Factor-analytic Inverse Regression for High-dimension, Small-sample Dimensionality Reduction

Aditi Jha [*1,2] **Michael J. Morais** [*1] **Jonathan W. Pillow** [1,3]

Abstract

Sufficient dimension reduction (SDR) methods are a family of supervised methods for dimensionality reduction that seek to reduce dimensionality while preserving information about a target variable of interest. However, existing SDR methods typically require more observations than the number of dimensions ($N > p$). To overcome this limitation, we propose Class-conditional Factor Analytic Dimensions (CFAD), a model-based dimensionality reduction method for high-dimensional, small-sample data. We show that CFAD substantially outperforms existing SDR methods in the small-sample regime, and can be extended to incorporate prior information such as smoothness in the projection axes. We demonstrate the effectiveness of CFAD with an application to functional magnetic resonance imaging (fMRI) measurements during visual object recognition and working memory tasks, where it outperforms existing SDR and a variety of other dimensionality-reduction methods.

1. Introduction

Dimensionality-reduction methods are important tools for analyzing noisy, high-dimensional data, which operate by mapping data to a low-dimensional space while preserving key features of interest. This low-dimensional projection of the original data allows for easier analysis, visualization and compression of data. Dimensionality reduction methods have been developed to exploit a wide variety of different data features, such as mean, covariance, class separation, and temporal or spatial structure (Cunningham & Yu, 2014; Cunningham & Ghahramani, 2015; Pang et al., 2016; Kobak et al., 2016). Another class of dimensionality reduction methods, known as Sufficient Dimension Reduction (SDR), reduce dimensions with the aim of preserving the statistical relationship between high-dimensional data (X) and an observed output variable (Y) (Globerson & Tishby, 2003; Cook, 2007; Cook & Forzani, 2009). SDR methods seek to find a subspace that captures the conditional distribution of $X \mid Y$. That is, they discard dimensions of X that are statistically independent of Y, preserving only the dimensions in which X depends on Y. So-called "inverse regression" methods, which seek to model the distribution of $X \mid Y$, provide a framework for estimating the dimension reduction subspace, leading to the popularity of inverse regression-based SDR methods (Cook & Ni, 2005).

Regression models for predicting scalar variables or class labels from high-dimensional data represent a common setting for the application of dimensionality reduction methods. In neuroscience, for example, such methods are frequently employed for quantifying the relationship between high-dimensional neural activity and the stimuli that elicit them (Cunningham & Yu, 2014; Kobak et al., 2016; Cowley et al., 2016; Aoi & Pillow, 2018).

However, existing inverse regression-based SDR methods are not well suited to such problems. Their performance severely degrades when the sample size N is small relative to the data-dimensionality p, and they completely collapse in the $N < p$ regime, as they often depend on the positive-definiteness of now-rank-deficient sample covariance matrices. As a result of these challenges, Principal Components Analysis or Independent Components Analysis continue to be the most commonly used method(s) for dimensionality reduction of neural data, despite the fact that they completely ignore the information provided by target labels. This motivates the need for a dimensionality reduction method for high-dimensional small-sample-size data, which preserves output-relevant information in the low-dimensional input.

In this paper, we introduce a model-based inverse regression method for high-dimensional data that is robust to sample size and does not collapse in the $N < p$ regime. Our method, Class-conditional Factor-Analytic Dimensions

[*]Equal contribution [1]Princeton Neuroscience Institute, Princeton University, NJ, USA [2]Department of Electrical and Computer Engineering, Princeton University, NJ, USA [3]Department of Psychology, Princeton University, NJ, USA. Correspondence to: Aditi Jha <aditijha@princeton.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

(CFAD)[1], relies on a model of the high-dimensional data in which the class-dependent variation in mean and covariance is restricted to a common low-dimensional subspace, specifying the conditional distribution of $X \mid Y$ with a form similar to that of factor analysis. To learn the CFAD subspace, we use Riemannian optimization over a product manifold (Stiefel $\times$ Euclidean) to maximize the model log-likelihood, ensuring that projection matrices obey semi-orthogonality constraints on the Stiefel manifold and the unconstrained parameters lie on the Euclidean manifold. We show the robustness of CFAD to sample-size compared to other inverse regression-based SDR methods on simulated data. The factor-analytic structure of CFAD allows for easy incorporation of known priors on the data. We show that the addition of smoothness prior, in particular in the small-sample regime, improves the performance of CFAD. We use CFAD with a graph Laplacian smoothing prior to reduce the dimensionality of fMRI activity in a visual object recognition task as well as a working memory task, and show an improvement in classification accuracy as compared to existing inverse regression SDR methods, PCA, Fisher Linear Discriminant (LDA), Reduced-rank Regression (RRR) and Linear Optimal Low-rank projection (LOL), a recently introduced method for dimensionality-reduction in classification settings (Vogelstein et al., 2017).

2. Background

Conventional linear regression of a $p-$dimensional vector X onto Y involves evaluating $\mathbb{E}(Y \mid X)$. Inverse regression, on the other hand, evaluates the curve $\mathbb{E}(X \mid Y)$ in $\mathbb{R}^p$, which consists of p one-dimensional regressions. This reduces the problem to estimating multiple one-dimensional regressions as opposed to a high-dimensional regression, circumventing the issues which conventional regression faces due to high dimensionality (p) of the data. Sufficient dimension reduction using inverse regression has so far employed two broad approaches: moment-based methods and parametric model-based methods. The goal of both of these approaches is to estimate a central subspace $\mathcal{S}_{Y|X}$ such that $\alpha \in \mathbb{R}^{p\times d}$ spans $\mathcal{S}_{Y|X}$ and $Y \perp\!\!\!\perp X \mid \alpha^\top X$. In other words, these methods find a low-dimensional subspace which preserves all information in X relevant to Y.

SIR and SAVE. Sliced Inverse Regression (SIR (Li, 1991)) and Sliced Average Variance Estimation (SAVE (Cook & Weisberg, 1991)) were the first moment-based inverse regression methods to be introduced, and have since found widespread application. They work under two mild conditions: (a) $\mathbb{E}(X \mid \alpha^\top X)$ is a linear function of X (Linearity condition); and (b) $\text{Var}(X \mid \alpha^\top X)$ is non-random (Li, 1991). In case of categorical data, each of the N observations of X is mapped to a label $Y \in \{1, 2, 3, ..h\}$. When the data is continuous, it is still possible to do this by "slicing" the range of Y into h equally spaced bins. Both SIR and SAVE standardise the data such that $Z = \widehat{\Sigma}_X^{-1}(X - \widehat{\mu})$. The SIR estimate uses the first-order moments $\widehat{m}_y = \mathbb{E}[Z \mid Y]$, and then computes the top d eigenvectors of a weighted covariance matrix:

$$\hat{\eta} = \text{eig}_{[1:d]}(\widehat{V}), \qquad \widehat{V} = \sum_{y=1}^{h} n_y \widehat{m}_y \widehat{m}_y^\top \tag{1}$$

where n_y refers to the number of samples in class y. It then transforms $\hat{\eta}$ back to the original space, yielding the following estimate of the central subspace: $\hat{\alpha}_{SIR} = \hat{\eta}\widehat{\Sigma}_X^{-1/2}$. SAVE extends this framework to second-order moments. It instead computes the top d eigenvectors of the following weighted combination of covariance matrices:

$$M = \sum_{y=1}^{h} n_y \left(I_p - \text{Cov}(Z \mid Y)\right) \left(I_p - \text{Cov}(Z \mid Y)\right)^\top \tag{2}$$

These eigenvectors are then similarly transformed to obtain the central subspace estimate: $\hat{\alpha}_{SAVE} = \text{eig}_{[1:d]}(M)\widehat{\Sigma}_X^{-1/2}$. While SIR and SAVE use class-means and class-covariance matrices, respectively, their inability to combine information from both moments makes them complementary, and inspired the introduction of directional regression.

DR. Directional Regression (DR (Li & Wang, 2007)) combines first-order and second-order moments to estimate the central subspace. It uses the top d eigenvectors of the following matrix:

$$M = \sum_{y=1}^{h} n_y \mathbb{E}^2[ZZ^\top - I_p \mid Y] + 2\widehat{V}^2 + 2\left(\sum_{y=1}^{h} n_y \widehat{m}_y^\top \widehat{m}_y\right)\widehat{V} \tag{3}$$

and then transforms them as in SIR and SAVE to obtain the estimate. An important drawback of the above eigenvector-based methods is that they work only when $N > p$ since they require inverting the sample covariance ($\widehat{\Sigma}_X \in \mathbb{R}^{p\times p}$) in order to compute $\hat{\alpha}$. They work well when $N \gg p$, but otherwise the covariance estimates in $\mathbb{R}^p$ are very noisy and performance degrades.

LAD. Likelihood acquired directions (LAD (Cook & Forzani, 2009)) is a model-based approach, which requires numerical optimization of a likelihood function to obtain an estimate of the central subspace. LAD assumes that the class-conditional distribution is a Gaussian such that $X \mid Y \sim \mathcal{N}(m_y, \Delta_y)$, where m_y and Δ_y denote class-conditional means and covariances. Using the linearity and constant-covariance conditions, the LAD model estimates $S_{Y|X}$ such that $\alpha^\top X$ contains all information about X and the projection of X on the subspaces orthogonal to $S_{Y|X}$ is independent of Y. However, since it models m_y and Δ_y in $\mathbb{R}^p$, it requires a large number of samples N relative to

[1]Code available at: https://github.com/97aditi/CFAD.git

the number of dimensions p. Formally, the LAD objective function requires each sample class-covariance matrix to be positive definite, meaning the method requires at least $N = (p+1)h$ samples, where h is the number of slices or classes in Y. This substantially limits its application to high-dimension, small-sample problems. Our method, CFAD, is motivated by similar modeling assumptions, but adapted to be tractable for the $N < p$ regime.

3. Class-conditional Factor-Analytic Dimensions (CFAD)

Motivated by the dearth of inverse regression SDR methods that are tractable for high-dimensional small-sample-size data, we develop a new framework for inverse regression tailored to such data. High-dimensional data often depends on only a small number of latent factors. For example, fMRI measurements are recorded at the level of voxels, but large correlations between nearby voxels reduce the effective number of data dimensions significantly. We exploit this assumption in our model-based method, which call Class-conditional Factor-Analytic Dimensions (CFAD).

The CFAD model describes high-dimensional data as arising from a generative model in which information about the target variable Y is contained entirely within a low-dimensional subspace (See Fig. 1). Specifically, CFAD models the class-dependent variation in high-dimensional data $X \in \mathbb{R}^p$ as arising from a mixture-of-Gaussians in $\mathbb{R}^d$ ($d < p$), projected up to $\mathbb{R}^p$. More precisely, every class or slice of the output variable Y is associated with a distinct d-dimensional Gaussian ($\mathcal{N}(\mu_y, \Lambda_y)$), which is projected using a semi-orthogonal matrix $\alpha \in \mathbb{R}^{p\times d}$ to the ambient space of X, such that

$$X_{CS} \mid Y \sim \mathcal{N}(\alpha\mu_y, \alpha\Lambda_y\alpha^\top), \tag{4}$$

where X_{CS} denotes the (noiseless) central-subspace component of $X \in \mathbb{R}^p$.

Of course, the distribution $X \mid Y$ may have additional dependencies that do not depend on Y. To capture this, we assume a separate q-dimensional Gaussian, $\mathcal{N}(0, \Lambda_0)$, with mean 0 and covariance Λ_0, which is projected through a different semi-orthogonal matrix $\alpha_0 \in \mathbb{R}^{p\times q}$ to $\mathbb{R}^p$. We require that $\alpha^\top\alpha_0 = 0 (\in \mathbb{R}^{d\times q})$, meaning the subspaces spanned by α and α_0 are orthogonal. This ensures that class-independent covariance—that is, correlation structure in X that is independent of Y—lies in a subspace orthogonal to the central subspace. Finally, we assume independent additive Gaussian noise along every dimension of X, with variances contained in the diagonal elements of a diagonal matrix Ψ). This leads to the following model:

$$X \mid Y \sim \mathcal{N}(\alpha\mu_y, \alpha\Lambda_y\alpha^\top + \alpha_0\Lambda_0\alpha_0^\top + \Psi) \tag{5}$$

This equation represents the CFAD model in its full generality. Note that when the targets Y are continuous, we can partition the range of Y into h classes, e.g. with percentiles or naive clustering, just as with other inverse regression SDR methods.

If $\mu_y = 0$ for each class in Y, and $q = 0$, meaning the class means are all zero and there are no additional Y-independent correlations in X, CFAD reduces to factor analysis on each class, with the added constraint that all sets of loading weights lie in the subspace spanned by α. If we fix $\Psi = \sigma^2 I_p$, CFAD fits probabilistic PCA to each class, constrained the same way. Further constraining all class-specific covariances Λ_y to be identical reduces CFAD to a single instance of factor analysis or probabilistic PCA based on the choice of Ψ (see Supplement for details). Ghahramani & Hinton (1997) proposed a mixture-of-factor analyzers model, which resembles CFAD but without imposing a shared low-dimensional subspace or a distinction between class-dependent and class-independent models of correlations in X. Thus, CFAD can capture the key intuitions of these other generative methods, but is extended to capture class-dependent variation in means and covariances.

Further, we can show that CFAD is provably an SDR method, as defined in Cook & Forzani, 2009; assuming that the target data, for some d and q, have the low-dimensional structure in eq. (5). We recall that the CFAD-discovered subspace α is a *sufficient* dimension reduction subspace if and only if $X \perp\!\!\!\perp Y \mid \alpha^\top X$, i.e. if conditioning on the projection onto that subspace leaves no class-dependent information in X. For the CFAD model in eq. (5), we can show using Gaussian identities (see Supplemental Materials) that

$$\alpha^\top X \mid Y \sim \mathcal{N}(\mu_y, \Lambda_y + \alpha^\top\Psi\alpha) \tag{6}$$

$$X \mid (\alpha^\top X, Y = y) \sim \mathcal{N}\left(0, \alpha_0\Lambda_0\alpha_0^\top + \Psi\right) \tag{7}$$

So we have that $Y \perp\!\!\!\perp X \mid \alpha^\top X$ and CFAD is an SDR method, with the requirement that the noise Ψ is diagonal.

3.1. Optimizing the CFAD likelihood function

We assume that the data X contains N samples and Y has h classes with π_y denoting the fraction of points belonging to a particular class y. As we are particularly interested in the scenario where N is small, imposing additional constraints on the covariance can be useful. Following this, we restrict the class-independent and class-dependent covariance (Λ_0, Λ_y) to be diagonal. We also assume that the noise is isotropic and hence set $\Psi = \sigma^2 I_p$.

We can easily obtain the MLE estimate of μ_y as $\alpha\widehat{\mu}_y$ where $\widehat{\mu}_y$ is the sample class mean. We define $\widehat{\Sigma}_{X|y}$ as the sample covariance of $X \mid (Y = y)$. Using these and the trace trick,

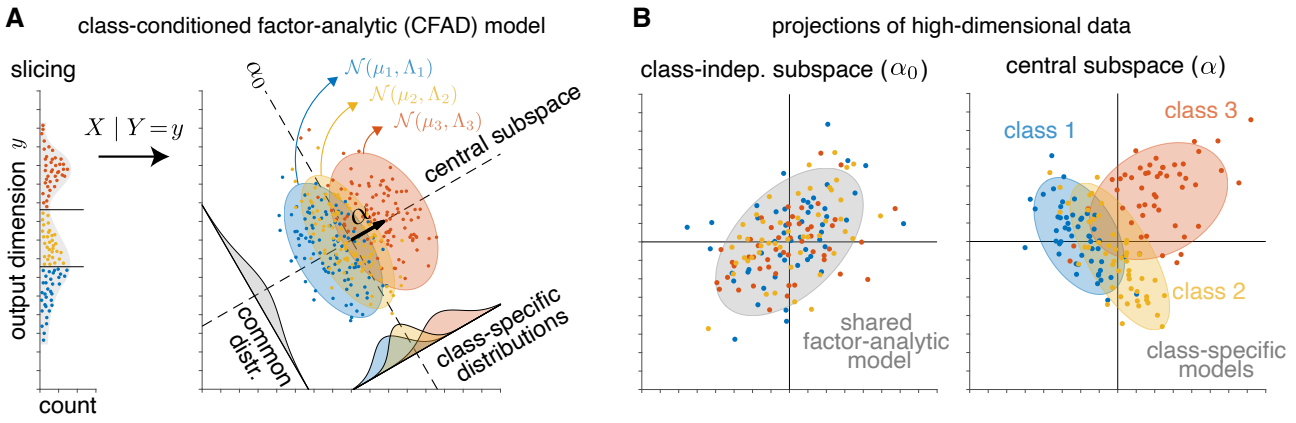


Figure 1. Schematic illustration of the CFAD model. **(A)** Left: slicing involves partitioning the dataset based on a discrete binning of the output variable Y. Right: the CFAD model describes class-conditional data $X \mid Y = y$, as multivariate normal, such that mean and covariance differs *only* within the central subspace (spanned by α). Data projected orthogonal to α have a common, class-independent distribution. **(B)** Bivariate projections of the CFAD model. Left: in the class-independent space spanned by α_0, data have a common Gaussian distribution that captures correlations in X that do not vary across classes. Right: in the central subspace, each class has its own multi-variate Gaussian distribution with unique mean and covariance. (Not shown: orthogonal to α and α_0, data have an independent axis-aligned normal distribution governed by the diagonal elements of Ψ.)

the log likelihood of the model in eq. 5 is

$$\mathcal{L}_{\text{CFAD}} = -\frac{Np}{2}\log(2\pi) - \frac{N}{2}\sum_{y=1}^{h}\pi_y \log|\Sigma_y| - \frac{N}{2}\sum_{y=1}^{h}\pi_y \operatorname{Tr}\left(\Sigma_y^{-1}\left(\widehat{\Sigma}_{X|(Y=y)} + B_y\right)\right) \quad (8)$$

$$\text{where } \Sigma_y = \alpha\Lambda_y\alpha^\top + \alpha_0\Lambda_0\alpha_0^\top + \sigma^2 I_p$$
$$B_y = (I_p - \alpha\alpha^\top)\widehat{\mu}_y\widehat{\mu}_y^\top(I_p - \alpha\alpha^\top)$$

We note that α and α_0 have semi-orthogonality constraints and they are also mutually orthogonal ($\alpha^\top\alpha_0 = 0$). Such constraints cannot be satisfied in the Euclidean space. Hence, to optimize the log likelihood, we resort to Riemannian Optimization. We combine α and α_0 such that $A = [\alpha\ \alpha_0]$ and $A^\top A = I_{d+q}$. A can then be optimized on the Stiefel manifold. The Stiefel manifold $St(n, d+q)$ is the set of matrices in $\mathbb{R}^{n\times(d+q)}$ whose columns are orthonormal. The other parameters $(\Lambda_y, \Lambda_0, \sigma)$ can be optimised in the Euclidean space $\mathbb{R}^k$ where $k = dh + q + 1$. Considering all parameters jointly, $\mathcal{L}$ can be optimised over a product manifold $\mathcal{M} = St(n, p+q)\times\mathbb{R}^k$, which fits $p(d+q)+dh+q+1$ total parameters. For reference, this is a decrease in parameter count from LAD on the same problem, which would fit $\mathcal{O}(p^2)$ parameters.

To perform this optimization over $\mathcal{M}$, we use Riemannian LBFGS proposed in Hosseini & Sra, 2020 that we implement using the pymanopt library (Townsend et al., 2016). To avoid getting stuck in local optima, we initialize the model parameters separately. α is initialized with the estimate obtained by any other dimensionality reduction method (like SIR, SAVE, etc). Once an initial guess for α has been obtained, α_0 and σ can be obtained by Probabilistic PCA of X projected on to the null space of α. The top $q-$dimensions form α_0 and the estimated noise variance acts as the initial guess for σ. The covariance matrices Λ_y and Λ_0 can be initialised as $\alpha^\top\widehat{\Sigma}_{X|(Y=y)}\alpha$ and $\alpha_0^\top\widehat{\Sigma}_X\alpha_0$ where $\widehat{\Sigma}_X$ is the sample covariance of X.

Also, CFAD requires an additional hyperparameter q (the dimensionality of the class independent subspace) as compared to other dimensionality reduction methods which only require d. In our experiments, we set $d + q$ to the number of components that capture 90% variance in the data, hence, eliminating the challenge of combinatorially optimizing d and q.

3.2. Smooth-CFAD

The factor-analytic structure of CFAD easily accommodates regularization in the form of priors over the model parameters. (Note that, due to the specific form of the profile likelihood it employs, priors are *not* easily incorporated into the LAD model (Cook & Forzani, 2009)). Smoothness is a common property of high-dimensional datasets, and fMRI data in particular. We, therefore, illustrate the benefits of regularization by adding a smoothness penalty on the columns of α. This encourages the axes defining the central subspace to be smooth along the natural dimensions of X. For fMRI data, the raw observations take the form of a 3D tensor defined by voxel locations in the brain. To encourage smoothness, we place a prior on α in the form

of a Gaussian, with inverse variance λ on the first-order differences between adjacent voxels, considered along each of the three cardinal dimensions of the tensor. This prior serves as a penalty on α which is constrained to lie on the Stiefel manifold. The resulting log posterior can be written as:

$$\mathcal{L}_{\text{sCFAD}} = \mathcal{L}_{\text{CFAD}} - \tfrac{1}{2}\lambda \operatorname{Tr}(\alpha^\top D\alpha) \tag{9}$$

where D is the graph Laplacian matrix incorporating that adjacency such that columns of α vary smoothly across the $p-$dimensions. The hyperparameter λ can be estimated by cross-validation over a range of values. We refer to this model as smooth-CFAD (sCFAD) in the rest of the paper.

4. Performance Evaluation

To test the effectiveness of CFAD as compared to other inverse regression SDR methods, we simulate observations from a CFAD model with $h = 3$ classes, $p = 100$, $d = 2$ and $q = 3$. The class means μ_y, of the latent mixture of Gaussians in $\mathbb{R}^d$, are randomly generated such that they satisfy the following:

$$c_{min} \max\{\operatorname{Tr}(\Lambda_i), \operatorname{Tr}(\Lambda_j)\} \geq ||\mu_i - \mu_j|| \geq c_{max} \max\{\operatorname{Tr}(\Lambda_i), \operatorname{Tr}(\Lambda_j)\}$$

$\forall i \neq j$, where c_{min} and c_{max} determine the degree of separation of the Gaussian mixture (Dasgupta, 1999; Hosseini & Sra, 2020). We use three different degrees of separation: $c \in [0.2, 0.5]$ (low separation), $c = [1, 3]$ (mid separation) and $c > 3$ (high separation). We set the class covariance, Λ_y, for $y = 1, 2, 3$ to $(2, 4)$, $(5, 3)$ and $(2, 2)$, and the class-independent covariance (Λ_0) to $(2, 8, 8)$. Further, the columns of projection matrix of α are sampled from $\mathcal{N}(0, I_p)$ such that they are orthonormal, and α_0 is set to any $q-$dimensional subspace in the null space of α.

Fig. 2 shows the the average principal subspace angle (between the subspace defined by α and its estimates) by different methods when N is varied from $[500, 10000]$. We see that CFAD is better than all the other methods at all three degrees of separation. The relative performance difference between CFAD and the other methods is higher at smaller values of N, hence showing its effectiveness in the low-sample regime where the performance of other methods degrades.

With the same parameters as the above experiment, and additionally smoothed α such that the columns of α are sampled from $\mathcal{N}(0, D^+)$ where $D^+ \in \mathbb{R}^{p\times p}$ is the graph Laplacian prior,we compare the performance of CFAD and sCFAD with other SDR methods in Fig. 3 at different degrees of separation of the latent Gaussian mixture. The hyperparameter λ for sCFAD is chosen by a grid search over $\lambda \in [10^{-3}, 10^3]$. We find that CFAD and sCFAD are better than all other methods (Fig. 3); in fact when $N \leq p$, the other methods break but CFAD and sCFAD still yield reasonably accurate subspaces. The smoothness prior in sCFAD is especially useful when N is small, leading to the drastic improvement in performance of sCFAD at small N.

We also illustrate the performance of CFAD in simple regressive relationships (as done by Cook & Forzani (2009)) by creating 4 simple regression models. In each case, $N = 500$, X is drawn from $\mathcal{N}(0, I_p)$ where $p = 8$ and Y is generated according to the following models: (a)$Y = 4X_1/a + \epsilon$, (b)$Y = X_1^2/(20a) + 0.1\epsilon$, (c)$Y = X_1/(10a) + aX_1^2/100 + 0.6\epsilon$, (d)$Y = 0.4a(\beta_1^\top X) + 3\sin(\beta_2^\top X/4) + 0.2\epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$. For models (a), (b) and (c), $\alpha = [1, 0, ...]^\top$. For model (d), $\beta_1 = [1, 1, 1, 0, ...]^\top$ and $\beta_2 = [1, 0, 0...., 1, 3]$. Hence, d=1 for the first three models and d=2 for the last model. The parameter a determines the strength of linear and quadratic relationships in the models and is varied over different ranges for each model. Except for the first model, all the other models are non-linear and their conditional distribution of $X \mid Y$ is not normal, an assumption that CFAD makes. Fig. 2 shows the principal subspace angle, averaged over 100 replications, between the estimated and true subspace for SIR, SAVE, DR, LAD and CFAD (with Y sliced into 5 equally spaced bins, $h = 5$). We find that CFAD is competitive with LAD in all the four models (with $q = 7$ for CFAD, which means α_0 is simply the null space of α). In this setting, both CFAD and LAD indeed are very similar at sufficiently high N to allow both the models to fit to the data. This also shows that while CFAD makes the normality assumption, its performance is robust to non-normality like that of LAD. (In all our experiments, we find that random initialization fails $< 0.1\%$ times.)

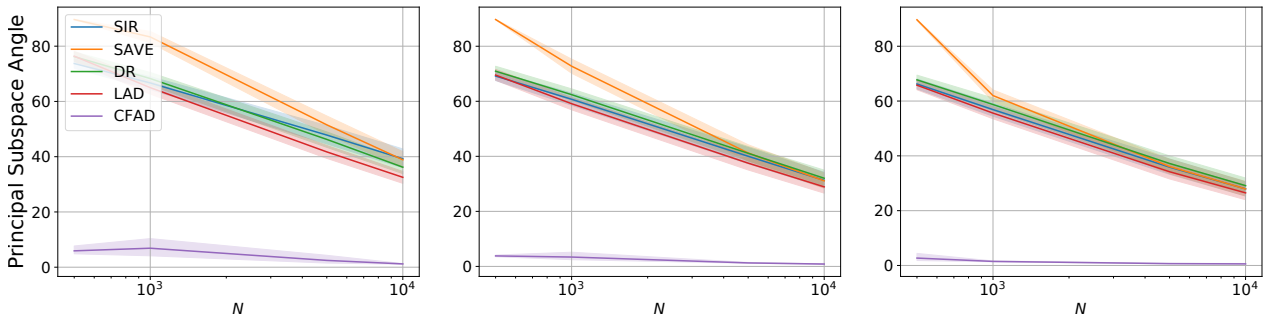
5. Application to fMRI data

To evaluate CFAD on real-world datasets, we used it to classify functional magnetic resonance imaging (fMRI) activity in two different tasks.

Visual Object Recognition

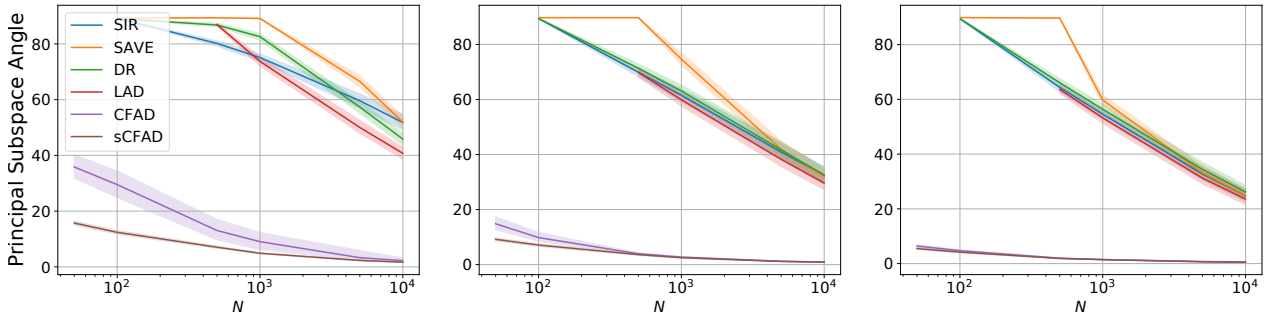
We first used CFAD to classify functional magnetic resonance imaging (fMRI) activity recorded during a visual object recognition task (Haxby et al., 2001). The fMRI data is pre-processed (using the nilearn package (Abraham et al., 2014)). Stimuli consisted of images from 8 classes: houses, chairs, bottles, scissors, shoes, faces, cats and nonsense patterns. fMRI activity was recorded in six subjects as they performed 1-back task; selectively extracting the responses for the ventral temporal cortex in each subject. More details about the dataset can be found in Haxby et al. 2001.

We compare sCFAD and CFAD with other inverse regression SDR methods, reduced-rank regression (RRR;,with an L2-penalty), Fisher Linear Discriminant (LDA) and PCA in



(a) Low separation (b) Medium separation (c) High separation

Figure 2. Principal Subspace angles between the true subspace and the subspace estimated by SDR methods at varying number of samples (averaged over 100 runs) when the data is simulated from the CFAD model ($p = 100, d = 2, q = 3, h = 3$). The latent class means in $\mathbb{R}^2$ are sampled at three degrees of separation.



(a) Low separation (b) Medium separation (c) High separation

Figure 3. Principal Subspace angles between the true subspace and the subspace estimated by SDR methods at varying number of samples (averaged over 100 runs) when the data is simulated from sCFAD ($p = 100, d = 2, q = 3, h = 3$). The class means in $\mathbb{R}^2$ are sampled at three degrees of separation. (SIR, SAVE and DR require at least $p + 1$ samples, while LAD requires every class to have $> (p + 1)$ samples)

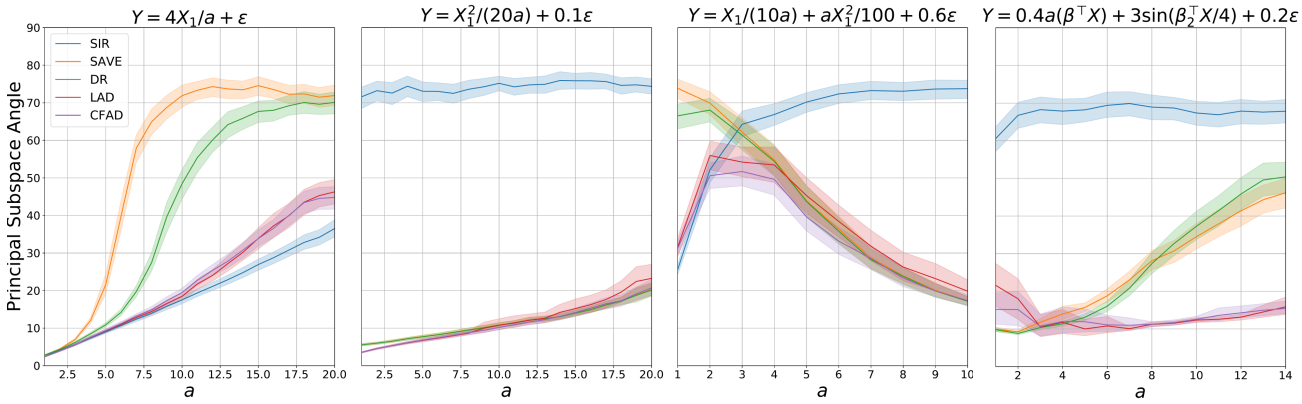


Figure 4. Principal subspace angles between the true and estimates dimension reduction subspaces for different SDR methods under varying input-output relationships.

terms of their accuracy in decoding stimuli labels from fMRI activity in all the six subjects. We also compare with Linear Optimal Low-rank projection (LOL) (Vogelstein et al., 2017), a dimensionality reduction method tailored to high-dimensional small-sample size data. LOL estimates α by performing a low-rank approximation of the class-centered covariances and appending mean information to it. We use the voxel locations in the recordings to form smoothness

priors for sCFAD, to help capture the smooth activations common throughout fMRI data.

fMRI recordings from the ventral temporal cortex in every subject form the data matrix, $X \in \mathbb{R}^{p \times N}$. The number of voxels in the ventral temporal cortex (p) varies between $(307, 675)$ across subjects, and the number of observations N per class (at different time points) is 108; and there are 8 stimuli classes. We perform a 5-fold cross-validation. In each fold, a dimension reduction method estimates α using the training set which is then used to project X to a lower-dimensional subspace. This is followed by classification on this low-dimensional data using a linear SVM and the average test accuracy across the folds in reported. For all methods, except RRR, we vary d in increments of 10 in $(0, p]$. In case of RRR, we use one-hot encoded class labels as the target output and select the optimal rank ($r \in [1, c]$ where $c = 8$ is the number of classes) as well as the L2-coefficient ($\in [10^{-3}, 10^4]$) by cross-validation. Now, since CFAD and sCFAD require d as well as q (dimensionality of the class-independent subspace), we fix $d + q$ to the number of principal components needed to explain 90% variance in X. We initialize sCFAD and CFAD with SIR, using the initialization scheme discussed in Section. 3.1 and choose the smoothness hyper-parameter $\lambda \in [10^{-3}, 10^4]$ by nested cross-validation within each of the 5 folds.

Table 1 shows the classification accuracy obtained by all the methods at the optimal d (based on average validation accuracy across folds) for sCFAD. We find that sCFAD is better than all other methods for every subject in terms of classification accuracy. In fact, even in the absence of the smoothness prior, CFAD performs better than remaining methods on 4 out of 6 subjects. Table 2 shows the classification accuracy for different methods at their respective best d (when d was varied in $[0, p]$ in increments of 10). We find that sCFAD performs best for 4 of the six subjects, while DR performs best for the remaining two subjects but at much higher d than sCFAD. Also, when we increase $d + q$ to include the number of principal components needed to explain 95% variance in X (instead of 90%), we find that sCFAD yields 89.3% accuracy at $d = 20$ for subject-1 and 78.4% at $d = 50$ for subject-6, surpassing DR (more details along with comparison to voxel selection in the Supplement).

Human Connectome Project Working Memory Task

Next, we considered the task of decoding stimuli from fMRI measurements in a working memory task. The dataset contains fMRI images from 10 human subjects performing the Human Connectome Project (HCP) working memory task, which is adapted from the classic n-back task (with two load levels: 0-back and 2-back). Subjects were presented with stimuli comprising of 4 categories of images: places, tools, faces and body parts with 132 TRs per category. fMRI measurements were obtained from a 3D brain volume of 59,412 voxels out of which 3093 voxels were extracted for analysis of the working memory task. Hence, the full dataset contains $N = 528$ samples and $d = 3093$ features per subject, falling in the $N < p$ regime that we are interested in. Further details about the dataset can be found in Barch et al., 2013; WU - Minn Consortium Human Connectome Project, 2017.

Similar to the visual object recognition task, we compare a set of dimensionality reduction methods with CFAD and sCFAD in term of their decoding accuracy in all 10 subjects. The inverse regression SDR methods cannot be applied in the $N < p$ regime, due to the non-invertibility of the sample covariance, hence we excluded them from this analysis. With each method, we perform a 3-fold cross validation analysis, using the dimensionality reduction methods to project the data X and subsequently classify the lower-dimensional projections using an SVM in each fold. We vary d in increments of 10 in the range of $(0, 500]$ and report the average cross-validation accuracy. For CFAD and sCFAD, we fix $d + q$ to the number of components that explain 90% variance in training data. All other hyperparameters (λ for sCFAD, L2-coefficient and rank for RRR) are set using cross-validation.

Table 3 shows the classification performance obtained by all methods at their respective optimal d (based on classification score), where we find that sCFAD outperforms all other methods for 9 out of 10 subjects while using only 10-20 dimensions. These results strongly suggest that sCFAD is highly suited for high-dimension small-sample-size dimensionality reduction.

6. Discussion

In this paper, we have proposed a linear dimensionality-reduction method for regression settings with high-dimensional, small-sample-size datasets. We showed that our method outperforms existing methods on simulated data in the $N < p$ regime, and achieves state-of-the-art classification performance on real fMRI datasets, proving its utility in real-world applications.

While dimensionality reduction methods have been extensively studied, methods for the undersampled ($N < p$) regime have received comparatively little attention (Candes & Tao, 2007). Previous work from Yata & Aoshima (2010) proposed a modified PCA for high-dimensional, small-sample-size data which uses a noise-reduction methodology to estimate the eigenvalues of the data covariance matrix. However, this is an unsupervised method, so it ignores the output label information that SDR and related supervised methods seek to capture. Linear-Optimal Low-rank projection (LOL) is an extension to PCA that incorporates class-

Table 1. 8-class classification accuracy on fMRI data after dimensionality reduction. d is optimal for sCFAD, 12.5% is chance performance

SUBJECT	d	sCFAD	CFAD	LDA	SIR	SAVE	DR	LAD	PCA	LOL	RRR
1	10	**62.4**	57.3	59.3	59.5	10.0	54.1	52.1	21.9	30.1	23.0
2	10	**71.8**	68.9	58.9	59.9	12.1	62.3	36.5	23.5	31.3	18.7
3	10	**66.4**	63.0	60.3	62.1	10.2	61.7	44.0	32.8	42.3	16.0
4	20	**62.2**	61.2	22.0	20.8	11.6	30.3	29.3	24.8	26.8	19.6
5	10	**72.8**	69.8	60.1	61.8	12.2	63.5	50.8	34.7	41.2	18.1
6	10	**73.1**	70.9	71.5	70.8	10.7	71.9	65.0	39.7	53.0	21.2

Table 2. 8-class classification accuracy on fMRI data after dimensionality reduction (at optimal d for the respective method)

SUBJECT	sCFAD		CFAD		LDA		SIR		SAVE		DR		LAD		PCA		LOL		RRR
	d	%	d	%	d	%	d	%	d	%	d	%	d	%	d	%	d	%	%
1	10	62.4	10	57.3	10	59.3	10	59.5	180	12.6	350	**75.8**	50	56.1	90	57.3	70	46.5	23.0
2	10	**71.8**	10	68.9	10	58.9	10	59.9	460	20.1	10	62.3	40	40.2	460	46.1	360	41.2	18.7
3	10	**66.4**	10	63.0	10	60.3	20	62.9	300	17.0	10	61.7	50	49.2	250	55.1	260	51.5	16.0
4	20	**62.2**	20	61.2	10	22.3	50	21.2	80	12.5	50	58.1	10	29.3	310	32.6	530	30.3	19.6
5	10	**72.8**	10	69.8	10	60.1	10	61.8	420	35.9	180	69.4	10	50.8	360	54.3	80	51.4	18.1
6	10	73.1	10	70.9	10	70.9	30	71.2	240	14.4	50	**74.0**	10	65.0	230	67.5	240	63.5	21.2

Table 3. 4-class classification accuracy on HCP working memory dataset after dimensionality reduction (at optimal d for the respective method; 25% is chance performance)

SUBJECT	sCFAD		CFAD		LDA		PCA		LOL		RRR
	d	%	d	%	d	%	d	%	d	%	%
1	10	**73.9**	10	70.9	10	64.7	10	67.4	20	72.7	20.3
2	10	**85.5**	10	83.3	10	84.8	40	77.3	30	82.7	24.4
3	10	**93.2**	10	91.8	10	88.6	40	80.1	50	85.0	20.8
4	20	**86.2**	20	86.0	10	82.2	30	82.6	20	84.8	24.5
5	10	85.1	10	83.5	10	**86.3**	140	82.2	40	82.0	25.2
6	10	**94.1**	10	93.0	10	87.1	70	85.2	30	87.5	25.0
7	10	**91.2**	10	89.9	10	88.4	20	88.2	40	91.0	26.7
8	10	**87.1**	10	83.3	10	81.6	30	82.6	10	82.7	25.2
9	10	**89.9**	10	87.3	10	87.8	50	79.2	30	79.9	22.5
10	10	**92.8**	10	89.8	10	92.6	70	86.6	50	87.9	29.2

label information, introducing a promising approach in this regime (Vogelstein et al., 2017). Recent work from Tan et al. (2018) proposed a modification of SIR for high-dimensional data by incorporating a lasso penalty such that the estimated subspace is sparse. This work succeeds several other approaches for high-dimensional data that combine variable selection with sliced inverse regression (Wang et al., 2018; Yin & Hilafu, 2015). Deleforge et al. (2015) propose an inverse regression framework for mapping high-dimensional data to a target output using a probabilistic mixture model, but the objective of this method is purely regression and it does not find a dimension reduction subspace (see Supplement for details).

However, unlike these existing methods, CFAD provides a generative model-based approach for estimating the dimension reduction subspace. Further, unlike the LAD model, CFAD allows for addition of known priors on data providing flexibility for incorporating data-specific information. Our approach thus makes an important contribution to the limited literature on high-dimensional, small-sample size dimensionality reduction methods. Finally, it opens a variety of promising avenues for future work, including the use of complex priors on the linear projections (incorporating structure beyond smoothness), automated methods for selecting dimensionality, and generalizations to heavy tailed and other non-Gaussian noise models.

Acknowledgements

We thank Anqi Wu for sharing the pre-processed HCP working memory dataset. This work was supported by grants from the Simons Collaboration on the Global Brain (SCGB AWD543027), the NIH BRAIN initiative (R01EB026946), an NSF GRFP, and a U19 NIH-NINDS BRAIN Initiative Award (5U19NS104648).

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8: 14, 2014. ISSN 1662-5196. doi: 10.3389/fninf.2014.00014. URL https://www.frontiersin.org/article/10.3389/fninf.2014.00014.

Aoi, M. C. and Pillow, J. W. Model-based targeted dimensionality reduction for neuronal population data. In *Advances in Neural Information Processing Systems*, 2018.

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., Nolan, D., Bryant, E., Hartley, T., Footer, O., Bjork, J. M., Poldrack, R., Smith, S., Johansen-Berg, H., Snyder, A. Z., and Van Essen, D. C. Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage*, 2013. ISSN 10538119. doi: 10.1016/j.neuroimage.2013.05.033.

Candes, E. and Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313 – 2351, 2007. doi: 10.1214/009053606000001523. URL https://doi.org/10.1214/009053606000001523.

Cook, R. D. Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22(1):1–26, 2007.

Cook, R. D. and Forzani, L. Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, 104(485):197–208, 2009. ISSN 01621459.

Cook, R. D. and Ni, L. Sufficient dimension reduction via inverse regression. *Journal of the American Statistical Association*, 100(470):410–428, 2005. doi: 10.1198/016214504000001501.

Cook, R. D. and Weisberg, S. Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):328–332, 1991. ISSN 01621459. URL http://www.jstor.org/stable/2290564.

Cowley, B. R., Smith, M. A., Kohn, A., and Yu, B. M. Stimulus-driven population activity patterns in macaque primary visual cortex. *PLoS computational biology*, 12 (12):e1005185, 2016.

Cunningham, J. P. and Ghahramani, Z. Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research*, 16(89):2859–2900, 2015. URL http://jmlr.org/papers/v16/cunningham15a.html.

Cunningham, J. P. and Yu, B. M. Dimensionality reduction for large-scale neural recordings. *Nature neuroscience*, 17(11):1500–1509, 2014.

Dasgupta, S. Learning mixtures of gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, FOCS '99, pp. 634, USA, 1999. IEEE Computer Society. ISBN 0769504094.

Deleforge, A., Forbes, F., and Horaud, R. High-dimensional regression with gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 25:893–911, 2015.

Ghahramani, Z. and Hinton, G. E. The em algorithm for mixtures of factor analyzers. Technical report, 1997.

Globerson, A. and Tishby, N. Sufficient dimensionality reduction. *Journal of Machine Learning Research*, 3 (Mar):1307–1331, 2003.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001. ISSN 0036-8075. doi: 10.1126/science.1063736. URL https://science.sciencemag.org/content/293/5539/2425.

Hosseini, R. and Sra, S. An alternative to em for gaussian mixture models: batch and stochastic riemannian optimization. *Mathematical Programming*, 181(1):187–223, May 2020. ISSN 1436-4646. doi: 10.1007/s10107-019-01381-4. URL https://doi.org/10.1007/s10107-019-01381-4.

Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C. E., Kepecs, A., Mainen, Z. F., Qi, X.-L., Romo, R., Uchida, N., and Machens, C. K. Demixed principal component analysis of neural population data. *Elife*, 5:e10989, 2016.

Li, B. and Wang, S. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008, September 2007. ISSN 0162-1459. doi: 10.1198/016214507000000536.

Li, K.-C. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414): 316–327, 1991. ISSN 01621459. URL http://www.jstor.org/stable/2290563.

Pang, R., Lansdell, B. J., and Fairhall, A. L. Dimensionality reduction in neuroscience. *Current Biology*, 26(14):R656–R660, 2016.

Tan, K. M., Wang, Z., Zhang, T., Liu, H., and Cook, R. D. A convex formulation for high-dimensional sparse sliced inverse regression. *Biometrika*, 105(4):769–782, 10 2018.

ISSN 0006-3444. doi: 10.1093/biomet/asy049. URL https://doi.org/10.1093/biomet/asy049.

Townsend, J., Koep, N., and Weichwald, S. Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *Journal of Machine Learning Research*, 17(137):1–5, 2016. URL http://jmlr.org/papers/v17/16-177.html.

Vogelstein, J. T., Bridgeford, E., Tang, M., Zheng, D., Burns, R., and Maggioni, M. Geometric dimensionality reduction for subsequent classification, 2017.

Wang, T., Chen, M., Zhao, H., and Zhu, L. Estimating a sparse reduction for general regression in high dimensions. *Statistics and Computing*, 28(1): 33–46, Jan 2018. ISSN 1573-1375. doi: 10.1007/s11222-016-9714-6. URL https://doi.org/10.1007/s11222-016-9714-6.

WU - Minn Consortium Human Connectome Project. WU-Minn HCP 1200 Subjects Data Release: Reference Manual. 2017(June):1–169, 2017. URL http://www.humanconnectome.org/documentation/S1200/HCP_S1200_Release_Reference_Manual.pdf.

Yata, K. and Aoshima, M. Effective pca for high-dimension, low-sample-size data with singular value decomposition of cross data matrix. *Journal of multivariate analysis*, 101(9):2060–2077, 2010.

Yin, X. and Hilafu, H. Sequential sufficient dimension reduction for large p, small n problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):879–892, 2015. doi: 10.1111/rssb.12093. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12093.