A Data-Dependent Method for Incorporating Known Structure into Regression Analysis with Applications in Computer Vision

David A. Shaw* Rama Chellappa†

Abstract
Various issues arise in performing regression and classification when one uses images as predictors, chief among them a high dimensionality. Considering a naive approach in which each pixel is entered into the model as a separate predictor, the analysis quickly becomes unwieldy, since modern cameras can output pictures with pixels numbering on the order of millions. However, images are highly constrained; the intrinsic dimension of an image is often far less than its ambient dimension. Moreover, various techniques to obtain new predictors (e.g., via feature extraction or by considering sets of predefined salient points in each image as variables in a regression model) can further reduce the dimension of the predictors, giving rise to an inherent lower-dimensional, manifold structure. We show that, for some regression problems in computer vision, using a data-dependent regularization that implicitly considers this manifold structure yields improvements over numerous alternatives. We extend this method to cases in which the structure is known a priori, in which it outperforms alternative methods, including those which explicitly take the known structure into account.

Key Words: Variable selection, regularization, computer vision, Grassmannian, age estimation

1. Introduction

Many problems in the field of computer vision involve predicting a variety of attributes of an individual pictured in a given image (e.g., age, gender, etc.) using the pixels in the image. For highly constrained settings, it is typical to formulate the problem with a simple, though somewhat naïve, approach in which the grayscale level of each pixel is considered as a separate predictor in a model. The predictors have both large dimensionality and inherent structure that cause trouble when building models and making predictions. Mitigating the adverse effects of this large dimensionality is an important issue and has recently received much attention in the literature; some of these problems are made less apparent using methods such as landmark extraction – that is, using locations of predefined, salient points on an individual's face as predictors as opposed to the raw pixel data. However, though the resulting predictors will have a smaller dimension and overcome some of the issues related to this high dimensionality, the problems due to the structure of the predictors is often still evident. This issue with the dependencies among the predictors – often called *collinearity* or *near-collinearity* – is a well-known problem in many areas of statistics. An early example is given in [11] in which the physical properties of pit props – lengths of lumber used to buttress walls in a mine – are estimated with numerous predictors that are highly correlated; principal component analysis (PCA) was used to investigate the effect that a new set of uncorrelated predictors has on the regression model. The issue of collinearity is also evident in a large number of problems in economics, and attempts to ameliorate its effects have been sought for years [6]. In this paper we will see the benefit to incorporating learned structure into a regression problem and also develop a method to handle the case in which the structure is known.

*University of Maryland, College Park, Department of Mathematics, College Park, MD 20742
†University of Maryland, College Park, Center for Automation Research, College Park, MD 20742

2. Methodology

2.1 Regression on Manifolds

Assume predictors $X_1, \ldots, X_n$ are independent, identically distributed (i.i.d.) in $\mathcal{X} \subset \mathbb{R}^D$. The response variables $Y_1, \ldots, Y_n \in \mathbb{R}$ are assumed to satisfy, for each $i = 1, \ldots, n$,

$$Y_i = m(X_i) + \sigma(X_i) \cdot \epsilon_i$$

with ϵ_i i.i.d., $E(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = 1$. Interest lies in finding the regression function m at a point x_0 defined as $m(x_0) = E(Y|X = x_0)$. In order to utilize the structure in the predictors, a common formalization is to assume that this structure arises from the fact that the predictors lie on a d-dimensional *manifold* $\mathcal{X}$ with $d < D$. For the purposes of this paper, a manifold will not be defined rigorously; the interested reader should consult [18] for an introduction to manifolds and differential geometry. Hereafter, a d-dimensional manifold will simply be thought of as a metric space which locally looks and acts like $\mathbb{R}^d$. The manifold assumption can simplify matters on the theoretical level, but there are still two issues. First, finding an embedding is not necessarily a trivial task. A large body of work exists to develop methods that perform this dimension reduction in order to learn an approximation to $\mathcal{X}$, with many popular techniques (e.g., LLE [16], ISOMAP [17]) working locally in an attempt to take advantage of local Euclidean properties of the manifold. Second, once an embedding is found, using these lower-dimensional points to build models that can be accurately interpreted in the higher-dimensional ambient space is not always possible with these projection methods; that is, if points are explicitly embedded into $\mathcal{X}$, some information that may be useful in the regression may be lost.

The manifold assumption arises quite naturally in computer vision. Purely data- dependent methods such as learning a face subspace using PCA on the difference between each data point and a test image have shown promising results [14], as have methods that incorporate prior knowledge of the structure through parameterization of small patches of each image in a database [15]. Analysis using landmark points can also be translated into the language of manifolds: after removing all affine transformations from the landmark points, the resulting predictors can be shown to lie on a *Grassmannian* $\mathcal{G}(2, B)$ – that is, the space of all 2-dimensional linear subspaces of $\mathbb{R}^B$ [19].

2.2 Related Work

There exists much prior work on estimating the regression function m when collinearities are present in the predictors $X_1, \ldots, X_n$. We use ordinary least squares (OLS) as a basis for comparison to alternative methods. OLS works by assuming the conditional mean of Y depends on X linearly, i.e., we seek solutions of the form $m(x_0) = \beta_0 + x_0^T \beta_1$ minimizing the residual sum of squares

$$\hat{\beta} = \arg\min_{\beta} ||Y - X\beta||_2^2$$

where $Y = [Y_1, \ldots, Y_n]^T$, $\beta = (\beta_0, \beta_1) \in \mathbb{R} \times \mathbb{R}^D$ and $X \in \mathbb{R}^{n \times (D+1)}$ is the design matrix. The solution is obtained by computing $\hat{\beta} = (X^T X)^{-1} X^T Y$. The presence of collinearities in the predictors X causes problems when taking the inverse of $X^T X$, requiring the use of alternative methods.

One such method to mitigate the effect of collinearities is to add a regularization parameter to ensure that the inverse of $X^T X$ is well-defined. This can be incorporated into the minimization above by adding an ℓ_2-penalty on some projection of the parameters β, resulting in the optimization

Algorithm 1 Calculation of the projection matrix in the regularization for the EDE [1]. When used in conjunction with an ℓ_2 penalty, this matrix will penalize regression coefficients for not lying parallel to the tangent space formed by $X_1, \ldots, X_n$.

1: **Estimate:** d with $\hat{d}$ using maximum likelihood [12]
2: $\hat{C} \leftarrow \sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T/n$
3: $\hat{C} \leftarrow [\hat{R}\ \hat{N}] \cdot \hat{\Lambda} \cdot [\hat{R}\ \hat{N}]^T$ eigenvalue decomposition of $\hat{C}$ with $\hat{R} \in \mathbb{R}^{D\times\hat{d}}$, $\hat{N} \in \mathbb{R}^{D\times(D-\hat{d})}$, $\hat{\Lambda}$ a diagonal matrix
4: $\hat{\Pi} \leftarrow \hat{N}\hat{N}^T$
5: $\hat{P} \leftarrow \mathrm{diag}(0, \hat{\Pi})$

Algorithm 2 Calculation of the projection matrix for EDE with Grassmann prior. This regularization assumes the data $X_1, \ldots, X_n$ lie on a Grassmannian $\mathcal{G}(2, B)$ and will penalize a projection of the regression coefficients into the horizontal space of $\mathcal{G}(2, B)$.

1: **Compute:** orthogonalization $\bar{X}_*$ of $\bar{X} \in \mathbb{R}^{B\times 2}$ using singular value decomposition
2: $\hat{\Pi^M} \leftarrow (I - \bar{X}_*\bar{X}_*^T)$
3: $\hat{\Pi} \leftarrow \mathrm{diag}(\hat{\Pi}^M, \hat{\Pi}^M)$, a block-diagonal matrix such that $\hat{\Pi} \in \mathbb{R}^{2B\times 2B}$
4: $\hat{P} \leftarrow \mathrm{diag}(0, \hat{\Pi})$

$$\hat{\beta} = \arg\min_{\beta} ||Y - X\beta||_2^2 + \lambda \cdot ||P\beta||_2^2$$

for $\lambda > 0$ a parameter. This yields the solution $\hat{\beta} = (X^TX + \lambda \cdot P^TP)^{-1}X^TY$, and the singularity of X^TX is no longer an issue. This method (known as *ridge regression* [10] when P is the identity matrix) can also handle the case in which $D \gg n$.

A slightly different attempt at removing the effect of collinearities on the predictors is found in principal components regression (PCR) [13]. PCR works by finding the d largest principal components of X and performing regression on these components. The benefit of PCR can be largely problem-dependent since the handling of collinearities and the problem of $D \gg n$ will depend on how many components are included in the final regression. By design PCR gives a sparse model – in fact, the model obtained is as sparse as the practitioner desires because it will always contain d predictors.

2.3 Exterior Derivative Estimator (EDE)

A more recent method given in [1] considers the manifold structure by locally penalizing the regression coefficients for not falling onto the d largest principal components (cf. PCR in which predictors are projected directly onto these components). The motivation for this projection is to penalize the coefficient vector for not lying parallel to the tangent space formed by the data points $X_1, \ldots, X_n$; this arises intuitively from the desire to interpret the regression coefficients as partial derivatives of the regression function m. For a globally linear manifold, the estimation of the projection matrix $\hat{P}$ for the EDE is given in Algorithm 1. In the case of nonlinear manifolds, structure can be taken into account by localizing the regression about each point of interest x_0.

2.4 Extensions of the EDE

2.4.1 Incorporation of Prior Structure

The EDE method given in [1] is largely concerned with the case where predictors lie on an *unknown* manifold; however, as was seen earlier, there are often problems in which the structure of the predictors is in fact known *a priori*, as in the case in which predictors lie on a Grassmannian $\mathcal{G}(2, B)$. We extend the EDE method to cases in which prior knowledge is available, with only the regularization needing modification. In constructing the EDE, the projection orthogonal to the tangent space is estimated with the data due to the structure in the predictors being unknown. In the case of a *known* manifold, we know the structure and can directly project coefficient vectors perpendicular to the tangent space for performing the regularization.

We think of a Grassmannian $\mathcal{G}(2, B)$ as the set of all 2-dimensional subspaces of $\mathbb{R}^B$, i.e.,

$$\mathcal{G}(2, B) = \mathcal{R}(2, B)/\sim$$

where $\mathcal{R}(2, B)$ is the space of all $B \times 2$ matrices of rank 2, and, for $U, V \in \mathbb{R}^{B\times 2}$, $U \sim V$ if there exists a nonsingular $L \in \mathbb{R}^{2\times 2}$ such that $V = UL$ [2]. The tangent structure of $\mathcal{G}(2, B)$ is slightly different from that of a manifold formed by data points $X_1, \ldots, X_n$ due to this quotient space representation. Rather than tangent spaces to points on $\mathcal{G}(2, B)$, we seek tangent spaces to equivalence classes of points, which for $\mathcal{G}(2, B)$ amount to orthogonal matrices $U \in \mathbb{R}^{B\times 2}$. The tangent space to the equivalence class of a point is known as the *vertical space*, and its orthogonal complement is called the *horizontal space* [4]. For two orthogonal matrices $U, V \in \mathcal{G}(2, B) \subset \mathbb{R}^{B\times 2}$, projection of a matrix U into the horizontal space at a point V can be done with the operator

$$\pi_V(U) = (I - VV^T)U$$

where I is the $B \times B$ identity matrix. In this case, if the predictors $X_1, \ldots, X_n \in \mathcal{G}(2, B)$, we think of the regression coefficients β_1^M as lying in $\mathbb{R}^{B\times 2}$ to allow for a projection of β_1^M into the horizontal space using π_V. In order to perform the regression, we reshape predictors X and coefficients β by concatenating column-wise so that $X, \beta \in \mathbb{R}^{2B}$. We did not consider the use of localization when performing regression on these points, in effect assuming that this manifold is approximately globally linear; this assumption was not shown to be too restrictive in practice. The estimation of the projection matrix $\hat{P}$ for the regularization is given in Algorithm 2.

2.4.2 Linearity Assumption

The Grassmannian $\mathcal{G}(2, B)$ is a nonlinear manifold; however, a useful property of manifolds is that locally they behave like Euclidean space. Since for our purposes the Grassmannian structure comes from predictors as landmark points on a face, it can be assumed that they do not have a high variability: an individual's eyes will typically appear above the nose and mouth and not be spaced arbitrarily far apart or close together.

To test the linearity assumption, 1000 pairs of points were chosen with replacement at random from the dataset and the quantity $X_i^T X_j$ was computed. The means and standard deviations of each element of this matrix are given in Table 1. For comparison, the same was done with data generated uniformly at random on $\mathcal{G}(2, B)$, with the method for obtaining these random observations Y_i outlined in Table 2 [2].

.9987 (.0012) -.0010 (.0321)	.0009 (.0037) .0001 (.0035)
.0031 (.0320) .9661 (.0396)	-.0000 (.0035) .0009 (.0033)

Table 1: Comparison between means and standard deviations of $X_i^T X_j$ for dataset (left) and randomly generated observations (right). Note for $X_i \in \mathcal{G}(2, B)$ we have $X_i^T X_i = I$ for I the 2×2 identity matrix.

1. Generate $2B$ random standard normal variates $u_1, \ldots, u_{2B} \sim N(0, 1)$;
2. Form random variates into matrix $U = [\mathbf{u}_1\ \mathbf{u}_2]$ where $\mathbf{u}_1 = [u_1, \ldots, u_B]^T$ and $\mathbf{u}_2 = [u_{B+1}, \ldots, u_{2B}]^T$;
3. Compute matrix $Z = U(U^T U)^{-1} U^T$;
4. Form $Y_i = [\mathbf{z}_1\ \mathbf{z}_2]$ where $\mathbf{z}_1$ and $\mathbf{z}_2$ are the first and second columns of Z, respectively.

Table 2: Method for generating a single uniform random variate on $\mathcal{G}(2, B)$.

Algorithm 3 Computation of the Karcher mean of a set of points [**?**].

given $X_1, \ldots, X_n \in \mathcal{G}(2, B) \subset \mathbb{R}^{B \times 2}$
initialize $\mu_0 = X_i$ (i random), $\epsilon = .5$, $\delta \in (0, 1)$, $j = 0$, and $d = 1$
while $d > \delta$ **do**
  for $i = 1, \ldots, n$ **do**
    $\nu_i \leftarrow \exp_{\mu_j}^{-1}(X_i)$
  end for
  $\bar{\nu} \leftarrow \sum_i \nu_i / n$
  $\mu_{j+1} \leftarrow \exp_{\mu_j}(\epsilon \bar{\nu})$
  $d \leftarrow ||\mu_j - \mu_{j+1}||$
  $j \leftarrow j + 1$
end while
return $\bar{X}_1 = \mu_j$

2.4.3 The Karcher Mean

A benefit to assuming linearity is that the computation of the mean of a set of observations is simplified on a computational level. The general procedure for computing a mean of a set of values on a Riemannian manifold is to use an iterative procedure: each point is projected into the tangent space about a candidate mean value, the sample mean in this tangent space is computed, and then this sample mean is projected back some distance along the geodesic between it and the previous candidate mean value. This procedure is outlined in Algorithm 3.

Since observations corresponding to normalized landmark points are contained within a small (read: approximately Euclidean) subset of the Grassmannian, a simpler computation of an approximation to the Karcher mean can be done as given in Algorithm 4. Instead of using an iterative procedure that relies on projecting and reprojecting sample points (using the inverse exponential and exponential map, respectively), the sample mean of the data can be taken and then orthogonalized to ensure it lies on the Grassmannian. This greatly improves computation time and additionally requires fewer tuning parameters than computation of the Karcher mean using Algorithm 3.

An empirical comparison between these two methods was performed using the FG-Net

Algorithm 4 Computation of the orthogonalized sample mean.

given $X_1, \ldots, X_n \in \mathcal{G}(2, B) \subset \mathbb{R}^{B\times 2}$
compute $\bar{X} \leftarrow \sum_i X_i/n$
let $\mathbf{v}_k$ be such that $\bar{X}\mathbf{v}_k = \lambda_k \mathbf{u}_k$ and $\bar{X}^*\mathbf{u}_k = \lambda_k \mathbf{v}_k$ with $\lambda_1 \geq \ldots \geq \lambda_B$
return $\bar{X}_2 = [\mathbf{u}_1\ \mathbf{u}_2]$

		2 iterations	**4 iterations**	**6 iterations**	**8 iterations**	**10 iterations**
Error between Algorithm 3 and proposed						
$n = 30$	Error	.012 (.006)	.011 (.005)	.009 (.004)	.008 (.004)	.007 (.003)
$n = 100$	Error	.012 (.006)	.010 (.005)	.009 (.005)	.007 (.004)	.007 (.003)
$n = 1000$	Error	.012 (.006)	.011 (.005)	.008 (.004)	.007 (.003)	.007 (.003)
Computation times						
$n = 30$	Algorithm 3	.035 (.000)	.070 (.001)	.105 (.001)	.139 (.002)	.174 (.002)
	Proposed	**.000** (.000)	**.000** (.000)	**.000** (.000)	**.000** (.000)	**.000** (.000)
$n = 100$	Algorithm 3	.096 (.009)	.187 (.002)	.278 (.002)	.369 (.002)	.460 (.003)
	Proposed	**.000** (.000)	**.000** (.000)	**.000** (.000)	**.000** (.000)	**.000** (.000)
$n = 1000$	Algorithm 3	.837 (.009)	1.67 (.011)	2.49 (.003)	3.32 (.010)	4.15 (.022)
	Proposed	**.003** (.001)	**.003** (.001)	**.003** (.001)	**.002** (.001)	**.002** (.000)

Table 3: Comparison between Algorithm 3 and the orthogonalized sample mean. For $n = 30, 100, 1000$ samples with replacement from the dataset, both Algorithm 3 and the proposed mean were computed, and the MSE between both and computation times are reported.

database [5], the results of which are given in Table 3. A random selection of 30, 100, and 1000 observations were chosen from the FG-Net database, and Algorithm 3 was performed using 2, 4, 6, 8, and 10 iterations. Meanwhile, the sample mean was also computed using Algorithm 4. The Frobenius norm between the two computed means was calculated, along with the computation times of both algorithms. It is interesting to note that, as the number of iterations increases, Algorithm 3 approaches the value obtained by simply orthogonalizing the sample mean, and as these iterations increased, the gap between computation times widened. On 1000 points using 10 iterations, it takes four seconds to compute the mean using Algorithm 3, compared with .002 seconds using the alternative method. Figure 1 shows a graphical comparison between the landmark points of the sample mean of the entire dataset computed using Algorithm 3 with 10 iterations and the proposed, simpler method, showing these methods obtain similar configurations.

2.4.4 Extension to Generalized Linear Models

It is worthwhile to consider the extension of both the EDE and the EDE with prior methods to generalized linear models (GLMs), viz. logistic regression, so that classification can be performed. This extension is straightforward, the chief assumption being that for a response variable $Y \in \{-1, +1\}$ the *logit* link function is used and the linear dependence on the parameters and predictors is through this function, i.e.,

$$\log \frac{\mathrm{P}(Y = +1|X = x_0)}{\mathrm{P}(Y = -1|X = x_0)} = \beta_0 + x_0^T \beta_1$$

where now instead of the conditional mean of Y depending linearly on X, some transformation of Y's conditional mean depends linearly on X. The estimation of β in this case can be performed using coordinate descent [7] or iterated reweighted least squares.

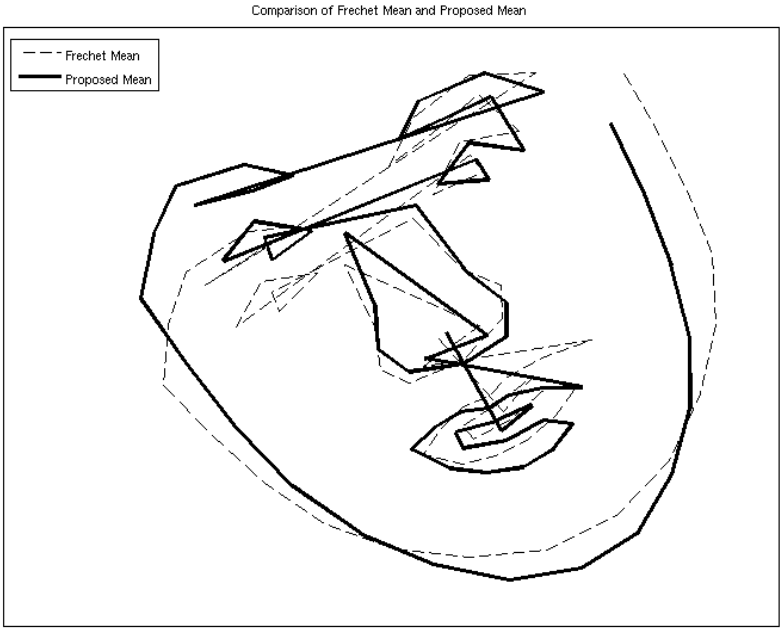

Figure 1: Comparison of mean face obtained via Algorithm 3 (Frechet mean with 10 iterations) and proposed mean

3. Applications

3.1 Experimental Set Up

We compare various methods applied to the FG-NET database [5]. This database consists of images of 82 separate individuals' faces; a total of 1002 images are in the database, 571 of which correspond to males and 431 females. The alternative methods tested for comparison were ordinary least squares (OLS), regression performed on points projected to tangent space about the orthogonalized sample mean (REM), principal components regression (PCR) [13], regression on points embedded using locality preserving projections (RLPP) [9], and ridge regression (RR), and these methods were compared to the exterior derivative estimator (EDE) and the exterior derivative estimator with prior (EDEwP). All forms of regression were performed on either the predictors or the embedded predictors. For age estimation, improvements in prediction can be gained by additionally including the square of each predictor in the model [8], but this was not considered in this analysis.

3.1.1 Unknown Structure

Feature extraction was used to obtain predictors whose structure is not explicitly known in advance. To obtain features, each image was converted to normalized grayscale taking values between 0 and 1, and the Viola-Jones face detection algorithm [20] was used to discard much of the noise and unwanted information contained in the background. Finally, a histogram of oriented gradients (HOG) [3] feature extraction method with 9 bins on 8×8 patches was used to generate predictors $X_1, \ldots, X_{1002} \in \mathbb{R}^{576}$.

3.1.2 Known Structure

Utilizing landmark points as predictors results in a manifold that is in fact known, so prior knowledge can be incorporated to aid estimation. For this dataset, 68 predefined landmark points are given for each image in $\mathbb{R}^2$ resulting in each predictor $X \in \mathbb{R}^{68\times 2}$. As was stated earlier, normalizing the predictors to remove all affine transformations by performing a singular value decomposition on each observation is a useful pre-processing step resulting in $X \in \mathcal{G}(2, 68)$. Predictors were concatenated column-wise to obtain vectors $X_1, \ldots, X_{1002} \in \mathbb{R}^{136}$.

3.2 Age Estimation

Age estimation is a popular problem in computer vision that has seen numerous solutions utilizing the manifold assumption on the predictors. Here each observation is labeled with ages Y ranging from 0 to 69. In various experiments, performing regression on $\sqrt{Y}$ yielded more accurate predictions; using this as a response variable has the added benefit that predictions of an individual's age will always be nonnegative.

A popular objective in the age estimation literature for assessing algorithm performance is to use a hold-one-person-out cross-validation and report the mean absolute error (MAE). In other words, 82 separate trials are performed where for each trial, the test dataset consists of all images of one specific individual while the training dataset is composed of the remaining 81 individuals. This method of assessment, hereafter referred to as Framework 3, gives a good indication as to how well methods are performing, but as an objective for both parameter tuning and performance assessment it leaves something to be desired. This cross-validation framework is closer in spirit to a jackknife cross-validation, and obtaining a randomized split between training and testing data will give a better idea of how the methods are performing relative to one another. We propose two alternative frameworks. Framework 1 chooses 5 test points at random for testing and uses the remaining observations for training the model. Framework 2, to be more consistent with hold-one-person-out cross-validation, does the same as Framework 1 but instead of training on the remaining individuals, each observation corresponding to a person in the testing set is removed from the training set and models are then built on this modified dataset.

Both of these tests (Framework 1 and Framework 2) are performed 100 times and the average and standard deviation of the MAE for each method is reported in Table **??**. In the case in which the structure of the predictors is unknown, the EDE outperforms the alternatives, with RR coming in a close second. Using the landmark data gives an overall improvement in performance of the non-projection methods (OLS, RR, EDE, EDEwP) while the projection methods (PCR, RLPP) actually perform worse. In this case, the EDEwP outperforms all alternatives with the EDE coming in a close second in Frameworks 1 and 2 and RR coming second in the hold-one-person-out validation.

4. Discussion

4.1 Bayesian Interpretation

Note that for the EDE with a Grassmann prior, estimates for β can be found by performing

$$\arg\min_{\beta} ||Y - X\beta||_2^2 + \lambda \cdot ||(I - \bar{X}_*\bar{X}_*^T)\beta^M||_F^2 \tag{1}$$

where as before $\bar{X}_*$ is the orthogonalized sample mean of the predictors, $\beta^M \in \mathbb{R}^{B\times 2}$ is the "matrix" version of β_1, and $||\cdot||_F^2$ is the squared Frobenius norm. This penalization

Model	Framework 1 MAE(SD)	Framework 2 MAE(SD)	Framework 3 MAE
OLS	11.24 (4.242)	11.94 (4.513)	10.84
PCR	10.18 (3.402)	9.750 (3.521)	10.21
RLPP	9.856 (3.411)	9.711 (3.411)	10.00
RR	7.883 (3.223)	8.249 (3.394)	7.788
EDE	**7.440** (3.091)	**7.709** (3.198)	**7.597**

Table 4: Age estimation results for various testing frameworks performed on HOG data in which the structure is unknown (optimal λ obtained for RR = 149, EDE = 723). Minimum mean absolute errors (MAEs) are given in bold.

Model	Framework 1 MAE(SD)	Framework 2 MAE(SD)	Framework 3 MAE
OLS	5.733 (3.004)	6.147 (3.258)	6.465
PCR	11.52 (4.777)	12.12 (4.949)	11.78
RLPP	10.26 (5.299)	9.642 (4.816)	10.10
RR	5.674 (3.095)	5.923 (3.280)	6.104
REM	5.674 (3.095)	5.923 (3.280)	6.614
EDE	5.668 (3.083)	5.920 (3.271)	6.465
EDEwP	**5.664** (3.077)	**5.919** (3.265)	**6.102**

Table 5: Age estimation results for various testing frameworks performed on landmark data in which the structure is known (optimal λ obtained for RR = .0053, EDE = .0047, EDEwP = .0044). Minimum mean absolute errors (MAEs) are given in bold.

term can be interpreted as placing a "Procrustean" prior on the parameters β^M. In other words, the estimate for β^M obtained by optimizing (1) above can be obtained as the Bayes posterior mode under the prior

$$f(\beta^M; \lambda) = K \cdot \exp\{-\lambda \cdot g(\bar{X}_*, \beta^M)\}$$

where $g(U, V) = \mathrm{tr}(V^T V - U^T V V^T U)$, $\mathrm{tr}(\cdot)$ denotes matrix trace, and K is a normalizing constant. This is called a "Procrustean" prior due to the fact that g above is similar to the Procrustes distance metric $g_P(U, V) = \mathrm{tr}(I - U^T V V^T U)$ given in [2]; in fact, it will hold locally that $(\beta^M)^T(\beta^M) \approx I$ since β^M lies on the tangent space to the manifold on which $X_1, \ldots, X_n$ reside (i.e., $\mathcal{G}(2, B)$), and any point X_i on this manifold satisfies $X_i^T X_i = I$.

5. Conclusion

Problems with dimensionality and collinearity abound in the field of computer vision. Datasets consisting of images have, by their design, an intrinsic structure regardless of their construction: whether raw pixel data are used, features are extracted, or landmark points are chosen, strong interdependencies between predictors will be unavoidable. Projection methods such as PCR attempt to sidestep this issue by obtaining new, uncorrelated predictors that are projections of the given data. This has been shown to be too restrictive; it is impossible to know whether useful information is being discarded when these projections are performed. By adopting a more flexible approach, both in the case in which the underlying manifold is known in advance or the case where the manifold is unknown, we have

obtained improvements in regression and classification. Posing the problem as an optimization and incorporating prior knowledge into the objective function results in improvements in performance and coefficient estimates that have an attractive interpretation in terms of the manifold structure of the predictors. While the data considered were assumed to have a globally linear structure, localization can be used to obtain better results on data that are assumed to lie on nonlinear manifolds.

References

[1] Anil Aswani, Peter J. Bickel, and Claire Tomlin. Regression on manifolds: Estimation of the exterior derivative. *The Annals of Statistics*, 39(1):pp. 48–81, 2011.

[2] Y. Chikuse. *Statistics on Special Manifolds*. Lecture notes in statistics. Springer, 2003.

[3] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages pp. 886–893, 2005.

[4] A. Edelman, T. Arias, and S. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Anaysis and Applications*, 20:pp. 303–353, 1998.

[5] Face and Gesture Recognition Research Network. FG-NET aging database, Accessed Apr 2011. http://www.fgnet.rsunit.com/.

[6] Donald E. Farrar and Robert R. Glauber. Multicollinearity in regression analysis: The problem revisited. *The Review of Economics and Statistics*, 49(1):pp. 92–107, 1967.

[7] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent, 2008. Available at http://www-stat.stanford.edu/˜jhf/ftp/glmnet.pdf.

[8] Yun Fu and T.S. Huang. Human age estimation with regression on discriminative aging manifold. *IEEE Transactions on Multimedia*, 10(4):pp. 578–584, June 2008.

[9] Xiaofei He and Partha Niyogi. Locality Preserving Projections. In Sebastian Thrun, Lawrence Saul, and Bernhard Scholkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.

[10] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*, 12:pp. 55–67, 1970.

[11] J.N.R. Jeffers. Two case studies in the application of principal component analysis. *Applied Statistics*, 16(3):pp. 225–236, 1967.

[12] E. Levina and P.J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in NIPS 17*, 2005.

[13] William F. Massy. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60(309):pp. 234–256, 1965.

[14] B. Moghaddam. Principal manifolds and probabilistic subspaces for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(6):pp. 780–788, June 2002.

[15] G Peyre. Manifold models for signals and images. *Computer Vision and Image Understanding*, 113(2):pp. 249–260, February 2009.

[16] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:pp. 2323–2326, 2000.

[17] J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science (New York, N.Y.)*, 290(5500):pp. 2319–23, December 2000.

[18] J.A. Thorpe. *Elementary Topics in Differential Geometry*. New York: Springer-Verlag, 1979.

[19] Pavan Turaga, Soma Biswas, and Rama Chellappa. The role of geometry for age estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages pp. 946–949, March 2010.

[20] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. *IEEE Conference on Computer Vision and Pattern Recognition*, 1, 2001.