

# Wide Supervised Manifold Learning by Low-Rank Optimal Linear Discriminant Analysis (LOL)

Joshua T. Vogelstein, Brett Mensh,  
Minh Tang, Da Zheng, Randal Burns, Mauro Maggioni

## Abstract

In the 21st century high-dimensional observations abound. In classical statistics, it is well known that as the dimensionality of a dataset increase, the number of samples required to estimate properties thereof increases even faster. A number of techniques have been developed to stave off this “curse of dimensionality”, including dimensionality reduction and regularization. Mitigating this curse is especially demanding in supervised learning problems, such as classification. The challenge in finding a low-dimensional discriminant boundary in the face of high-dimensional observations is that direct searches are computationally infeasible. Therefore, previous art have taken one of two approaches. First, one can ignore the discriminant problem, and try to reduce the dimensionality of the data—for example, using principal components analysis or manifold learning techniques—in hopes that one has not discarded the discriminant information. Second, one directly solve the problem after making a number of overly restrictive assumptions of the data, such as independence and sparsity. Here, we demonstrate via a simple geometric argument that one can algebraically compute a low-dimensional embedding that is approximately optimal with high probability in many settings. The intuition is that we can project the data both in the directions of maximal variance (ignoring the classification task) and the linear direction of maximal discrimination (ignoring the variance). The result is an extremely computationally efficient supervised manifold learning method that we call “Linear Optimal Low-Rank” (LOL). LOL outperforms state-of-the-art algorithms in a wide variety of settings, including both those for which theoretical arguments predict it should, and in much more general settings for which we are unable as yet to obtain solid theoretical footing. This includes both simulated settings, where our intuition enables us to characterize a number of simple variants designed for more complex settings, as well as several benchmark problems. In particular, LOL outperforms several reference algorithms in terms of both efficiency and accuracy on said benchmarks. We therefore believe that LOL could be useful in myriad applications, and many extensions are readily available.

Supervised learning (SL)—the art and science of estimating statistical relationships using labeled training data—is a crucial tool in scientific discovery. SL has been enabled a wide variety of basic and applied findings, ranging from discovering biomarkers in omics data [? ], to object recognition from images [? ], and includes celebrated methods such as support vector machines, random forests, and deep networks [1]. A special case of SL is classification; a classifier predicts the ‘class’ of a novel observation via partitioning the space of observations into different classes. (for example, predicting male vs. female from MRI scans). In this big data age, the ambient (or observed) dimensionality of the observations is quickly ballooning, and with it, the ambient 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 scientists or analysts obtain new datasets, they must make a number of decisions, by either implicitly or explicitly weighting considerations across several levels of analysis, including both statistical and computational considerations.

As suggested above, the goal of such problems is to estimate a discriminant boundary that partitions the space as optimally and efficiently as possible. If we knew the “true data generating mechanism”, then writing down the optimal discriminant boundary would be relatively trivial. However, in all real data science problems, at least some parts of the true data generating mechanism remain unknown, because the “Truth” is likely infinitely complex. Because we do not know the truth, we seek tools that will estimate boundaries as accurately and efficiently as possible given the training data available to us. To do so, one must make a series of decisions. These decisions have a variety of ramifications; the art of supervised learning is making choices that balance those ramifications appropriately for the problem at hand. The choices can reasonably be organized into four levels of analysis: (i) model, (ii) algorithm, (iii) implementation, and (iv) platform, each have different implications, both quantitative and qualitative. The quantitative impact can be further subdivided into statistical and computational implications, with statistical implications largely determined by the top two levels, and computational implications governed by the bottom two. Below, we list in italics the *choices* one must make, and in bold the **design considerations** associated with each. These considerations directly determine the design goals of our machine learning system, as will be elaborated upon below.

First, one must choose a *discriminant model* (often called feasible region or action space in other contexts, and often this choice is implicit in the algorithm choice). This modeling assumptions determine geometric constraints on the shape of the estimated boundary. For example, a linear boundary is a potential discriminant model, and the particular feasible optimal boundary for a given problem is then specified by (a potentially multidimensional) parameter: the slope and intercept. **Model bias** is the minimum distance between the optimal discriminant boundary in the discriminant model and true boundary (determined by the data generating mechanism). As mentioned above, in real data problems, we always expect some model bias unless the discriminant model is “non-parametric”, meaning that it can account for any discriminant boundary.

Second, one must choose an *algorithm* to estimate the boundary, such as Fishers Linear Discriminant (FLD), linear Support Vector Machines (SVM), or logistic regression. Each such algorithm has different convergence properties as sample size increases. Importantly, for different data generating mechanisms, different algorithms may converge on different solutions. The additional distance between the expected estimated boundary and the optimal boundary is called the **estimator bias**, which we also want to minimize by choosing an algorithm that will select an estimate as close as possible to the truth. Different algorithms will also differ in their **estimator variance**, which is a function both of the size of the discriminant model, and the idiosyncrasies of the algorithm. Of note, in certain circumstances, the two bias terms and the variance together determine expected mean square error, so one also desires to minimize the variance of the estimated boundary.

In real data problems, there are two additional quantitative considerations impacted by the algorithm choice: **hyper-parameter searches** and **robustness**. As mentioned above, we desire to minimize both the bias and the variance; alas, decreasing one often comes at the cost increasing the other. For any given data generating mechanism and dataset, the optimal trade-off is unknown. Fortunately, many algorithms are equipped with a hyper-parameter, that enables the practitioner to “tune” the bias vs. variance to obtain an

“optimal” fit in terms of expected error on other data. The ease and efficiency with which those hyper-parameters can be tuned can substantially impact performance. The robustness of an algorithm quantifies its ability to estimate nearly optimal boundaries when there are outliers. This is another consideration that plays a crucial role in real data.

The third and fourth choices impact computational considerations. Third, one must choose an *implementation*. First, implementations should be **numerically stable**, over the range of the number of samples and dimensions under consideration. Second, one desires an implementation that minimizes computational complexity, both in terms of **space** and **time**. Fourth, one must choose a *platform* upon which the implementation operates, including both software dependencies and hardware choice. In the big data age, both **scale-up** and **scale-out** are crucial design considerations. Scale-up, generically speaking, refers to the ability of utilizing all available computational resources, rather than allowing computers to sit idly. Scale-out refers to the ability to incorporate additional available resources. Together, these terms determine the scalability and resource efficiency of a system, which can have dramatic consequences in real world big data problems.

All of the above quantitative considerations can be evaluated **theoretically**, but typically only under (overly) restrictive assumptions. Even when the assumptions are met, theory often merely provides bounds, without concrete numbers in any particular situations. **Simulated experiments** can therefore buttress the theoretical results. By knowing the truth, one can compare the performance of different approaches to provide insight into when the theories are working as expected, including in settings that extend beyond the theoretical constraints. Moreover, **real data performance**, provide further insight into different methodologies in real data settings.

The design or choice of a machine learning system for solving a given real world problem reflects each of the above design considerations. Regardless of the problem, we want a discriminant model that exhibits low bias for that problem, an algorithm with low estimator bias and variance, with easy hyper-parameter searches that is robust to outliers, an implementation that is numerically stable with small space and time requirements, deployed on a platform that scales both up and out. And we want all the properties demonstrated via theory, buttressed by simulations, and corroborated by real data experiments. Obviously, it is quite a challenge to optimize along all of the above dimensions for any particular kind of exploitation task. We therefore focus on the simplest supervised learning task we can think of: a high-dimensional two-class classification problem. To our knowledge, there does not yet exist a machine learning system that performs well along each of the above design considerations for even this problem.

The first century of statistical pattern recognition focused on elegant SL approaches for low-dimensional problems (e.g., Fisher’s Linear Discriminant (FLD) [? ]), and dimensionality reduction techniques for high-dimensional problems (e.g., Principal Components Analysis (PCA) [? ]), but did not seriously investigate high-dimensional SL approaches. More recent work has focused on high-dimensional SL approaches, including sparse methods (e.g., Lasso [? ]) and supervised dictionary learning techniques [2]. However, theoretical guarantees for sparse methods require very strict assumptions that are often far from accurate [? ], meaning that sparse methods tend to exhibit significant bias. Moreover, their susceptibility to correlations of the predictor variables can lead to substantial variance, insofar that which variables are selected can be highly noise dependent [? ]. Dictionary learning, on the other hand, has scant theoretical guarantees in any setting [? ]. Moreover, because dictionary learning tends to consider a large complex discriminant model, its resulting estimates are highly dependent on initialization choice. Finally, computationally, both are problematic for big data, in that they require large amounts of time, and implementations on scalable platforms are currently unavailable.

In complementary work, generalizations of principal components to include various kinds of nonlinear dimensionality reduction techniques, collectively referred to as “manifold learning”, have been developing recently [? ]. These tools are predominantly “unsupervised”, and therefore, do not explicitly search for a discriminant boundary, and therefore may add significant bias to the problem at hand. Adding supervision to these techniques so far has been difficult; methods proposed thus far embed the data in an unsupervised fashion first, again, potentially adding bias [? ], and require substantial computations. Thus, a supervised manifold learning algorithm designed for one of the simplest possible high-dimensional classification problems remains amiss.

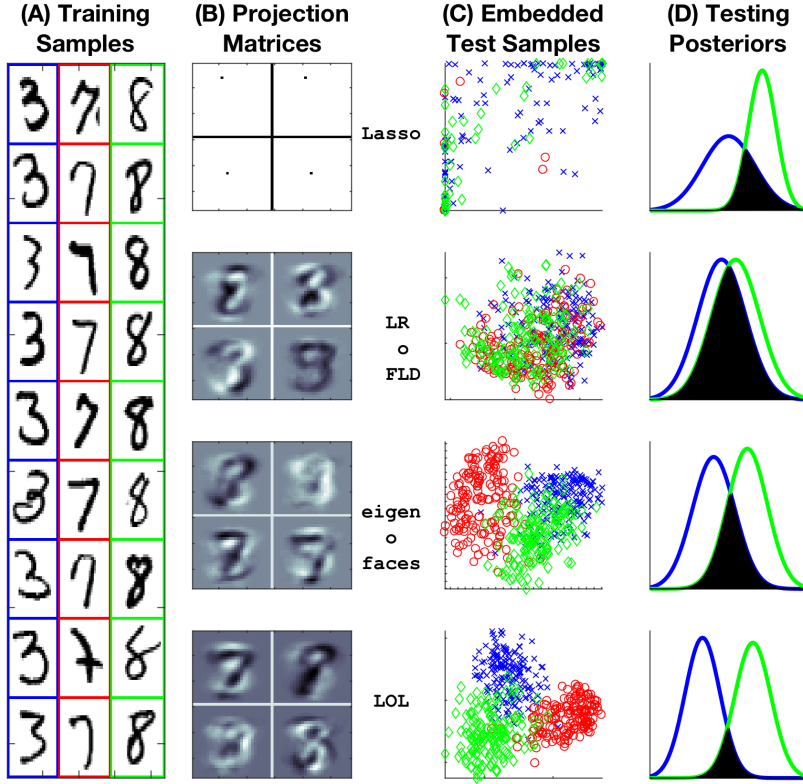
The computer science community has been developing tools for big data for several decades now. Specifically, algorithm complexity has lowered significantly, and ability to scale-up and scale-out has increased significantly. This is true both for simple statistical primitives, such as computing a mean, and more complicated primitives like PCA [? ]. More recently, a large number of distributed machine learning libraries have become available, including Apache Spark’s mllib, H2O, Dato, and Vowpal Wabbit [? ]. These focus almost entirely on large sample size and low dimensionality regime, whereas the motivating problems of interest for this work are small sample size and high-dimensionality. Moreover, they utilize a distributed platform that is susceptible to low bandwidth communication between nodes, meaning that as one adds resources, the computations cannot scale out optimally.

Thus, a gap remains in the literature for a statistical tool that, under relatively general assumptions, achieves both small bias and variance, can be made robust, and has easily tunable hyper-parameters. And complementarily, an implementation of that algorithm that is numerically stable, requires minimal space and time, is highly scalable for high-dimensional problems. Importantly, the method should perform well both on simulated and real data experiments. And ideally, the software is intuitive, free and open source, and easy to use.

Here, we describe a novel solution, called Low-Rank Optimal Linear Discriminant Analysis (LOL), that satisfies all of the above desiderata. We developed LOL by considering the above described simple scenario, and realizing that the optimal dimension reduction for such a setting would have to combine both the means and the covariances, because together those two parameters determine the optimal decision boundary. While other methods have previously had a similar insight (in particular the “First Two Moment” (F2M) methods [3–7]), F2M methods only operate in low-dimensional settings. Therefore, we decided to combine the means and covariances differently, essentially by first projecting the data onto the means, and then effectively computing PCA on the “residuals”. Running FLD after projecting the data onto the subspace spanned jointly by the means and the top principal components comprises the LOL method. In addition to MATLAB and R, we also implemented LOL using FlashX, a recently developed big data analytics library [? ]. We were therefore able to take advantage of matrix parallelizations optimized for machines with solid state drives.

We demonstrate LOL outperforms—or at least performs no worse than—several reference benchmarks along every statistical consideration mentioned above. This includes proving the improvement over several standards, along both statistical considerations (bias and variance) for the simple Gaussian setting mentioned above. Moreover, we provide a distributed implementation that improves on previous computational capabilities in terms of: stability, time, space, and scalability under more general settings. Moreover, numerical experiments quantitatively demonstrate the improvement of LOL over reference methods on a wide range of simulated experiments that satisfy our theoretical assumptions. By virtue of LOL being built upon geometric intuition, it is easy to extend it along various dimensions. Additional numerical experiments demonstrate that practice matches our intuition, in particular, generalizations of LOL can easily be constructed to match different geometric assumptions of the data. Indeed, even when data are sampled from distributions outside our theory and intuition, we see that LOL achieves smaller error than its competitors. In fact, we can extend our intuition outside of classification problems to other problems, including wide regression and hypothesis testing. A trivial modification of LOL leads to regression and testing procedures that also outperform the natural competitors. Computationally, LOL is always numerically stable, and requires no more computational space, and smaller computational time, than both Fisherfaces and sparse methods. Moreover, we provide a highly scalable implementation that efficiently runs on terabyte datasets with hundreds of thousands or millions of features. Finally, we test LOL against a set of standard methods on four different benchmark real datasets. For any of the four different settings, there was not a single method that outperformed LOL on both accuracy and time.

In addition to the above quantitative properties, LOL is intuitive, only has a single hyper-parameter which is exceedingly easy to tune, and we provide free and open source implementations, in MATLAB, R, and FlashX, as well as pre-configured environments that will enable anybody to run LOL on commodity machines, including laptops, workstations, and cloud instances. LOL is not likely to necessarily be the best algorithm on any given new high-dimensional challenge problem. Rather, the arguments and methodology developed herein provide a framework for developing big supervised manifold learning algorithms to tackle



**Figure 1:** Illustrating three different classifiers—**LASSO** (top), **FLDPCA** (middle), and **LOL** (bottom)—for embedding images of the digits 3, 7, and 8 (from MNIST), each of which is  $28 \times 28 = 784$  dimensional. **(A)** A subset of training samples (boundary colors are only for visualization purposes). **(B)** The first four projection matrices learned by the three different approaches on 300 training samples. Note that **LASSO** is sparse and supervised, **PCA** is dense and unsupervised, and **LOL** is dense and supervised. **(C)** Embedding 500 test samples into the top 2 dimensions using each approach. Digits color coded as in A. **(D)** The estimated posterior distribution of test samples after projecting data into 5-dimensions via each method. We show only 3 vs. 8 for simplicity. The vertical line shows the classification threshold. The filled area is the estimated error rate; the goal of any classification algorithm is to minimize that area. Clearly, **LOL** exhibits the best separation after embedding, which results in the best classification performance.

the data science problems of the next century.

## Results

### An Illustrative Real Data Example of Supervised Linear Manifold Learning

A general strategy for high-dimensional classification proceeds as schematized in Figure 1: A obtain/select  $n$  training samples of the data, B learn a low dimensional projection, C project  $m$  testing samples onto the lower dimensional space, D classify the embedded testing samples using some classifier. We consider three different linear dimensionality reduction methods—**LASSO**, **PCA**, and **LOL**—each of which we compose with a classifier to form high-dimensional classifiers.<sup>1</sup>

To demonstrate the utility of **LOL**, we first consider one of the most popular benchmark datasets ever, the MNIST dataset [9]. This dataset consists of  $n = 60,000$  examples of images of the digits 0 through 9. Each such image is represented by a  $28 \times 28$  matrix, which means that the observed (or ambient) dimensionality of the data is  $p = 28^2 = 784$ . Because we are motivated by the  $n \ll p$  scenario, we subsample the data to

<sup>1</sup>Although **LASSO** is not a 2-step method (where embedding is learned first, and then a classifier is applied), adaptive lasso [8] and its variants improve on lasso's theoretical and empirical properties, so we consider such an approach here.



select  $n = 300$  examples of the numbers 3, 7, and 8. We then apply all three approaches to this subsample of the MNIST dataset, learning a projection, and embedding  $m = 500$  testing samples, and classifying the resulting embedded data.

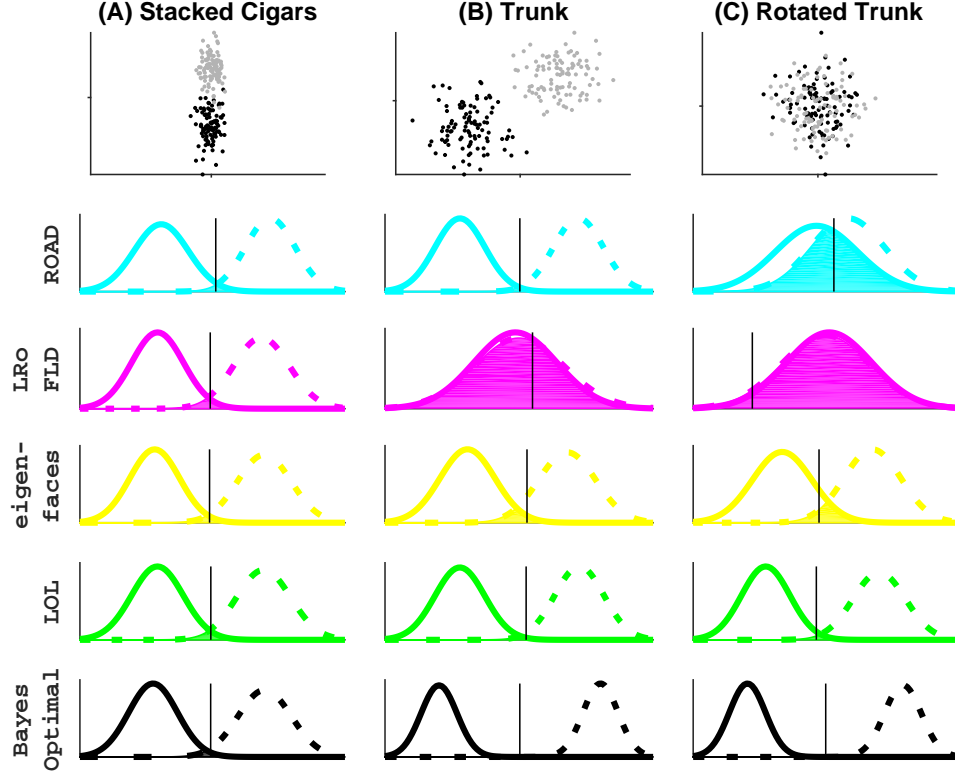
**LASSO**, by virtue of being a sparse method, finds the pixels that most discriminate the 3 classes. The resulting embeddings mostly live along the boundaries, because these images are close to binary, and therefore, images either have or do not have a particular pixel. Indeed, although the images themselves are nearly sparse (over 80% of the pixels in the dataset have intensity  $\leq 0.05$ ), a low-dimensional discriminant boundary does not seem to be sparse. **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 variability. 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 our newly proposed 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. The result of these embeddings yields classifiers whose performance is obvious from looking at the embeddings: **LOL** achieves significantly smaller error than the other two approaches. This numerical experiment justifies the use of supervised linear manifold learning, we next build a geometric intuition for these methods in simple simulated examples, to better illustrate when we can expect **LOL** to outperform other methods, and perhaps more importantly, when we expect **LOL** to fail.

## Linear Gaussian Intuition

The above real data example suggests the geometric intuition for when **LOL** outperforms its sparse and unsupervised counterparts. To further investigate, both theoretically and numerically, we consider the simplest setting that illustrates the relevant geometry. In particular, we consider a two-class classification problem, where both classes are distributed according to a multivariate normal distribution, where the two classes have the same covariance matrix. so that the only difference between the classes is their means (we call this the Linear Discriminant Analysis (**LDA**) model; see Methods for details).

Consider what the optimal projection would be in this scenario. The optimal low-dimensional projection is analytically available as the dot product of the difference of means and the inverse covariance matrix,  $A_* = \delta^T \Sigma^{-1}$  [10] (see Methods for derivation). **PCA**, the dominant unsupervised manifold learning method, utilizes only the covariance structure of the data, and ignores the difference between the means (because **PCA** for classification operates on the class-conditionally centered data, see Methods for details). In particular, **PCA** would project the data on the top  $d$  eigenvectors of the covariance matrix. **The key insight of our work is the following: we can combine the difference of the means and the covariance matrix in a simple fashion, rather than just the covariance matrix, to find a low dimensional projection.** Naïvely, this should typically improve performance, because in this stylized scenario, both are important. The **F2M** literature has a similar insight, but a different construction that requires the dimensionality to be smaller than the sample size [3–7]. We implement our idea by simply concatenating the difference of the means with the top  $d$  eigenvectors of the covariance matrix. This is equivalent to first projecting onto the difference of the means vector, and then projecting the residuals onto the first  $d$  principle components. Thus, it requires almost no additional computational time or complexity over that of **PCA**, rather, merely estimates the difference of the means. In this sense, **LOL** can be thought of as a very simple “supervised **PCA**”.

Figure 2 shows three different examples of data sampled from the **LDA** model to geometrically illustrate this intuition. In each, we sample  $n = 100$  training samples in  $p = 1000$  dimensional space, so  $n \ll p$ . Figure 2A shows an example we call “stacked cigars”. In this example the covariance matrix is diagonal, so all ambient dimensions are independent of one another. Moreover, the difference between the means and direction of maximum variance are both large along the same dimensions (they are highly correlated with one another). This is an idealized setting for **PCA**, because **PCA** finds the direction of maximal variance, which happens to correspond to the direction of maximal separation. However, in the face of high-dimensional data, **PCA** does



**Figure 2:**  $\text{LOL}$  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  $n = 100$  points in  $p = 1000$  dimensions. And for each approach, we embed into the top 20 dimensions. Note that we use the sample estimates, rather than the true population values of the parameters. The five columns show (in decreasing order): **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. **Row 2**  $\text{FLD} \circ \text{PCA}$ . **Row 3**  $\text{ROAD}$ , a sparse method designed specifically for this model [11]. **Row 4**  $\text{LOL}$ , our newly proposed method. **Row 5** the Bayes optimal classifier, which is what all classifiers strive to achieve. **(A)** The mean difference vector is aligned with the direction of maximal variance, making it ideal for both  $\text{PCA}$  to discover the discriminant direction and a sparse solution. In this setting, the results are similar for all methods, and essentially optimal. **(B)** The mean difference vector is orthogonal to the direction of maximal variance, making  $\text{PCA}$  fail, but sparse methods and  $\text{LOL}$  can still recover the correct dimensions, achieving nearly optimal performance. **(C)** Same as B, but the data are rotated, in this case, only  $\text{LOL}$  performs well. Note that  $\text{LOL}$  is closest to Bayes optimal in all three settings.

not weight the discriminant directions sufficiently, and therefore performs only moderately well.<sup>2</sup> Because all dimensions are independent, this is a good scenario for sparse methods. Indeed, ROAD, a sparse classifier designed for precisely this scenario, does an excellent job finding the most useful ambient dimensions. LOL does the best of all three approaches, by using both the difference of the means and the covariance. This is obvious upon comparing each approach to the Bayes optimal solution for this setting, which is analytically available due to its simplicity.

Figure 2B shows an example which is a worst case scenario for using PCA to find the optimal projection for classification. In particular, the variance is getting larger for subsequent dimensions,  $\sigma_1 < \sigma_2 < \dots < \sigma_p$ , while the magnitudes of the difference between the means are decreasing with dimension,  $\delta_1 > \delta_2 < \dots > \delta_p$ . Thus, for any truncation level, PCA finds exactly the *wrong* directions. ROAD is not hampered by this problem, it is also able to find the directions of maximal discrimination, rather than those of maximal variance. Again, LOL, by using both the means and the covariance, does extremely well.

Figure 2C is exactly the same as B, except the data have been randomly rotated in all 1000 dimensions. This means that none of the original coordinates have much information, rather, linear combinations of them do. This is evidenced by observing the scatter plot, which shows that the first two dimensions fail to disambiguate the two classes. PCA, being rotationally invariant, fails in this scenario as it did in B. Now, there is no small number of ambient dimensions that separate the data well, so ROAD also fails. LOL is unperturbed by this rotation; in particular, it is able to “unrotate” the data, to find dimensions that optimally separate the two classes.

## Statistical Theory

The above numerical experiments provide the intuition to guide our theoretical developments.

**Theorem 1.** *LOL better than (or as good as) PCA, under the LDA model when the parameters are provided.*

In words, it is better to incorporate the mean difference vector into the projection matrix. The *degree* of improvement is a function of the embedding dimension  $d$ , the ambient dimensionality  $p$ , and the parameters (see Methods for details and proof), but the *existence* of an improvement, or at least no worse performance, is independent of those factors. Note that better here includes both bias and variance. In the population version of these two methods, there is no variance, so only the bias is important. Moreover, we know the conditions under which LOL is strictly better than PCA, they are formally stated in Assumption ?? (in Methods). Informally, LOL is better than PCA whenever the angle between the mean vector and the top  $d$  principal components is large enough. Under reasonable assumptions, in the appendix, we show that this happens 84% of the time.

## Numerical Experiments Extending Our Theoretical Results

In the previous section, we proved that LOL is often better than PCA, and never worse under the LDA model, regardless of observed dimensionality, embedded dimensionality, and parameters, assuming the parameters are given. In this section, we numerically investigate what happens when the parameters must be estimated from training data, and in particular, from  $n = 100$  training samples, when the observed dimensionality is  $p = 100$ . These results strengthen our theoretical results by providing exact magnitudes of performance, rather than merely better than guarantees, in addition to operating in a finite sample setting, more akin to real data scenarios.

**LDA Model** We begin by investigating three scenarios that satisfy the LDA model assumptions required by our proofs. First, consider the rotated trunk example from Figure 2C as well as a “Toeplitz” example, as depicted in Figures 3A and B, respectively. In both cases, we compare LOL to FLDOPCA and ROAD. In both scenarios, for all dimensions, LOL achieves a lower error rate than either of its competitors.

**Multiple Classes** LOL can trivially be extended to  $> 2$  class situations. Naïvely it may seem like we would

<sup>2</sup>When having to estimate the eigenvector from the data, PCA performs even worse. This is because when  $n \ll p$ , PCA is an inconsistent estimator with large variance [12, 13].



need to keep all pairwise differences between means. However, given  $k$  classes, the set of all  $k^2$  differences is only rank  $k - 1$ . In other words, we can equivalently compute the distance of each mean to a single mean (in practice, for minimizing variance reasons, we use the class which has the maximum number of samples (breaking ties randomly)). Figure 3C shows a 3-class generalization of the rotated Trunk example from Figure 3A. While LOL uses the additional class naturally, many previously proposed high-dimensional classifiers, such as ROAD, natively only work for 2-classes. Thus, we only compare LOL to FLDOPCA here. As before, LOL significantly outperforms FLDOPCA for all embedding dimensions.

## Generalizations of LOL

The above simulations were all under the LDA model, for which we have theoretical guarantees that LOL should outperform FLDOPCA. In this section, we explore simulation settings that extend beyond the geometric constraints we required for our proofs, to develop an even deeper understanding of LOL.

**Fat Tails** Figure 3D shows a sparse example with “fat tails” to mirror real data settings better. More specifically, each class is the sum of multiple Gaussians, with the same mean, but different covariances (see Methods for details). The qualitative results are consistent with those of A and B, even though the setting is no longer exactly the setting under which we have theoretical confirmation. More specifically, LOL outperforms FLDOPCA and ROAD for all dimensions.

**QDA** Sometimes, it makes more sense to model each class as having a unique covariance matrix, rather than a shared covariance matrix. Assuming everything is Gaussian, the optimal classifier in this scenario is called Quadratic Discriminant Analysis (QDA) [1]. Intuitively then, we can modify LOL to compute the eigenvectors separately for each class, and concatenate them (sorting them according to their singular values). Moreover, rather than classifying the projected data with LDA, we can then classify the projected data with QDA. Indeed, simulating data according to such a model (Figure 3E), LOL performs slightly better than chance, regardless of the number of dimensions we use to project, whereas QQQ (for Quadratic Optimal QDA) performs significantly better regardless of how many dimensions it keeps. This demonstrates a straightforward generalization of LOL, available to us because of the simplicity and intuitiveness of LOL.

**Outliers** Outliers persist in many real data sets. Finding outliers, especially in high-dimensional data, is both tedious and difficult. Therefore, it is often advantageous to have estimators that are robust to certain kinds of outliers [14–16]. PCA and eigenvector computation are particularly sensitive to outliers [17]. Because LOL is so simple and modular, we can replace typical eigenvector computation with a robust variant thereof, such as the geometric median subspace embedding [18]. Figure 3F shows an example where we generated  $n/2$  training samples according to the simple LDA model, but then added another  $n/2$  training samples from a noise model. LRL (Linear Robust Low-Rank), performs better than LOL regardless of the number of dimensions we keep. This simulation setting further demonstrates the flexibility of the LOL framework to be extensible to other, more complicated scenarios.

**XOR** XOR is perhaps the simplest nonlinear problem, the problem that led to the demise of the perceptron, prior to its resurgence after the development of multi-layer perceptrons [19]. Thus, in our opinion, it is warranted to check whether any new classification method can perform well in this scenario. The classical (two-dimensional) XOR problem is quite simple: the output of a classifier is zero if both inputs are the same (00 or 11), and the output is one if the inputs differ (01 or 10). Figure 3G shows a high dimensional and stochastic variant of XOR (see Methods for details). This simulation was designed such that standard classifiers, such as support vector machines and random forests, achieve chance levels (not shown). LOL, performs moderately better than chance, and QQQ performs significantly better than chance, regardless of the chosen dimensionality. This demonstrates that our classifiers developed herein, though quite simple and intuition, can perform well even in settings where the data are badly modeled by our underlying assumptions. This mirrors previous findings where the so-called “idiots’s Bayes” classifier outperforms more sophisticated classifiers [10]. In fact, we think of our work as finding intermediate points between idiot’s Bayes (or naïve Bayes) and FLD, by enabling degrees of regularization by changing the dimensionality used.

## Computational Efficiency

In many applications, the main quantifiable consideration in whether to use a particular method, other than accuracy, is computational efficiency. Because implementing `LOL` requires only highly optimized linear algebraic routines—including computing moments and singular value decomposition—rather than the costly iterative programming techniques currently required for sparse or dictionary learning type problems. Moreover, because `LOL` only requires computing means and covariances, we can leverage recent advances in parallel matrix operations [20] on fast solid-state drives (SSDs). We implement `LOL` with the R interface of FlashX [20–22]. Given the FlashX programming framework, we can run `LOL` on essentially arbitrarily large data. For these examples, we simulated data satisfying the `LDA` model; specifically, we used a spherically symmetric covariance matrix, with means just  $\pm$  the one vector (see Methods for simulation details).

Figure 4A demonstrates both the in memory and semi-external memory [23] computing models of our implementation. For the in memory implementation (light green line), we see that the run time increases linearly with the number of dimensions, requiring about 11 minutes to run `LOL` on a  $p = 32,000,000$  dimensional problem with  $n = 2000$  samples. This linear increase is the optimal scale up according to Ahmdel’s Law [24]. This example already requires 477 gigabyte (GB) to store the data, but  $<1$  GB to perform the computations. For larger data, we developed a semi-external memory implementation, which stores the data matrix on solid state drives, and the low-rank estimates in RAM [25]. This strategy allows us to run `LOL` as long as the data can fit on disk, and the low-rank approximation can fit in RAM. The semi-external memory implementation (dark green line) achieves the same performance as the in memory implementation whenever the data are small enough for in memory, and continues scaling optimally (linearly) as the dimensionality further increases to  $p = 128,000,000$  dimensions, a multi-terabyte scenario.

Another option for making `LOL` scale up better is to utilize randomized algorithms. In particular, random projections—for which the data are multiplied by a lower-dimensional random matrix—have been shown to provide excellent approximation eigenvectors [26]. Moreover, very sparse random projections, in which the elements of the matrix are mostly zero, with  $\pm 1$  randomly distributed, have been shown to be effective, and have significant computational benefits [27]. We therefore further modified FlashX to incorporate random projections and very sparse random projections, and let `LAL` denote Linear Approximate Low-rank. Figure 4A shows an order of magnitude improvement in both the in-memory and semi-external memory implementations of `LAL`.

Figure 4B shows the error for `LOL` and `FLDOPCA` in this scenario, simply to demonstrate that even in this extremely high dimensional setting, `LOL` still outperforms `FLDOPCA`. As expected, `LAL` performs as well as `LOL` for the low dimensional settings.

## Computational Theory

We reify the above computational experiments with theoretical statements. Computing the mean for each class requires  $\mathcal{O}(n_j p)$  floating point operations (flops), where  $n_j$  is the number of samples per class. Subtracting the means requires  $\mathcal{O}((J-1)p)$  flops, where  $J$  is the total number of classes. Computing the first  $d$  singular triples (left and right singular vectors, and singular value) requires  $\mathcal{O}(npd)$  flops. The sparse random projection however only requires  $\mathcal{O}(npd/c)$ , where  $c \ll 1$  is the sparsity of the sparse random projection matrix. Since,  $d \ll n, p$ , this can substantially reduce computational complexity. Regardless, `LOL` computation is clearly bottlenecked by the `PCA` computation or random projection for single threaded operations.

Our parallel in memory implementation scales-up optimally (linearly) with the number of threads, requiring  $\mathcal{O}(npd/T)$  flops for `LOL`, and  $\mathcal{O}(npd/cT)$  flops for `LAL`, where  $T$  is the number of threads. Our parallel semi-external memory implementation scales-out optimally (linearly) as well. This is in agreement with the linear fit of the lines in Figure 4A.

## Benchmark Real Data Applications

Although `LOL` both statistically and computationally outperforms its natural competitors both in theory and in a variety of simulation settings, real data often tell a different story. We have therefore selected four commonly used high-dimensional datasets to compare `LOL` to several state-of-the-art algorithms (see Methods

for details). For each dataset, we compare  $\text{LOL}$  to (i) support vector machines (SVM), (ii)  $\text{ROAD}$ , (iii) lasso, (iv) and random forest (RF). Because in practice all these approaches have “hyperparameters” to tune, we consider several possible values for SVM, lasso, and  $\text{LOL}$  (but not RF, as its runtime was too high). Figure 5 shows the results for all four datasets. For each, we use MATLAB implementations, partially because  $\text{ROAD}$  provides MATLAB code, and also because of MATLAB’s ubiquity in certain communities. Because we also have R and FlashX implementations of  $\text{LOL}$ , comparisons using other languages would be straightforward.

Qualitatively, the results are similar across datasets:  $\text{LOL}$  achieves high accuracy and computational efficiency as compared to the other methodologies. Considering Figure 5A and B, two popular sparse settings, we find that  $\text{LOL}$  can find very low dimensional projections with very good accuracy. For the prostate data, with a sufficiently non-sparse solution for  $\text{ROAD}$ , it slightly outperforms  $\text{LOL}$ , but at substantial computational cost; in particular,  $\text{ROAD}$  takes about 100 times longer to run on this dataset. Figure 5C and D are 10-class problems, so  $\text{ROAD}$  is no longer possible. Here, SVM can again slightly outperform  $\text{LOL}$ , but again, requiring 100 fold additional computational time. In all cases, the beloved random forest classifier performs subpar. In all four scenarios, there is never an algorithm that achieves smaller error in less time, as indicated by the dark gray box being empty in the lower panels of Figure 5A-D.

## Extensions to Other Supervised Learning Problems

The utility of incorporating the mean difference vector into supervised machine learning for big and wide data extends beyond merely classification. In particular, hypothesis testing can be considered as a special case of classification, with a particular loss function. Therefore we apply the same idea to a hypothesis testing scenario. The multivariate generalization of the t-test, called Hotelling’s Test, suffers from the same problem as does the classification problem; namely, it requires inverting an estimate of the covariance matrix whose estimate would be low-rank and therefore singular, in the high-dimensional setting. To mitigate this issue in the hypothesis testing scenario, prior art applied similar tricks as they have done in the classification setting. One particularly nice and related example is that of Lopes et al. [28], who addresses this dilemma by using random projections to obtain a low-dimensional representation, following by applying Hotelling’s Test in the lower dimensional subspace. Figure 6A and B shows the power of their test alongside the power of the same approach, but using the  $\text{LOL}$  projection rather than random projections. The two different simulations include the simulated settings considered in their manuscript (see Methods for details). The results make it clear that the  $\text{LOL}$  test has higher power for essentially all scenarios. Moreover, it is not merely the replacing random projections with  $\text{PCA}$  (solid magenta line), nor simply incorporating the mean difference vector (dashed green line), but rather, it appears that  $\text{LOL}$  for testing uses both modifications to improve performance.

High-dimensional regression is another supervised learning method that can utilize the  $\text{LOL}$  idea. Linear regression, like classification and Hotelling’s Test, requires inverting a matrix as well. By projecting the data only a lower dimensional subspace first, followed by linear regression on the low-dimensional data, we can mitigate the curse of high-dimensions. To choose the projection matrix, we partition the data into  $K$  partitions, based on the percentile of the target variable, we obtain a  $K$  class classification problem. Then, we can apply  $\text{LOL}$  to learn the embedding. Figure 6C shows an example of this approach, contrasted with  $\text{LASSO}$  and partial least squares, in a sparse simulation setting (see Methods for details).  $\text{LOL}$  is able to find a better low-dimensional projection than  $\text{LASSO}$ , and performs significantly better than partial least squares, for essentially all choices of number of dimensions to embed into.

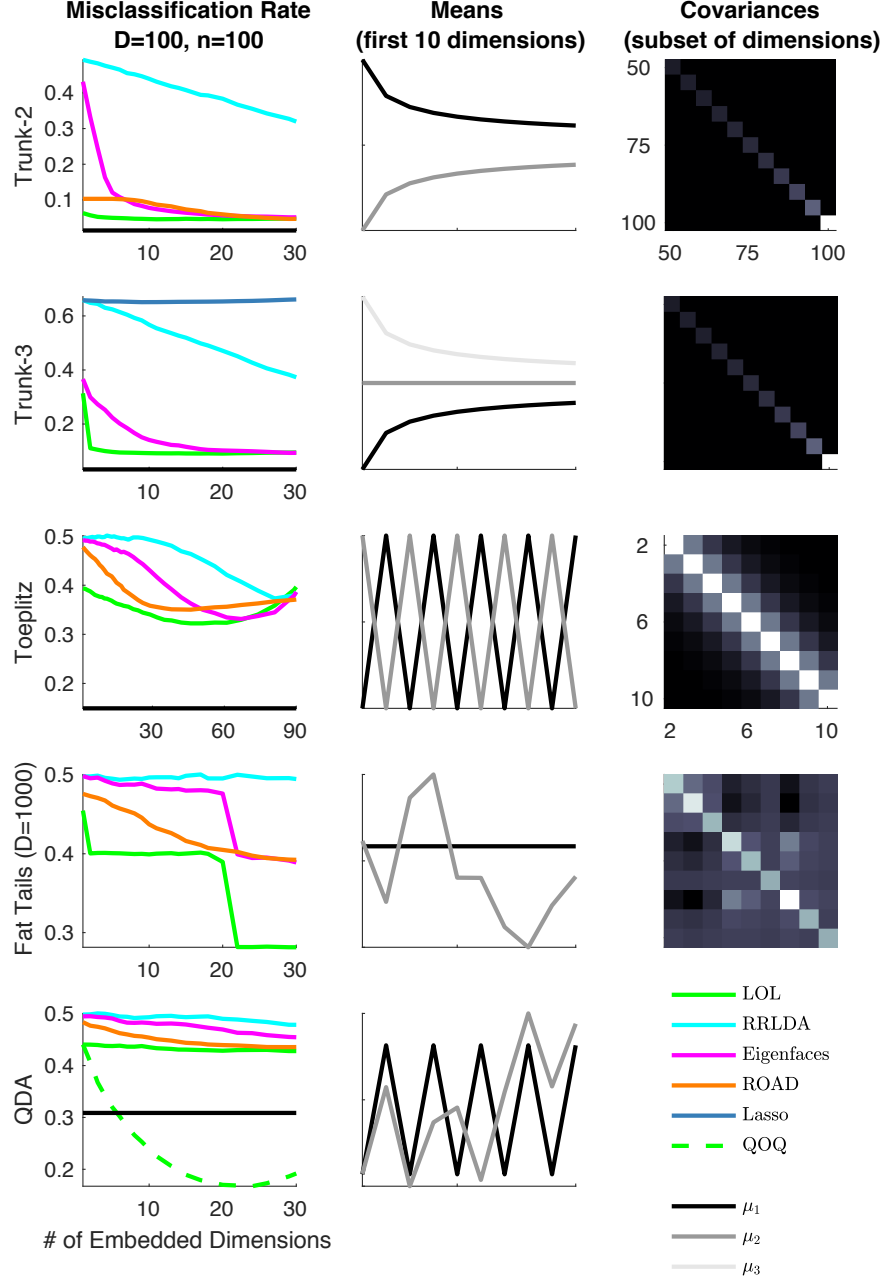
## Discussion

We have introduced a very simple, yet new, methodology to improve performance on supervised learning problems with big and wide data. In particular, we have proposed a supervised manifold learning procedure that utilizes both the difference of the means, and the covariance matrices, and proved that it performs better than first applying  $\text{PCA}$  to the data under reasonable assumptions. This is in stark contrast to most previous approaches, which only utilize the covariance matrices (or kernel variants thereof), or try to solve a difficult

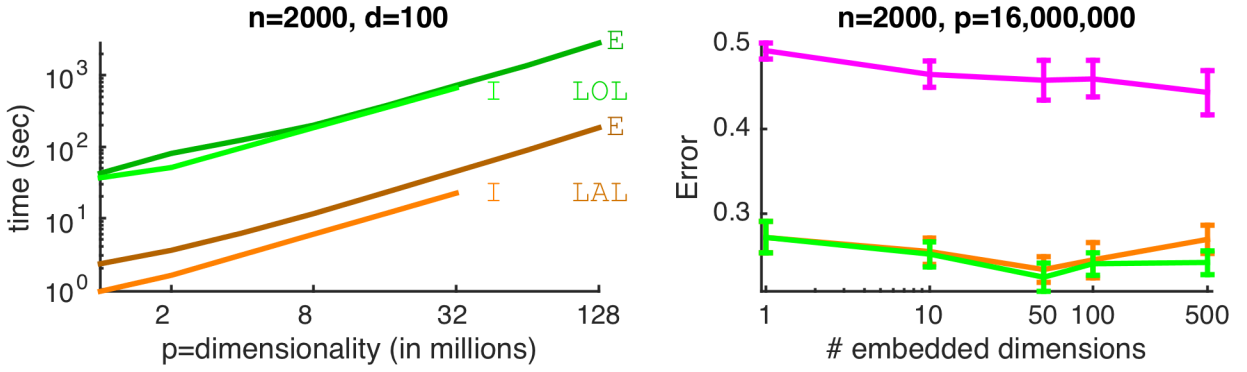
optimization theoretic problem. In addition to demonstrating the statistical accuracy and computational efficiency of `LOL` on simulated and real classification problems, we also demonstrate how the same idea can also be used for other kinds of supervised learning problems, including regression and hypothesis testing. Theoretical guarantees suggest that this line of research is promising and can be extended to other more general settings and tasks.

**Related Work** One of the first publications to compose `FLD` with an unsupervised learning method was the celebrated Fisherfaces paper [29]. The authors showed via a sequence of numerical experiments the utility of embedding with `PCA` prior to classifying with `FLD`. We extend this work by adding a supervised component to the initial embedding. Moreover, we provide the geometric intuition for why and when this is advantageous, as well as show numerous examples demonstrating its superiority. We also prove that Fisherfaces can be implemented very efficiently (see Methods), while shedding light on why it is performing relatively well in general.

**Next Steps** The `LOL` idea, appending the mean difference vector to convert unsupervised manifold learning to supervised manifold learning, has many potential applications. We have presented the first few. Incorporating additional nonlinearities via kernel methods [30], ensemble methods such as random forests [31], and multiscale methods [32] are all of immediate interest. MATLAB, R, and FlashR code for the experiments performed in this manuscript is available from <http://docs.neurodata.io/LOL/>.

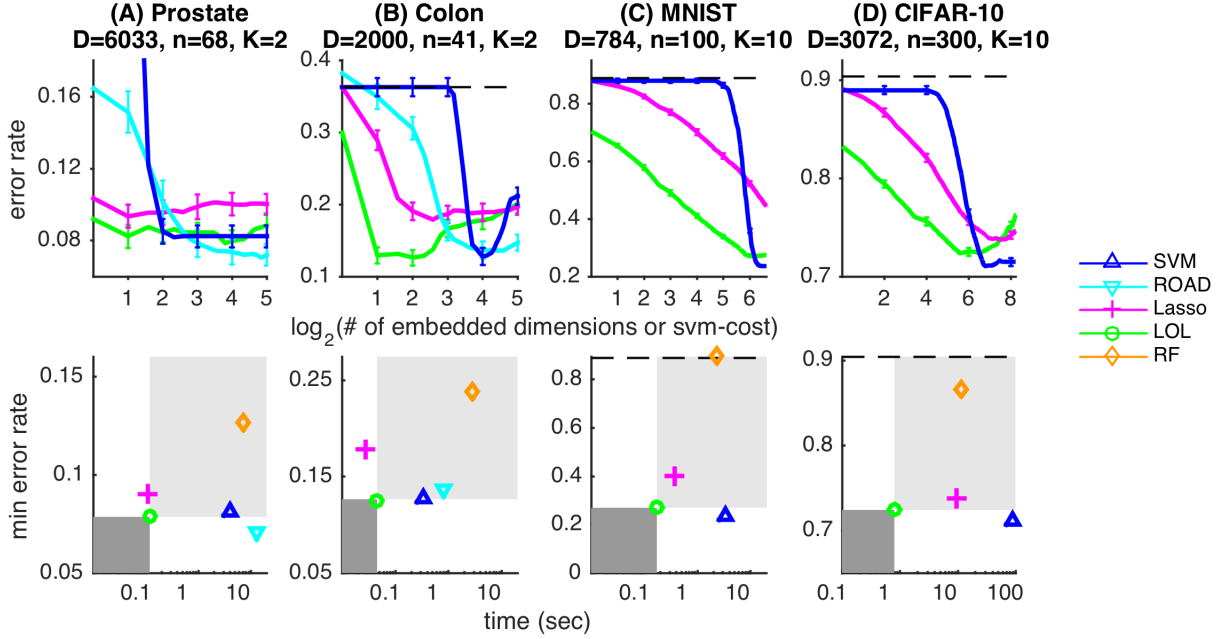


**Figure 3:** Seven simulations demonstrating  $L_{OL}$  achieves superior finite sample performance over competitors both in settings for which we have asymptotic theoretical guarantees, and those for which we do not. For the first three, the top panels depict the means (top), the shared covariance matrix (middle). For the next three, the top panels depict a 2D scatter plot (left), mean and level set of one standard deviation of covariance matrix (right). For all seven simulations, the bottom panel shows misclassification rate as a function of the number of embedded dimensions, for several different classifiers. The simulations settings are as follows: **(A)** Rotated Trunk: same as Figure 2C. **(B)** Toeplitz: another setting where mean difference is not well correlated with any eigenvector, and no ambient coordinate is particularly useful on its own. **(C)** 3 Class variant of the rotated Trunk example to demonstrate that  $L_{OL}$  naturally adapts, and excels in, multi-class problems. **(D)** Fat Tails: a common phenomenon in real data that is more general than our theory supports. **(E)** QDA:  $Q_{OQ}$ , a variant of  $L_{OL}$  when each class has a unique covariance, outperforms  $L_{OL}$ , as expected, when the true discriminant boundary is a quadratic, rather than linear, function. **(F)** Outliers: adding high-dimensional outliers degrades performance of standard eigensolvers, but those can easily be replaced in  $L_{OL}$  for a robust variants (called  $L_{RL}$ ). **(G)** XOR: a high-dimensional stochastic generalization of XOR, demonstrating that  $Q_{OQ}$  works even in scenarios that are quite distinct from the original motivating problems. In all 7 cases,  $L_{OL}$ , or the appropriate generalization thereof, outperforms unsupervised or sparse methods. Moreover, the optimal embedding dimension is never the true discriminant dimension, but rather, a smaller number jointly determined by parameter settings and sample size.

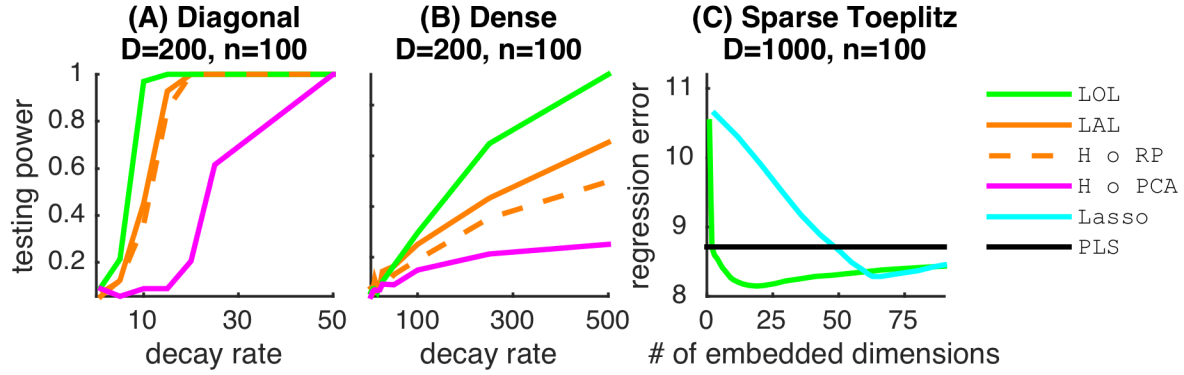


**Figure 4:** Computational efficiency of various low-dimensional projection methods. In all cases,  $n = 2000$ , and we used spherically symmetric simulation parameters (see Methods for details). We compare  $\text{PCA}$  with the projection step of  $\text{LOL}$  (light green for in memory, dark green for semi-external memory) and  $\text{LFL}$  (light orange for in-memory, dark orange for semi-external memory) for different observed dimensions ( $p$ ). **(A)**  $\text{LOL}$  exhibits optimal (linear) scale up and scale out, requiring only 46 minutes to find the embedding on a 2TB dataset, and only 3 minutes using  $\text{LAL}$  (the sparse constant of sparse random projection  $c = \frac{1}{\sqrt{p}}$ ). **(B)** Error for  $\text{LAL}$  is the same as  $\text{LOL}$  in this setting, and both are significantly better than  $\text{FLDOPCA}$  for all choices of embedding dimension.





**Figure 5:** For four standard datasets, we benchmark `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`). Bottom panels show the minimum error rate achieved by each of the five algorithms versus time. The lower left dark gray (upper right light gray) rectangle is the area in which any algorithm is *better* (worse) than `LOL` in terms of both accuracy and efficiency. **(A)** Prostate: a standard sparse dataset. 1-dimensional `LOL` does very well, although keeping  $2^5$  ambient coordinates slightly improves performance, at a significant cost of compute time (two orders of magnitude), with minimal additional interpretability. **(B)** Colon: another standard sparse dataset. Here, 2-4 dimensions of `LOL` outperforms all other approaches considered regardless of how many dimensions they keep. **(C)** MNIST: 10 image categories, so `ROAD` is not possible. `LOL` does very well regardless of the number of dimensions kept. `SVM` marginally improves on `LOL` accuracy, at a significant cost in computation (two orders of magnitude). **(D)** CIFAR-10: a higher dimensional and newer 10 category image classification problem. Results are qualitatively similar to C. Note that, for all four of the problems, there is no algorithm performing better and faster than `LOL`; rather, most algorithms typically perform worse and slower (though some are more accurate and much more computationally expensive).



**Figure 6:** The intuition of including the mean difference vector is equally useful for other supervised manifold learning problems, including testing and regression. **(A)** and **(B)** show two different high-dimensional testing settings, as described in Methods. Power is plotted against the decay rate of the spectrum, which approximates the effective number of dimensions. **LOL** composed with Hotelling’s test outperforms the random projections variants described in [28], as well as several other variants. **(C)** A sparse high-dimensional regression setting, as described in Methods, designed for sparse methods to perform well.  $\log_{10}$  mean squared error is plotted against the number of embedded dimensions. **LOL** composed with linear regression outperforms **LASSO** (cyan), the classic sparse regression method, as well as partial least squares (PLS; black). These three simulation settings therefore demonstrate the generality of this technique.

## A Simulations

For most simulation settings, each class is Gaussian:  $f_{x|y} = \mathcal{N}(\mu_y, \Sigma_y)$ ,  $f_y = \mathcal{B}(\pi)$ . We typically assume that both classes are equally like,  $\pi = 0.5$ , and the covariance matrices are the same,  $\Sigma_0 = \Sigma_1 = \Sigma$ . Under such assumptions, we merely specify  $\theta = \{\mu_0, \mu_1, \Sigma\}$ .

### Stacked Cigars

- $\mu_0 = \mathbf{0}$ ,
- $\mu_1 = (a, b, a, \dots, a)$ ,
- $\Sigma$  is a diagonal matrix, with diagonal vector,  $d = (1, b, 1, \dots, 1)$ ,

where  $a = 0.15$  and  $b = 4$ .

### Trunk

- $\mu_0 = b/\sqrt{(1, 3, 5, \dots, 2p)}$ ,
- $\mu_1 = -\mu_0$ ,
- $\Sigma$  is a diagonal matrix, with diagonal vector,  $d = 100/\sqrt{(p, p-1, p-2, \dots, 1)}$ ,

where  $b = 4$ .

### Rotated Trunk

Same as Trunk, but the data are randomly rotated, that is, we sample  $Q$  uniformly from the set of  $p$ -dimensional rotation matrices, and then set:

- $\mu_0 \leftarrow Q\mu_0$ ,
- $\mu_1 \leftarrow Q\mu_1$ ,
- $\Sigma \leftarrow Q\Sigma Q^\top$ .

### Toeplitz

- $\mu_0 = b \times (1, -1, 1, -1, \dots, 1)$ ,
- $\mu_1 = -\mu_0$ ,
- $\Sigma$  is a Toeplitz matrix, where the top row is  $\rho^{(0,1,2,\dots,p-1)}$ ,

where  $b$  is a function of the Toeplitz matrix such that the noise stays constant as dimensionality increases, and  $\rho = 0.5$ .

### 3 Classes

Same as Trunk, but with a third mean equal to the zero vector,  $\mu_2 = \mathbf{0}$ .

### Fat Tails

For this setting, each class is actually a mixture of two Gaussians with the same mean (the two classes have the same covariances):

- $\mu_0 = \mathbf{0}$ ,
- $\mu_1 = (0, \dots, 0, 1, \dots, 1)$ , where the first  $s = 10$  elements are zero,
- $\Sigma_0$  is a matrix with one's on the diagonal, and 0.2 on the off diagonal,
- $\Sigma_1 = 15 \times \Sigma_0$ ,

and then we randomly rotated as in the rotated Trunk example.

459 **QDA**

460 A generalization of the Toeplitz setting, where the two classes have two different covariance matrices,  
 461 meaning that the optimal discriminant boundary is quadratic.

- 462 •  $\mu_0 = b \times (1, -1, 1, -1, \dots, 1),$
- 463 •  $\mu_1 = -Q \times (\mu_0 + 0.1),$
- 464 •  $\Sigma_0$  is the same Toeplitz matrix as described above, and
- 465 •  $\Sigma_1 = Q \Sigma_0 Q^T.$

466 **Outliers**

467 In this dataset, we generate  $n/2$  samples from an inlier model, and the remaining  $n/2$  samples from an  
 468 outlier model. For the inlier model, we first generate a random  $d \times p$  dimensional orthonormal matrix,  $V$ ,  
 469 where  $d = p/10$ . Then, the first half of the inlier points are generated by  $f_{x|0}$ , the next half by  $f_{x|1}$ , and the  
 470 remaining points generated by  $f_{x|\emptyset}$ . For the outliers, we sampled their class randomly from a fair Bernoulli  
 471 distribution:

- 472 •  $f_{x|0} = \mathcal{N}_d(0, \sigma^2) \times V^T,$
- 473 •  $f_{x|1} = \mathcal{N}_d(0, \sigma^2) \times V^T + b,$
- 474 •  $f_{x|\emptyset} = \mathcal{N}_p(\mathbf{0}, \sigma^2 \mathbf{I}),$
- 475 •  $f_{\emptyset} = \mathcal{B}(0.5).$

476 where we set  $\sigma = 0.1$  and  $b = 0.5$ .

ALGORITHMS	$p < n$	2-Class LDA MODELS, $p > n$			GENERALIZED MODELS, $p > n$					OTHER TASKS, $p > n$	ALGORITHM PROPERTIES							ALGORITHMS	REF	NOTES
	LDA model	delta & cov aligned	& cov misaligned	delta & cov misaligned rotated	Fat Tails	>2 Class	QDA Model	Outlier Model	Non- linear	Testing Regress	Fast	Simple	Theory	Open Source	Dense	2-step/ DL	Super- vised			
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	
LDA o Trace-Ratio	X	X	X	X							X	X	X		X	X	X	LDA o Trace-Ratio	4	
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	Deep Learning	11	best when $n \gg p$
Vowpal Wabbit	X				X	X		X		X	X			X	X	X	X	Vowpal Wabbit	5	best when $n \gg p$
Ridge Regression	X										X	X	X	X	X		X	Ridge Regression	1	regression
Partial Least Squares										X	X						X	Partial Least Squares	1	regression
Naive Bayes	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	X	LASSO	3	
kSVM	X	X	X	X	X	X	X	X	X				X	X	X		X	kSVM	1	
Random Forest	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	1/2	LDA o PCA	2	
LOL	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	LOL	*	
FIGURE		2A	2B, 3A	2C, 3B	3C	3D	3E	3F	3G	5A, 5D	5C, 5D	6								

**Figure 7:** Table of algorithms and their properties for high-dimensional data. Gray elements indicate that results are demonstrated in the Figure labeled in the bottom row. 'X' denotes relatively good performance for a given setting, or has the particular property.

## B Theoretical Background

### II.A The Classification Problem

Let  $(X, Y)$  be a pair of random variables, jointly sampled from  $F := F_{X,Y} = F_{X|Y}F_Y$ . Let  $X$  be a multivariate vector-valued random variable, such that its realizations live in  $p$  dimensional Euclidean space,  $x \in \mathbb{R}^p$ . Let  $Y$  be a categorical random variable, whose realizations are discrete,  $y \in \{0, 1, \dots, C\}$ . The goal of a classification problem is to find a function  $g(x)$  such that its output tends to be the true class label  $y$ :

$$g^*(x) := \operatorname{argmax}_{g \in \mathcal{G}} \mathbb{P}[g(x) = y].$$

When the joint distribution of the data is known, then the Bayes optimal solution is:

$$g^*(x) := \operatorname{argmax}_y f_{y|x} = \operatorname{argmax}_y f_{x|y}f_y = \operatorname{argmax}_y \{\log f_{x|y} + \log f_y\} \quad (1)$$

Denote expected misclassification rate of classifier  $g$  for a given joint distribution  $F$ ,

$$L_g^F := \mathbb{E}[g(x) \neq y] := \int \mathbb{P}[g(x) \neq y] f_{x,y} dx dy,$$

where  $\mathbb{E}$  is the expectation, which in this case, is with respect to  $F_{X,Y}$ . For brevity, we often simply write  $L_g$ , and we define  $L_* := L_{g^*}$ .

### II.B Linear Discriminant Analysis ( $LDA$ )

Linear Discriminant Analysis ( $LDA$ ) is an approach to classification that uses a linear function of the first two moments of the distribution of the data. More specifically, let  $\mu_j = \mathbb{E}[F_{X|Y=j}]$  denote the class conditional mean, and let  $\Sigma = \mathbb{E}[F_X^2]$  denote the joint covariance matrix, and  $\pi_j = \mathbb{P}[Y = j]$ . Using this notation, we can define the  $LDA$  classifier:

$$g_{LDA}(x) := \operatorname{argmin}_y \frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0) + \mathbb{I}\{Y = y\} \log \pi_y,$$

where  $\mathbb{I}\{\cdot\}$  is one when its argument is true, and zero otherwise. Let  $L_{LDA}^F$  be the misclassification rate of the above classifier for distribution  $F$ . Assuming equal class prior and centered means,  $\pi_0 = \pi_1$  and  $(\mu_0 + \mu_1)/2 = 0$ , re-arranging a bit, we obtain

$$g_{LDA}(x) := \operatorname{argmin}_y x^T \Sigma^{-1} \mu_y.$$

In words, the  $LDA$  classifier chooses the class for whom the projection of an input vector  $x$ , onto  $\Sigma^{-1} \mu_y$ , is maximized. When there are only two classes, this further simplifies to

$$g_{2-LDA}(x) := \mathbb{I}\{x^T \Sigma^{-1} \delta > 0\},$$

where  $\delta = \mu_0 - \mu_1$ . Note that the equal class prior and centered means assumptions merely changes the threshold constant from 0 to something else.

### II.C $LDA$ Model

A statistical model is a family of distributions indexed by a parameter  $\theta \in \Theta$ ,  $\mathcal{F}_\theta = \{F_\theta : \theta \in \Theta\}$ . Consider the special case of the above where  $F_{X|Y=y}$  is a multivariate Gaussian distribution,  $\mathcal{N}(\mu_y, \Sigma)$ , where each class has its own mean, but all classes have the same covariance. We refer to this model as the  $LDA$  model. Let  $\theta = (\pi, \mu, \Sigma)$ , and let  $\Theta_{C-LDA} = (\Delta_C, \mathbb{R}^{p \times C}, \mathbb{R}_{>0}^{p \times p})$ , where  $\mu = (\mu_1, \dots, \mu_C)$ ,  $\Delta_C$  is the  $C$  dimensional simplex, that is  $\Delta_C = \{x : x_i \geq 0 \forall i, \sum_i x_i = 1\}$ , and  $\mathbb{R}_{>0}^{p \times p}$  is the set of positive definite  $p \times p$  matrices. Denote  $\mathcal{F}_{LDA} = \{F_\theta : \theta \in \Theta_{LDA}\}$ , dropping the superscript  $C$  for brevity where appropriate. The following lemma is well known:



492 **Lemma 1.**  $L_{Lda}^F = L_*^F$  for any  $F \in \mathcal{F}_{Lda}$ .

493 *Proof.* Under the LDA model, the Bayes optimal classifier is available by plugging the explicit distributions  
494 into Eq. (1).  $\square$

## 495 C Projection Based Classifiers

Let  $A \in \mathbb{R}^{d \times p}$  be an orthonormal matrix, that is, a matrix that projects  $p$  dimensional data into a  $d$  dimensional subspace, where  $AA^T$  is the  $d \times d$  identity matrix, and  $A^T A$  is symmetric  $p \times p$  matrix with rank  $d$ . The question that motivated this work is: what is the best projection matrix that we can estimate, to use to “pre-process” the data prior to applying LDA. Projecting the data  $x$  onto a low-dimensional subspace, and the classifying via LDA in that subspace is equivalent to redefining the parameters in the low-dimensional subspace,  $\Sigma_A = A \Sigma A^T \in \mathbb{R}^{d \times d}$  and  $\delta_A = A \delta \in \mathbb{R}^d$ , and then using  $g_{Lda}$ . When  $C = 2$ ,  $\pi_0 = \pi_1$ , and  $(\mu_0 + \mu_1)/2 = 0$ , this amounts to:

$$g_A^d(x) := \mathbb{I}\{(Ax)^T \Sigma_A^{-1} \delta_A > 0\}, \text{ where } A \in \mathbb{R}^{d \times p}. \quad (2)$$

496 Let  $L_A^d := \int \mathbb{P}[g_A(x) = y] f_{x,y} dx dy$ . Our goal therefore is to be able to choose  $A$  for a given parameter  
497 setting  $\theta = (\pi, \delta, \Sigma)$ , such that  $L_A$  is as small as possible (note that  $L_A$  will never be smaller than  $L_*$ ).

498 Formally, we seek to solve the following optimization problem:

$$\begin{aligned} & \underset{A}{\text{minimize}} && \mathbb{E}[\mathbb{I}\{x^T A^T \Sigma_A^{-1} \delta_A > 0\} \neq y] \\ & \text{subject to} && A \in \mathbb{R}^{p \times d}, \quad AA^T = I_{d \times d}, \end{aligned} \quad (3)$$

499 where  $I_{u \times v}$  is the  $u \times v$  identity matrix identity, that is,  $I(i, j) = 1$  for all  $i = j \leq \min(u, v)$ , and zero otherwise.  
500 Let  $\mathcal{A}^d = \{A : A \in \mathbb{R}^{d \times p}, AA^T = I_{d \times d}\}$ , and let  $\mathcal{A}_* \subset \mathcal{A}$  be the set of  $A$  that minimize Eq. (3), and let  
501  $A_* \in \mathcal{A}_*$  (where we dropped the superscript  $d$  for brevity). Let  $L_A^* = L_{A_*}$  be the misclassification rate for  
502 any  $A \in \mathcal{A}_*$ , that is,  $L_A^*$  is the Bayes optimal misclassification rate for the classifier that composes  $A$  with  
503 LDA.

504 In our opinion, Eq. (3) is the simplest supervised manifold learning problem there is: a two-class classification  
505 problem, where the data are multivariate Gaussians with shared covariances, the manifold is linear,  
506 and the classification is done via LDA. Nonetheless, solving Eq. (3) is difficult, because we do not know  
507 how to evaluate the integral analytically, and we do not know any algorithms that are guaranteed to find the  
508 global optimum in finite time. This has led to previous work using a surrogate function [? ]. We proceed by  
509 studying a few natural choices for  $A$ .

### 510 III.A Bayes Optimal Projection

511 **Lemma 2.**  $\delta^T \Sigma^{-1} \in \mathcal{A}_*$

*Proof.* Let  $B = (\Sigma^{-1} \delta)^T = \delta^T (\Sigma^{-1})^T = \delta^T \Sigma^{-1}$ , so that  $B^T = \Sigma^{-1} \delta$ , and plugging this in to Eq. (2), we  
obtain

$$\begin{aligned} g_B(x) &= \mathbb{I}\{xB^T \Sigma_B^{-1} \delta_B > 0\} \\ &= \mathbb{I}\{x^T \Sigma^{-1} \delta \times (\Sigma_B^{-1} \delta_B) > 0\} && \text{plugging in } B \\ &= \mathbb{I}\{x^T \Sigma^{-1} \delta > 0\} && \text{because } \Sigma_B^{-1} \delta_B > 0. \end{aligned}$$

512 In other words, letting  $B$  be the Bayes optimal projection recovers the Bayes classifier, as it should. Or,  
513 more formally, for any  $F \in \mathcal{F}_{Lda}$ ,  $L_{\delta^T \Sigma^{-1}} = L_*$   $\square$

### III.B Principle Components Analysis (PCA) Projection

Principle Components Analysis (PCA) finds the directions of maximal variance in a dataset. PCA is closely related to eigendecompositions and singular value decompositions (SVD). In particular, the top principle component of a matrix  $X \in \mathbb{R}^{p \times n}$ , whose columns are centered, is the eigenvector with the largest corresponding eigenvalue of the centered covariance matrix  $XX^T$ . SVD enables one to estimate this eigenvector without ever forming the outer product matrix, because SVD factorizes a matrix  $X$  into  $USV^T$ , where  $U$  and  $V$  are orthonormal  $p \times n$  matrices, and  $S$  is a diagonal matrix, whose diagonal values are decreasing,  $s_1 \geq s_2 \geq \dots \geq s_n$ . Defining  $U = [u_1, u_2, \dots, u_n]$ , where each  $u_i \in \mathbb{R}^p$ , then  $u_i$  is the  $i^{\text{th}}$  eigenvector, and  $s_i$  is the square root of the  $i^{\text{th}}$  eigenvalue of  $XX^T$ . Let  $A_d^{\text{PCA}} = [u_1, \dots, u_d]$  be the truncated PCA orthonormal matrix.

The PCA matrix is perhaps the most obvious choice of a orthonormal matrix for several reasons. First, truncated PCA minimizes the squared error loss between the original data matrix and all possible rank  $d$  representations:

$$\operatorname{argmin}_{A \in \mathbb{R}^{d \times p} : AA^T = I_{d \times d}} \|X - A^T A\|_F^2.$$

Second, the ubiquity of PCA has led to a large number of highly optimized numerical libraries for computing PCA (for example, LAPACK [33]).

Moreover, let  $U_d = [u_1, \dots, u_d] \in \mathbb{R}^{p \times d}$ , and note that  $U_d^T U_d = I_{d \times p}$  and  $U_d^T U_d = I_{p \times d}$ . Similarly, let  $USU^T = \Sigma$ , and  $US^{-1}U^T = \Sigma^{-1}$ . Let  $S_d$  be the matrix whose diagonal entries are the eigenvalues, up to the  $d^{\text{th}}$  one, that is  $S_d(i, j) = s_i$  for  $i = j \leq d$  and zero otherwise. Similarly,  $\Sigma_d = US_dU^T = U_d S_d U_d^T$ .

Let  $g_{\text{PCA}}^d := g_{A_{\text{PCA}}^d}$ , and let  $L_{\text{PCA}}^d := L_{A_{\text{PCA}}^d}$ . And let  $g_{\text{LDA}}^d := \mathbb{I}\{x \Sigma_d^{-1} \delta > 0\}$  be the regularized LDA classifier, that is, the LDA classifier, but sets the bottom  $p - d$  eigenvalues to zero.

**Lemma 3.**  $L_{\text{PCA}}^d = L_{\text{LDA}}^d$ .

*Proof.* Plugging  $U_d$  into Eq. (2) for  $A$ , and considering only the left side of the operand, we have

$$\begin{aligned} (Ax)^T \Sigma_A^{-1} \delta_A &= x^T A^T A \Sigma^{-1} A^T A \delta, \\ &= x^T U_d U_d^T \Sigma^{-1} U_d U_d^T \delta, \\ &= x^T U_d U_d^T U S^{-1} U U_d U_d^T \delta, \\ &= x^T U_d I_{d \times p} S^{-1} I_{p \times d} U_d^T \delta, \\ &= x^T U_d S_d^{-1} U_d^T \delta, \\ &= x^T \Sigma_d^{-1} \delta. \end{aligned}$$

□

The implication of this lemma is that if one desires to implement Fisherfaces, rather than first learning the eigenvectors and then learning LDA, one can instead directly implement regularized LDA by setting the bottom  $p - d$  eigenvalues to zero.

### III.C Linear Optimal Low-Rank (LOL) Projection

The basic idea of LOL is to use both  $\delta$  and the top  $d$  eigenvectors. Most naïvely, we could simply concatenate the two,  $A_{\text{LOL}}^d = [\delta, A_{\text{PCA}}^{d-1}]$ . Recall that eigenvectors are orthonormal. To maintain orthonormality, we could easily apply Gram-Schmidt,  $A_{\text{LOL}}^d = \text{ORTH}([\delta, A_{\text{PCA}}^{d-1}])$ . Both in practice and in theory (as will be shown below), this orthogonalization step does not matter much.

to ensure that they are balanced appropriately, we normalize  $\delta$

each vector in  $\delta$  to have norm unity. Formally, let  $\tilde{\delta}_j = \delta_j / \|\delta_j\|$ , where  $\delta_j$  is the  $j^{\text{th}}$  difference of the mean vector (remember, the number of vectors is equal to  $C - 1$ , where  $C$  is the total number of classes),

and let  $A_{\text{LoI}}^d = [\tilde{\delta}, A_{\text{Pca}}^{d-(C-1)}]$ . The eigenvectors are all normalized and orthogonal to one another; to impose orthogonality between  $\tilde{\delta}$  and the eigenvectors, we could use any number of numerically optimized algorithms. However, in practice, orthogonalizing does not matter very much, so we do not bother. We formally demonstrate this below.

## D Theoretical Properties of LDA based Classifiers

### IV.A LDA is rotationally invariant

For certain classification tasks, the ambient coordinates have intrinsic value, for example, when simple interpretability is desired. However, in many other contexts, interpretability is less important [? ]. When the exploitation task at hand is invariant to rotations, then we have no reason to restrict our search space to be sparse in the ambient coordinates, rather, for example, we can consider sparsity in the eigenvector basis. Fisherfaces is one example of a rotationally invariant classifier, under certain model assumptions. Let  $W$  be a rotation matrix, that is  $W \in \mathcal{W} = \{W : W^T = W^{-1} \text{ and } \det(W) = 1\}$ . Moreover, let  $W \circ F$  denote the distribution  $F$  after transformation by an operator  $W$ . For example, if  $F = \mathcal{N}(\mu, \Sigma)$  then  $W \circ F = \mathcal{N}(W\mu, W\Sigma W^T)$ .

**Definition 1.** A rotationally invariant classifier has the following property:

$$L_g^F = L_g^{W \circ F}, \quad F \in \mathcal{F}.$$

In words, the Bayes risk of using classifier  $g$  on distribution  $F$  is unchanged if  $F$  is first rotated, for any  $F \in \mathcal{F}$ .

Now, we can state the main lemma of this subsection: LDA is rotationally invariant.

**Lemma 4.**  $L_{\text{Lda}}^F = L_{\text{Lda}}^{W \circ F}$ , for any  $F \in \mathcal{F}$ .

*Proof.* LDA simply becomes thresholding  $x^T \Sigma^{-1} \delta$ . Thus, we can demonstrate rotational invariance by demonstrating that  $x^T \Sigma^{-1} \delta$  is rotationally invariant.

$$\begin{aligned}
 (Wx)^T (W\Sigma W^T)^{-1} W\delta &= x^T W^T (WUSU^T W^T)^{-1} W\delta && \text{by substituting } USU^T \text{ for } \Sigma \\
 &= x^T W^T (\tilde{U}\tilde{S}\tilde{U}^T)^{-1} W\delta && \text{by letting } \tilde{U} = WU \\
 &= x^T W^T (\tilde{U}S^{-1}\tilde{U}^T)W\delta && \text{by the laws of matrix inverse} \\
 &= x^T W^T WUS^{-1}U^T W^T W\delta && \text{by un-substituting } WU = \tilde{U} \\
 &= x^T US^{-1}U^T \delta && \text{because } W^T W = I \\
 &= x^T \Sigma^{-1} \delta && \text{by un-substituting } US^{-1}U^T = \Sigma
 \end{aligned}$$

□

One implication of this lemma is that we can reparameterize without loss of generality. Specifically, defining  $W := U^T$  yields a change of variables:  $\Sigma \mapsto S$  and  $\delta \mapsto U^T \delta := \delta''$ , where  $S$  is a diagonal covariance matrix. Moreover, let  $d = (\sigma_1, \dots, \sigma_D)^T$  be the vector of eigenvalues, then  $S^{-1} \delta' = d^{-1} \odot \tilde{\delta}$ , where  $\odot$  is the Hadamard (entrywise) product. The LDA classifier may therefore be encoded by a unit vector,  $\tilde{d} := \frac{1}{m} d^{-1} \odot \tilde{\delta}'$ , and its magnitude,  $m := \|d^{-1} \odot \tilde{\delta}\|$ . This will be useful later.

### IV.B Rotation of Projection Based Linear Classifiers $g_A$

By a similar argument as above, one can easily show that:

$$\begin{aligned}
(AWx)^T(AW\Sigma W^T A^T)^{-1}AW\delta &= x^T(W^T A^T)(AW)\Sigma^{-1}(W^T A^T)(AW)\delta \\
&= x^T Y^T Y \Sigma^{-1} Y^T Y \delta \\
&= x^T Z \Sigma^{-1} Z^T \delta \\
&= x^T (Z \Sigma Z^T)^{-1} \delta = x^T \tilde{\Sigma}_d^{-1} \delta,
\end{aligned}$$

where  $Y = AW \in \mathbb{R}^{d \times p}$  so that  $Z = Y^T Y$  is a symmetric  $p \times p$  matrix of rank  $d$ . In other words, rotating and then projecting is equivalent to a change of basis. The implications of the above is:

**Lemma 5.**  $g_A$  is rotationally invariant if and only if  $\text{span}(A) = \text{span}(\Sigma_d)$ . In other words, PCA is the only rotationally invariant projection.

## IV.C Chernoff information

We now introduce the notion of the Chernoff information, which serves as our surrogate measure for the Bayes error of any classification procedure given the *projected* data – in the context of this paper the projection is via LOL or PCA. Our discussion of the Chernoff information is under the context of decision rules for hypothesis testing, nevertheless, as evidenced by the fact that the Maximum A Posterior decision rule – equivalently the Bayes classifier – achieves the Chernoff information rate, this distinction between hypothesis testing and classification is mainly for ease of exposition.

Let  $F_0$  and  $F_1$  be two absolutely continuous multivariate distribution in  $\Omega \subset \mathbb{R}^d$  with density function  $f_0$  and  $f_1$ , respectively. Suppose that  $Y_1, Y_2, \dots, Y_m$  are independent and identically distributed random variables, with  $Y_i$  distributed either  $F_0$  or  $F_1$ . We are interested in testing the simple null hypothesis  $\mathbb{H}_0: F = F_0$  against the simple alternative hypothesis  $\mathbb{H}_1: F = F_1$ . A test  $T$  is a sequence of mapping  $T_m: \Omega^m \mapsto \{0, 1\}$  such that given  $Y_1 = y_1, Y_2 = y_2, \dots, Y_m = y_m$ , the test rejects  $\mathbb{H}_0$  in favor of  $\mathbb{H}_1$  if  $T_m(y_1, y_2, \dots, y_m) = 1$ ; similarly, the test rejects  $\mathbb{H}_1$  in favor of  $\mathbb{H}_0$  if  $T_m(y_1, y_2, \dots, y_m) = 0$ . The Neyman-Pearson lemma states that, given  $Y_1 = y_1, Y_2 = y_2, \dots, Y_m = y_m$  and a threshold  $\eta_m \in \mathbb{R}$ , the likelihood ratio test which rejects  $\mathbb{H}_0$  in favor of  $\mathbb{H}_1$  whenever

$$\left( \sum_{i=1}^m \log f_0(y_i) - \sum_{i=1}^m \log f_1(y_i) \right) \leq \eta_m$$

is the most powerful test at significance level  $\alpha_m = \alpha(\eta_m)$ , i.e., the likelihood ratio test minimizes the type-II error  $\beta_m$  subject to the constraint that the type-I error is at most  $\alpha_m$ .

Assuming that  $\pi \in (0, 1)$  is a prior probability that  $\mathbb{H}_0$  is true. Then, for a given  $\alpha_m^* \in (0, 1)$ , let  $\beta_m^* = \beta_m(\alpha_m^*)$  be the type-II error associated with the likelihood ratio test when the type-I error is at most  $\alpha_m^*$ . The quantity  $\inf_{\alpha_m^* \in (0, 1)} \pi \alpha_m^* + (1 - \pi) \beta_m^*$  is then the Bayes risk in deciding between  $\mathbb{H}_0$  and  $\mathbb{H}_1$  given the  $m$  independent random variables  $Y_1, Y_2, \dots, Y_m$ . A classical result of Chernoff [34] states that the Bayes risk is intrinsically linked to a quantity known as the *Chernoff information*. More specifically, let  $C(F_0, F_1)$  be the quantity

$$\begin{aligned}
C(F_0, F_1) &= -\log \left[ \inf_{t \in (0, 1)} \int_{\mathbb{R}^d} f_0^t(x) f_1^{1-t}(x) dx \right] \\
&= \sup_{t \in (0, 1)} \left[ -\log \int_{\mathbb{R}^d} f_0^t(x) f_1^{1-t}(x) dx \right]
\end{aligned} \tag{4}$$

Then we have

$$\lim_{m \rightarrow \infty} \frac{1}{m} \inf_{\alpha_m^* \in (0, 1)} \log(\pi \alpha_m^* + (1 - \pi) \beta_m^*) = -C(F_0, F_1). \tag{5}$$

Thus  $C(F_0, F_1)$  is the *exponential* rate at which the Bayes error  $\inf_{\alpha_m^* \in (0, 1)} \pi \alpha_m^* + (1 - \pi) \beta_m^*$  decreases as  $m \rightarrow \infty$ ; we also note that the  $C(F_0, F_1)$  is independent of  $\pi$ . We also define, for a given  $t \in (0, 1)$  the Chernoff divergence  $C_t(F_0, F_1)$  between  $F_0$  and  $F_1$  by

$$C_t(F_0, F_1) = -\log \int_{\mathbb{R}^d} f_0^t(x) f_1^{1-t}(x) dx.$$

591 The Chernoff divergence is an example of a  $f$ -divergence as defined in [35]. When  $t = 1/2$ ,  $C_t(F_0, F_1)$  is  
 592 the Bhattacharyya distance between  $F_0$  and  $F_1$ .

593 The result of Eq. (5) can be extended to  $K + 1 \geq 2$  hypothesis, with the exponential rate being the min-  
 594 imum of the Chernoff information between any pair of hypothesis. More specifically, let  $F_0, F_1, \dots, F_K$  be  
 595 distributions on  $\mathbb{R}^d$  and let  $Y_1, Y_2, \dots, Y_m$  be independent and identically distributed random variables with  
 596 distribution  $F \in \{F_0, F_1, \dots, F_K\}$ . Our inference task is in determining the distribution of the  $Y_i$  among the  
 597  $K + 1$  hypothesis  $\mathbb{H}_0: F = F_0, \dots, \mathbb{H}_K: F = F_K$ . Suppose also that hypothesis  $\mathbb{H}_k$  has *a priori* probability  
 598  $\pi_k$ . For any decision rule  $\delta$ , the risk of  $\delta$  is  $r(\delta) = \sum_k \pi_k \sum_{l \neq k} \alpha_{lk}(\delta)$  where  $\alpha_{lk}(\delta)$  is the probability of  
 599 accepting hypothesis  $\mathbb{H}_l$  when hypothesis  $\mathbb{H}_k$  is true. Then we have [36]

$$\inf_{\delta} \lim_{m \rightarrow \infty} \frac{r(\delta)}{m} = -\min_{k \neq l} C(F_k, F_l). \quad (6)$$

600 where the infimum is over all decision rule  $\delta$ , i.e., for any  $\delta$ ,  $r(\delta)$  decreases to 0 as  $m \rightarrow \infty$  at a rate no  
 601 faster than  $\exp(-m \min_{k \neq l} C(F_k, F_l))$ .

When the distributions  $F_0$  and  $F_1$  are multivariate normal, that is,  $F_0 = \mathcal{N}(\mu_0, \Sigma_0)$  and  $F_1 = \mathcal{N}(\mu_1, \Sigma_1)$ ; then, denoting by  $\Sigma_t = t\Sigma_0 + (1-t)\Sigma_1$ , we have

$$C(F_0, F_1) = \sup_{t \in (0,1)} \left( \frac{t(1-t)}{2} (\mu_1 - \mu_0)^\top \Sigma_t^{-1} (\mu_1 - \mu_0) + \frac{1}{2} \log \frac{|\Sigma_t|}{|\Sigma_0|^t |\Sigma_1|^{1-t}} \right).$$

## 602 IV.D Projecting data and Chernoff information

603 We now discuss how the Chernoff information characterizes the effect a linear transformation  $A$  of the data  
 604 has on classification accuracy. We start with the following simple result whose proof follows directly from  
 605 Eq. (6).

606 **Lemma 6.** Let  $F_0 = \mathcal{N}(\mu_0, \Sigma)$  and  $F_1 \sim \mathcal{N}(\mu_1, \Sigma)$  be two multivariate normals with equal covariance  
 607 matrices. For any linear transformation  $A$ , let  $F_0^{(A)}$  and  $F_1^{(A)}$  denotes the distribution of  $AX$  when  $X \sim F_0$   
 608 and  $X \sim F_1$ , respectively. We then have

$$\begin{aligned} C(F_0^{(A)}, F_1^{(A)}) &= \frac{1}{8} (\mu_1 - \mu_0)^\top A^\top (A \Sigma A^\top)^{-1} A (\mu_1 - \mu_0) \\ &= \frac{1}{8} (\mu_1 - \mu_0)^\top \Sigma^{-1/2} \Sigma^{1/2} A^\top (A \Sigma A^\top)^{-1} A \Sigma^{1/2} \Sigma^{-1/2} (\mu_1 - \mu_0) \\ &= \frac{1}{8} \|P_{\Sigma^{1/2} A^\top} \Sigma^{-1/2} (\mu_1 - \mu_0)\|_F^2 \end{aligned} \quad (7)$$

609 where  $P_Z = Z(Z^\top Z)^{-1} Z^\top$  denotes the matrix corresponding to the orthogonal projection onto the columns  
 610 of  $Z$ .

Thus for a classification problem where  $X|Y=0$  and  $X|Y=1$  are distributed multivariate normals with mean  $\mu_0$  and  $\mu_1$  and the same covariance matrix  $\Sigma$ , Lemma 6 then states that for any two linear transformations  $A$  and  $B$ , the transformed data  $AX$  is to be preferred over the transformed data  $BX$  if

$$(\mu_1 - \mu_0)^\top A^\top (A \Sigma A^\top)^{-1} A (\mu_1 - \mu_0) > (\mu_1 - \mu_0)^\top B^\top (B \Sigma B^\top)^{-1} B (\mu_1 - \mu_0).$$

As an example, suppose  $\Sigma$  is diagonal with distinct eigenvalues where the diagonal entries of  $\Sigma$  are in non-increasing order. Denote by  $\delta = \mu_1 - \mu_0$  and let  $A = \delta^\top$  and  $B = e_1^\top = (1, 0, 0, \dots, 0)$  be the linear transformations for LOL and PCA of  $X$  into  $\mathbb{R}$ . We then have

$$C(F_0^{(A)}, F_1^{(A)}) = \frac{(\delta^\top \delta)^2}{\delta^\top \Sigma \delta} = \frac{(\sum_i \delta_i^2)^2}{\sum_i \delta_i^2 \lambda_i}; \quad C(F_0^{(B)}, F_1^{(B)}) = \frac{\delta_1^2}{\lambda_1}$$

where  $\lambda_1$  is the largest eigenvalue of  $\Sigma$ . Suppose furthermore that  $\delta_1 \leq \delta_2 \leq \dots \leq \delta_p$  and  $\lambda_1 > \lambda_2 > \dots > \lambda_p$ . Then  $C(F_0^{(A)}, F_1^{(A)})$  can be lower-bounded as

$$C(F_0^{(A)}, F_1^{(A)}) = \frac{(\sum_i \delta_i^2)^2}{\sum_i \delta_i^2 \lambda_i} \geq \frac{(p\delta_1^2)^2}{p\delta_1^2 \lambda_1}$$

611 and hence  $C(F_0^{(A)}, F_1^{(A)}) > C(F_0^{(B)}, F_1^{(B)})$  provided  $p\delta_1^2 \geq \delta_p^2$ .

When  $A = [\delta \mid e_1 \mid e_2 \cdots \mid e_{d-1}]^\top \in \mathbb{R}^{d \times p}$  and  $B = [e_1 \mid e_2 \mid \dots \mid e_d]^\top \in \mathbb{R}^{d \times p}$  are the linear transformation for LOL and PCA of  $X$  into  $\mathbb{R}^d$ , we have

$$C(F_0^{(A)}, F_1^{(A)}) = \frac{(\sum_{i=d}^p \delta_i^2)^2}{\sum_{i=d}^p \delta_i^2 \lambda_i} + \sum_{i=1}^{d-1} \frac{\delta_i^2}{\lambda_i}; \quad C(F_0^{(B)}, F_1^{(B)}) = \sum_{i=1}^d \frac{\delta_i^2}{\lambda_i}.$$

612 This can be seen as follows. Let  $\xi_{d-1} = [e_1 \mid e_2 \mid \dots \mid e_{d-1}] \in \mathbb{R}^{p \times (d-1)}$  and  $\zeta_{d-1} = (\lambda_1 \delta_1, \lambda_2 \delta_2, \dots, \lambda_{d-1} \delta_{d-1})^\top \in \mathbb{R}^{d-1}$ . Then

$$A \Sigma A^\top = [\delta \mid \xi_{d-1}]^\top \Sigma [\delta \mid \xi_{d-1}] = \begin{bmatrix} \delta^\top \Sigma \delta & \delta^\top \Sigma \xi_{d-1} \\ \xi_{d-1}^\top \Sigma \delta & \xi_{d-1}^\top \Sigma \xi_{d-1} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^p \delta_i^2 \lambda_i & \zeta_{d-1}^\top \\ \zeta_{d-1} & \Sigma_{d-1} \end{bmatrix} \quad (8)$$

614 where  $\Sigma_{d-1} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_{d-1})$  is the submatrix of  $\Sigma$  corresponding to the eigenvalues  $\lambda_1, \lambda_2, \dots, \lambda_{d-1}$ .

615 Using a formula for the inverse of a partitioned matrix, we have

$$\begin{aligned} (A \Sigma A^\top)^{-1} &= \begin{bmatrix} \sum_{i=1}^p \delta_i^2 \lambda_i & \zeta_{d-1}^\top \\ \zeta_{d-1} & \Sigma_{d-1} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} (\sum_{i=1}^p \delta_i^2 \lambda_i - \zeta_{d-1}^\top \Sigma_{d-1}^{-1} \zeta_{d-1})^{-1} & -(\sum_{i=1}^p \delta_i^2 \lambda_i - \zeta_{d-1}^\top \Sigma_{d-1}^{-1} \zeta_{d-1})^{-1} \zeta_{d-1}^\top \Sigma_{d-1}^{-1} \\ -\Sigma_{d-1}^{-1} \zeta_{d-1} (\sum_{i=1}^p \delta_i^2 \lambda_i - \zeta_{d-1}^\top \Sigma_{d-1}^{-1} \zeta_{d-1})^{-1} & \Sigma_{d-1}^{-1} + \Sigma_{d-1}^{-1} \zeta_{d-1} (\sum_{i=1}^p \delta_i^2 \lambda_i - \zeta_{d-1}^\top \Sigma_{d-1}^{-1} \zeta_{d-1})^{-1} \zeta_{d-1}^\top \Sigma_{d-1}^{-1} \end{bmatrix} \end{aligned} \quad (9)$$

616 Now,  $\sum_{i=1}^p \delta_i^2 \lambda_i - \zeta_{d-1}^\top \Sigma_{d-1}^{-1} \zeta_{d-1} = \sum_{i=1}^p \delta_i^2 - \sum_{i=1}^{d-1} \delta_i^2 \lambda_i = \sum_{i=d}^p \delta_i^2 \lambda_i$ . Therefore,

$$(A \Sigma A^\top)^{-1} = \begin{bmatrix} (\sum_{i=d}^p \delta_i^2 \lambda_i)^{-1} & -\frac{\zeta_{d-1}^\top \Sigma_{d-1}^{-1}}{\sum_{i=d}^p \delta_i^2 \lambda_i} \\ -\frac{\Sigma_{d-1}^{-1} \zeta_{d-1}}{\sum_{i=d}^p \delta_i^2 \lambda_i} & \Sigma_{d-1}^{-1} + \frac{\Sigma_{d-1}^{-1} \zeta_{d-1} \zeta_{d-1}^\top \Sigma_{d-1}^{-1}}{\sum_{i=d}^p \delta_i^2 \lambda_i} \end{bmatrix}. \quad (10)$$

In addition,  $A(\mu_1 - \mu_0) = (\delta^\top \delta, \delta_1, \delta_2, \dots, \delta_{d-1})^\top = (\delta^\top \delta, \zeta_{d-1}^\top \Sigma_{d-1}^{-1})^\top \in \mathbb{R}^d$ . Hence

$$\begin{aligned} (\mu_1 - \mu_0)^\top A^\top (A \Sigma A^\top)^{-1} A(\mu_1 - \mu_0) &= [\delta^\top \delta \mid \zeta_{d-1}^\top \Sigma_{d-1}^{-1}] \begin{bmatrix} (\sum_{i=d}^p \delta_i^2 \lambda_i)^{-1} & -\frac{\zeta_{d-1}^\top \Sigma_{d-1}^{-1}}{\sum_{i=d}^p \delta_i^2 \lambda_i} \\ -\frac{\Sigma_{d-1}^{-1} \zeta_{d-1}}{\sum_{i=d}^p \delta_i^2 \lambda_i} & \Sigma_{d-1}^{-1} + \frac{\Sigma_{d-1}^{-1} \zeta_{d-1} \zeta_{d-1}^\top \Sigma_{d-1}^{-1}}{\sum_{i=d}^p \delta_i^2 \lambda_i} \end{bmatrix} \begin{bmatrix} \delta^\top \delta \\ \Sigma_{d-1}^{-1} \zeta_{d-1} \end{bmatrix} \\ &= \frac{(\delta^\top \delta)^2}{\sum_{i=d}^p \delta_i^2 \lambda_i} - 2\delta^\top \delta \frac{\zeta_{d-1}^\top \Sigma_{d-1}^{-2} \zeta_{d-1}}{\sum_{i=d}^p \delta_i^2 \lambda_i} + \left( \zeta_{d-1}^\top \Sigma_{d-1}^{-3} \zeta_{d-1} + \frac{(\zeta_{d-1}^\top \Sigma_{d-1}^{-2} \zeta_{d-1})^2}{\sum_{i=d}^p \delta_i^2 \lambda_i} \right) \\ &= \frac{(\delta^\top \delta - \zeta_{d-1}^\top \Sigma_{d-1}^{-2} \zeta_{d-1})^2}{\sum_{i=d}^p \delta_i^2 \lambda_i} + \zeta_{d-1}^\top \Sigma_{d-1}^{-3} \zeta_{d-1} \\ &= \frac{(\sum_{i=1}^p \delta_i^2 - \sum_{i=1}^{d-1} \delta_i^2)^2}{\sum_{i=d}^p \delta_i^2 \lambda_i} + \sum_{i=1}^{d-1} \frac{\delta_i^2}{\lambda_i} = \frac{(\sum_{i=d}^p \delta_i^2)^2}{\sum_{i=d}^p \delta_i^2 \lambda_i} + \sum_{i=1}^{d-1} \frac{\delta_i^2}{\lambda_i}. \end{aligned}$$

The derivation of  $C(F_0^{(B)}, F_1^{(B)})$  is straightforward and will be omitted. We therefore have

$$C(F_0^{(A)}, F_1^{(A)}) - C(F_0^{(B)}, F_1^{(B)}) = \frac{(\sum_{i=d}^p \delta_i^2)^2}{\sum_{i=d}^p \delta_i^2 \lambda_i} - \frac{\delta_d^2}{\lambda_d}.$$



From our assumption that  $\lambda_d \geq \lambda_{d+1} \geq \dots \geq \lambda_p$ , we have  $\sum_{i=d}^p \delta_i^2 \lambda_i \leq \lambda_d \sum_{i=d}^p \delta_i^2$  and hence

$$\frac{(\sum_{i=d}^p \delta_i^2)^2}{\sum_{i=d}^p \delta_i^2 \lambda_i} \geq \frac{(\sum_{i=d}^p \delta_i^2)^2}{\lambda_d \sum_{i=d}^p \delta_i^2} = \frac{1}{\lambda_d} \sum_{i=d}^p \delta_i^2 \geq \frac{\delta_d^2}{\lambda_d}$$

and hence  $C(F_0^{(A)}, F_1^{(A)}) \geq C(F_0^{(B)}, F_1^{(B)})$  always, and the inequality is strict provided that  $\sum_{d+1}^p \delta_d^2 > 0$ .

Finally we consider the case where  $\Sigma$  is an arbitrary covariance matrix. Let  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$  be the eigenvalues of  $\Sigma$  and let  $u_1, u_2, \dots, u_p$  be the corresponding eigenvectors. For  $d \leq p$ , let  $U_{d-1} = [u_1 | u_2 | \dots | u_{d-1}] \in \mathbb{R}^{p \times (d-1)}$  be the matrix whose columns are the eigenvectors  $u_1, u_2, \dots, u_{d-1}$  of  $\Sigma$ . Then the LOL projection matrix into  $\mathbb{R}^d$  is given by  $A = [\delta | U_{d-1}]^\top$ . We first have

$$A \Sigma A^\top = [\delta | U_{d-1}]^\top \Sigma [\delta | U_{d-1}] = \begin{bmatrix} \delta^\top \Sigma & \delta^\top \Sigma U_{d-1} \\ U_{d-1}^\top \Sigma \delta & U_{d-1}^\top \Sigma U_{d-1} \end{bmatrix} = \begin{bmatrix} \delta^\top \Sigma & \delta^\top \Sigma U_{d-1} \\ U_{d-1}^\top \Sigma \delta & \Lambda_{d-1} \end{bmatrix}$$

where  $\Lambda_{d-1} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_{d-1})$  is the  $(d-1) \times (d-1)$  diagonal matrix formed by the eigenvalues  $\lambda_1, \lambda_2, \dots, \lambda_{d-1}$ . Therefore, letting  $\gamma = \delta^\top \Sigma \delta - \delta^\top \Sigma U_{d-1} \Lambda_{d-1}^{-1} U_{d-1}^\top \Sigma \delta$ , we have

$$\begin{aligned} (A \Sigma A^\top)^{-1} &= \begin{bmatrix} \delta^\top \Sigma \delta & \delta^\top \Sigma U_{d-1} \\ U_{d-1}^\top \Sigma \delta & U_{d-1}^\top \Sigma U_{d-1} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \gamma^{-1} & -\delta^\top \Sigma U_{d-1} \Lambda_{d-1}^{-1} \gamma^{-1} \\ -\Lambda_{d-1}^{-1} U_{d-1}^\top \Sigma \delta \gamma^{-1} & (\Lambda_{d-1} - \frac{U_{d-1}^\top \Sigma \delta \delta^\top \Sigma U_{d-1}}{\delta^\top \Sigma \delta})^{-1} \end{bmatrix}. \end{aligned}$$

The Sherman-Morrison-Woodbury formula then implies

$$\begin{aligned} \left( \Lambda_{d-1} - \frac{U_{d-1}^\top \Sigma \delta \delta^\top \Sigma U_{d-1}}{\delta^\top \Sigma \delta} \right)^{-1} &= \Lambda_{d-1}^{-1} + \frac{\Lambda_{d-1}^{-1} U_{d-1}^\top \Sigma \delta \delta^\top \Sigma U_{d-1} \Lambda_{d-1}^{-1} / (\delta^\top \Sigma \delta)}{1 - \delta^\top \Sigma U_{d-1} \Lambda_{d-1}^{-1} U_{d-1}^\top \Sigma \delta / (\delta^\top \Sigma \delta)} \\ &= \Lambda_{d-1}^{-1} + \frac{\Lambda_{d-1}^{-1} U_{d-1}^\top \Sigma \delta \delta^\top \Sigma U_{d-1} \Lambda_{d-1}^{-1}}{\delta^\top \Sigma \delta - \delta^\top \Sigma U_{d-1} \Lambda_{d-1}^{-1} U_{d-1}^\top \Sigma \delta} \\ &= \Lambda_{d-1}^{-1} + \gamma^{-1} \Lambda_{d-1}^{-1} U_{d-1}^\top \Sigma \delta \delta^\top \Sigma U_{d-1} \Lambda_{d-1}^{-1} \end{aligned}$$

We note that  $\Sigma U_{d-1} = U_{d-1} \Lambda_{d-1}$  and  $U_{d-1}^\top \Sigma = \Lambda_{d-1} U_{d-1}^\top$  and hence

$$\begin{aligned} \gamma &= \delta^\top \Sigma \delta - \delta^\top \Sigma U_{d-1} \Lambda_{d-1}^{-1} U_{d-1}^\top \Sigma \delta = \delta^\top \Sigma \delta - \delta^\top U_{d-1} \Lambda_{d-1} \Lambda_{d-1}^{-1} \Lambda_{d-1} U_{d-1}^\top \delta \\ &= \delta^\top \Sigma \delta - \delta^\top U_{d-1} \Lambda_{d-1} U_{d-1}^\top \delta = \delta^\top (\Sigma - \Sigma_{d-1}) \delta \end{aligned}$$

where  $\Sigma_{d-1} = U_{d-1} \Lambda_{d-1} U_{d-1}^\top$  is the best rank  $d-1$  approximation to  $\Sigma$  with respect to any unitarily invariant norm. In addition,

$$\Lambda_{d-1}^{-1} U_{d-1}^\top \Sigma \delta \delta^\top \Sigma U_{d-1} \Lambda_{d-1}^{-1} = \Lambda_{d-1}^{-1} \Lambda_{d-1} U_{d-1}^\top \delta \delta^\top U_{d-1} \Lambda_{d-1}^{-1} = U_{d-1}^\top \delta \delta^\top U_{d-1}.$$

We thus have

$$(A \Sigma A^\top)^{-1} = \begin{bmatrix} \gamma^{-1} & -\delta^\top \Sigma U_{d-1} \Lambda_{d-1}^{-1} \gamma^{-1} \\ -\Lambda_{d-1}^{-1} U_{d-1}^\top \Sigma \delta \gamma^{-1} & (\Lambda_{d-1} - \frac{U_{d-1}^\top \Sigma \delta \delta^\top \Sigma U_{d-1}}{\delta^\top \Sigma \delta})^{-1} \end{bmatrix} = \begin{bmatrix} \gamma^{-1} & -\gamma^{-1} \delta^\top U_{d-1} \\ -\gamma^{-1} U_{d-1}^\top \delta & \Lambda_{d-1}^{-1} + \gamma^{-1} U_{d-1}^\top \delta \delta^\top U_{d-1} \end{bmatrix}.$$

Therefore,

$$\begin{aligned} \delta^\top A^\top (A \Sigma A^\top)^{-1} A \delta &= \delta^\top [\delta | U_{d-1}] \begin{bmatrix} \gamma^{-1} & -\gamma^{-1} \delta^\top U_{d-1} \\ -\gamma^{-1} U_{d-1}^\top \delta & \Lambda_{d-1}^{-1} + \gamma^{-1} U_{d-1}^\top \delta \delta^\top U_{d-1} \end{bmatrix} [\delta | U_{d-1}]^\top \delta \\ &= [\delta^\top \delta | \delta^\top U_{d-1}] \begin{bmatrix} \gamma^{-1} & -\gamma^{-1} \delta^\top U_{d-1} \\ -\gamma^{-1} U_{d-1}^\top \delta & \Lambda_{d-1}^{-1} + \gamma^{-1} U_{d-1}^\top \delta \delta^\top U_{d-1} \end{bmatrix} \begin{bmatrix} \delta^\top \delta \\ U_{d-1}^\top \delta \end{bmatrix} \\ &= \gamma^{-1} (\delta^\top \delta)^2 - 2\gamma^{-1} \delta^\top \delta \delta^\top U_{d-1} U_{d-1}^\top \delta + \delta^\top U_{d-1} (\Lambda_{d-1}^{-1} + \gamma^{-1} U_{d-1}^\top \delta \delta^\top U_{d-1}) U_{d-1}^\top \delta \\ &= \gamma^{-1} (\delta^\top \delta - \delta^\top U_{d-1} U_{d-1}^\top \delta)^2 + \delta^\top U_{d-1} \Lambda_{d-1}^{-1} U_{d-1}^\top \delta \\ &= \gamma^{-1} (\delta^\top (I - U_{d-1} U_{d-1}^\top) \delta)^2 + \delta^\top \Sigma_{d-1}^\dagger \delta \end{aligned}$$

where  $\Sigma_{d-1}^\dagger$  is the Moore-Penrose pseudo-inverse of  $\Sigma_{d-1}$ . The PCA projection matrix into  $\mathbb{R}^d$  is given by  $B = U_d^\top$  and hence

$$\delta^\top B^\top (B \Sigma B^\top)^{-1} B \delta = \delta^\top U_d \Lambda_d^{-1} U_d^\top \delta = \delta^\top \Sigma_d^\dagger \delta. \quad (11)$$

We thus have

$$\begin{aligned} C(F_0^{(A)}, F_1^{(A)}) - C(F_0^{(B)}, F_1^{(B)}) &= \gamma^{-1} (\delta^\top (I - U_{d-1} U_{d-1}^\top) \delta)^2 - \delta^\top (\Sigma_d^\dagger - \Sigma_{d-1}^\dagger) \delta \\ &= \frac{(\delta^\top (I - U_{d-1} U_{d-1}^\top) \delta)^2}{\delta^\top (\Sigma - \Sigma_{d-1}) \delta} - \delta^\top (\Sigma_d^\dagger - \Sigma_{d-1}^\dagger) \delta \\ &\geq \frac{(\delta^\top (I - U_{d-1} U_{d-1}^\top) \delta)^2}{\lambda_d \delta^\top (I - U_{d-1} U_{d-1}^\top) \delta} - \frac{1}{\lambda_d} \delta^\top u_d u_d^\top \delta \\ &= \frac{1}{\lambda_d} \delta^\top (I - U_{d-1} U_{d-1}^\top) \delta - \frac{1}{\lambda_d} \delta^\top (U_d U_d^\top - U_{d-1} U_{d-1}^\top) \delta \geq 0 \end{aligned}$$

where we recall that  $u_d$  is the  $d$ -th column of  $U_d$ . Thus  $C(F_0^{(A)}, F_1^{(A)}) \geq C(F_0^{(B)}, F_1^{(B)})$  always, and the inequality is strict whenever  $\delta^\top (I - U_d U_d^\top) \delta > 0$ .

Next we consider the case when  $A = [\delta | U_{d-1}]^\top$  and  $B = U_d$  where  $U_d$  is an arbitrary  $p \times d$  matrix with  $U_d^\top U_d = I$ , i.e.,  $U_d$  has  $d$  orthonormal columns, and  $U_{d-1}$  is the first  $d-1$  columns of  $U_d$ . A similar derivation to the above yields

$$C(F_0^{(A)}, F_1^{(A)}) = \frac{(\delta^\top \Sigma^{-1/2} (I - V_{d-1} V_{d-1}^\top) \Sigma^{1/2} \delta)^2}{\delta^\top \Sigma^{1/2} (I - V_{d-1} V_{d-1}^\top) \Sigma^{1/2} \delta} + \delta^\top \Sigma^{-1/2} V_{d-1} V_{d-1}^\top \Sigma^{-1/2} \delta \quad (12)$$

$$C(F_0^{(B)}, F_1^{(B)}) = \delta^\top \Sigma^{-1/2} V_d V_d^\top \Sigma^{-1/2} \delta \quad (13)$$

where  $V_d V_d^\top = \Sigma^{1/2} U_d (U_d^\top \Sigma U_d)^{-1} U_d^\top \Sigma^{1/2}$  is the orthogonal projection onto the column space of  $\Sigma^{1/2} U_d$ . Hence  $C(F_0^{(A)}, F_1^{(A)}) > C(F_0^{(B)}, F_1^{(B)})$  if and only if

$$\frac{(\delta^\top \Sigma^{-1/2} (I - V_{d-1} V_{d-1}^\top) \Sigma^{1/2} \delta)^2}{\delta^\top \Sigma^{1/2} (I - V_{d-1} V_{d-1}^\top) \Sigma^{1/2} \delta} > \delta^\top \Sigma^{-1/2} (V_d V_d^\top - V_{d-1} V_{d-1}^\top) \Sigma^{-1/2} \delta.$$

Let  $C(F_0^{(A)}, F_1^{(A)})$  and  $C(F_0^{(B)}, F_1^{(B)})$  be the Chernoff informations when  $A = [\delta | U_{d-1}]^\top$  and  $B = [\delta | U_d]^\top$  where  $U_{d-1}$  and  $U_d$  contain the eigenvectors of the population covariance matrices  $\Sigma$ ; similarly, let  $\hat{C}(F_0^{(A)}, F_1^{(A)})$  and  $\hat{C}(F_0^{(B)}, F_1^{(B)})$  be the Chernoff informations when  $A = [\delta | \hat{U}_{d-1}]^\top$  and  $B = [\delta | \hat{U}_d]^\top$  where  $\hat{U}_{d-1}$  and  $\hat{U}_d$  contain the eigenvectors of the sample covariance matrices  $\hat{\Sigma}$ . Suppose furthermore that  $C(F_0^{(A)}, F_1^{(A)}) > C(F_0^{(B)}, F_1^{(B)})$ . Then for sufficiently large  $n$ , with high probability,  $\hat{C}(F_0^{(A)}, F_1^{(A)}) > \hat{C}(F_0^{(B)}, F_1^{(B)})$ .

Finally we consider the case when  $B = \tilde{U}_d^\top$  where  $\tilde{U}_d$  is the  $p \times d$  matrix whose columns are the  $d$  largest eigenvectors of the *pooled* covariance matrix  $\tilde{\Sigma} = \mathbb{E}[(X - \frac{\mu_0 + \mu_1}{2})(X - \frac{\mu_0 + \mu_1}{2})^\top]$ . Assume, without loss of generality, that  $\mu_1 = -\mu_0 = \mu$ . We then have

$$\tilde{\Sigma} = \mathbb{E}[X X^\top] = \pi \Sigma + \pi \mu_0 \mu_0^\top + (1 - \pi) \Sigma + (1 - \pi) \mu_1 \mu_1^\top = \Sigma + \mu \mu^\top = \Sigma + \frac{1}{4} \delta \delta^\top.$$

Therefore

$$(B \Sigma B^\top)^{-1} = (\tilde{U}_d^\top \Sigma \tilde{U}_d)^{-1} = (\tilde{U}_d^\top (\tilde{\Sigma} - \frac{1}{4} \delta \delta^\top) \tilde{U}_d)^{-1} = (\tilde{S}_d - \frac{1}{4} \tilde{U}_d^\top \delta \delta^\top \tilde{U}_d)^{-1} = \tilde{S}_d^{-1} + \frac{\tilde{S}_d^{-1} \tilde{U}_d^\top \delta \delta^\top \tilde{U}_d \tilde{S}_d^{-1}}{4 - \delta^\top \tilde{U}_d \tilde{S}_d^{-1} \tilde{U}_d^\top \delta}$$

634 where  $\tilde{S}_d$  is the diagonal matrix containing the  $d$  largest eigenvalues of  $\tilde{\Sigma}$ . Hence

$$\begin{aligned} C(F_0^{(B)}, F_1^{(B)}) &= \delta^\top B^\top (B \Sigma B^\top)^{-1} B \delta = \delta^\top \tilde{U}_d \left( \tilde{S}_d^{-1} + \frac{\tilde{S}_d^{-1} \tilde{U}_d^\top \delta \delta^\top \tilde{U}_d \tilde{S}_d^{-1}}{4 - \delta^\top \tilde{U}_d \tilde{S}_d^{-1} \tilde{U}_d^\top \delta} \right) \tilde{U}_d^\top \delta \\ &= \delta^\top \tilde{U}_d \tilde{S}_d^{-1} \tilde{U}_d^\top \delta + \frac{(\delta^\top \tilde{U}_d \tilde{S}_d^{-1} \tilde{U}_d^\top \delta)^2}{4 - \delta^\top \tilde{U}_d \tilde{S}_d^{-1} \tilde{U}_d^\top \delta} \\ &= \delta^\top \tilde{\Sigma}_d^\dagger \delta + \frac{(\delta^\top \tilde{\Sigma}_d^\dagger \delta)^2}{4 - \delta^\top \tilde{\Sigma}_d^\dagger \delta} = \frac{4 \delta^\top \tilde{\Sigma}_d^\dagger \delta}{4 - \delta^\top \tilde{\Sigma}_d^\dagger \delta}. \end{aligned} \quad (14)$$

635 where  $\tilde{\Sigma}_d = \tilde{U}_d \tilde{S}_d \tilde{U}_d^\top$  is the best rank  $d$  approximation to  $\tilde{\Sigma} = \Sigma + \frac{1}{4} \delta \delta^\top$ .

We recall that the LOL projection  $A = [\delta \mid U_{d-1}]^\top$  yields

$$C(F_0^{(A)}, F_1^{(A)}) = \frac{(\delta^\top (I - U_{d-1} U_{d-1}^\top) \delta)^2}{\delta^\top (\Sigma - \Sigma_{d-1}) \delta} + \delta^\top \Sigma_{d-1}^\dagger \delta.$$

636 To illustrate the difference between the LOL projection and that based on the eigenvectors of the *pooled*  
 637 covariance matrix, consider the following simple example. Let  $\Sigma = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$  be a diagonal matrix  
 638 with  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ . Also let  $\delta = (0, 0, \dots, 0, s)$ . Suppose furthermore that  $\lambda_p + s^2/4 < \lambda_d$ . Then we  
 639 have  $\tilde{\Sigma}_d = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_d, 0, 0, \dots, 0)$ . Thus  $\tilde{\Sigma}_d^\dagger = \text{diag}(1/\lambda_1, 1/\lambda_2, \dots, 1/\lambda_d, 0, 0, \dots, 0)$  and  $\delta^\top \tilde{\Sigma}_d^\dagger \delta = 0$ .  
 640 Therefore,  $C(F_0^{(B)}, F_1^{(B)}) = 0$ .

On the other hand, we have

$$C(F_0^{(A)}, F_1^{(A)}) = \frac{(\delta^\top (I - U_{d-1} U_{d-1}^\top) \delta)^2}{\delta^\top (\Sigma - \Sigma_{d-1}) \delta} + \delta^\top \Sigma_{d-1}^\dagger \delta = \frac{s^4}{s^2 \lambda_p} + 0 = s^2 / \lambda_p.$$

We can generalize the previous example as follows. Let  $\Sigma$  be a  $p \times p$  covariance matrix of the form

$$\Sigma = \begin{bmatrix} \Sigma_d & 0 \\ 0 & \Sigma_d^\perp \end{bmatrix}$$

where  $\Sigma_d$  is a  $d \times d$  matrix. Let  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$  be the eigenvalues of  $\Sigma$  and suppose furthermore that the eigenvalues of  $\Sigma_d$  are  $\lambda_1, \lambda_2, \dots, \lambda_d$ . Let  $\gamma = \lambda_d - \lambda_{d+1}$ . We now assume that  $\delta$  are generated randomly as follows. The entries of  $\delta$  are i.i.d. random variable sampled according to the following distribution. Given an index  $i$ , with probability  $\epsilon$ ,  $\delta_i = 0$  and with probability  $1 - \epsilon$ ,  $\delta_i$  is distributed according to a normal distribution with mean  $\tau > 0$  and variance  $\sigma^2$ . Then with probability at least  $\epsilon^d$ , the covariance matrix for  $\tilde{\Sigma}$  is of the form

$$\tilde{\Sigma} = \begin{bmatrix} \Sigma_d & 0 \\ 0 & \Sigma_d^\perp + \frac{1}{4} (\tilde{\delta} \tilde{\delta}^\top) \end{bmatrix}$$

where  $\tilde{\delta} \in \mathbb{R}^{p-d}$  is formed by excluding the first  $d$  elements of  $\delta$ . Now, if  $\lambda_{d+1} + \frac{1}{4} \|\tilde{\delta}\|^2 < \lambda_d$ , then the  $d$  largest eigenvalues of  $\tilde{\Sigma}$  is still  $\lambda_1, \lambda_2, \dots, \lambda_d$ , and thus the eigenvectors corresponding to the  $d$  largest eigenvalues of  $\tilde{\Sigma}$  is the same as that for the  $d$  largest eigenvalues of  $\Sigma$ . That is to say,

$$\lambda_{d+1} + \frac{1}{4} \|\tilde{\delta}\|^2 < \lambda_d \implies \tilde{\Sigma}_d^\dagger = \Sigma_d^\dagger \implies \delta^\top \tilde{\Sigma}_d^\dagger \delta = 0 \implies C(F_0^{(B)}, F_1^{(B)}) = 0.$$

We now compute the probability that  $\lambda_{d+1} + \frac{1}{4} \|\tilde{\delta}\|^2 < \lambda_d$ . Suppose for the moment that  $\epsilon > 0$  is fixed and do not varies with  $p$ . We then have

$$\frac{\sum_{i=d+1}^p \delta_i^2 - (p-d)(1-\epsilon)\tau^2}{\sqrt{(p-d)(2(1-\epsilon)(2\tau^2\sigma^2 + \sigma^4) + \epsilon(1-\epsilon)(\tau^4 + 2\tau^2\sigma^2 + \sigma^4))}} \xrightarrow{d} N(0, 1).$$

Thus, as  $p \rightarrow \infty$ , the probability that  $\lambda_{d+1} + \frac{1}{4} \|\tilde{\delta}\|^2 < \lambda_d$  converges to that of

$$\Phi \left( \frac{4(\lambda_d - \lambda_{d+1}) - (p-d)(1-\epsilon)\tau^2}{\sqrt{(p-d)(2(1-\epsilon)(2\tau^2\sigma^2 + \sigma^4) + \epsilon(1-\epsilon)(\tau^4 + 2\tau^2\sigma^2 + \sigma^4))}} \right).$$

This probability can be made arbitrarily close to 1 provided that  $\lambda_d - \lambda_{d+1} \geq Cp(1 - \epsilon)\tau^2$  for all sufficiently large  $p$  and for some constant  $C > 1/4$ . Since the probability that  $\delta_1 = \delta_2 = \dots = \delta_d$  is at least  $\epsilon^d$ , we thus conclude that for sufficiently large  $p$ , with probability at least  $\epsilon^d$ ,

$$C(F_0^{(B)}, F_1^{(B)}) = 0 < C(F_0^{(A)}, F_1^{(A)}).$$

In the case where  $\epsilon = \epsilon(p) \rightarrow 1$  as  $p \rightarrow \infty$  such that  $p(1 - \epsilon) \rightarrow \theta$  for some constant  $K$ , then the probability that  $\lambda_{d+1} + \frac{1}{4}\|\tilde{\delta}\|^2 < \lambda_d$  converges to the probability that

$$\frac{1}{4} \sum_{i=1}^K \sigma^2 \chi_1^2(\tau) \geq \lambda_d - \lambda_{d+1}$$

where  $K$  is Poisson distributed with mean parameter  $\theta$  and  $\chi_i^2(\tau)$  is the non-central chi-square distribution with one degree of freedom and non-centrality parameter  $\tau$ . Thus if  $\lambda_d - \lambda_{d+1} \geq C\theta\tau^2 \log p$  for sufficiently large  $p$  and for some constant  $C$ , then this probability can also be made arbitrarily close to 1.

## IV.E Finite Sample Performance

We now consider the finite sample performance of LOL and PCA-based classifiers in the high-dimensional setting with small or moderate sample sizes, e.g., when  $p$  is comparable to  $n$  or when  $p \gg n$ . Once again we assume that  $X|Y = i \sim \mathcal{N}(\mu_i, \Sigma)$  for  $i = 0, 1$ . Furthermore, we also assume that  $\Sigma$  belongs to the class  $\Theta(p, r, k, \tau, \lambda)$  as defined below.

**Definition** Let  $\lambda > 0$ ,  $\tau \geq 1$  and  $k \leq p$  be given. Denote by  $\Theta(p, r, k, \tau, \lambda, \sigma^2)$  the collection of matrices  $\Sigma$  such that

$$\Sigma = V\Lambda V^\top + \sigma^2 I$$

where  $V$  is a  $p \times r$  matrix with orthonormal columns and  $\Lambda$  is a  $r \times r$  diagonal matrix whose diagonal entries  $\lambda_1, \lambda_2, \dots, \lambda_r$  satisfy  $\lambda \geq \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r \geq \lambda/\tau$ . In addition, assume also that  $|\text{supp}(V)| \leq k$  where  $\text{supp}(V)$  denote the non-zero rows of  $V$ , i.e.,  $\text{supp}(V)$  is the subset of  $\{1, 2, \dots, p\}$  such that  $V_j \neq 0$  if and only if  $j \in \text{supp}(V)$ .

We note that in general  $r \leq k \ll p$  and  $\lambda/\tau \gg \sigma^2$ . We then have the following result.

**Theorem [37]** Suppose there exists constants  $M_0$  and  $M_1$  such that  $M_1 \log p \geq \log n \geq M_0 \log \lambda$ . Then there exists a constant  $c_0 = c_0(M_0, M_1)$  depending on  $M_0$  and  $M_1$  such that for all  $n$  and  $p$  for which

$$\frac{\tau k}{n} \log \frac{ep}{k} \leq c_0,$$

there exists an estimate  $\hat{V}$  of  $V$  such that

$$\sup_{\Sigma \in \Theta(p, r, k, \tau, \lambda, \sigma^2)} \mathbb{E} \|\hat{V}\hat{V}^\top - VV^\top\|^2 \leq \frac{Ck(\sigma\lambda + \sigma^2)}{n\lambda^2} \log \frac{ep}{k}$$

where  $C$  is a universal constant not depending on  $p, r, k, \tau, \lambda$  and  $\sigma^2$ .

We can therefore show that provided that  $M_0$  and  $M_1$  is large enough, and  $n$  and  $p$  satisfies the condition in the preceding theorem, then provided that the Chernoff information of the population version of LOL is larger than the Chernoff information of the population version of PCA, the expected Chernoff information for the sample version of LOL is also larger than the expected Chernoff information of the sample version of PCA. We emphasize that it is necessary that the LOL and the PCA version both projected into the top  $d \leq r$  dimension of the sample covariance matrices.

## 664 E The R implementation of LOL

665 Figure 8 shows the R implementation of LOL for binary classification using FlashMatrix [20]. The imple-  
666 mentation takes a  $D \times I$  matrix, where each column is a training instance and each instance has D features,  
667 and outputs a  $D \times k$  projection matrix.

```

LOL <- function(m, labels, k) {
  counts <- fm.table(labels)
  num.labels <- length(counts$val)
  num.features <- dim(m)[1]
  nv <- k - (num.labels - 1)
  gr.sum <- fm.groupby(m, 1, fm.as.factor(labels, 2), fm.bo.add)
  gr.mean <- fm.mapply.row(gr.sum, counts$Freq, fm.bo.div, FALSE)
  diff <- fm.get.cols(gr.mean, 1) - fm.get.cols(gr.mean, 2)
  svd <- fm.svd(m, nv=0, nu=nv)
  fm.cbind(diff, svd$u)
}

```

**Figure 8:** The R implementation of LOL.



## F Bibliography

- [1] T. Hastie, R. Tibshirani, and J. H. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," *Beijing: Publishing House of Electronics Industry*, 2004.
- [2] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach, "Supervised Dictionary Learning," in *Advances in Neural Information Processing Systems*, 2009, pp. 1033–1040. [Online]. Available: <http://papers.nips.cc/paper/3448-supervised>
- [3] 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>
- [4] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method arXiv : physics / 0004057v1 [ physics . data-an ] 24 Apr 2000," *Neural Computation*, pp. 1–16, 1999.
- [5] 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](http://www.crossref.org/jmlr_DOI.html)
- [6] 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>
- [7] 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.
- [8] H. Zou, "The Adaptive Lasso and Its Oracle Properties," pp. 1418–1429, 2006.
- [9] Y. LeCun, C. Cortes, and C. Burges, "MNIST handwritten digit database." [Online]. Available: <http://yann.lecun.com/exdb/mnist/>
- [10] 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>
- [11] 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>
- [12] J. Baik and J. W. Silverstein, "Eigenvalues of large sample covariance matrices of spiked population models," *Journal of Multivariate Analysis*, vol. 97, no. 6, pp. 1382–1408, 2006.
- [13] D. Paul, "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model," *Statistica Sinica*, vol. 17, no. 4, p. 1617, 2007. [Online]. Available: <http://www3.stat.sinica.edu.tw/statistica/oldpdf/A17n418.pdf>
- [14] P. J. Huber, *Robust Statistics*, ser. Wiley Series in Probability and Statistics. Wiley, 1981, vol. 82, no. 3. [Online]. Available: <http://doi.wiley.com/10.1002/9780470434697>
- [15] P. J. Rousseeuw and K. V. Driessen, "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, Mar. 1999. [Online]. Available: <http://amstat.tandfonline.com/doi/abs/10.1080/00401706.1999.10485670#.VTejtq1Viko>
- [16] D. Ferrari and Y. Yang, "Maximum L<sub>q</sub>-likelihood estimation," *The Annals of Statistics*, vol. 38, no. 2, pp. 753–783, apr 2010. [Online]. Available: <http://projecteuclid.org/euclid.aos/1266586613>
- [17] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust Principal Component Analysis?" *Journal of the ACM*, vol. 58, no. 3, pp. 1–37, Dec. 2009. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1970392.1970395http://arxiv.org/abs/0912.3599>

- [18] T. Zhang and G. Lerman, "A Novel M-Estimator for Robust PCA," *Journal of Machine Learning Research*, vol. 15, pp. 749–808, 2014. [Online]. Available: <http://jmlr.org/papers/v15/zhang14a.html>
- [19] C. M. Bishop, *Pattern Recognition and Machine Learning*, ser. Information science and statistics. Springer, 2006. [Online]. Available: <http://www.library.wisc.edu/selectedtocs/bg0137.pdf>
- [20] D. Zheng, D. Mhembere, J. T. Vogelstein, C. E. Priebe, and R. Burns, "Flashmatrix: Parallel, scalable data analysis with generalized matrix operations using commodity ssds," *arXiv preprint arXiv:1604.06414*, 2016.
- [21] D. Zheng, D. Mhembere, R. Burns, J. Vogelstein, C. E. Priebe, and A. S. Szalay, "FlashGraph: Processing billion-node graphs on an array of commodity SSDs," in *13th USENIX Conference on File and Storage Technologies (FAST 15)*, Santa Clara, CA, 2015.
- [22] D. Zheng, R. Burns, J. Vogelstein, C. E. Priebe, and A. S. Szalay, "An ssd-based eigensolver for spectral analysis on billion-node graphs," *CoRR*, vol. abs/1602.01421, 2016.
- [23] D. Zheng, D. Mhembere, V. Lyzinski, J. Vogelstein, C. E. Priebe, and R. Burns, "Semi-external memory sparse matrix multiplication on billion-node graphs in a multicore architecture," *CoRR*, vol. abs/1602.02864, 2016.
- [24] G. M. Amdahl, "Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities," in *AFIPS Spring Joint Computer Conference, 1967. AFIPS '67 (Spring). Proceedings of the*, vol. 30, 1967, pp. 483–485. [Online]. Available: <http://delivery.acm.org/10.1145/1470000/1465560/p483-amdahl.pdf?ip=202.189.127.238&id=1465560&acc=ACTIVESERVICE&key=CDD1E79C27AC4E65.DE0A32330AE3471B.4D4702B0C3E38B35.4D4702B0C3E38B35&CFID=646862329&CFTOKEN=33418015&acm{-}=1426882996{-}a666a43ab4250317fe0>
- [25] J. Abello, A. L. Buchsbaum, and J. R. Westbrook, "A functional approach to external graph algorithms," in *Algorithms ESA98*. Springer, 1998, pp. 332–343.
- [26] 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>
- [27] T. Hastie, K. W. Church, P. Li, and K. C. Kdd, "Very sparse random projections," *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*, p. 287, 2006. [Online]. Available: <http://portal.acm.org/citation.cfm?doid=1150402.1150436>
- [28] M. Lopes, L. Jacob, and M. J. Wainwright, "A More Powerful Two-Sample Test in High Dimensions using Random Projection," in *Neural Information Processing Systems*, 2011, pp. 1206–1214. [Online]. Available: <http://papers.nips.cc/paper/4260-a-more-powerful-two-sample-test-in-high-dimensions-using-random-projection>
- [29] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.
- [30] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Mullers, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No.98TH8468)*. IEEE, 1999, pp. 41–48. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=788121>
- [31] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [32] 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>

- 756 [33] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. D. Croz, A. Greenbaum,  
 757 S. Hammerling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide: Third Edition*. SIAM, 1999.  
 758 [Online]. Available: <https://books.google.com/books?hl=en&lr=&id=AZlvEnr9gCgC&pgis=1>
- 759 [34] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of obser-  
 760 vations," *Annals of Mathematical Statistics*, vol. 23, pp. 493–507, 1952.
- 761 [35] I. Csizár, "Information-type measures of difference of probability distributions and indirect observa-  
 762 tions," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 229–318, 1967.
- 763 [36] C. C. Leang and D. H. Johnson, "On the asymptotics of M-hypothesis bayesian detection," *IEEE Trans-*  
 764 *actions on Information Theory*, vol. 43, pp. 280–282, 1997.
- 765 [37] T. T. Cai, Z. Ma, and Y. Wu, "Optimal estimation and rank detection for sparse spiked covariance  
 766 matrices," 2013, arxiv preprint at <http://arxiv.org/abs/1305.3235>.