


Supervised dimensionality reduction for big data

Joshua T. Vogelstein ^{1,2}✉, Eric W. Bridgeford^{1,2}, Minh Tang¹, Da Zheng¹, Christopher Douville¹, Randal Burns¹ & Mauro Maggioni¹

To solve key biomedical problems, experimentalists now routinely measure millions or billions of features (dimensions) per sample, with the hope that data science techniques will be able to build accurate data-driven inferences. Because sample sizes are typically orders of magnitude smaller than the dimensionality of these data, valid inferences require finding a low-dimensional representation that preserves the discriminating information (e.g., whether the individual suffers from a particular disease). There is a lack of interpretable supervised dimensionality reduction methods that scale to millions of dimensions with strong statistical theoretical guarantees. We introduce an approach to extending principal components analysis by incorporating class-conditional moment estimates into the low-dimensional projection. The simplest version, Linear Optimal Low-rank projection, incorporates the class-conditional means. We prove, and substantiate with both synthetic and real data benchmarks, that Linear Optimal Low-Rank Projection and its generalizations lead to improved data representations for subsequent classification, while maintaining computational efficiency and scalability. Using multiple brain imaging datasets consisting of more than 150 million features, and several genomics datasets with more than 500,000 features, Linear Optimal Low-Rank Projection outperforms other scalable linear dimensionality reduction techniques in terms of accuracy, while only requiring a few minutes on a standard desktop computer.

¹Johns Hopkins University, Baltimore, MD, USA. ²These authors contributed equally: Joshua T. Vogelstein and Eric W. Bridgeford. ✉email: jovo@jhu.edu

Supervised learning—the art and science of estimating statistical relationships using labeled training data—has enabled a wide variety of basic and applied findings, ranging from discovering biomarkers in omics data¹ to recognizing objects from images². A special case of supervised learning is classification, where a classifier predicts the “class” of a novel observation (for example, by predicting sex from an MRI scan). One of the most foundational and important approaches to classification is Fisher’s Linear Discriminant Analysis (LDA)³. LDA has a number of highly desirable properties for a classifier. First, it is based on simple geometric reasoning: when the data are Gaussian, all the information is in the means and variances, so the optimal classifier uses both the means and the variances. Second, LDA can be applied to multiclass problems. Third, theorems guarantee that when the sample size n is large and the dimensionality p is relatively small, LDA converges to the optimal classifier under the Gaussian assumption. Finally, algorithms for implementing it are highly efficient.

Modern scientific datasets, however, present challenges for classification that were not addressed in Fisher’s era. Specifically, the dimensionality of datasets is quickly ballooning. Current raw data can consist of hundreds of millions of features or dimensions; for example, an entire genome or connectome. Yet, the sample sizes have not experienced a concomitant increase. This “large p , small n ” problem is a non-starter for many classical statistical approaches because they were designed with a “small p , large n ” situation in mind. Running LDA when $p \geq n$ is like trying to fit a line to a point: there are infinitely many equally good fits (all lines that pass through the point), and no way to know which of them is “best”. Therefore, without further constraints these algorithms will overfit, meaning they will choose a classifier based on noise in the data, rather than discarding the noise in favor of the desired signal. We also desire methods that can adapt to the complexity of the data, are robust to outliers, and are computationally efficient. Several complementary strategies have been pursued to address these $p \geq n$ problems.

First, and perhaps the most widely used method, is Principal Components Analysis (PCA)⁴. According to PubMed, PCA has been referenced over 40,000 times, and nearly 4000 times in 2018 alone. This is in contrast to other methods that receive much more attention in the media, such as deep learning, random forests, and sparse learning, which received ~2000, ~1200, and ~500 hits, respectively. This suggests that PCA remains the most popular workhorse for high-dimensional problems. PCA “pre-processes” the data by reducing its dimensionality to those dimensions whose variance is largest in the dataset. While highly successful, PCA is a wholly unsupervised dimensionality reduction technique, meaning that PCA does not use the class labels while learning the low-dimensional representation, resulting in suboptimal performance for subsequent classification. Nonlinear manifold learning techniques generalize PCA⁵, but also typically do not incorporate class label information; moreover, they scale poorly. Deep learning provides the most recent version of nonlinear manifold learning, for example, using (supervised) auto-encoders, but these methods remain poorly understood, have many parameters to tune, and typically do not provide interpretable results⁶. Further, deep learning tends to suffer in the wide data problem, where the number of samples is far less than the dimensionality.

The second set of strategies regularize or penalize a supervised method, such as regularized LDA⁷ or canonical correlation analysis (CCA)⁸. Such approaches can drastically overfit in the $p > n$ setting, tend to lack theoretical support in these contexts, and have multiple “knobs” to tune that are computationally taxing. Partial least squares (PLS) is another popular method in this set that often achieves impressive empirical performance, though it

lacks strong theoretical guarantees and a scalable implementation^{9,10}. Sparse methods are the third common strategy to mitigate this “curse of dimensionality”^{11–13}. Unfortunately, exact solutions are computationally intractable, and approximate solutions have theoretical guarantees only under very restrictive assumptions, and are quite fragile to those assumptions¹⁴. Thus, there is a gap: no existing approach can classify multi-class wide data with millions of features while obtaining strong theoretical guarantees, favorable and interpretable empirical performance, and a flexible, robust, and scalable implementation.

To address these issues, we developed a technique for incorporating class-conditional moment estimates, XOX, the simplest example of which is LOL. The key intuition behind LOL is that we can jointly use the means and variances from each class (like LDA and CCA), but without requiring more dimensions than samples (like PCA), or restrictive sparsity assumptions. Using random matrix theory, we are able to prove that when the data are sampled from a Gaussian, LOL finds a better low-dimensional representation than PCA, LDA, CCA, and other linear methods. Under relatively relaxed assumptions, this is true regardless of the dimensionality of the features, the number of samples, or the number of dimensions in which we project. We then demonstrate the superiority of techniques derived using the XOX approach—including (i) LOL, (ii) a variant of XOX which allows greater flexibility of the class-conditional covariances called QOQ, and (iii) a robust variant of LOL called RLOL—over other methods numerically on a variety of simulated settings including several not following the theoretical assumptions. Finally, we show that on several 500 gigabyte neuroimaging datasets, and several multi-gigabyte genomics datasets, LOL achieves superior accuracy at lower dimensions while requiring only a few minutes of time on a single workstation.

Results

Flexibility and accuracy of XOX framework. We empirically investigate the flexibility and accuracy of XOX using simulations that extend beyond theoretical claims. For three different scenarios, we sample 100 training samples each with 100 features; therefore, Fisher’s LDA cannot solve the problem (because there are infinitely many ways to overfit). We consider a number of different methods, including PCA, rrLDA, PLS, random projections (RP), and CCA to project the data onto a low dimensional space. After projecting the data, we train either LDA (for the first two scenarios) or quadratic discriminant analysis (QDA, for the third scenario), which generalizes LDA by allowing each class to have its own covariance matrix¹⁵. For each scenario, we evaluate the misclassification rate on held-out data.

Figure 1 shows a two-dimensional scatterplot (left) and misclassification rate versus dimensionality (right) for each simulation. Hereafter, LOL will refer to the version of LOL with a robust estimate of the location (the class medians, related to the central moment when the population has a symmetric distribution), and a truncated singular value decomposition to estimate of the second moment. A robust location estimate tends to make little difference when a robust estimate was not necessary, and empirically improves performance in simulations and real-data examples when a robust estimate was warranted. Alternative strategies would have been to use robust estimates of the first moment or second moment directly^{16–18}. We do not use a robust estimate of the second moment, as typical robust estimates of the second moment available in standard numerical packages require $d < n$, which is unsuitable for wide data. The top $C-1$ embedding dimensions for LOL correspond to the performance after projection onto the class-conditional means, and rrLDA

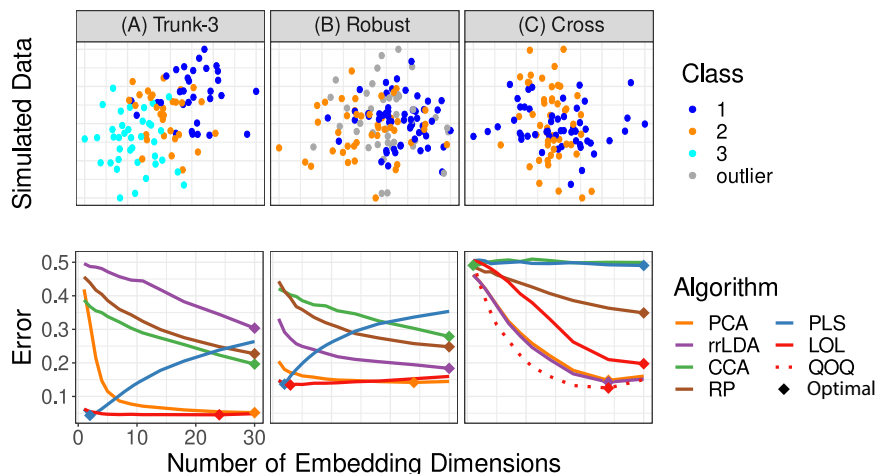


Fig. 1 Three simulations demonstrating the flexibility and accuracy of XOX in settings beyond current theoretical claims. For all cases, training sample size and dimensionality were both 100. The top row depicts the values of the sampled points for two of the 100 dimensions to illustrate the classification task. The bottom row misclassification rate as a function of the number of projected dimensions, for several different embedding approaches. Classification is performed on the embedded data using the LDA classifier for (A) and (B), and using QDA for (C). The simulation settings are: **A** Trunk-3 A variation of Fig. 5b in which three classes are present. **B** Robust Outliers are prominent in the sample while estimating the projection matrix. LOL is robust to the outliers due to the robust estimate of the first moment. **C** Cross The two classes have the same mean but orthogonal covariances. Points are classified using the QDA classifier after projection. QOQ, a variant of LOL where each class' covariance is incorporated into the projection matrix, outperforms other methods, as expected. In essentially all cases and dimensions, LOL, or the appropriate generalization thereof, outperforms other approaches.

corresponds to the performance of projection onto the class-conditional covariance matrix. Figure 1a shows a three class generalization of the Trunk example from Fig. 5b. LOL can trivially be extended to more than two classes (see Supplementary Note 2 for details), unlike ROAD which only operates in a two-class setting. Figure 1b shows a two-class example with many outliers, as is typical in modern biomedical datasets. Both LOL and PLS perform well, despite the outliers, and efficiently identify embedding dimensions despite the outliers. Figure 1c shows an example which should be adversarial for LOL in comparison to PCA or rrLDA. This is because the difference of means is utterly uninformative, so LOL utilizes additional dimensions which are noise compared to PCA. Further, the class-conditional covariances are orthogonal, whereas LOL assumes the class-conditional covariance is the same across both classes. While LOL cannot possibly do as well as PCA in this situation, its performance is only slightly worse. Further, another XOX variant, quadratic optimal QDA (QOQ), uses the same difference of means as LOL and then computes the eigenvectors separately for each class, concatenates them (sorting them according to their singular values), and then classifies with QDA instead of LDA. QOQ is able to identify a slightly more efficient projection for classification than PCA. This is due to the fact that while the first few dimensions are uninformative (those spanned by the difference of the means), the successive dimensions are far more efficient (the class-conditional covariances). For all three scenarios, either LOL—or its extended variant QOQ—achieves a misclassification rate comparable to or lower than other methods, for all dimensions. These three results demonstrate how straightforward generalizations of LOL under the XOX framework which incorporate alternate or robust moment estimates can dramatically improve performance over other projection methods. This is in marked contrast to other approaches, for which such flexibility is either not available, or otherwise problematic.

XOX is computationally efficient and scalable. When the dimensionality is large (e.g., millions or billions), the main bottleneck is sometimes merely the ability to run anything on the data, rather than its predictive accuracy. We evaluate the

computational efficiency and scalability of LOL in the simplest setting: two classes of spherically symmetric Gaussians (see Supplementary Note 3 for details) with dimensionality varying from 2 million to 128 million, and 1000 samples per class. Because LOL admits a closed form solution, it can leverage highly optimized linear algebra routines rather than the costly iterative programming techniques currently required for sparse or dictionary learning type problems¹⁹. To demonstrate these computational capabilities, we built FlashLOL, an efficient scalable LOL implementation with R bindings, to complement the R package used for the above figures.

Four properties of LOL enable its scalable implementation. First, LOL is linear in both sample size and dimensionality (Fig. 2a, solid red line). Second, LOL is easily parallelizable using recent developments in “semi-external memory”^{20–22} (Fig. 2a, dashed red line demonstrates that LOL is also linear in the number of cores). Also note that LOL does not incur any meaningful additional computational cost over PCA (orange dashed line). Third, LOL can use randomized approximate algorithms for eigendecompositions to further accelerate its performance^{23,24} (Fig. 2a, orange lines). FlashLFL, short for Flash Low-rank Fast Linear embedding, achieves an order of magnitude improvement in speed when using very sparse RP instead of the eigenvectors. Fourth, hyper-parameter selection for LOL is nested, meaning that once estimating the d -dimensional projection, every lower dimensional projection is automatically available. This is in contrast to tuning the weight of a penalty term, which leads to a new optimization problem for each different parameter values. Thus, the computational complexity of LOL is $O(npd/Tc)$, where n is sample size, p is the dimension of the data, d is the dimension of the projection, T is the number of threads, and c is the sparsity of the projection.

Finally, note that this simulation setting is ideal for PCA and rrLDA, because the first principal component includes the mean difference vector. Nonetheless, both LOL and LFL achieve near optimal accuracy, whereas rrLDA is at chance, and PCA requires 500 dimensions to even approach the same accuracy that LOL achieves with only one dimension. While PCA would also benefit efficiency wise from a randomized approach, we emphasize that

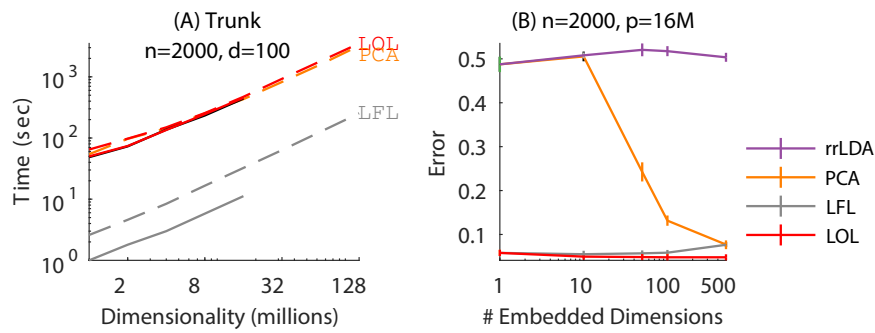


Fig. 2 Computational efficiency and scalability of LOL using $n = 2000$ samples from spherically symmetric Gaussian data (see Supplementary Note 3 for details). **A** LOL exhibits optimal (linear) scale up, requiring only 46 min to find the projection on a 500 gigabyte dataset, and only 3 min using LFL (dashed lines show semi-external memory performance). **B** Error for LFL is the same as LOL in this setting, and both are significantly better than PCA and rrLDA for all choices of projection dimension, regardless of whether a randomized approach is used to compute the projection dimensions. Note that while similar scalability enhancements can be made to PCA in (A), our focus is to highlight that LFL maintains the high performance of LOL in comparison to PCA in (B) despite the randomization technique.

LFL maintains the high performance of LOL in comparison to PCA despite the randomization technique, with the benefit of greater computational efficiency compared to LOL.

Real data benchmarks and applications. Real data often break the theoretical assumptions in more varied ways than the above simulations, and can provide a complementary perspective on the performance properties of different algorithms. We describe two sets of problems, one from brain imaging, and the other from genomics. In both cases we consider a classification problem. To classify participants, researchers typically employ substantive preprocessing pipelines²⁵ to reduce the dimensionality of the data. Unfortunately, as debates persist about the validity of preprocessing approaches, there is no defacto “standard” for the optimal strategies to preprocess the data. Traditional approaches typically include a deep processing chain, with many steps of parametric modeling and downsampling^{26–28}. We therefore investigate the possibility of directly classifying on the nearly raw, high-dimensional data.

The Consortium for Reliability and Reproducibility (CoRR)²⁹ has generated anatomical and diffusion magnetic resonance imaging scans from $n > 800$ participants from five processing sites, each featuring participant-specific annotations for the sex of each individual. At the native resolution, each brain volume is over 150 million dimensions, and each dataset consists of between 42 (60 GB of data) and >400 samples (600 GB of data).

We then also consider a large genomics dataset³⁰ consisting of 340 individuals: 144 patients with nonmetastatic cancer and 196 healthy controls, of which 198 are male and 142 are female. Samples are aligned to $> 750,000$ amplicons distributed throughout the genome to investigate the presence of aneuploidy (abnormal chromosomal counts) in samples from cancer patients (see Supplementary Note 5 for details). The raw amplicon counts are then used with no further preprocessing. We have two tasks of interest: classification on the basis of either sex or age.

For each of the above described problems, we first compute an embedding matrix to project the training data using LOL, PCA, rrLDA, and RP, and then train LDA to classify the resulting low-dimensional representations. The held-out set is then projected and classified using the embedding matrix and trained classifier respectively, and the average cross-validated error is computed over all folds of the data. For each problem, the optimal dimensionality for each strategy is selected to be the number of embedding dimensions with the lowest average cross-validated error. We compute Cohen’s Kappa κ to compare performance across methods because it normalizes the performance of the

classification strategy between zero (the classifier is equivalent to the random chance classifier) and one (the classifier performs perfectly). Finally, for each projection technique, we measure the effect size for each strategy as the difference $\kappa(\text{PCA}) - \kappa(\text{embed})$. See Supplementary Table 1 for a table detailing the datasets employed.

Our FlashLOL implementations are the only algorithms that could successfully run on these data with a single core on a standard desktop computer. In Fig. 3a, LOL is the only technique to outperform PCA on all problems. Figure 3b shows the relative ranks of the average cross-validated misclassification rates for the LDA classifier on each dataset after projection with the specified embedding technique. For all problems, LOL is the technique with the lowest average cross-validated misclassification rate. Further, LOL performs significantly better than all other techniques (Wilcoxon signed-rank statistic, all p values = 0.008). The average misclassification rate achieved at the optimal number of embedding dimensions via LOL is between 5% and 15% across all datasets, which is the same performance we and others obtain using extensively processed and downsampled data that is typically required on similar datasets^{31,32}. LOL therefore enables researchers to side-step hotly debated preprocessing issues by hardly preprocessing at all, and instead simply applying LOL to the data in its native dimensionality.

Discussion

We have introduced a very simple methodology to improve performance on supervised learning problems with wide data (that is, big data where dimensionality is at least as large as sample size) by using class-conditional moments to estimate a low rank projection under a generalized framework, XOX. In particular, LOL uses both the difference of the means and the class-centered covariance matrices, which enables it to outperform PCA, as well as existing supervised linear classification schemes, in a wide variety of scenarios without incurring any meaningful additional computational cost. Straightforward generalizations enable robust and nonlinear variants by using robust estimators and/or class specific covariance estimators. Our open source implementation optimally scales to terabyte datasets. Moreover, the intuition can be extended for both hypothesis testing and regression (see Supplementary Note 6 for additional numerical examples in these settings).

Two commonly applied approaches in these settings are PLS and CCA. CCA is equivalent to rrLDA whenever $p < n$, which is not of interest here. When $p \geq n$, CCA and rrLDA are not equivalent; however, in such settings, CCA exhibits the “maximal

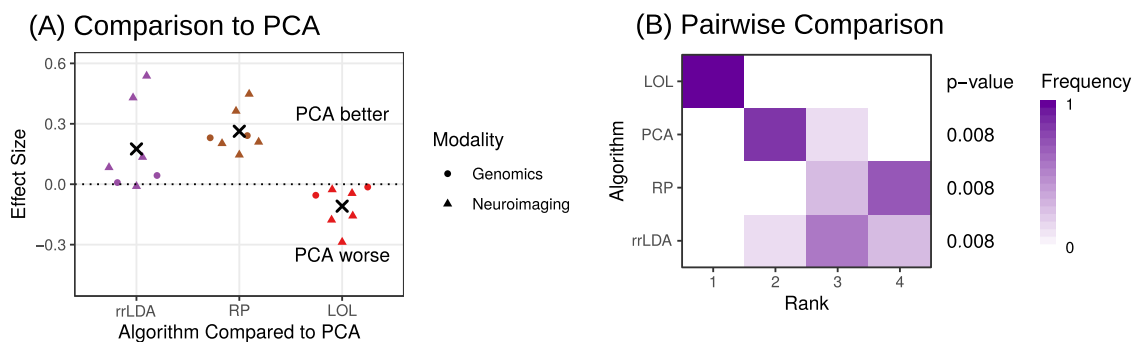


Fig. 3 Comparing various dimensionality reduction algorithms on two real datasets: neuroimaging and genomics. **A** Beeswarm plots show the classification performance of each technique with respect to PCA at the optimal number of embedding dimensions, the number of embedding dimensions with the lowest misclassification rate. Performance is measured by the effect size, defined as $\kappa(\text{LDA} \circ \text{PCA}) - \kappa(\text{LDA} \circ \text{embed})$, where κ is Cohen's Kappa, and embed is one of the embedding techniques compared to PCA. Each point indicates the performance of PCA relative the other technique on a single dataset, and the sample size-weighted average effect is indicated by the black "x." LOL always outperforms PCA and all other techniques. **B** Frequency histograms of the relative ranks of each of the embedding techniques on each dataset after classification, where a 1 indicates the best relative classification performance and a 4 indicates the worst relative classification performance, after embedding with the technique indicated. Projecting first with LOL provides a significant improvement over competing strategies (Wilcoxon signed-rank test, $n = 7$, p value = 0.008) on all benchmark problems.

data piling problem"³³ (see Supplementary Note 2.6 for details). Specifically, all the points in each class are projected onto the exact same point. This results in severe overfitting of the data, yielding poor empirical performance in essentially all settings we considered here (the first dimension of CCA is typically worse even than the difference of the means). While PLS does not exhibit these problems, it lacks strong theoretical guarantees and simple geometric intuition. In contrast to XOX, neither CCA nor PLS enable straightforward generalizations, such as when there are outliers or the discriminant boundary is quadratic (see Fig. 1). Further, across all simulations, XOX outperforms both of these approaches, sometimes quite dramatically (for example, XOX outperforms CCA on over all of the simulations considered). Finally, no scalable nor parallelized implementations are readily available for these methods (see Fig. 2). One could use stochastic gradient descent with penalties to solve these other optimization problems, but they would still need to tune the penalty parameter which would be quite computationally costly. Neither PLS nor CCA could be successfully run on the massive neuroimaging dataset nor the amplicon-level genomics dataset using readily-available tools.

Many previous investigations have addressed similar challenges. The celebrated Fisherfaces paper was the first to compose Fisher's LDA with PCA (equivalent to PCA in this manuscript)³⁴. The authors showed via a sequence of numerical experiments the utility of projecting the data using PCA prior to classifying with LDA. We extend this work by adding a supervised component to the initial projection. Moreover, we provide the geometric intuition for why and when incorporating supervision is advantageous, with numerous examples demonstrating its superiority, and theoretical guarantees formalizing when LOL outperforms PCA. The "sufficient dimensionality reduction" literature has similar insights, but a different construction that typically requires the dimensionality to be smaller than the sample size^{35–39} (although see⁴⁰ for some promising work). More recently, communication-inspired classification approaches have yielded theoretical bounds on linear and affine classification performance⁴¹; they do not, however, explicitly compare different projections, and the bounds we provide are more general and tighter. Moreover, none of the above strategies have implementations that scale to millions or billions of features. Recent big data packages are designed for millions or billions of samples^{42,43}. In biomedical sciences, however, it is far more common to have

tens or hundreds of samples, and millions or billions of features (e.g., genomics or connectomics).

Most manifold learning methods, while exhibiting both strong theoretical^{44–46} and empirical performance, are typically fully unsupervised. Thus, in classification problems, they discover a low-dimensional representation of the data, ignoring the labels. This approach can be highly problematic when the discriminant dimensions and the directions of maximal variance in the learned manifold are not aligned (see Fig. 4 for some examples). Moreover, nonlinear manifold learning techniques tend to learn a mapping from the original samples to a low-dimensional space, but do not learn a projection, meaning that new samples cannot easily be mapped onto the low-dimensional space, a requirement for supervised learning. Deep learning methods⁶ can easily be supervised, but they tend to require huge sample sizes, lack theoretical guarantees, or are opaque "black-boxes" that are insufficient for many biomedical applications. This yields a dearth of "out of the box" supervised scalable dimensionality reduction techniques with strong theoretical guarantees with respect to classification performance bounds designed for wide datasets. Random forests circumvent many of these problems, but implementations that operate on millions of dimensions do not exist⁴⁷, and often produce embeddings that perform no better than PCA on wide datasets (Fig. 3).

Other approaches formulate an optimization problem, such as projection pursuit⁴⁸ and empirical risk minimization⁴⁹. These methods are limited because they are prone to fall into local minima, require costly iterative algorithms, lack any theoretical guarantees on classification accuracy⁴⁹. Feature selection strategies, such as higher criticism thresholding⁵⁰ effectively filter the dimensions, possibly prior to performing PCA on the remaining features⁵¹. These approaches could be combined with LOL in ultrahigh-dimensional problems. Similarly, another recently proposed supervised PCA variant builds on the elegant Hilbert–Schmidt independence criterion⁵² to learn an embedding⁵³. Our theory demonstrates that under the Gaussian model, composing this linear projection with the difference of the means will improve subsequent performance under general settings, implying that this will be a fertile avenue to pursue. A natural extension to this work would therefore be to estimate a Gaussian mixture model per class, rather than simply a Gaussian per class, and project onto the subspace spanned by the collection of all Gaussians.

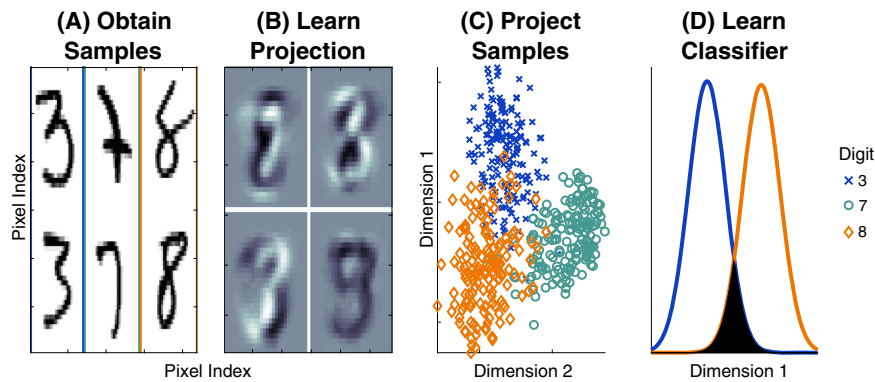


Fig. 4 Schematic illustrating linear optimal low-rank (LOL) as a supervised manifold learning technique. **A** 300 training samples of the numbers 3, 7, and 8 from the MNIST dataset (100 samples per digit); each sample is a $28 \times 28 = 784$ dimensional image (boundary colors are for visualization purposes). **B** The first four projection matrices learned by LOL. Each is a linear combination of the sample images. **C** Projecting 500 new (test) samples into the top two learned dimensions; digits color coded as in (A). LOL-projected data from three distinct clusters. **D** Using the low-dimensional data to learn a classifier. The estimated distributions for 3 and 8 of the test samples (after projecting data into two dimensions and using LDA to classify) demonstrate that 3 and 8 are easily separable by linear methods after LOL projections (the color of the line indicates the digit). The filled area is the estimated error rate; the goal of any classification algorithm is to minimize that area. LOL is performing well on this high-dimensional real data example.

In conclusion, the key XOX idea, appending class-conditional moment estimates to convert unsupervised manifold learning to supervised manifold learning, has many potential applications and extensions. We have presented the first few, including LOL, QOQ, and RLOL, which demonstrated the flexibility of XOX under both theoretical and benchmark settings. Incorporating additional nonlinearities via higher order moments, kernel methods⁵⁴, ensemble methods⁵⁵ such as random forests⁵⁶, and multiscale methods⁴⁶ are all of immediate interest.

Methods

Supervised manifold learning. A general strategy for supervised manifold learning is schematized in Fig. 4, and outlined here. Step (A): Obtain or select n training samples of high-dimensional data. For concreteness, we use one of the most popular benchmark datasets, the MNIST dataset⁵⁷. This dataset consists of images of hand-written digits 0 through 9. Each image is represented by a 28×28 matrix, which means that the observed dimensionality of the data is $p = 28^2 = 784$. Because we are motivated by the $n \ll p$ scenario, we subsample the data to select $n = 300$ examples of the numbers 3, 7, and 8 (100 of each). Step (B): Learn a “projection” that maps the high-dimensional data to a low-dimension representation. One can do so in a way that ignores which images correspond to which digit (the “class labels”), as PCA and most manifold learning techniques do, or try to use the labels, as LDA and sparse methods do. LOL is a supervised linear manifold learning technique that uses the class labels to learn projections that are linear combinations of the original data samples. Step (C): Use the learned projections to map high-dimensional data into the learned lower-dimensional space. This step requires having learned a projection that can be applied to new (test) data samples for which we do not know the true class labels. Nonlinear manifold learning methods typically cannot be applied in this way (though see⁵⁸). LOL, however, can project new samples in such a way as to separate the data into classes. Step (D): Using the low-dimensional representation of the data, learn a classifier. A good classifier correctly identifies as many points as possible with the correct label. For these data, when LDA is used on the low-dimensional data learned by LOL, the data points are mostly linearly separable, yielding a highly accurate classifier.

The geometric intuition of LOL. To build intuition for situations when LOL performs well, and when it does not, we consider the simplest high-dimensional classification setting. We observe n samples (x_i, y_i) , where x_i are p dimensional feature vectors, and y_i is the binary class label, that is, y_i is either 0 or 1. We assume that both classes are distributed according to a multivariate Gaussian distribution, the two classes have the same identity covariance matrix (all features are uncorrelated with unity variance), and data from either class is equally likely, so that the only difference between the classes is their means. In this scenario, the optimal low-dimensional projection is analytically available: it is the dot product of the difference of means and the inverse covariance matrix, commonly referred to as Fisher’s Linear Discriminant Analysis (LDA)⁵⁹ (see Supplementary Note 1.2 for derivation). When the distribution of the data is unavailable, as in all real data problems, machine learning methods can be used to estimate the parameters. Unfortunately, when $n < p$, the estimated covariance matrix will not be invertible (because the solution to the underlying mathematical problem is under specified),

so some other approach is required. As mentioned above, PCA is commonly used to learn a low-dimensional representation. PCA uses the pooled sample mean and the pooled sample covariance matrix. The PCA projection is composed of the top d eigenvectors of the pooled sample covariance matrix, after subtracting the pooled mean (thereby completely ignoring the class labels).

In contrast, LOL uses the class-conditional means and class-centered covariance. This approach is motivated by Fisher’s LDA, which uses the same two terms, and should therefore improve performance over PCA. More specifically, for a two-class problem, LOL is constructed as follows:

1. Compute the sample mean of each class.
2. Estimate the difference between means.
3. Compute the class-centered covariance matrix, that is, compute the covariance matrix after subtracting the class mean from each point.
4. Compute the eigenvectors of this class-conditionally centered covariance.
5. Concatenate the difference of the means with the top $d - 1$ eigenvectors of class-centered covariance.

Note that the sample class-centered covariance matrix estimates the population covariance, whereas the sample pooled covariance matrix is distorted by the difference of the class means. Further, as discussed in Methods, the class-centered covariance matrix is equivalent to “Reduced Rank LDA”⁶⁰ (r rLDA hereafter, which is simply LDA but truncating the covariance matrix). For the theoretical background on LDA and r rLDA, a formal definition of LOL, and detailed description of the simulation settings that follow, see Supplementary Notes 1, 2, and 3, respectively. Figure 5 shows three different examples of 100 data points sampled from a 1000 dimensional Gaussian to geometrically illustrate the intuition that motivated LOL. In each case, all dimensions are uncorrelated with one another, and all classes are equally likely with the same covariance; the only difference between the classes are their means.

Figure 5a shows “stacked cigars”, in which the difference between the means and the direction of maximum variance are large and aligned 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 of the classes. r rLDA performs well here too, for the same reason that PCA does. Because all dimensions are uncorrelated, and one dimension contains most of the information discriminating between the two classes, this is also an ideal scenario for sparse methods. Indeed, ROAD, a sparse classifier designed for precisely this scenario, does an excellent job finding the most useful dimensions¹². LOL, using both the difference of means and the directions of maximal variance, also does well. To calibrate all of these methods, we also show the performance of the optimal classifier.

Figure 5b shows an example that is worse for PCA. In particular, the variance is getting larger for subsequent dimensions, while the magnitude of the difference between the means is decreasing with dimension. Because PCA operates on the pooled sample covariance matrix, the dimensions with the maximum difference are included in the estimate, and therefore, PCA finds some of them, while also finding some of the dimensions of maximum variance. The result is that PCA performs fairly well in this setting. r rLDA, however, by virtue of subtracting out the difference of the means, is now completely at chance performance. 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 5c is exactly the same as Fig. 5b, except the data have been randomly rotated in all 1000 dimensions. This means that none of the original features have

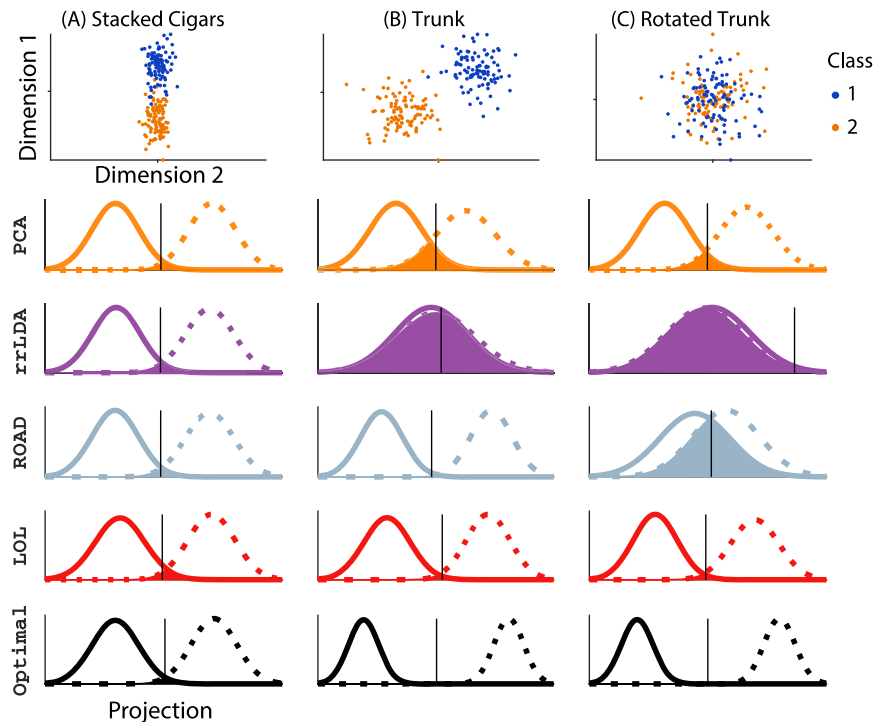


Fig. 5 LOL achieves near-optimal performance for three different multivariate Gaussian distributions, each with 100 samples in 1000 dimensions. For each approach, we project into the top three dimensions, and then use LDA to classify 10,000 new samples. The six rows show (from top to bottom): Row 1: A scatter plot of the first two dimensions of the sampled points, with class 0 and 1 as orange and blue dots, respectively. The next rows each show the estimated posterior for class 0 and class 1, in solid and dashed lines, respectively. The overlap of the distributions—which quantifies the magnitude of the error—is filled. The black vertical line shows the estimated threshold for each method. The techniques include: PCA; reduced rank LDA (rRLDA), a method that projects onto the top d eigenvectors of sample class-conditional covariance; ROAD, a sparse method designed specifically for this model; LOL, our proposed method; and the Bayes optimal classifier. **A** Stacked Cigars The mean difference vector is aligned with the direction of maximal variance, and is mostly concentrated in a single dimension, making it ideal for PCA, rRLDA, and sparse methods. In this setting, the results are similar for all methods, and essentially optimal. **B** Trunk The mean difference vector is orthogonal to the direction of maximal variance; PCA performs worse and rRLDA is at chance, but sparse methods and LOL can still recover the correct dimensions, achieving nearly optimal performance. **C** Rotated Trunk Same as (B), but the data are rotated; in this case, only LOL performs well. Note that LOL is closest to Bayes optimal in all three settings.

much information, but 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 performs even worse in this scenario than in the previous one. rRLDA is rotationally invariant (see Supplementary Note 2.4 for details), so still performs at chance levels. Because there is no small number of features that separate the data well, ROAD fails. LOL performs as well here as it does in the other examples.

When is LOL better than PCA and other supervised linear methods? We desire theoretical confirmation of the above numerical results. To do so, we investigate when LOL is “better” than other linear dimensionality reduction techniques. In the context of supervised dimensionality reduction or manifold learning, the goal is to obtain low dimensional representation that maximally separates the two classes, making subsequent classification easier. Chernoff information quantifies the dissimilarity between two distributions. Therefore, we can compute the Chernoff information between distribution of the two classes after embedding to evaluate the quality of a given embedding strategy. As it turns out, Chernoff information is the exponential convergence rate for the Bayes error⁶¹, and therefore, the tightest possible theoretical bound. The use of Chernoff information to theoretically evaluate the performance of an embedding strategy is novel, to our knowledge, and leads to the following main result:

Main theoretical result. LOL is always better than or equal to rRLDA under the Gaussian model when $p \geq n$, and better than or equal to PCA (and many other linear projection methods) with additional (relatively weak) conditions. This is true for all possible observed dimensionalities of the data, and the number of dimensions into which we project, for sufficiently large sample sizes. Moreover, under relatively weak assumptions, these conditions almost certainly hold as the number of dimensions increases.

Formal statements of the theorems and proofs required to substantiate the above result are provided in Methods. The condition for LOL to be better than PCA is essentially that the d^{th} eigenvector of the pooled sample covariance matrix has less information about classification than the difference of the means vector. The

implication of the above theorem is that it is better to incorporate the mean difference vector into the projection matrix, rather than ignoring it, under basically the same assumptions that motivate PCA. The degree of improvement is a function of the dimensionality of the feature set p , the number of samples n , the projection dimension d , and the parameters, but the existence of an improvement—or at least no worse performance—is independent of those factors.

Data availability

Data used within this manuscript are available from <https://neurodata.io/lol/> and <https://neurodata.io/mri>.

Code availability

MATLAB, R, and Python code for the experiments performed in this manuscript and a docker container for FlashLOL are available from <https://neurodata.io/lol/>, and an R package is available on the Comprehensive R Archive Network (CRAN)⁶².

Received: 9 August 2020; Accepted: 26 March 2021;

Published online: 17 May 2021

References

1. Vogelstein, J. T. et al. Discovery of brainwide neural-behavioral maps via multiscale unsupervised structure learning. *Science* **344**, 386–392 (2014).
2. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Proc. Advances in Neural Information Processing Systems* (eds. Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) 1097–1105 (Curran Associates, Inc. 2012).
3. Fisher, R. A. Theory of statistical estimation. *Math. Proc. Cambridge Philos. Soc.* **22**, 700–725 (1925).

4. Jolliffe, I. T. in *Principal Component Analysis*, Springer Series in Statistics Ch. 1 (Springer, 1986).
5. Lee, J. A. & Verleysen, M. *Nonlinear Dimensionality Reduction* (Springer, 2007).
6. Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. *Deep Learning* (MIT press, 2016).
7. Witten, D. M. & Tibshirani, R. Covariance-regularized regression and classification for high-dimensional problems. *J. R. Stat. Soc. Series B Stat. Methodol.* **71**, 615–636 (2009).
8. Shin, H. & Eubank, R. L. Unit canonical correlations and high-dimensional discriminant analysis. *J. Stat. Comput. Simulation* **81**, 167–178 (2011).
9. ter Braak, C. J. F. & de Jong, S. The objective function of partial least squares regression. *J. Chemom.* **12**, 41–54 (1998).
10. Brereton, R. G. & Lloyd, G. R. Partial least squares discriminant analysis: taking the magic away: PLS-DA: taking the magic away. *J. Chemom.* **28**, 213–225 (2014).
11. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Series B* **58**, 267–288 (1996).
12. Fan, J., Feng, Y. & Tong, X. A road to classification in high dimensional space: the regularized optimal affine discriminant. *J. R. Stat. Soc. Series B Stat. Methodol.* **74**, 745–771 (2012).
13. Hastie, T., Tibshirani, R. & Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations* (Chapman and Hall/CRC, 2015).
14. Weijie, S. et al. False discoveries occur early on the Lasso path. *Ann. Stat.* **45**, 2133–2150 (2017).
15. Hastie, T., Tibshirani, R. & Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Publishing House of Electronics Industry, 2004).
16. Fan, J., Wang, W. & Zhu, Z. A shrinkage principle for heavy-tailed data: high-dimensional robust low-rank matrix recovery. Preprint at [arXiv:1603.08315](https://arxiv.org/abs/1603.08315) (2016).
17. Ke, Y., Minsker, S., Ren, Z., Sun, Q. & Zhou, W.-X. User-friendly covariance estimation for heavy-tailed distributions. *Statist. Sci.* **34**, 454–471 (2019).
18. Minsker, S., and Wei, X. Estimation of the covariance structure of heavy-tailed distributions. Preprint at <https://arxiv.org/abs/1708.00502v3> (2017).
19. Mairal, J., Ponce, J., Sapiro, G., Zisserman, A. & Bach, F. R. Supervised dictionary learning. In *Proc. Advances in Neural Information Processing Systems* (eds. Koller, D., Schuurmans, D., Bengio, Y. & Bottou, L.) 1033–1040 (Curran Associates Inc. 2009).
20. Zheng, D. et al. FlashGraph: Processing billion-node graphs on an array of commodity SSDs. In *Proc. 13th USENIX Conference on File and Storage Technologies (FAST 15)* 45–58 (USENIX Association 2015).
21. Zheng, D., Mhembe, D., Vogelstein, J. T., Priebe, C. E. & Burns, R. Flashmatrix: parallel, scalable data analysis with generalized matrix operations using commodity ssds. Preprint at [arXiv:1604.06414](https://arxiv.org/abs/1604.06414) (2016b).
22. Zheng, D., Burns, R., Vogelstein, J., Priebe, C. E. & Szalay, A. S. An ssd-based eigensolver for spectral analysis on billion-node graphs. Preprint at [arxiv:1602.01421](https://arxiv.org/abs/1602.01421) (2016a).
23. Candès, E. J. & Tao, T. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inf. Theory* **52**, 5406–5425 (2006).
24. Li, P., Hastie, T. J. & Church, K. W. Very sparse random projections. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* 287–296 (Association for Computing Machinery, 2006).
25. Bridgeford, E. W. et al. Eliminating accidental deviations to minimize generalization error and maximize reliability: applications in connectomics and genomics. Preprint at [bioRxiv https://doi.org/10.1101/802629](https://doi.org/10.1101/802629) (2020).
26. Gray, W. R. et al. Magnetic resonance connectome automated pipeline. *IEEE Pulse* **3**, 42–48 (2011).
27. Roncal, W. G. et al. MIGRAINE: MRI graph reliability analysis and inference for connectomics In *Proc. 2013 IEEE Global Conference on Signal and Information Processing* 313–316 (IEEE, 2013).
28. Kiar, G. et al. Science in the cloud (sic): a use case in MRI connectomics. *GigaScience* <https://doi.org/10.1093/gigascience/gix013> (2017).
29. Zuo, X.-N. et al. An open science resource for establishing reliability and reproducibility in functional connectomics. *Sci. Data* **1**, 140049 (2014).
30. Douville, C. et al. Assessing aneuploidy with repetitive element sequencing. *Proc. Natl Acad. Sci. USA* **117**, 4858–4863 (2020).
31. Vogelstein, J. T., Roncal, W. G., Vogelstein, R. J. & Priebe, C. E. Graph classification using signal-subgraphs: applications in statistical connectomics. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1539–1551 (2013).
32. Duarte-Carvajalino, J. M. & Jahanshad, N. Hierarchical topological network analysis of anatomical human brain connectivity and differences related to sex and kinship. *Neuroimage* **59**, 3784–3804 (2011).
33. Ahn, J. & Marron, J. S. The maximum data piling direction for discrimination. *Biometrika* **97**, 254–259 (2010).
34. Belhumeur, P. N., Hespanha, J. P. & Kriegman, D. J. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**, 711–720 (1997).
35. Li, K.-C. Sliced inverse regression for dimension reduction. *J. Am. Stat. Assoc.* **86**, 316–327 (1991).
36. Naftali, T., Fernando, C. P. & William, B. The Information Bottleneck Method. *The 37th annual Allerton Conference on Communication, Control, and Computing*. pp. 368–377 (1999).
37. Globerson, A. & Tishby, N. Sufficient dimensionality reduction. *J. Mach. Learn. Res.* **3**, 1307–1331 (2003).
38. Cook, R. D. & Ni, L. Sufficient dimension reduction via inverse regression. *J. Am. Stat. Assoc.* **100**, 410–428 (2005).
39. Fukumizu, K., Bach, F. R. & Jordan, M. I. Dimensionality reduction for supervised learning with reproducing Kernel Hilbert spaces. *J. Mach. Learn. Res.* **5**, 73–99 (2004).
40. Cook, R. D., Forzani, L. & Rothman, A. J. Prediction in abundant high-dimensional linear regression. *Electron. J. Stat.* **7**, 3059–3088 (2013).
41. Nokleby, M., Rodrigues, M. & Calderbank, R. Discrimination on the grassmann manifold: Fundamental limits of subspace classifiers. *IEEE Trans. Inf. Theory* **61**, 2133–2147 (2015).
42. Agarwal, A., Chapelle, O., Dudík, M. & Langford, J. A reliable effective terascale linear learning system. *J. Mach. Learn. Res.* **15**, 1111–1133 (2014).
43. Abadi, M. et al. Tensorflow: large-scale machine learning on heterogeneous distributed systems. Preprint at [arXiv:1603.04467](https://arxiv.org/abs/1603.04467) (2016).
44. Eckart, C. & Young, G. The approximation of one matrix by another of lower rank. *Psychometrika* **1**, 211–218 (1936).
45. de Silva, V. & Tenenbaum, J. B. Global versus local methods in nonlinear dimensionality reduction. In *Proc. 15th International Conference on Neural Information Processing Systems* 721–728 (eds. Becker, S., Thrun, S. & Obermayer, K.) (MIT Press 2003).
46. Allard, W. K., Chen, G. & Maggioni, M. Multi-scale geometric methods for data sets II: geometric multi-resolution analysis. *Appl. Comput. Harmon. Anal.* **32**, 435–462 (2012).
47. Tomita, T., Maggioni, M. & Vogelstein, J. ROFLMAO: robust oblique forests with linear Matrix operations. In *Proc. 2017 SIAM International Conference on Data Mining* 498–506 (eds. Chawla, N. & Wang, W.) (Society for Industrial and Applied Mathematics, 2017).
48. Huber, P. J. Projection pursuit. *Ann. Stat.* **13**, 435–475 (1985).
49. Belkin, M., Niyogi, P. & Sindhvani, V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.* **7**, 2399–2434 (2006).
50. Donoho, D. L. & Jin, J. Higher criticism thresholding: optimal feature selection when useful features are rare and weak. *Proc. Natl Acad. Sci. USA* **105**, 14790–5 (2008).
51. Bair, E., Hastie, T., Paul, D. & Tibshirani, R. Prediction by supervised principal components. *J. Am. Stat. Assoc.* **101**, 119–137 (2006).
52. Gretton, A., Herbrich, R., Smola, A., Bousquet, O. & Schölkopf, B. Kernel methods for measuring independence. *J. Mach. Learn. Res.* **6**, 2075–2129 (2005).
53. Barshan, E., Ghodsi, A., Azimifar, Z. & Jahromi, M. Z. Supervised principal component analysis: visualization, classification and regression on subspaces and submanifolds. *Pattern Recognit.* **44**, 1357–1371 (2011).
54. Mika, S., Ratsch, G., Weston, J., Schölkopf, B. & Mullers, K. R. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX: Pro. 1999 IEEE Signal Processing Society Workshop (Cat. No. 98TH8468)* (eds. Hu, Y.-H., Larsen, J., Wilson, E. & Douglas, S.) 41–48 (IEEE, 1999).
55. Cannings, T. I. & Samworth, R. J. Random-projection ensemble classification. Preprint at [arXiv:1504.04595](https://arxiv.org/abs/1504.04595) (2015).
56. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
57. LeCun, Y., Cortes, C. & Burges, C. MNIST Handwritten Digit Database <http://yann.lecun.com/exdb/mnist/> (2015).
58. Bengio, Y. et al. Out-of-Sample extensions for LLE, isomap, MDS, eigenmaps, and spectral clustering. In *Advances in Neural Information Processing Systems* (eds Thrun, S., Saul, L. K. & Schölkopf, P. B.) 177–184 (MIT Press, 2004).
59. Bickel, P. J. & Levina, E. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **10**, 989–1010 (2004).
60. Hastie, T. & Tibshirani, R. Discriminant analysis by gaussian mixtures. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**, 155–176 (1996).
61. Chernoff, H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.* **23**, 493–507 (1952).
62. Bridgeford, E. W., Tang, M., Yim, J. & Vogelstein, J. T. Linear optimal low-rank projection. *Zenodo* <https://doi.org/10.5281/zenodo.1246979> (2018).

Acknowledgements

The authors are grateful for the support by the XDATA program of the Defense Advanced Research Projects Agency (DARPA) administered through Air Force Research Laboratory contract FA8750-12-2-0303; DARPA GRAPHS contract N66001-14-1-4028;

and DARPA SIMPLEX program through SPAWAR contract N66001-15-C-4041 and DARPA Lifelong Learning Machines program through contract FA8650-18-2-7834.

Author contributions

M.T. and M.M. contributed theoretical results, D.Z. and R.B. devised the semi-external memory implementation, C.D. procured relevant genomics datasets, J.T.V. and E.W.B. wrote the paper, E.W.B. developed the experiments and R package, J.T.V. supervised.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-23102-2>.

Correspondence and requests for materials should be addressed to J.T.V.

Peer review information *Nature Communications* thanks Andrew Patterson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021