifold Learning for Wide Data

Joshua T. Vogelstein, Mauro Maggioni

Abstract

Supervised learning (SL)—the art and science of discovering statistical relationships from data between multiple different measurement types—is one of the most useful tools in the scientific toolbox. SL has been enabled a wide variety of basic and applied findings, ranging from discovering biomarkers in omics data, to object recognition from images. A special case of SL is classification; a classifier can predict the 'class' of a novel observation via training on a set of paired observations and class labels (for example, predicting male vs. female from MRI scans). In such problems, the goal is to find the optimal discrimnant boundary, which partitions the space of observations into the different classes. In the data age, the dimensionality of the observations is quickly ballooning, and with it, the dimensionality of the discriminant boundary. While historical data may have consisted of only a few dimensions (e.g., height and weight), modern scientific datasets often consist of hundreds, thousands, or even millions of dimensions (e.g. genetics, neuroscience, omics). Regardless of the dimensionality, when a scientist or analyst obtains a new dataset consisting of some observations and labels, she must decide which of the mryiad avaiable tools to use. *Reference* algorithms for datasets with low dimensionality include linear and quadratic discriminant analysis, support vector machines, and random forests ([1]).

Classical methods, however, often rely on very restrictive assumptions. In particular, the theoretical guarantees upon which many classical methods rest require that the number of samples is much larger than the dimensionality of the problem ($n \gg D$). While the dimensionality of datasets is booming, the sample is not witnessing a concomittant increase (see, for example, Table 2 of ([2]) in connectomics). When the number of dimensions is orders of magntidue larger than the sample size, as is now typical, this $n \gg D$ assumption is woefully inadequate. This inadequacy is not a mere theoretical footnote; rather, the implementation of the algorithm itself sometimes fails or crashes if this assumption is not met. Worse, often times the algorithm will run to completion, but the answer will be essentially random, with little or no predicive power or accuracy (e.g., ([3])).

To combat these issues, the fields of statistics and machine learning have developed a large collection of new methods that relax these assumptions, and exhibit significantly improved performance characteristics. Each such approach makes some "structural" assumptions about the data, therefore potentially adding some bias, while reducing variance. The best approach, for a given problem, is the approach that wins this bias/variance trade-off; that is, the approach whose assumptions best correspond to the properties of the data (minimizing bias), and yields estimates have that low error under those assumptions (minimizing variance). These approaches complement "feature engineering" approaches, where the practitioner designs new features based on prior domain specific knowledge.

One of the earliest approaches to discovering low-dimensional representations of data utilizes regularization or shrinkage ([4], [5]). Many shrinkage methods mitigate the dimensionality problem by smoothing ([6]), for example, by regressing parameters to the mean. More recently, a special case of regularization has risen to prominence, called *sparsity* ([7]), in which it is assumed that a small number of dimensions can encode the discriminant boundary ([8]). This assumption, when accurate, can lead to substantial improvements in accuracy with relatively moderate increase in computational cost ([9]). This framework includes methods such as Lasso ([8]), higher criticism thresholding ([10]), and sparse variants of linear discriminant analysis ([6], [11]–[15]).

However, for a wide class of problems, such as image classification, sparisty in the ambient space is an overly restrictive, and therefore, bias-inducing assumption (see Figure [1] for an example on the classic MNIST dataset). A generalization of sparsity is "low-rank", in which a small number of linear combinations of the ambient dimensions characterize the data. Unsupervised low-rank methods date back over a century, including multidimensional scaling ([16], [17]) and principal components analysis ([18], [19]). More recent non-linear versions of unsupervised dimensionality reduction, or manifold learning, include developments from

1

neural network theory such as self-organizing maps (20), generative topographic mapping (21). In this century, manifold learning became more popular, including isomap (22), local linear embedding (23), Laplacian eigenmaps (24), local tangest space alignment (25), diffusion maps (26), and geometric multi-resolution analysis (27).

Such manifold learning methods, while exhibiting both strong theoretical (27–29) and empirical performance, are fully unsupervised. Thus, they discover a low-dimensional representation of the data, ignoring the labels. This can be highly problematic when the discriminant dimensions and the directions of maximal variance are not aligned (see Figure 1 for an example). Supervised dimensionality reduction techniques, therefore, combine the best of both worlds, search for low-dimensional discriminant boundaries. A set of methods from the statistics community is collectively referred to as "sufficient dimensionality reduction" (SIR) methods (30–35). 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 (36) or supervised dictionary learning (37). These methods are limited because they are prone to fall into local minima, they require costly iterative algorithms, and lack any theoretical guarantees (**?**). Thus, there remains a gap in the literature: a supervised learning method with theoretical convergence guarantees appropriate when the dimensionality is orders of magnitude larger than the sample size.

The challenge lies is posing the problem in such a way that efficient numerical algorithms can be brought to bear, without costly iterations or tuning

Figure 1: Illustrating three different embedding strategies—Lasso, PCA, and LOL—for embedding images of digits (from MNIST), each of which is 28 × 28 = 784 dimensional. **Col 1**: Examplars of 3, 7, and 8. **Col 2**: The first four projection matrices learned by the three different approaches on 300 training samples. Note that Lasso is a sparse method, and PCA is unsupervised. **Col 3**: Embedding 500 test samples using each approach. Digits color coded as in Col 1. Note that LOL separates the three different digits very well, unlike Lasso and PCA.

parameters. Our approach, which we call "Linear Optimal Low-rank" (LOL) embedding (see Figure 1), utilizes the first two moments, as do SIR, spectral decompositions, and high-dimensional discriminant analysis methods (35). The intuition for our method comes from a simple geometric intuition as provided in Figure 2. Indeed, we provide both theoretical insight explaining why our method is more general than previous approachs (low-bias), as well as finite sample guarantees (low-variance). A variety of simulations provide further evidence that LOL efficiently finds a better low-dimensional representation than competing methods, not just under the provable model assumptions, but also under much more general contexts (Figure 3). Moreover, LOL can also be used to improve high-dimensional regression and testing (Figure 5). Finally, we demonstrate that LOL achieves better performance, in less time, as compared to several reference classifiers, on several benchmark datasets, including genomics, connectomics, and image processing problems (Figure 4). Therefore, we suggest that LOL be considered as one of the reference method for supervised manifold learning for wide data.

Results

Embedding by Supervised Linear Manifold Learning

To demonstrate the utility of LOL, we first consider one of the most popular benchmark datasets ever, the MNIST dataset. This dataset consists of many thousands of examples of images of the digits 0 through
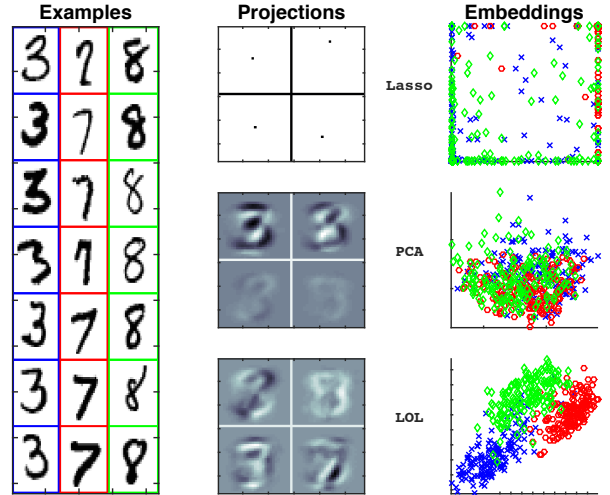
9. Each such image is represented by a 28×28 matrix, which means that the observed (or ambient) dimensionality of the data is $D =784$. Because we are motivated by the $n \ll D$ scenario, we subsample the data to select $n =300$ examples of the numbers 3, 7, and 8.The leftmost column of Figure 1 shows several examples of each, the background of each digit is colored for visualization purposes only. We then consider three different linear dimensionality reduction methods: Lasso, PCA, and LOL. The middle column of Figure 1 shows the estimated projection matrices for each approach (see Methods for details). These are the matrices that transform the data into a low-dimensional representation. The rightmost column of Figure 1 shows 500 randomly selected training samples from the digits 3, 7, and 8, embedded into the first two dimensions using each of the three approaches. The colors of the embedded points correspond to those in the leftcolumn, that is, 3 is blue, 7 is red, and 8 is green.

Lasso, by virtue of being a sparse method, finds the pixels that most discriminate the 3 classes. The resuling embeddings mostly live along the boundaries, because these images are close to binary, and therefore, images either have or do not have a particular pixel. Indeed, although the images themselves are nearly sparse (over 80% of the pixels in the dataset have intensity ≤ 0.05), the discriminant boundary does not seem to be so. PCA, on the other hand, finds the linear combinations of training samples that maximize the variance. This unsupervised linear manifold learning method results in projection matrices that indeed look like linear combinations of the three different digits. The goal here, however, is separating classes, not maximizing dimensionality. The resulting embeddings are not particularly well separated, suggesting the the directions of discriminability are not the same as the directions of maximum variance. LOL is a supervised linear manifold learning method (see below for details). The projection matrices it learns look qualitatively much like those of PCA. This is not surprising, as both are linear combinations of the training examples. The resulting embeddings however, look quite different. The three different classes are very clearly separated by even the first two dimensions. This numerical experiment justifies the use of supervised linear manifold learning, but it does not quantify the effects of performance on downstream inference.

Classification by Supervised Linear Manifold Learning

Here we address classification performance after embedding data from the $n \ll D$ regime. For illustration and theoretical purposes, we consider the simplest $n \ll D$ setting we can imagine. We have a two-class classification problem, where each class is governed by a multivariate Gaussian distribution, and the only difference between the two classes is their means.
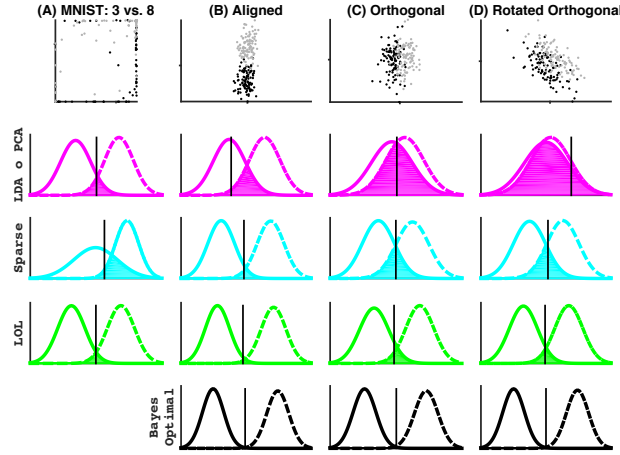
Figure 2: LOL achieves near optimal performance for a wide variety of distributions. Each point is sampled from a multivariate Gaussian; the three columns correspond to different simulation parameters (see §Methods for details). In each of 3 simulations, we sample 100 points in 1000 dimensions. **Row 1**: A scatter plot of the first two dimensions of the sampled points, with class 0 and 1 as black and gray dots, respectively. Rows 2-5: The estimated distribution of test samples after projection onto the discriminant boundary using one of four different methods. For each panel, class 0 and class 1 distributions are shown as solid and dashed lines, respectively. The vertical line shows the classification threshold. The filled area is the expected error rate: the goal of any classification algorithm is to minimize that area. Specifically: **Row 2**: PCA composed with linear discriminant analysis. **Row 3**: ROAD, a sparse method designed specifically for this model. **Row 4**: LOL, our newly proposed method. **Row 5**: the Bayes optimal classifier, which is what all classifiers strive to acheive. **Column (A)**: The mean difference vector is aligned with the direction of maximal variance, maxing it ideal for both PCA to discover the discriminant direction and a sparse solution. **Column (B)**: The mean difference vector is orthogonal to the direction of maximal variance, making both PCA and ROAD struggle to find the discriminant direction. **Column (C)**: Same as (B), but the data are rotated. Note that LOL is closest to Bayes optimal in all three settings.
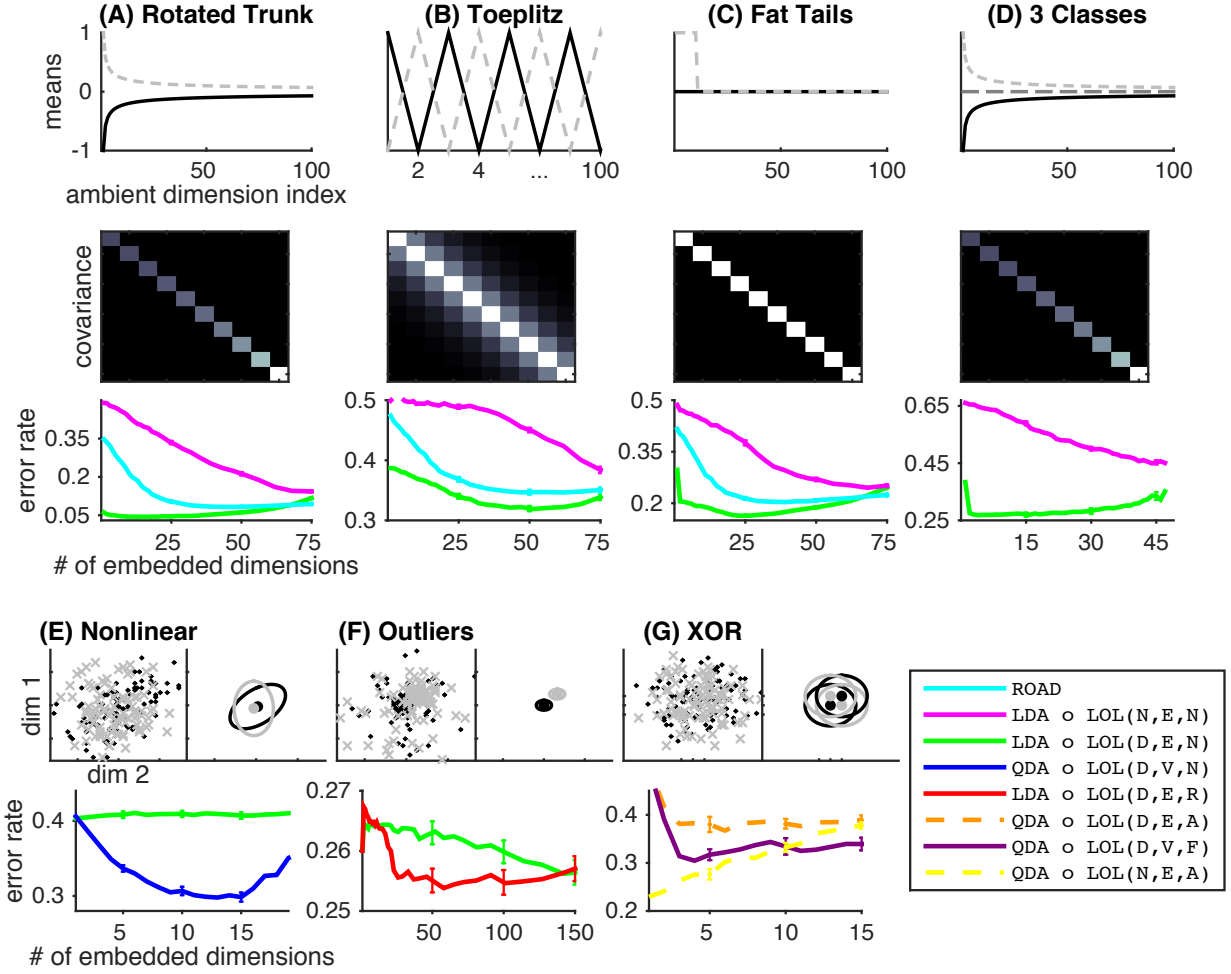
Figure 3: The goal of supervised manifold learning is *not* to learn the *true* manifold, but rather, to learn the *optimal* one, given the amount and complexity of the data. This figure shows seven such examples. For the first three, we show the two means (top panels), the shared covariance matrix (second panels), and the error rate as a function of the number of embedded dimensions (third panels). While (A) and (B) are both Gaussian models with a shared covariance matrix, (C) is a sparse example with fat tails, that is, tails decreasing less quickly than a Gaussian, which is a more general setting. In all three settings, the vanilla LOL method (equivalent to LDA ∘ δ-PCA) outperforms the PCA and spare methods for nearly all settings. Simulations (D) through (G) demonstrate additional generalization, including (D) >2 classes, (E) each class having a unique covariance matrix, (F) data includes outliers, and (G) highly nonlinear discrimination boundary. For each, we show a 2D scatter plot (top panel) and error rate as a function of embedded dimension. Scatter plots also show the Gaussian means (large dot) and 1 standard deviation level sets (circles). A version of LOL designed for each setting is show to improve over the vanilla LOL, as expected. In all 7 cases, the optimal embedding dimension is never the true discriminant dimension, but rather, a smaller number partially determined by sample size.

5
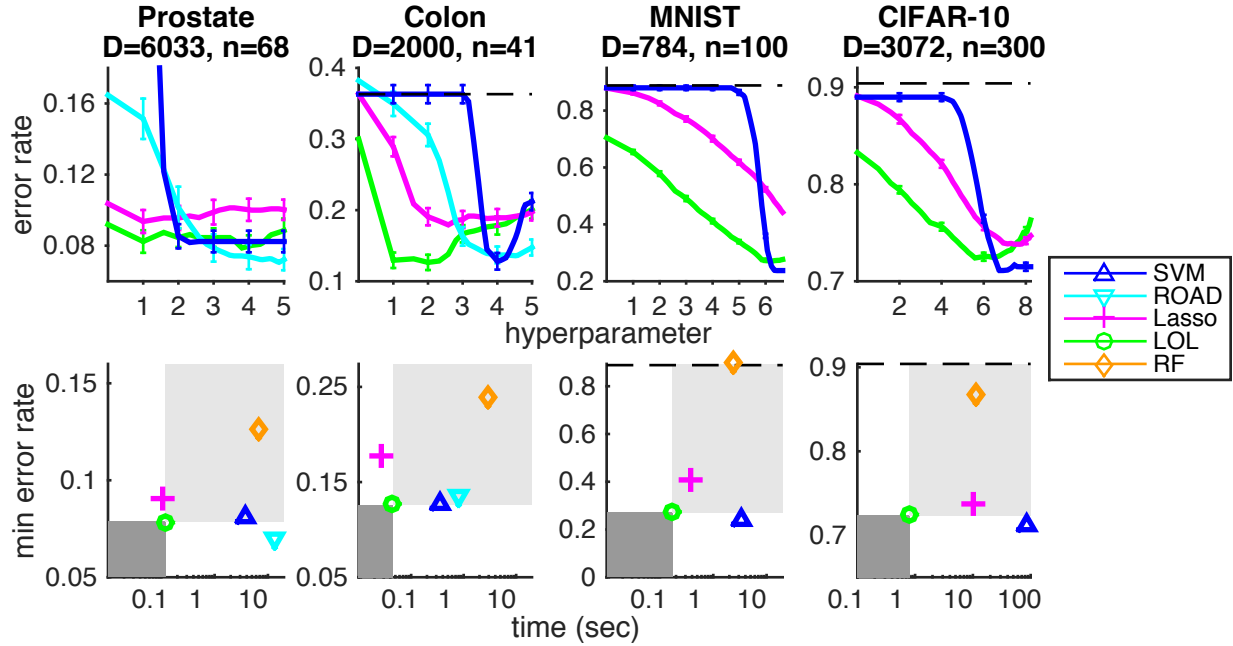
Figure 4: For four standard datasets, we benchmark comparison of LOL (green circles) versus standard classification methods, including support vector machines (blue up triangles), ROAD (cyan down triangles), Lasso (magenta pluses), and random forest (orange diamonds). Top panels show error rate as a function of $\log_2$ number of embedded dimensions (for LOL, ROAD, and Lasso) or cost (for SVM). LOL consistent achieves lower error for fewer dimensions. Bottom panels show the minimum error rate achieved by each of the five algorithms plotted against $\log_{10}$ number of seconds to run. The lower right filled rectangle is the area in which any algorithm is *better* than LOL in terms of both accuracy and efficiency. The upper right filled rectangle is the area in which any algorithm is *worse* than LOL in terms of both accuracy and efficiency. Note that no algorithm is ever in the lower left rectangle; rather, most typically reside in the upper right. This suggests that regardless of how one subjectively weights time versus accuracy, LOL is the best default algorithm in a variety of settings.
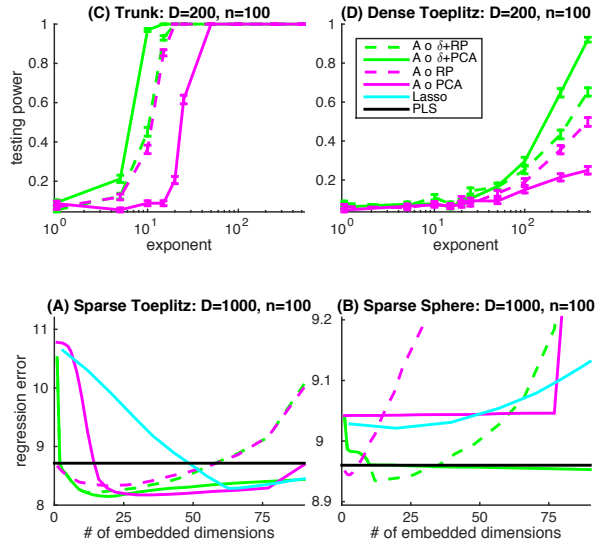
Figure 5: The intuition of including the mean difference vector is equally useful for other supervised manifold learning problems, including regression and testing. (A) and (B) show two regression settings, as described in Methods. In both cases, $\log_{10}$ mean squared error is plotted against the number of embedded dimensions. In both cases, LOL outperforms Lasso (cyan), the classic sparse regression method, as well as partial least squares (PLS; black). (C) and (D) show two different hypothesis testing scenarios. For essentially all settings, vanilla LOL outperforms the other approaches, including that from (**?**). In the legend, 'A' denote either 'linear regression' (in the top panels), or 'Hotelling' (in the bottom panels). These four simulation settings therefore demonstrate the generality of this technique.

ALGORITHMS	p < n	PERFORMANT LDA MODELS, p > n			PERFORMANT GENERALIZED MODELS, p > n					OTHER TASKS, p > n		ALGORITHM PROPERTIES								ALGORITHMS	REF	NOTES
	LDA model	delta & cov aligned	delta & cov misaligned	delta & cov misaligned rotated	Fat Tails	>2 Classes	Different Covariances	Robust to Outliers	Nonlinear Discriminant Boundary	Regression	Testing	Fast	Simple	Theory	Open Source	Abundant	2-step/DL	Sup	Can run on data that does not fit in RAM			
kNN	X	X	X	X	X	X		X	X	X			X	X	X	X		X		kNN	1	not considered
HCT	X	X	X									X	X	X		X		X		HCT	14	no code
LDA o Trace-Ratio	X	X	X	X								X	X	X		X		X		LDA o Trace-Ratio	4	no code
F2M / SDR	X			X		X							X	X	X	X		X		F2M / SDR	10	needs n>p
Deep Learning	X				X	X	X	X	X	X			X	X	X	X		X	X	Deep Learning	11	wants n>>p
Vowpal Wabbit	X				X	X		X		X		X			X	X		X	X	Vowpal Wabbit	5	wants n>>p
Ridge Regression	X									X		X	X	X	X	X		X		Ridge Regression	1	regression
Partial Least Squares										X		X								Partial Least Squares	1	regression
Naive Bayes	X	X	X			X				X		X	X	X	X	X		X	?	Naive Bayes	1,12	
HDDA	X	X	X	X		X									X	X		X		HDDA	6	
ROAD	X	X											X	X	X			X		ROAD	7,8,9	
LASSO	X	X				X				X		X	X	X	X			X		LASSO	3	
kSVM	X	X	X	X	X	X	X	X	X				X	X	X	X		X	?	kSVM	1	
Random Forest	X												X	X	X	X		X	?	Random Forest	1, 15	
LDA	X											X	X	X	X			X		LDA	1	
LDA o PCA	X	X				X				X		X	X	X	X	X	X		X	LDA o PCA	2	
LOL	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	LDA o d+PCA	*	
FIGURE		1A	1B	1C	2C	2D	2E	2F	2G	3A, 3B	3C, 3D	4										

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 (2) 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 (38**?**). Sparse methods include screening methods such as higher criticism screening (10), L1 penalized methods such as LASSO (8), and more recent sparse discriminant analysis methods (13, 15, 39). While very useful because they admit readily interpretable results, they are guaranteed to work only in very restricted situations (38). 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) (18**?**). In the >110 years since its original development, many extensions and generalizations of PCA have been proposed, including random projections (40), NMF (41**?**), and kernel variants of PCA such as isomap (22), LLE (23), and Laplacian Eigenmaps (24). 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 (27, 29), 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 (30, 32, 35**? ?**). 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 (36) or supervised dictionary learning (37). These methods are limited because they are prone to fall into local minima, they require costly iterative algorithms, and lack any theoretical guarantees (42). 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 covarianced, $\hat{\boldsymbol{\Sigma}}$. 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 2(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 (43, 44). 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 (15)), 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. Unfortunately 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 ∘ 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
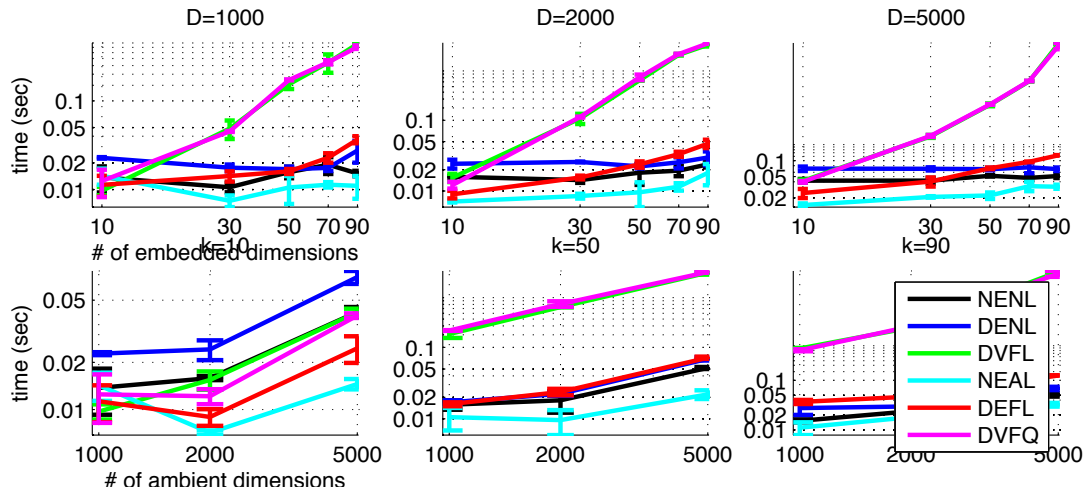
Discussion

Figure 7

B ig 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 (**?**). The top row of Figure 1 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 (10), L1 penalized methods such as LASSO (8), and more recent sparse discriminant analysis methods (13, 15, 39). 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 (19), random projections (40), NMF (41**?**), and LLE (23). 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 (30, 32, 35**? ?**), 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 (36) or supervised dictionary learning (37). 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 2, top left panel).

Figure 2 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, $\boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}$. 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.

1 Results

1.1 Main Approach

1.2 Numerical Results

Illustrative example on real data

Generalizations The simplicity of LOL enables a wide range of applications and extensions (Figure 5).

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∘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) (45), sparse optimal scoring (SOS) (13), direct sparse discriminant analysis (DSDA) (**?**), and regularized optimal affine discriminant (ROAD) (15). (**?**) proved that the first three are equivalent, and (14) numerically compares them with ROAD, as well as nearest shrunken centroids (PAM) (11) 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 (**?**). (14) 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 (15), 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∘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 (46) 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.

Scalability

Novel Applications

1.3   Theoretical Results

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.*

2   Discussion

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

Figure 8: top: regression, bottom: two sample testing

Related work   (4, 14, 47**? ? ? ? ? ?** –54)

A   Approaches for dealing with Wide Data

A.1   Unsupervised Manifold Learning

A.2   Sparse Methods

A.3   Supervised Manifold Learning

B   Theory

C Supplement

C.1   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 $[\boldsymbol{S}, \boldsymbol{V}] = \mathtt{eig}(\boldsymbol{\Sigma})$, such that $\boldsymbol{V}_j$ is the jth 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 $[\boldsymbol{Q}, \boldsymbol{R}] = \mathtt{qr}(\boldsymbol{Y})$ 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 $[\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{V}] = \mathtt{svd}(\boldsymbol{X})$ 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 $\mathbb{R}^D$.

C.2 P **is Given,** $k = 2$

C.2.1   LDA

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

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

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 $\widetilde{\boldsymbol{\Sigma}} = \psi_{D,d}\boldsymbol{\Sigma}\psi_{D,d} \in \mathbb{R}^{d\times d}_{\preceq 0}$, where $\mathbb{R}^{d\times d}_{\preceq 0}$ is the set of $d \times d$ positive definite matrices.

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

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

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

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

Remark 1. *When $P \in \mathcal{P}^{LDA}_{D2}$, $d^* = 1$ 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.*

C.2.2 QDA

The QDA model is given by the following

Model 1.

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

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:

$$\phi^Q_D(\boldsymbol{x}; \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1) = \underset{y \in \mathcal{Y}}{\arg\max} \frac{\pi_y}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_y|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_y)^\mathsf{T} \boldsymbol{\Sigma}_y^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_y)\right\} \tag{15}$$

Let $\psi^Q_{D,d}$ 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

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

Let $\widetilde{\phi}^Q_{D,d} := \phi_d \circ \psi^Q_{D,d}$. Let $R^Q_{D,d} := R_{0-1,P}(\widetilde{\phi}^L_{D,d})$. Let $R^Q_{D,d^*} = \min_{d \in [D]} R^Q_{D,d}$.
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).

C.3 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 $\hat{x}$ denote an estimate of x. We use the following estimators:

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $\hat{\boldsymbol{\delta}} \to \boldsymbol{\delta}$, $\hat{\boldsymbol{V}}$ converges to a $D \times D$ orthonormal matrix, and $\hat{\boldsymbol{\Sigma}} \to \boldsymbol{\Sigma}$.

$\square$

Remark 2. *This theorem means that LDA $\circ$ PCA and LDA $\circ$ 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 $spect(\boldsymbol{\Sigma}) = k\boldsymbol{1}$, then δ contains all the information, and $R^L < R^P = R^R$ with high probability.*

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

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

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

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

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

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 $P \in \mathcal{P}_{D2}^L$,*

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

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

C.3.1 don't forget this stuff

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

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

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

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

C.4 subsection name

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

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

If $\mathbb{E}[Z_i] = 0$ 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 $\boldsymbol{B}$

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

$\mathfrak{D}(\boldsymbol{B})$ 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

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

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

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

C.4.1 Applications

Empirical mean: Let x_i be bounded: $\|x_i\| \leq \frac{R}{2}$ 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,

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

We obtain

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

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

Empirical covariance:

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

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

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

So,

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

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

C.5 Numerical Results

C.5.1 Simulations

C.5.2 Cancer

C.5.3 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:

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

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:

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

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*:

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

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 (55). 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 $q < \infty$ (56, 57).

C.5.4 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^*

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

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 Σ_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] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*.   Springer, 2001. 1

[2] A. DiMartino, D. Fair, C. Kelly, T. Satterthwaite, F. Castellanos, M. Thomason, R. C. Craddock, B. Luna, B. Leventhal, X.-N. Zuo, and M. Milham, "Unraveling the Miswired Connectome: A Developmental Perspective," *Neuron*, vol. 83, no. 6, pp. 1335–1353, Sep. 2014. [Online]. Available: http://www.cell.com/article/S0896627314007806/fulltext 1, 9

[3] A. Eklund, M. Andersson, C. Josephson, M. Johannesson, and H. Knutsson, "Does parametric fMRI analysis with SPM yield valid results?-An empirical study of 1484 rest datasets." *NeuroImage*, vol. 61, no. 3, pp. 565–578, Apr. 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22507229 1

[4] J. H. Friedman, "Regularized Discriminant Analysis," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165–175, Mar. 1989. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01621459.1989.10478752 1, 18

[5] P. J. Bickel and E. Levina, "Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations," *Bernoulli*, vol. 10, no. 6, pp. 989–1010, Dec. 2004. [Online]. Available: http://projecteuclid.org/euclid.bj/1106314847 1

[6] D. M. Witten and R. Tibshirani, "Covariance-regularized regression and classification for high-dimensional problems." *Journal of the Royal Statistical Society. Series B, Statistical methodology*, vol. 71, no. 3, pp. 615–636, Feb. 2009. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2806603&tool=pmcentrez&rendertype=abstract 1

[7] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vision Research*, vol. 37, no. 23, pp. 3311–3325, Dec. 1997. [Online]. Available: http://dx.doi.org/10.1016/S0042-6989(97)00169-7 1

[8] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B*, vol. 58, pp. 267–288, 1996. 1, 9, 15

[9] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data - Methods, Theory and Applications*.   Springer, 2011. [Online]. Available: http://www.springer.com/statistics/statistical+theory+and+methods/book/978-3-642-20191-2 1

[10] D. L. Donoho and J. Jin, "Higher criticism thresholding: Optimal feature selection when useful features are rare and weak." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 39, pp. 14 790–5, Sep. 2008. [Online]. Available: http://www.pnas.org/content/105/39/14790http://www.pnas.org/content/105/39/14790.short 1, 9, 15

[11] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 10, pp. 6567–6572, May 2002. [Online]. Available: http://www.pubmedcentral. nih.gov/articlerender.fcgi?artid=124443&tool=pmcentrez&rendertype=abstract 1, 16

[12] J. Fan and Y. Fan, "High-dimensional classification using features annealed independence rules," *The Annals of Statistics*, vol. 36, no. 6, pp. 2605–2637, Dec. 2008. [Online]. Available: http://projecteuclid.org/euclid.aos/1231165181

[13] L. Clemmensen, T. Hastie, D. M. Witten, and B. Ersbø ll, "Sparse Discriminant Analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, Nov. 2011. [Online]. Available: http://pubs.amstat.org/doi/ abs/10.1198/TECH.2011.08118 9, 15, 16

[14] Q. Mai and H. Zou, "A Note On the Connection and Equivalence of Three Sparse Linear Discriminant Analysis Methods," *Technometrics*, vol. 55, pp. 243–246, 2013. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/00401706.2012.746208 16, 18

[15] J. Fan, Y. Feng, and X. Tong, "A road to classification in high dimensional space: the regularized optimal affine discriminant," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 74, no. 4, pp. 745–771, Sep. 2012. [Online]. Available: http: //doi.wiley.com/10.1111/j.1467-9868.2012.01029.x 1, 9, 10, 15, 16, 17

[16] A. S. Householder and G. Young, "Discussion of a set of points in terms of their mutual distances," *Psychometrika*, vol. 3, no. 1, pp. 19–22, 1938. [Online]. Available: http://www.springerlink.com/index/10.1007/BF02287916 1

[17] I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling: Theory and Applications (Springer Series in Statistics)*. Springer, 2010. [Online]. Available: http://www.amazon.com/ Modern-Multidimensional-Scaling-Applications-Statistics/dp/1441920463 1

[18] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine Series 6*, vol. 2, no. 11, pp. 559–572, Nov. 1901. [Online]. Available: http: //www.tandfonline.com/doi/abs/10.1080/14786440109462720#.VMlOwl7F-oU 1, 9

[19] I. T. Jolliffe, *Principal Component Analysis*, second edi ed.   New York: Springer, 2002. 1, 15

[20] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, 1982. [Online]. Available: http://link.springer.com/10.1007/BF00337288 2

[21] C. M. Bishop, M. Svensén, and C. K. I. Williams, "GTM: The Generative Topographic Mapping," *Neural Computation*, vol. 10, no. 1, pp. 215–234, Jan. 1998. [Online]. Available: http://www.mitpressjournals.org/doi/abs/10.1162/089976698300017953#.VMIRX17F-oU 2

[22] J. B. Tenenbaum, V. de Silva, J. C. Langford, and V. D. Silva, "A global geometric framework for nonlinear dimensionality reduction." *Science*, vol. 290, no. 5500, pp. 2319–23, Dec. 2000. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11125149 2, 9

[23] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding." *Science (New York, N.Y.)*, vol. 290, no. 5500, pp. 2323–6, Dec. 2000. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11125150 2, 9, 15

[24] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data," *Neural Computation*, vol. 15, pp. 1373–1396, 2003. 2, 9

[25] Z. Zhang and H. Zha, "Principal Manifolds and Nonlinear Dimensionality Reduction via Tangent Space Alignment," *SIAM Journal on Scientific Computing*, vol. 26, no. 1, pp. 313–338, Jan. 2004. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/S1064827502419154 2

[26] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, Jul. 2006. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1063520306000546 2

[27] W. K. Allard, G. Chen, and M. Maggioni, "Multi-scale geometric methods for data sets II: Geometric Multi-Resolution Analysis," *Applied and Computational Harmonic Analysis*, vol. 32, no. 3, pp. 435–462, May 2012. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1063520311000868 2, 9

[28] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, Sep. 1936. [Online]. Available: http://www.springerlink.com/content/9v4274h33h75lq24/

[29] V. De˜Silva and J. B. Tenenbaum, "Global Versus Local Methods in Nonlinear Dimensionality Reduction," in *Advances in Neural Information Processing Systems*, 2003, pp. 721–728. 2, 9

[30] K.-C. Li, "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, vol. 86, no. 414, pp. 316–327, Jun. 1991. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01621459.1991.10475035 2, 9, 15

[31] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method." in *Proceedings 37th Allerton Conference on Communication, Control, and Computing*, 1999.

[32] A. Globerson and N. Tishby, "Sufficient Dimensionality Reduction," *Journal of Machine Learning Research*, vol. 3, no. 7-8, pp. 1307–1331, Oct. 2003. [Online]. Available: http://www.crossref.org/jmlr_DOI.html 9, 15

[33] R. D. Cook and L. Ni, "Sufficient Dimension Reduction via Inverse Regression," *Journal of the American Statistical Association*, vol. 100, no. 470, pp. 410–428, Jun. 2005. [Online]. Available: http://amstat.tandfonline.com/doi/abs/10.1198/016214504000001501#.U6tH3Y1dUts

[34] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces," *Journal of Machine Learning Research*, vol. 5, pp. 73–99, 2004.

[35] 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 2, 9, 15

[36] P. J. Huber, "Projection Pursuit," *The Annals of Statistics*, vol. 13, no. 2, pp. 435–475, Jun. 1985. [Online]. Available: http://projecteuclid.org/euclid.aos/1176349519 2, 9, 15

[37] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised Dictionary Learning," Sep. 2008. [Online]. Available: http://arxiv.org/abs/0809.3083 2, 9, 15

[38] P. Bühlmann, M. Kalisch, and L. Meier, "High-Dimensional Statistics with a View Toward Applications in Biology," *Annual Review of Statistics and Its Application*, vol. 1, no. 1, pp. 255–278, 2014. [Online]. Available: http://www.annualreviews.org/doi/full/10.1146/annurev-statistics-022513-115545 9

[39] 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 9, 15

[40] E. J. Candès and T. Tao, "Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4016283 9, 15

[41] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization." *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999. [Online]. Available: http://dx.doi.org/10.1038/44565http://www.ncbi.nlm.nih.gov/pubmed/10548103 9, 15

[42] M. Belkin and P. Niyogi, "Convergence of Laplacian Eigenmaps," *Statistics*, pp. 1–31, 2008. 9

[43] T. T. Cai, Z. Ma, and Y. Wu, "Sparse PCA: Optimal rates and adaptive estimation," *The Annals of Statistics*, vol. 41, no. 6, p. 46, Nov. 2013. [Online]. Available: http://arxiv.org/abs/1211.1309http://projecteuclid.org/euclid.aos/1388545679 10

[44] D. E. Fishkind, C. Shen, and C. E. Priebe, "On the Incommensurability Phenomenon," *Arxiv preprint 1301.1954v1*, Jan. 2013. [Online]. Available: http://arxiv.org/abs/1301.1954 10

[45] C.-L. Wu and A.-S. Chiang, "Genes and Circuits for Olfactory-Associated Long-Term Memory in Drosophila." *Journal of neurogenetics*, no. June, pp. 1–28, Nov. 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19012072 16

[46] 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. 17

[47] J. Duchene and S. Leclercq, "An optimal transformation for discriminant and principal component analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, 1988. 18

[48] S. Aeberhard, D. Coomans, and O. de Vel, "Comparative analysis of statistical pattern recognition methods in high dimensional settings," *Pattern Recognition*, vol. 27, no. 8, pp. 1065–1077, Aug. 1994. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0031320394901457

[49] H. Bensmail and G. Celeux, "Regularized Gaussian Discriminant Analysis through Eigenvalue Decomposition," *Journal of the American Statistical Association*, vol. 91, no. 436, pp. 1743–1748, Dec. 1996. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476746

[50] J. J. Rodríguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A new classifier ensemble method." *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 10, pp. 1619–30, Oct. 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16986543

[51] R. Caruana, N. Karampatziakis, and A. Yessenalina, "An empirical evaluation of supervised learning in high dimensions," in *Proceedings of the 25th international conference on Machine learning - ICML '08*.   New York, New York, USA: ACM Press, Jul. 2008, pp. 96–103. [Online]. Available: http://dl.acm.org/citation.cfm?id=1390156.1390169

[52] A. Kaban and R. J. Durrant, "Dimension-Adaptive Bounds on Compressive FLD Classification," *Algorithmic Learning Theory*, vol. 8139, pp. 294–308, 2013. [Online]. Available: http://link.springer.com/10.1007/978-3-642-40935-6

[53] W. R. Carson, M. Chen, M. R. D. Rodrigues, R. Calderbank, and L. Carin, "Communications-Inspired Projection Design with Application to Compressive Sensing," *SIAM Journal on Imaging Sciences*, vol. 5, no. 4, pp. 1185–1212, Oct. 2012. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/120878380

[54] M. Chen, W. R. Carson, M. Rodrigues, R. Calderbank, and L. Carin, "Communications Inspired Linear Discriminant Analysis," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 2012, pp. 919–926. [Online]. Available: http://machinelearning.wustl.edu/mlpapers/papers/ICML2012Chen_482 18

[55] P. J. Bickel and K. A. Doksum, *Mathematical Statistics: Basic Ideas And Selected Topics, Volume 1*. Prentice Hall, 2006. [Online]. Available: http://books.google.com/books?id=wImDQgAACAAJ&pgis=1 25

[56] V. Vapnik, *Statistical Learning Theory*, ser. Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications and Control, S. Haykin, Ed.   Wiley, 1998, vol. 2, no. 4. [Online]. Available: http://www.amazon.com/Statistical-Learning-Theory-Vladimir-Vapnik/dp/0471030031 25

[57] S. Villa, L. Rosasco, and T. Poggio, "On Learnability, Complexity and Stability," Mar. 2013. [Online]. Available: http://arxiv.org/abs/1303.5976 25