REGRESSION ON MANIFOLDS: ESTIMATION OF THE EXTERIOR DERIVATIVE[1]

BY ANIL ASWANI, PETER BICKEL AND CLAIRE TOMLIN

University of California, Berkeley

Collinearity and near-collinearity of predictors cause difficulties when doing regression. In these cases, variable selection becomes untenable because of mathematical issues concerning the existence and numerical stability of the regression coefficients, and interpretation of the coefficients is ambiguous because gradients are not defined. Using a differential geometric interpretation, in which the regression coefficients are interpreted as estimates of the exterior derivative of a function, we develop a new method to do regression in the presence of collinearities. Our regularization scheme can improve estimation error, and it can be easily modified to include lasso-type regularization. These estimators also have simple extensions to the "large p, small n" context.

1. Introduction. Variable selection is an important topic because of its wide set of applications. Amongst the recent literature, lasso-type regularization [13, 30, 54, 59] and the Dantzig selector [10] have become popular techniques for variable selection. It is known that these particular tools have trouble handling collinearity. This has prompted work on extensions [60], though further developments are still possible.

Collinearity is a geometric concept: it is equivalent to having predictors which lie on manifolds of dimension lower than the ambient space, and it suggests the use of manifold learning to regularize ill-posed regression problems. The geometrical intuition has not been fully understood and exploited, though several techniques [5, 8, 30, 60] have provided some insight. Though it is not strictly necessary to learn the manifold for prediction [8], doing so can improve estimation in a min–max sense [43].

This paper considers variable selection and coefficient estimation when the predictors lie on a lower-dimensional manifold, and we focus on the case where this manifold is nonlinear; the case of a global, linear manifold is a simple extension, and we include a brief discussion and numerical results for this case. Prediction of function value on a nonlinear manifold was first studied in [8], but the authors

did not study estimation of derivatives of the function. We do not consider the case of global estimation and variable selection on nonlinear manifolds because [21] showed that learning the manifold globally is either poorly defined or computationally expensive.

1.1. *Overview.* We interpret collinearities in the language of manifolds, and this provides the two messages of this paper. This interpretation allows us to develop a new method to do regression in the presence of collinearities or near-collinearities. This insight also allows us to provide a novel interpretation of regression coefficients when there is significant collinearity of the predictors.

On a statistical level, our idea is to learn the manifold formed by the predictors and then use this to regularize the regression problem. This form of regularization is informed by the ideas of manifold geometry and the exterior derivative [34, 40]. Our idea is to learn the manifold either locally (in the case of a local, nonlinear manifold) or globally (in the case of a global, linear manifold). The regression estimator is posed as a least-squares problem with an additional term which penalizes for the regression vector lying in directions perpendicular to the manifold.

Our manifold interpretation provides a new interpretation of the regression coefficients. The gradient describes how the function changes as each predictor is changed independently of other predictors. This is impossible to do when there is collinearity of the predictors, and the gradient does not exist [51]. The exterior derivative of a function [34, 40] tells us how the function value changes as a predictor and its collinear terms are simultaneously changed, and it has applications in control engineering [47], physics [40] and mathematics [51]. In particular, most of our current work is in high-dimensional system identification for biological and control engineering systems [3, 4]. We interpret the regression coefficients in the presence of collinearities as the exterior derivative of the function.

The exterior derivative interpretation is useful because it says that the regression coefficients only give derivative information in the directions parallel to the manifold, and the regression coefficients do not give any derivative information in the directions perpendicular to the manifold. If we restrict ourselves to computing regression coefficients for only the directions parallel to the manifold, then the regression coefficients are unique and they are uniquely given by the exterior derivative.

This is not entirely a new interpretation. Similar geometric interpretations are found in the literature [16, 18, 19, 30, 37, 58], but our interpretation is novel because of two main reasons. The first is that it is the first time the geometry is interpreted in the manifold context, and this is important for many application domains. The other reason is that this interpretation allows us to show that existing regularization techniques are really estimates of the exterior derivative, and this has important implications for the interpretation of estimates calculated by existing techniques. We do not explicitly show this relationship; rather, we establish a link from our estimator to both principal components regression (PCR) [16, 37]

and ridge regression (RR) [26, 30]. Links between PCR, RR and other regularization techniques can be shown [18, 22, 24, 25].

1.2. *Previous work.* Past techniques have recognized the importance of geometric structure in doing regression. Ordinary least squares (OLS) performs poorly in the presence of collinearities, and this prompted the development of regularization schemes. RR [26, 30] provides proportional shrinkage of the OLS estimator, and elastic net (EN) [60] combines RR with lasso-type regularization. The Moore–Penrose pseudoinverse (MP) [30] explicitly considers the manifold. MP works well in the case of a singular design matrix, but it is known to be highly discontinuous in the presence of near-collinearity caused by errors-in-variables. PCR [16, 37] and partial least squares (PLS) [2, 16, 55] are popular approaches which explicitly consider geometric structure.

The existing techniques are for the case of a global, linear manifold, but these techniques can easily be extended to the case of local, nonlinear manifolds. The problem can be posed as a weighted, linear regression problem in which the weights are chosen to localize the problem [46]. Variable selection in this context was studied by RODEO [32], but this tool requires a heuristic form of regularization which does not explicitly consider collinearity.

Sparse estimates can simultaneously provide variable selection and improved estimates, but producing sparse estimates is difficult when the predictors lie on a manifold. Lasso-type regularization, the Dantzig selector and the RODEO cannot deal with such situations. The EN produces sparse estimates, but it does not explicitly consider the manifold structure of the problem. One aim of this paper is to provide estimators that can provide sparse estimates when the regression coefficients are sparse in the original space and the predictors lie on a manifold.

If the coefficients are sparse in a rotated space, then our estimators admit extensions which consider rotations of the predictors as another set of tunable parameters which can be chosen with cross-validation. In variable selection applications, interpretation of selected variables is difficult when dealing with rotated spaces, and so we only focus on sparsity in the original space. Numerical results show that our sparse estimators without additional rotation parameters do not seem to significantly worsen estimation when there is no sparsity in the unrotated space.

The estimators we develop learn the manifold and then use this to regularize the regression problem. As part of the manifold learning, it is important to estimate the dimension of the manifold. This can either be done with dimensionality estimators [12, 23, 35] or with resampling-based approaches. Though it is known that cross-validation performs poorly when used with PCR [31, 41], we provide numerical examples in Section 7 which show that bootstrapping, to choose dimension, works well with our estimators. Also, it is worth noting that our estimators only work for manifolds with integer dimensions, and our approach cannot deal with fractional dimensions.

Learning the manifold differs in the case of the local, nonlinear manifold as opposed to the case of the global, linear manifold. In the local case, we use kernels to localize the estimators which (a) learn the manifold and (b) do the nonparametric regression. For simplicity, we use the same bandwidth for both, but we could also use separate bandwidths. In contrast, the linear case has faster asymptotic convergence because it does not need localization. We consider a linear case with errors-in-variables where the noise variance is identifiable [28, 31], and this distinguishes our setup from that of other linear regression setups [13, 30, 54, 59, 60].

2. Problem setup. We are interested in prediction and coefficient estimation of a function which lies on a local, nonlinear manifold. In the basic setup, we are only concerned with local regression. Consequently, in order to prove results on the pointwise-convergence of our estimators, we only need to make assumptions which which hold locally. The number of predictors is kept fixed. Note that it is possible that the dimension of the manifold varies at different points in the predictor space; we do not prohibit such behavior. We cannot do estimation at the points where the manifold is discontinuous, but we can do estimation at the remaining points.

Suppose that we would like to estimate the derivative information of the function about the point $X_0 \in \mathbb{R}^p$, where there are p predictors. The point X_0 is the choice of the user, and varying this point allows us to compute the derivative information at different points. Because we do local estimation, it is useful to select small portions of the predictor-space; we define a ball of radius R centered at X in p-dimensions using the notation: $\mathcal{B}^p_{x,R} = \{v \in \mathbb{R}^p : \|v - x\|_2^2 < R\}$.[2]

We assume that the predictors form a d-dimensional manifold $\mathcal{M}$ in a small region surrounding X_0, and we have a function which lies on this manifold $f(\cdot) : \mathcal{M} \to \mathbb{R}$. Note that $d \le p$, and that d is in general a function of X_0; however, implicit in our assumptions is that the manifold $\mathcal{M}$ is continuous within the ball. We can more formally define the manifold at point X_0 as the image of a local chart:

$$(2.1) \qquad \mathcal{M} = \{\phi(u) \in \mathcal{B}^p_{x_0,\mu} \subset \mathbb{R}^p : u \in \mathcal{B}^d_{0,r} \subset \mathbb{R}^d\},$$

for small $\mu, r > 0$. An example of this setup for $p = 2$ and $d = 1$ can be seen in Figure 1.

We make n measurements of the predictors $X_i \in \mathbb{R}^p$, for $i = \{1, \ldots, n\}$, where the X_i are independent and identically distributed. We also make n noisy measurements of the function $Y_i = f(X_i) + \varepsilon_i$, where the ε_i are independent and identically distributed with $\mathbb{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$. Let $\kappa, M > 0$ be finite constants, and assume the following:

---

[2]In our notation, we denote subscripts in lower case. For instance, the ball surrounding the point X_0 is denoted in subscripts with the lower case x_0.
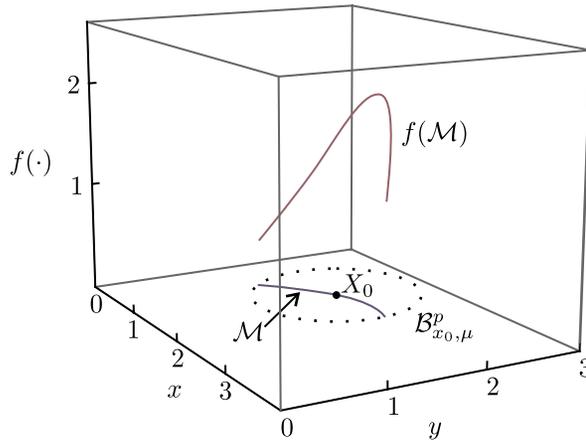
FIG. 1.   *In a small ball $\mathcal{B}_{x_0,\mu}^p$ about the point X_0, the predictors form the manifold $\mathcal{M}$. The response variable is a function of the predictors, which lie on this manifold. Here the manifold is of dimension $d = 1$, and the number of predictors is $p = 2$.*

1. The kernel function $K(\cdot)$, which is used to localize our estimator by selecting points within a small region of predictor-space, is three-times differentiable and radially symmetric. These imply that $K(\cdot)$ and $K''(\cdot)$ are even functions, while $K'(\cdot)$ is an odd function.

2. The bandwidth h is the radius of predictor points about X_0 which are used by our estimator, and it has the following asymptotic rate: $h = \kappa n^{-1/(d+4)}$.

3. The kernel $K(\cdot)$ either has exponential tails or a finite support [8]. Mathematically speaking,

$$\mathbb{E}\big[K^\gamma\big((X-x)/h\big)w(X)\mathbb{1}\big(X \in (\mathcal{B}_{x,h^{1-\varepsilon}}^p)^c\big)\big] = o(h^{d+4}),$$

   for $\gamma \in \{1, 2\}$, $0 < \varepsilon < 1$ and $|w(x)| \leq M(1 + |x|^2)$.

4. The local chart $\phi(\cdot)$ which is used to define the manifold in (2.1) is invertible and three-times differentiable within its domain. The manifold $\mathcal{M}$ is a differentiable manifold, and the function $f(\cdot)$ is three-times differentiable: $\|\partial_i \partial_j \partial_k (f \circ \phi)\|_\infty \leq M$.

5. The probability density cannot be defined in the ambient space because the Lebesgue measure of a manifold is generally zero. We have to define the probability density with respect to a d-dimensional measure by inducing the density with the map $\phi(\cdot)$ [8]. We define:

$$\mathbb{P}(X \in \mathcal{S}) = \mathbb{Q}\big(Z \in \phi^{-1}(\mathcal{S})\big),$$

   where $\mathcal{S} \subseteq \mathbb{R}^p$. The density $\mathbb{Q}(\cdot)$ is denoted by $F(z)$, and we assume that it is three-times differentiable and strictly positive at $(z = 0) \in \phi^{-1}(X_0)$.

6. The Tikhonov-type regularization parameter λ_n is nondecreasing and satisfies the following rates: $\lambda_n/nh^{d+2} \to \infty$ and $h\lambda_n/nh^{d+2} \to 0$. The lasso-type regularization parameter μ_n is nonincreasing and satisfies the following rates: $\mu_n(nh^{d+2})^{-1/2} \to 0$ and $\mu_n(nh^{d+2})^{(\gamma-1)/2} \to \infty$.

The choice of the local chart $\phi(\cdot)$ is not unique; we could have chosen a different local chart $\psi(\cdot)$. Fortunately, it can be shown that our results are invariant under the change of coordinates $\psi^{-1} \circ \phi$ as long as the measure $\mathbb{Q}(\cdot)$ is defined to follow suitable compatibility conditions under arbitrary, smooth coordinate changes. This is important because it tells us that our results are based on the underlying geometry of the problem.

3. Change in rank of local covariance estimates. To localize the regression problem, we use kernels, bandwidth matrices and weight matrices. We define the scaled kernel $K_h(U) = h^{-p}K(U/h)$, where h is a bandwidth. The weight matrix centered at X_0 with bandwidth h is given by

$$W_{x_0} = \text{diag}(K_h(X_1 - X_0), \ldots, K_h(X_n - X_0)),$$

and the augmented bandwidth matrix is given by $H = H^{1/2}H^{1/2}$, where

$$H^{1/2} = \sqrt{nh^d}\begin{bmatrix} 1 & 0 \\ 0 & h\mathbb{I}_{p\times p} \end{bmatrix}.$$

If we define the augmented data matrix as

$$X_{x_0} = \begin{bmatrix} 1 & (X_1 - X_0)' \\ \vdots & \vdots \\ 1 & (X_n - X_0)' \end{bmatrix},$$

then the weighted Gram matrix of X_{x_0} is

$$(3.1) \qquad \hat{C}_n \triangleq \begin{bmatrix} \hat{C}_n^{11} & \hat{C}_n^{12} \\ \hat{C}_n^{21} & \hat{C}_n^{22} \end{bmatrix} = h^p \cdot H^{-1/2}X'_{x_0}W_{x_0}X_{x_0}H^{-1/2}.$$

A formal statement is given in the Appendix, but the weighted Gram matrix (3.1) converges in probability to the following population parameters:

$$C^{11} = F(0)\int_{\mathbb{R}^d} K(d_u\phi \cdot u)\,du,$$

$$(3.2) \qquad C^{21} = C^{12'} = 0,$$

$$C^{22} = F(0)d_u\phi \cdot \left[\int_{\mathbb{R}^d} K(d_u\phi(0) \cdot u)uu'\,du\right] \cdot d_u\phi'.$$

If we expand the $\hat{C}_n^{22}$ term from the weighted Gram matrix (3.1) into

$$(3.3) \qquad \hat{C}_n^{22} = \frac{1}{nh^{d+2-p}}\sum_{i=1}^n K_h(X_i - X_0)(X_i - X_0)(X_i - X_0)',$$

it becomes apparent that $\hat{C}_n^{22}$ can be interpreted as a local covariance matrix. The localizing terms associated with the kernel add a bias, and this causes problems when doing regression. The $\hat{C}_n^{11}$ term does not cause problems because it is akin to the denominator of the Nadaraya–Watson estimator [15] which does not need any regularizations.

The bias of the local covariance estimate $\hat{C}_n^{22}$ causes problems in doing regression, because the bias can cause the rank of $\hat{C}_n^{22}$ to be different than the rank of C_n^{22}. The change in rank found in the general case of the local, nonlinear manifold causes problems with MP which is discontinuous when the covariance matrix changes rank [1]. In the special case of a global, linear manifold, a similar change in rank can happen because of errors-in-variables. It is worth noting that MP works well in the case of a singular design matrix.

4. Manifold regularization. To compensate for this change in rank, we use a Tikhonov-type regularization similar to RR and EN. The distinguishing feature of our estimators is the particular form of the regularizing matrix used. Our approach is to estimate the tangent plane at X_0 of the manifold $\mathcal{M}$ and then constrain the regression coefficients to lie close to the principal components of the manifold. The idea for this type of regularization is informed by the intuition of exterior derivatives.[3] An advantage of this regularization is that it makes it easy to apply lasso-type regularizations, and this combination of the two types of regularization is similar to EN.

To constrain the regression coefficients to lie close to the manifold, we pose the problem as a weighted least-squares problem with Tikhonov-type regularization:

$$(4.1) \qquad \hat{\beta} = \arg\min \|W(Y - X\beta)\|_2^2 + \lambda \|\Pi\beta\|_2^2.$$

The matrix Π is a projection matrix chosen to penalize β for lying off of the manifold. Contrast this to RR and EN which choose Π to be the identity matrix. Thus, RR and EN do not fully take the manifold structure of the problem into consideration.

Stated in another way, Π is a projection matrix which is chosen to penalize the components of β which are perpendicular to the manifold. The cost function we are minimizing has the term $\|\Pi\beta\|_2^2$, and this term is large if β has components perpendicular to the manifold. Components of β parallel to the manifold are not penalized because the projection onto these directions is zero.

Since we do not know the manifold a priori, we must learn the manifold from the sample local covariance matrix $\hat{C}_n^{22}$. We do this by looking at the principal components of $\hat{C}_n^{22}$, and so our estimators are very closely related to PCR. Suppose that we do an eigenvalue decomposition of $\hat{C}_n^{22}$:

$$(4.2) \qquad \hat{C}_n^{22} = [\hat{U}^R \quad \hat{U}^N] \operatorname{diag}(\lambda^1, \ldots, \lambda^P) [\hat{U}^R \quad \hat{U}^N]',$$

---

[3]We specifically use the intuition that the exterior derivative lies in the cotangent space of the manifold, and this statement can be mathematically written as: $d_x f \in T_p^* \mathcal{M}$.

where $\hat{U}^R \in \mathbb{R}^{p \times d}$, $\hat{U}^N \in \mathbb{R}^{p \times (p-d)}$ and $\lambda^1 \geq \lambda^2 \geq \cdots \geq \lambda^p$. Note that the eigenvalue decomposition always exists because $\hat{C}_n^{22}$ is symmetric. The estimate of the manifold is given by the d most relevant principal components, and the remaining principal components are perpendicular to the estimated manifold.

Because we want the projection matrix Π to project β onto the directions perpendicular to the estimated manifold, we define the following projection matrices

(4.3)
$$\hat{\Pi} \triangleq \hat{U}^N \hat{U}^{N'},$$
$$\hat{P} \triangleq \operatorname{diag}(0, \hat{\Pi}).$$

The choice of d is a tunable parameter that is similar to the choice in PCR. These matrices act as a regularizer because d can always be chosen to ensure that $\operatorname{rank}(\hat{C}_n^{22} + \lambda \hat{\Pi}_n) = p$. Furthermore, we have the following theorem regarding the full regularizing matrix $\hat{P}$:

THEOREM 4.1 [Lemma A.2, part (d)]. *Under the assumptions given in Section 2, the following holds with probability one*:

(4.4)
$$\operatorname{rank}(\hat{C}_n + \lambda_n \hat{P}_n / n h^{d+2}) = p + 1.$$

Our estimators can perform better than PCR because of a subtle difference. PCR requires that the estimate lies on exactly the first d most relevant principal components; however, our estimator only penalizes for deviation from the d most relevant principal components. This is advantageous because in practice d is not known exactly and because the principal components used are estimates of the true principal components. Thus, our regularization is more robust to errors in the estimates of the principal components. Also, our new regularization allows us to easily add additional lasso-type regularization to potentially improve the estimation. PCR cannot be easily extended to have lasso-type regularization.

We denote the function value at X_0 as $f|_{x_0}$. Also, we denote the exterior derivative of $f(\cdot)$ at X_0 as $d_x f|_{x_0}$. Then, the true regression coefficients are denoted by the vector

(4.5)
$$\beta' = [\, f|_{x_0} \quad d_x f|_{x_0} \,].$$

The nonparametric exterior derivative estimator (NEDE) is given by

(4.6)
$$\hat{\beta} = \arg\min_{\tilde{\beta}} \{ h^p \| W_{x_0}^{1/2} (Y - X_{x_0} \tilde{\beta}) \|_2^2 + \lambda_n \| \hat{P}_n \cdot \tilde{\beta} \|_2^2 \},$$

where $\hat{P}_n$ is defined using (4.3) with $\hat{C}_n$. We can also define a nonparametric adaptive lasso exterior derivative estimator (NALEDE) as

(4.7)
$$\hat{\beta} = \arg\min_{\tilde{\beta}} \left\{ h^p \| W_{x_0}^{1/2} (Y - X_{x_0} \tilde{\beta}) \|_2^2 + \lambda_n \| \hat{P}_n \cdot \tilde{\beta} \|_2^2 + \mu_n \sum_{j=1}^{p} \frac{1}{\hat{w}_j^{\gamma}} |\tilde{\beta}_j| \right\},$$

where $\hat{P}_n$ is define in (4.3) using $\hat{C}_n$, $\hat{w}$ is the solution to (4.6), and $\gamma > 0$.

Our estimators have nice statistical properties, as the following theorem shows.

THEOREM 4.2. *If the assumptions in Section* 2 *hold, then the NEDE* (4.6) *and NALEDE* (4.7) *estimators are consistent and asymptotically normal*:

$$H^{1/2}(\hat{\beta} - \beta) \xrightarrow{d} C^{\dagger}\mathcal{N}(B', \sigma^2 V),$$

where B and V are, respectively, given in (A.3) *and* (A.4), *and $C^{\dagger}$ denotes the Moore–Penrose psuedoinverse of C. Furthermore, we asymptotically have that $\hat{\beta}' \in \mathbb{R} \times T_P^*\mathcal{M}$. The NALEDE* (4.7) *estimator has the additional feature that*

$$\mathbb{P}\big(\mathrm{sign}(\hat{\beta}) = \mathrm{sign}(\beta)\big) \to 1.$$

Note that the asymptotic normality is biased because of the bias typical in nonparametric regression. This bias is seen in both the NEDE (4.6) and NALEDE (4.7) estimators, but examining B one sees that the bias only exists for the function estimate $\hat{f}_{x_0}$ and not for the exterior derivative estimate $d_x \hat{f}|_{x_0}$. This bias occurs because we are choosing h to converge at the optimal rate. If we were to choose h at a faster rate, then there would be no asymptotic bias, but the estimates would converge at a slower rate.

It is worth noting that the rate of convergence in Theorem 4.2 has an exponential dependence on the dimension of the manifold d, and our particular rates are due to the assumption of the existence of three derivatives. As is common with local regression, it is possible to improve the rate of convergence by using local polynomial regression which assumes the existence of higher-order derivatives [8, 46]. However, the general form of local polynomial regression on manifolds would require the choice of a particular chart $\phi(\cdot)$ and domain $\mathcal{U}$. Local linear regression on manifolds is unique in that one does not have to pick a chart and domain.

As a last note, recall that the rate of convergence in Theorem 4.2 depends on the dimension of the manifold d and does not depend on the dimension p of the ambient space. We might mistakenly think that this means that the estimator converges in the "large p, small n" scenario, but without additional assumptions these results are only valid for when p grows more slowly than n. Analogous to other "large p, small n" settings, if we assume sparsity then we can achieve faster rates of convergence, which is the subject of the next section.

5. Large p, small n. We consider extensions of our estimators to the "large p, small n" setting. The key difference in this case is the need to regularize the covariance matrix. Our NEDE (4.6) and NALEDE (4.7) estimators use the eigenvectors of the sample covariance matrix, and it is known [7, 33] that the sample covariance matrix is poorly behaved in the "large p, small n" setting. To ensure the sample

eigenvectors are consistent estimates, we must use some form of covariance regularization [7, 33, 61].

We use the regularization technique used in [7] for ease of analysis and because other regularization techniques [33, 61] do not work when the true covariance matrix is singular. The scheme in [7] works by thresholding the covariance matrix, which leads to consistent estimates as long as the threshold is correctly chosen. We define the thresholding operator as

$$T_t(m) = m\mathbb{1}(|m| > t),$$

and by abuse of notation $T_t(M)$ is $T_t(\cdot)$ applied to each element of M.

The setup and assumptions are nearly identical to those of the fixed p case described in Section 2. The primary differences are that (a) d, p, n increase at different rates toward infinity, and (b) there is some amount of sparsity in the manifold and in the function. The population parameters C_n, analogous to (3.2), are functions of n and are defined in nearly the same manner, except with $[C_n^{21}]_k = F(0)/2 \int_{\mathbb{R}^d} K(d_u\phi \cdot u)\partial_{ij}\phi^k u^i u^j \, du$. Their estimates are now defined

$$\hat{C}_n = H^{-1} X'_{x_0} W_{x_0} X_{x_0};$$

compare this to (3.1). Just as C_n can be interpreted as a local covariance matrix, we now define a local cross-covariance matrix:

$$R_n = \begin{bmatrix} R_n^1 \\ R_n^2 \end{bmatrix} = \begin{bmatrix} C_n^{11} \cdot f|_{x_0} \\ C_n^{21} \cdot f|_{x_0} + C_n^{22} \cdot d_x f|_{x_0} \end{bmatrix},$$

and the estimates are given by

$$\hat{R}_n = H^{-1/2} X'_{x_0} W_{x_0} Y.$$

For the sake of brevity, we summarize the other the differences from Section 2. The following things are different:

1. The manifold $\mathcal{M}_n$, local chart $\phi_n(\cdot)$, manifold dimension d_n, number of predictors p_n, and density function $F_n(\cdot)$ are all functions of n. We drop the subscript n when it is clear from the context. These objects are defined in the same manner as in Section 2, and we additionally assume that the density $F(\cdot)$ is Lipschitz continuous.
2. The asymptotic rates for d, p, n are given by $d = o(\log n)$,

$$h = o\big((c_n^4 n / \log p)^{-1/(4+d)}\big),$$

$$c_n \sqrt{\frac{\log p}{nh^d}} = o(1);$$

where c_n is a measure of sparsity that describes the number of nonzero entries in covariance matrices, exterior derivative, etc.

3. The kernel $K(\cdot)$ has finite support and is Lipschitz continuous, which implies that

$$K\left(\frac{\phi(hu) - \phi(0)}{h}\right) = K(d_u\phi \cdot u) = 0,$$

for $u \notin \mathcal{B}_{0,\Omega}^{d_n}$. Contrast this to the second assumption in Section 2.

4. The local chart $\phi_n(\cdot)$, function $f_n(\cdot)$ and local (cross-)covariance matrices C_n, R_n satisfy the following sparsity conditions:

(5.1)
$$\sum_{k=1}^{p} \mathbb{1}(Q^k \neq 0) \leq c_n \quad \text{and} \quad |Q^k| \leq M,$$

for (derivatives of the local chart; the index k denotes the kth component of the vector-valued ϕ) $Q^k = \partial_i\phi^k, \partial_{ij}\phi^k, \partial_{ijm}\phi^k, \partial_{ijmn}\phi^k$; for (derivatives of the function) $Q^k = [d_x f]_k, \partial_{ik}(f \circ \phi), \partial_{ijk}(f \circ \phi)$; and for (local covariance matrices) $Q^k = [C_n]_{ik}, [R_n]_{ik}$.

5. The smallest, nonzero singular value of the local covariance matrix is bounded. That is, there exists $\varepsilon > 0$ such that

(5.2)
$$\inf_{n>0}\left(\inf_{\sigma(\cdot)>0} \sigma(C_n)\right) > \varepsilon.$$

This condition ensures that the regularized inverse of the local covariance matrix is well defined in the limit; otherwise we could have a situation with ever-decreasing nonzero singular values.

6. The Tikhonov-type regularization parameter λ_n and the lasso-type regularization parameter μ_n have the following asymptotic rates:

$$\lambda_n = O_p\left(\sqrt{c_n}\left(\frac{nh^d}{\log p}\right)^{1/4}\right),$$

$$\mu_n c_n^{3/2}\left(\frac{\log p}{nh^d}\right)^{1/4} = o(1),$$

$$\mu_n\left(c_n^{3/2}\left(\frac{\log p}{nh^d}\right)^{1/4}\right)^{1-\gamma} \to \infty.$$

7. The threshold which regularizes the local sample covariance matrix is given by

(5.3)
$$t_n = K\sqrt{\frac{\log p}{n}},$$

where $\frac{\log p}{n} = o(1)$. This regularization will make the regression estimator consistent in the "large p, small n" case.

5.1. *Manifold regularization.* The idea is to regularize the local sample co-variance matrix by thresholding. If the true, local covariance matrix is sparse, this regularization will give consistent estimates. This is formalized by the following theorem.

THEOREM 5.1. *If the assumptions given in Section 5 are satisfied, then*

$$\|T_t(\hat{C}_n) - C_n\| = O_p(c_n\sqrt{\log p/nh^d}),$$

$$\|T_t(\hat{R}_n) - R_n\| = O_p(c_n\sqrt{\log p/nh^d}).$$

Once we have consistent estimates of the true, local covariance matrix, we can simply apply our manifold regularization scheme described in Section 4. The non-parametric exterior derivative estimator for the "large p, small n" case (NEDEP) is given by

$$(5.4) \qquad \hat{\beta}_n = \arg\min_{\tilde{\beta}} \|(T_t(\hat{C}_n) + \lambda_n \hat{P}_n)\tilde{\beta} - T_t(\hat{R}_n)\|_2^2,$$

where $\hat{P}_n$ is as defined in (4.3) except using $T_t(\hat{C}_n^{22})$. The nonparametric adaptive lasso exterior derivative estimator for the "large p, small n" case (NALEDEP) is given by

$$\hat{\beta} = \arg\min_{\tilde{\beta}} \arg\min_{\tilde{\beta}} \|(T_t(\hat{C}_n) + \lambda_n \hat{P}_n)\tilde{\beta} - T_t(\hat{R}_n)\|_2^2$$

$$(5.5)$$

$$+ \mu_n \sum_{j=1}^{p} \frac{1}{\hat{w}_j^{\gamma}} |\tilde{\beta}_j|,$$

where $\hat{P}_n$ is as defined in (4.3) except using $T_t(\hat{C}_n^{22})$ and $\hat{w}$ is the solution to (5.4). These estimators have nice statistical properties.

THEOREM 5.2. *If the assumptions given in Section 5 are satisfied, then the NALEDE (5.4) and NALEDEP (5.5) estimators are consistent:*

$$\|\hat{\beta} - \beta\|_2 = O_p\left(c_n^{3/2}\left(\frac{\log p}{nh^d}\right)^{1/4}\right).$$

The NALEDEP (5.5) estimator is also sign consistent:

$$\mathbb{P}(\text{sign}(\hat{\beta}) = \text{sign}(\beta)) \to 1.$$

We do not give a proof of this theorem, because it uses essentially the same argument as Theorem 5.3. One minor difference is that the proof uses our Theorem 5.1 instead of Theorem 1 from [7].

5.2. *Linear case.*   Our estimators admit simple extensions in the special case where predictors lie on a global, linear manifold and the response variable is a linear function of the predictors. We specifically consider the errors-in-variables situation with manifold structure in order to present our formal results, because: in principle, our estimators provide no improvements in the linear manifold case over existing methods when there are no errors-in-variables. In practice, our estimators sometimes provide an improvement in this case. Furthermore, our estimators provide another solution to the identifiability problem [19]; the exterior derivative is the unique set of regression coefficients because the predictors are only sampled in directions parallel to the manifold, and there is no derivative information about the response variable in directions perpendicular to the manifold.

Suppose there are n data points and p predictors, and the dimension of the global, linear manifold is d. We assume that d, n, p increase to infinity, and leaving d fixed is a special case of our results. We consider a linear model $\eta = \Xi\overline{\beta}$, where $\eta \in \mathbb{R}^{n \times 1}$ is a vector of function values, $\Xi \in \mathbb{R}^{n \times p}$ is a matrix of predictors, and $\overline{\beta} \in \mathbb{R}^p$ is a vector.

The Ξ are distributed according to the covariance matrix Σ_ξ, which is also a singular design matrix in this case. The exterior derivative of this linear function is given by $\beta = P_{\Sigma_\xi}\overline{\beta}$, where P_{Σ_ξ} is the projection matrix that projects onto the range space of Σ_ξ. We make noisy measurements of η and Ξ:

$$X = \Xi + \nu,$$
$$Y = \eta + \varepsilon.$$

The noise ν and ε are independent of each other, and each component of ν is independent and identically distributed with mean 0 and variance σ_ν^2. Similarly, each component of ε is independent and identically distributed with mean 0 and variance σ^2. In this setup, the variance σ_ν^2 is identifiable [28, 31], and an alternative that works well in practice for low noise situations is to set this quantity to zero.

Our setup with errors-in-variables is different from the setup of existing tools [10, 38], but it is important because in practice, many of the near-collinearities might be true collinearities that have been perturbed by noise. Several formulations explicitly introduce noise into the model [11, 14, 28, 31, 41]. We choose the setting of [28, 31], because the noise in the predictors is identifiable in this situation.

The exterior derivative estimator for the "large p, small n" case (EDEP) is given by

$$(5.6) \qquad \hat{\beta} = \arg\min_{\tilde{\beta}} \left\| (T_t(X'X/n) - \sigma_\nu^2\mathbb{I} + \lambda_n \hat{P}_n)\tilde{\beta} - T_t(X'Y/n) \right\|_2^2,$$

where $\hat{P}_n$ is as defined in (4.3) except applied to $\hat{C}_n^{22} = T_t(X'X/n) - \sigma_\nu^2\mathbb{I}$. This is essentially the NEDEP estimator, except the weighting matrix is taken to be the identity matrix and there are additional terms to deal with errors-in-variables.

We can also define an adaptive lasso version of our estimator. The adaptive lasso exterior derivative estimator for the "large p, small n" case (ALEDEP) is given by

$$\hat{\beta} = \arg\min_{\tilde{\beta}} \big\| (T_t(X'X/n) - \sigma_v^2 \mathbb{I} + \lambda_n \hat{P}_n)\tilde{\beta} - T_t(X'Y/n) \big\|_2^2$$

(5.7)

$$+ \mu_n \sum_{j=1}^{p} \frac{1}{\hat{w}_j^{\gamma}} |\tilde{\beta}_j|,$$

where $\hat{P}_n$ is as defined in (4.3) except applied to $\hat{C}_n^{22} = T_t(X'X) - \sigma_v^2 \mathbb{I}$ and $\hat{w}$ is the solution to (5.6). We can analogously define the EDE and ALEDE estimators which are the EDEP and ALEDEP estimators without any thresholding. In practice, setting the value of the σ_v^2 term equal to zero seems to work well with actual data sets.

The technical conditions we make are essentially the same as those for the case of the local, nonlinear manifold. The primary difference is that we ask that the conditions in Section 5 hold globally, instead of locally. This also means that we do not use any kernels to localize the estimators, and the W matrix in the estimators is simply the identity matrix. If the theoretical rates for the regularization and threshold parameters are compatibility redefined, then we can show that these estimators have nice statistical properties.

THEOREM 5.3. *If the assumptions in Sections 5 and 5.2 hold, then the EDEP (5.6) and ALEDEP (5.7) estimators are consistent. They asymptotically converge at the following rate*:

$$\|\hat{\beta} - \beta\|_2 = O_p\Big(c_n^{3/2} \Big(\frac{\log p}{n}\Big)^{1/4} \Big).$$

The ALEDE (5.7) estimator is also sign consistent:

(5.8) $$\mathbb{P}\big(\mathrm{sign}(\hat{\beta}) = \mathrm{sign}(\beta)\big) \to 1.$$

Our theoretical rate of convergence is slower than that of other techniques [10, 38] because we have applied our technique for local estimation to global estimation, and we have not fully exploited the setup of the global case. However, we do get faster rates of convergence in our global case versus our local case. Furthermore, our model has errors-in-variables, while the model used in other techniques [10, 38] assumes that the predictors are measured with no noise. Applying the various techniques to both real and simulated data shows that our estimators perform comparably to or better than existing techniques. It is not clear if the rates of convergence for the existing techniques [10, 38] would be slower if there were errors-in-variables, and this would require additional analysis.

6. Estimation with data. Applying our estimators in practice requires careful usage. The NEDE estimator requires the choice of two tuning parameters, while the NALEDE and NEDEP estimators require choosing three; the NALEDEP estimator requires choosing four. The extra tuning parameters—in comparison to existing techniques like MP or RR—make our method prone to over-fitting. This makes it crucial to select the parameters using methods, such as cross-validation or bootstrapping, that protect against overfitting. It is also important to select from a small number of different parameter-values to protect against overfitting caused by issues related to multiple-hypothesis testing [42, 44, 45].

Bootstrapping is one good choice for parameter selection with our estimators [7, 9, 48–50]. Additionally, we suggest selecting parameters in a sequential manner; this is to reduce overfitting caused by testing too many models [42, 44, 45]. Another benefit of this approach is that it simplifies the parameter selection into a set of one-dimensional parameter searches—greatly reducing the computational complexity of our estimators. For instance, we first select the Tikhonov-regularization parameter λ for RR. Using the same λ value, we pick the dimension d for the NEDE estimator. The prior values of λ and d are used to pick the lasso-regularization parameter μ for the NALEDE estimator.

MATLAB implementations of both related estimators and our estimators can be found online.[4] The lasso-type regressions were computed using the coordinate descent algorithm [17, 57], and we used the "improved kernel PLS by Dayal" code given in [2] to do the PLS regression. The increased computational burden of our estimators, as compared to existing estimators, is reasonable because of: improved estimation in some cases, easy parallelization, and computational times of a few seconds to several minutes on a general desktop for moderate values of p.

7. Numerical examples. We provide three numerical examples: the first two examples use simulated data, and the third example uses real data. In the examples with simulated data, we study the estimation accuracy of various estimators as the amount of noise and number of data points vary. The third example uses the Isomap face data[5] used in [53]. In the example, we do a regression to estimate the pose of a face based on an image.

For examples involving linear manifolds and functions, we compare our estimators with popular methods. The exterior derivative is locally defined, but in the linear case it is identical at every point—allowing us to do the regression globally. This is in contrast to the example with a nonlinear manifold and function where we pick a point to do the regression at. Though MP, PLS, PCR, RR and EN are typically thought of as global methods, we can use these estimators to do local, nonparametric estimation by posing the problem as a weighted, linear regression

---

[4]http://hybrid.eecs.berkeley.edu/~NEDE/EDE_Code.zip.
[5]http://isomap.stanford.edu/datasets.html.

which can then be solved using either our or existing estimators. As a note, the MP and OLS estimators are equivalent in the examples we consider.

Some of the examples involve errors-in-variables, and this suggests that we should use an estimator that explicitly takes this structure into account. We compared these methods with Total Least Squares (TLS) [27] which does exactly this. TLS performed poorly with both the simulated data and experimental data, and this is expected because standard TLS is known to perform poorly in the presence of collinearities [27]. TLS performed comparably to or worse than OLS/MP, and so the results are not included.

Based on the numerical examples, it seems that the improvement in estimation error of our estimators is mainly due to the Tikhonov-type regularization, with lasso-type regularization providing additional benefit. Thresholding the covariance matrices did not make significant improvements, partly because bootstrap has difficulty with picking the thresholding parameter. Improvements may be possible by refining the parameter selection method or by changes to the estimator. We also observed the well-known tendency of lasso to overestimate the number of nonzero coefficients [39]; using stability selection [39] to select the lasso parameter would likely lead to better results.

7.1. *Simulated data.* We simulate data for two different models and use this to compare different estimators. One model is linear, and we do global estimation in this case. The other model is nonlinear, and hence we do local estimation in this case. In both models, there are p predictors and the dimension of the manifold is $d = \mathrm{round}(\frac{3}{4}p)$. The predictors ξ and response η are measured with noise:

$$x = \xi + \mathcal{N}(0, \sigma_v^2),$$

$$y = \eta + \mathcal{N}(0, \sigma^2).$$

And for notational convenience, let $q = \mathrm{round}(\frac{1}{2}p)$. Define the matrix

$$F_{ij} = \begin{cases} 0.3^{|i-j|}, & \text{if } 1 \le i, j \le d, \\ 0.3, & \text{if } d+1 \le i \le p \wedge j = q+i-d, \\ 0.3, & \text{if } d+1 \le i \le p \wedge j = q+i+1-d, \\ 0, & \text{o.w.} \end{cases}$$

The two models are given by:

1. Linear model: the predictors are distributed $\xi = \mathcal{N}(0, FF')$, and the function is

(7.1) $$\eta = f(\xi) = 1 + \sum_{\substack{i=1 \\ i \text{ is odd}}}^{q} \xi_i.$$

If $w = [1 \ 0 \ 1 \ \cdots]$ is a vector with ones in the odd index-positions and zeros elsewhere, then the exterior derivative of this linear function at every point on the manifold is given by the projection of w onto the range space of the matrix F.

2. Nonlinear model: the predictors are distributed $\xi = \sin(\mathcal{N}(0, FF'))$, and the function is

$$(7.2) \qquad \eta = f(\xi) = 1 + \sum_{\substack{i=1 \\ i \text{ is odd}}}^{q} \sin(\xi_i).$$

We are interested in local regression about the point $x_0 = [0 \;\cdots\; 0]$. If $w = [1 \; 0 \; 1 \;\cdots]$ is a vector with ones in the odd index-positions and zeros elsewhere, then the exterior derivative of this nonlinear function at the origin is given by the projection of w onto the range space of the matrix F.

Table 1 shows the average square-loss estimation error $\|\hat{\beta} - \beta\|_2^2$ for different estimators using data generated by the linear model and nonlinear model given above, over different noise variances and number of data points n. We conducted 100 replications of generating data and doing a regression, and this helped to provide standard deviations of square-loss estimation error to show the variability of

TABLE 1
Averages and standard deviations over 100 *replications of square-loss estimation error for different estimators using data generated by the linear model and nonlinear model given in Section* 7.1, *over different noise variances and number of data points n*

	$n = 10$		$n = 100$		$n = 1000$	
	Linear model: $\sigma_v^2 = 0.01, \sigma^2 = 1.00$					
OLS/MP	3.741	(2.476)	6.744	(3.602)	0.588	(0.368)
RR	2.523	(1.020)	0.369	(0.193)	0.167	(0.086)
EN	2.562	(1.054)	0.117	(0.197)	0.017	(0.008)
PLS	2.428	(0.536)	0.501	(0.207)	0.031	(0.013)
PCR	3.391	(0.793)	1.629	(0.144)	1.583	(0.047)
EDE	2.528	(1.030)	0.367	(0.185)	0.166	(0.086)
ALEDE	2.564	(1.061)	0.111	(0.177)	0.015	(0.006)
EDEP	2.527	(1.025)	0.370	(0.184)	0.167	(0.085)
ALEDEP	2.563	(1.057)	0.111	(0.177)	0.015	(0.006)
	Linear model: $\sigma_v^2 = 0.10, \sigma^2 = 1.00$					
OLS/MP	3.629	(1.488)	1.259	(0.598)	0.173	(0.050)
RR	2.717	(0.847)	0.892	(0.351)	0.260	(0.042)
EN	2.783	(0.895)	0.661	(0.523)	0.064	(0.013)
PLS	2.588	(0.566)	0.740	(0.252)	0.138	(0.037)
PCR	3.425	(0.747)	1.645	(0.144)	1.569	(0.047)
EDE	2.716	(0.846)	0.840	(0.305)	0.255	(0.042)
ALEDE	2.782	(0.893)	0.590	(0.459)	0.050	(0.013)
EDEP	2.720	(0.849)	0.841	(0.305)	0.255	(0.042)
ALEDEP	2.784	(0.895)	0.590	(0.459)	0.050	(0.013)

TABLE 1
(*Continued*)

	$n = 20$		$n = 100$		$n = 1000$	
Nonlinear model: $\sigma_\nu^2 = 0.01, \sigma^2 = 1.00$						
OLS/MP	657.7	(1842)	3.797	(2.172)	0.367	(0.223)
RR	2.188	(0.715)	1.160	(0.265)	0.300	(0.132)
EN	2.140	(0.768)	1.083	(0.341)	0.162	(0.081)
PLS	2.102	(0.757)	0.995	(0.256)	0.172	(0.120)
PCR	2.731	(0.439)	1.679	(0.244)	0.123	(0.043)
NEDE	2.184	(0.716)	1.104	(0.269)	0.288	(0.118)
NALEDE	2.136	(0.769)	1.009	(0.357)	0.144	(0.067)
NEDEP	2.184	(0.716)	1.103	(0.269)	0.288	(0.118)
NALEDEP	2.136	(0.769)	1.008	(0.357)	0.144	(0.067)
Nonlinear model: $\sigma_\nu^2 = 0.1, \sigma^2 = 1.00$						
OLS/MP	147.5	(338.3)	1.843	(0.975)	0.473	(0.110)
RR	2.693	(1.793)	1.260	(0.369)	0.672	(0.139)
EN	2.698	(1.927)	1.168	(0.455)	0.472	(0.197)
PLS	2.385	(0.629)	1.210	(0.318)	0.768	(0.133)
PCR	2.767	(0.450)	1.766	(0.293)	0.554	(0.348)
NEDE	2.694	(1.793)	1.233	(0.352)	0.641	(0.119)
NALEDE	2.702	(1.925)	1.126	(0.445)	0.407	(0.130)
NEDEP	2.693	(1.794)	1.231	(0.352)	0.641	(0.119)
NALEDEP	2.702	(1.926)	1.124	(0.445)	0.407	(0.130)

the estimators. Table 2 shows computation times in seconds for the different estimators; it shows that our estimators require more computation, but the computation time is still reasonable. The computation time does not depend on the noise level, and so we have averaged the computation times over the $100 + 100$ replications for $\sigma_\nu^2 = 0.01, 0.10$.

One curious phenomenon observed is that the estimation error goes down in some cases as the error variance of the predictors σ_ν^2 increases. To understand why, consider the sample covariance matrix in the linear case $\hat{S} = X'X/n$ with population parameter $S = FF' + \sigma_\nu^2 \mathbb{I}$. Heuristically, the OLS estimate will tend to $(FF' + \sigma_\nu^2 \mathbb{I})^{-1} X'Y/n$, and the error in the predictors actually acts as the Tikhonov-type regularization found in RR, with lower levels of noise leading to less regularization.

The results indicate that our estimators are not significantly more variable than existing ones, and our estimators perform competitively against existing estimators. Though our estimators are closely related to PCR, RR and EN, our estimators performed comparably to or better than these estimators. PLS also did quite well, and our estimators did better than PLS in some cases. Increasing the noise in the

TABLE 2
Averages and standard deviations over 200 *replications of computation times in seconds*
for the different estimators using data generated by the linear model and
nonlinear model given in Section 7.1

	$n = 10$		$n = 100$		$n = 1000$	
			Linear model			
OLS/MP	0.001	(0.000)	0.001	(0.000)	0.001	(0.000)
RR	0.055	(0.005)	0.063	(0.005)	0.110	(0.004)
EN	1.739	(1.325)	0.292	(0.032)	0.387	(0.042)
PLS	1.631	(0.029)	1.716	(0.077)	1.925	(0.051)
PCR	0.185	(0.006)	0.192	(0.010)	0.253	(0.007)
EDE	0.239	(0.007)	0.255	(0.016)	0.367	(0.021)
ALEDE	2.071	(1.363)	0.715	(0.043)	0.912	(0.059)
EDEP	0.411	(0.011)	0.523	(0.019)	0.731	(0.052)
ALEDEP	2.248	(1.366)	0.990	(0.042)	1.283	(0.083)
	$n = 20$		$n = 100$		$n = 1000$	
			Nonlinear model			
OLS/MP	0.001	(0.000)	0.001	(0.000)	0.001	(0.000)
RR	0.359	(0.039)	0.470	(0.028)	0.938	(0.040)
EN	0.916	(0.433)	1.005	(0.058)	1.917	(0.269)
PLS	1.980	(0.130)	2.114	(0.072)	2.879	(0.162)
PCR	1.014	(0.062)	1.118	(0.038)	1.911	(0.166)
NEDE	0.840	(0.078)	1.066	(0.058)	2.031	(0.092)
NALEDE	2.175	(0.594)	1.843	(0.090)	3.381	(0.401)
NEDEP	1.403	(0.120)	1.726	(0.087)	2.277	(0.175)
NALEDEP	2.671	(0.647)	2.513	(0.122)	4.631	(0.397)

predictors did not seem to significantly affect the qualitative performance of the estimators, except for OLS as explained above.

In Section 5.2, we discussed how the converge rate of our linear estimators is of order $n^{-1/4}$ which is in contrast to the typical convergence rate of $n^{-1/2}$ for lasso-type regression [38]. We believe that this theoretical discrepancy is because our model has errors-in-variables while the standard model used in lasso-type regression does not [38]. These theoretical differences do not seem significant in practice. As seen in Table 1, our estimators can be competitive with existing lasso-type regression.

7.2. *Isomap face data*. The Isomap face data[6] from [53] consists of images of an artificial face. The images are labeled with and vary depending upon three

---

[6]http://isomap.stanford.edu/datasets.html.

FIG. 2. *Sample images from the Isomap face data* [53]. *The images are labeled with and vary depending upon three variables*: *illumination-, horizontal- and vertical-orientation.*

variables: illumination-, horizontal- and vertical-orientation; sample images taken from this data set can be seen in Figure 2. Three-dimensional images of the face would form a three-dimensional manifold (each dimension corresponding to a variable), but this data set consists of two-dimensional projections of the face. Intuitively, a limited number of additional variables are needed for different views of the face (e.g., front, profile, etc.). This intuition is confirmed by dimensionality estimators which estimate that the two-dimensional images form a low-dimensional manifold [35].

To compare the different estimators, we did 100 replications of the following experiment: we randomly split the data ($n = 698$ data points) into a training set $n_t = 688$ and validation set $n_v = 10$, and then we used the training set to estimate the horizontal pose angle of images in the validation set. Since we are doing local linear estimation, the estimate for each image requires its own regression. The number of predictors p is large in this case because each data point X_i is a two-dimensional image. Estimation when p is large is computationally slow, and so we chose a small validation set size to ensure that the experiments completed in a reasonable amount of time. Replicating this experiment 100 times helps to prevent spurious results due to having a small validation set.

To speed up the computations further, we scaled the images from 64×64 pixels to 7×7 pixels in size. This is a justifiable approach because the images form a low-dimensional manifold, and so this resizing of the images does not lead to a loss in predictor information [56]. This leads to significantly faster computations, because this process reduces the number of predictors from $p = 4096$ to $p = 49$. In practice, our choice of $p = 49$ gives predictions that deviate from the true horizontal pose angle of images (which uniformly ranges between -75 to 75 degrees) in the validation set by a root-mean-squared error of three degrees or less.

Table 3 gives the prediction error of the models generated by different estimators on the validation set. The specific quantity provided is

$$(7.3) \qquad \sum_{X_i \in \mathcal{V}} \|\hat{\beta}^0(X_i) - Y_i\|_2^2 / n_v,$$

where $\mathcal{V}$ is the set of predictors in the validation set, $\hat{\beta}^0(X_i)$ is the first component of the estimated regression coefficients computed about the point $X_0 = X_i$, and Y_i is the corresponding horizontal pose angle of the image X_i. The regression is computed using only data taken from the training set. The results from this real data set shows that our estimators can provide improvements over existing tools,

TABLE 3
Averages and standard deviations over 100 *replications of validation set prediction error for different estimators using the Isomap face data* [53]. *The average computation time and its standard deviation is given in seconds*, *and it gives the time to estimate the horizontal pose angle of a single image*

	Prediction error		Computation time	
OLS/MP	5.276	(11.06)	0.002	(0.001)
RR	5.286	(6.156)	2.052	(0.166)
EN	5.168	(6.112)	17.79	(5.235)
PLS	10.12	(14.84)	16.22	(0.841)
PCR	5.813	(7.617)	8.877	(0.609)
NEDE	4.523	(4.926)	4.740	(0.347)
NALEDE	4.409	(4.889)	22.94	(5.349)
NEDEP	4.527	(4.925)	7.777	(0.511)
NALEDEP	4.406	(4.900)	25.77	(5.221)

because our estimators have the lowest prediction errors. Table 3 also provides the computation times for estimating the horizontal pose, and it again shows that our estimators require more computation but not by an excessively larger amount.

8. Conclusion. By interpreting collinearity as predictors on a lower-dimensional manifold, we have developed a new regularization, which has connections to PCR and RR, for linear regression and local linear regression. This viewpoint also allows us interpret the regression coefficients as estimates of the exterior derivative. We proved the consistency of our estimators in both the classical case and the "large p, small n" case and this is useful from a theoretical standpoint.

We provided numerical examples using simulated and real data which show that our estimators can provide improvements over existing estimators in estimation and prediction error. Though our estimators provide modest improvements over existing ones, these improvements are consistent over the different examples. Specifically, the Tikhonov-type and lasso-type regularizations provided improvements, and the thresholding regularization did not provide major improvements. This is not to say that thresholding is not a good regularization, because as we showed: from a theoretical standpoint, thresholding does provide consistency in the "large p, small n" situation. This leaves open the possibility of future work on how to best select this thresholding parameter value.

There is additional future work possible on extending our set of estimators. There is some benefit provided by shrinkage from the Tikhonov-type regularization which is independent of the manifold structure. Exploring more fully the relationship between manifold structure and shrinkage will likely lead to improved estimators.

APPENDIX

In this section, we provide the proofs of our theorems. We also give a few lemmas, which are needed for the proofs, that were not stated in the main text.

LEMMA A.1. *If the assumptions in Section 2 hold, then*:

(a) $\|\mathbb{E}(\hat{C}_n^{22}) - C^{22}\|_2^2 = O(h^4)$;

(b) $\|\mathbb{E}[(\hat{C}_n^{22} - C^{22})(\hat{C}_n^{22} - C^{22})']\|_2^2 = O(1/nh^d)$;

(c) $\hat{C}_n \xrightarrow{P} C$.

PROOF. This proof follows the techniques of [8, 46]. We first prove part (c). Note that

$$\hat{C}_n^{11} = \frac{1}{nh^{d-p}} \sum_{i=1}^n K_h(x_i - x_0),$$

and consider its expectation

$$\mathbb{E}(\hat{C}_n^{11}) = \mathbb{E}\left(\frac{1}{h^{d-p}} K_h(x_i - x_0)\mathbb{1}\left(X \in (\mathcal{B}_{x,h^{1-\varepsilon}}^p)\right)\right)$$
$$+ \mathbb{E}\left(\frac{h^p}{h^d} K_h(x_i - x_0)\mathbb{1}\left(X \in (\mathcal{B}_{x,h^{1-\varepsilon}}^p)^c\right)\right)$$
$$= \int_{B_{0,h^{1-\varepsilon}}^d} \frac{1}{h^d} K\left(\frac{\phi(z) - \phi(0)}{h}\right) F(z)\,dz + o(h^{2+p})$$
$$= \int_{\mathbb{R}^n} K(d_u\phi \cdot u) F(0)\,du + O(h^2),$$

where we have used the assumption that $K(\cdot)$ is an even function, $K'(\cdot)$ is an odd function, and $K''(\cdot)$ is an even function.

A similar calculation shows that for

$$\hat{C}_n^{21} = \frac{1}{nh^{d+1-p}} \sum_{i=1}^n K_h(x_i - x_0)(x_i - x_0),$$

we have that the expectation is

$$\mathbb{E}(\hat{C}_n^{21}) = O(h) = o(1).$$

And, a similar calculation shows that for

$$\hat{C}_n^{22} = \frac{1}{nh^{d+2-p}} \sum_{i=1}^n K_h(x_i - x_0)(x_i - x_0)(x_i - x_0)',$$

we have that the expectation is

$$\mathbb{E}(\hat{C}_n^{22}) = F(0)\,d_u\phi \cdot \left[\int_{\mathbb{R}^d} K(d_u\phi(0) \cdot u)uu'\,du\right] \cdot d_u\phi' + O(h^2).$$

The result in part (c) follows from the weak law of large numbers. The last calculation also proves part (a).

Next, we prove part (b). For notational simplicity, let

$$T_i = K_h(x_i - x_0)(x_i - x_0)(x_i - x_0)'.$$

The variance is

$$\mathrm{Var}(\hat{C}_n^{22}) = \frac{1}{n^2 h^{4+2d-2p}} \mathrm{Tr}\big(n\big(\mathbb{E}(T_i T_i') - \mathbb{E}(T_i)\mathbb{E}(T_i)'\big)$$
$$+ n(n-1)\big(\mathbb{E}(T_i T_j') - \mathbb{E}(T_i)\mathbb{E}(T_j)'\big)\big).$$

Since T_i and T_j are independent, it follows that $\mathbb{E}(T_i T_j') - \mathbb{E}(T_i)\mathbb{E}(T_j)' = 0$. Next, note that

$$\mathbb{E}(T_i T_i') = h^{-2p}\bigg(\int_{B_{0,h^{1-\varepsilon}}^d} \bigg(K\bigg(\frac{\phi(z) - \phi(0)}{h}\bigg)\bigg)^2 [\phi(z) - \phi(0)][\phi(z) - \phi(0)]'$$
$$\times [\phi(z) - \phi(0)][\phi(z) - \phi(0)]' F(z)\, dz + o(h^{d+2})\bigg)$$
$$= h^{d+4-2p}\bigg(F(0)\int_{\mathbb{R}^d} \big(K(d_u\phi \cdot u)\big)^2 d_u\phi \cdot uu' \cdot d_u\phi' \cdot d_u\phi$$
$$\times uu' \cdot d_u\phi' \cdot du + O(h^2)\bigg).$$

Thus, the variance is given by

$$\mathrm{Var}(\hat{C}_n^{22}) = \frac{1}{nh^d} \mathrm{Tr}\bigg(F(0)\int_{\mathbb{R}^d} \big(K(d_u\phi \cdot u)\big)^2 d_u\phi \cdot uu' \cdot d_u\phi' \cdot d_u\phi$$
$$\times uu' \cdot d_u\phi' \cdot du + o_p(1)\bigg). \qquad \square$$

LEMMA A.2. *If the assumptions in Section* 2 *hold, then the matrices* $\hat{C}_n$, C^{22}, Π, $\hat{\Pi}_n$, $\hat{P}_n$, *and* P *have the following properties:*

(a) $\mathrm{rank}(C^{22}) = d$ *and* $\mathcal{R}(C^{22}) = T_p\mathcal{M}$;
(b) $\mathcal{R}(\Pi) = \mathcal{N}(C^{22})$, $\mathcal{N}(\Pi) = \mathcal{R}(C^{22})$ *and* $\mathcal{N}(\Pi) \cap \mathcal{N}(C^{22}) = \{0\}$;
(c) $\|\hat{P}_n - P\|_2^2 = \|\hat{\Pi}_n - \Pi\|_2^2 = O_p(1/nh^d)$;
(d) $\mathbb{P}(\mathrm{rank}(\hat{C}_n + \lambda_n \hat{P}_n/nh^{d+2}) = p+1) \to 1$.

PROOF.   To show property (a), we first show that for $M \in \mathbb{R}^{d \times d}$, where

$$M = \int_{\mathbb{R}^d} K(d_u\phi(0) \cdot u)uu'\, du,$$

we have that rank$(M) = d$. To prove this, choose any $v \in \mathbb{R}^d \setminus \{0\}$ and then consider the quantity

$$v'Mv = \int_{\mathbb{R}^d} K\big(d_u\phi(0) \cdot u\big) v'uu'v \, du.$$

By construction, $v'uu'v > 0$ almost everywhere. Additionally, since ϕ is three times differentiable, we have that $K(d_u\phi(0) \cdot u) > 0$ on a set of nonzero measure and $K(d_u\phi(0) \cdot u) \geq 0$ elsewhere. Thus, $v'Mv > 0$ for all $v \in \mathbb{R}^d \setminus \{0\}$. It follows that M is symmetric and positive definite with rank$(M) = d$. Since $\mathcal{M}$ is a d-dimensional manifold, we have that rank$(d_u\phi) = d$ by Corollary 8.4 of [34]. The Sylvester Inequality [47] implies that

$$\text{rank}(C_{22}) = \text{rank}(d_u\phi M d_u\phi') = d,$$

and this implies that

$$\mathcal{R}(C_{22}) = \mathcal{R}(d_u\phi M d_u\phi') = \mathcal{R}(d_u\phi).$$

However, $\mathcal{R}(d_u\phi) = T_p\mathcal{M}$, where we take $p = x_0$. This proves the result.

We next consider property (b). We have that

$$\sigma_1, \ldots, \sigma_d \neq 0 \quad \text{and} \quad \sigma_{d+1} = \cdots = \sigma_p = 0,$$

because rank$(C_{22}) = d$ by property (a). Thus, the null-space of C_{22} is given by the column-span of U_N; however, the construction of P implies that the column-span of U_N is the range-space of P. Ergo, $\mathcal{R}(P) = \mathcal{N}(C_{22})$. Note that the column-span of U_R belongs to the null-space of P, because each column in U_R is orthogonal—by property of the SVD—to each column in U_N. Thus, we have the dual result that $\mathcal{N}(P) = \mathcal{R}(C_{22})$. The orthogonality of U_R and U_N due to the SVD implies that $\mathcal{N}(P) \cap \mathcal{N}(C_{22}) = \{0\}$.

Now, we turn to property (c). For $h = \kappa n^{-1/(d+4)}$, Lemma A.1 says that $\|\hat{C}_n^{22} - C_{22}\|_F^2 = O_p(1/nh^d)$. The result follows from Corollary 3 of [29], by the fact that $\mathbb{I} - P_X$ is the projection matrix onto the null-space of X, and by the equivalence:

$$\|P_X - P_Z\|_2^2 \equiv \|\sin\Theta[\mathcal{R}(X), \mathcal{R}(Z)]\|,$$

where P_X is a projection matrix onto the range space of X [52].

Lastly, we deal with property (d). Lemma A.1 shows that

$$C_n \xrightarrow{p} C = \begin{bmatrix} C_{11} & 0 \\ 0 & C_{22} \end{bmatrix}.$$

Since $F(0) \neq 0$ by assumption, $C_{11} \neq 0$; thus, rank$(C) = 1 + \text{rank}(C_{22})$. Since $\mathcal{N}(P) \cap \mathcal{N}(C_{22}) = \{0\}$, we have that rank$(C_{22} + \lambda_n P/nh^{d+2}) = p$. Consequently, rank$(C) = p + 1$. Next, consider the expression

$$\|C_n + \lambda_n \tilde{P}_n/nh^{d+2} - C - \lambda_n \tilde{P}/nh^{d+2}\|_2^2$$

$$\leq \|\hat{C}_n - C\|_2^2 + \frac{\lambda_n}{nh^{d+2}}\|\tilde{P}_n - \tilde{P}\|_2^2$$

$$\leq O_p(h^2) + O_p\left(\frac{\lambda_n}{n^2 h^{2d+2}}\right)$$

$$\leq o_p(h).$$

Weyl's theorem [6] implies that

$$(A.1) \qquad \|\sigma_i(C_n + \lambda_n \tilde{P}_n/nh^{d+2}) - \sigma_i(C + \lambda_n \tilde{P}/nh^{d+2})\|_2^2 \leq o_p(h).$$

Note that $\sigma_i(C + \lambda_n \tilde{P}/nh^{d+2})$ is nondecreasing because λ_n/nh^{d+2} is nondecreasing. Define

$$\eta = \min(\sigma_i(C + \lambda_n \tilde{P}/n^{2/(d+4)})),$$

and consider the probability

$$(A.2)\qquad \begin{aligned} \mathbb{P}&\big(\mathrm{rank}(C_n + \lambda_n \tilde{P}_n/nh^{d+2}) = p + 1\big) \\ &\geq \mathbb{P}\big(|\sigma_i(C_n + \lambda_n \tilde{P}_n/nh^{d+2}) \\ &\qquad - \sigma_i(C + \lambda_n \tilde{P}/nh^{d+2})| \leq \eta, \forall i\big) \\ &\geq \sum_{i=1}^{p+1} \mathbb{P}\big(|\sigma_i(C_n + \lambda_n \tilde{P}_n/nh^{d+2}) \\ &\qquad - \sigma_i(C + \lambda_n \tilde{P}/nh^{d+2})| \leq \eta\big) - p. \end{aligned}$$

The result follows from (A.1) and (A.2). $\quad\square$

For notational convenience, we define

$$B_n = h^p(f(X) - \beta X_x)' W_{x_0} X_x H^{-1/2}$$

and

$$M = \tfrac{1}{2}[\partial_i \partial_j (f \circ \phi) - d_x f \cdot \partial_i \partial_j \phi].$$

Then, we have the following result concerning the asymptotic bias of the estimator:

LEMMA A.3. *If $h = \kappa n^{-1/(d+4)}$, then $B_n \xrightarrow{p} B$, where*

$$(A.3) \qquad B = \left[\kappa F(0) \int_{\mathbb{R}^d} K(d_u\phi \cdot u) uu' \, du \, M' \quad \mathbb{O}_{p\times 1}\right].$$

PROOF. First, recall the Taylor polynomial of the pullback of f to z:

$$f(\phi(z)) = f(\phi(0)) + d_x f \cdot d_u\phi \cdot z + \tfrac{1}{2}\partial_i \partial_j(f \circ \phi) \cdot zz' + o(\|z\|^2),$$

where we have performed a pullback of $d_x f$ from $T_x^* \mathcal{M}$ to $T_x^* \mathcal{U}$. In the following expression, we set $z = hu$:

$$f(\phi(z)) - f(x_0) - d_x f \cdot [\phi(z) - \phi(0)]$$

$$= \frac{h^2}{2} [\partial_i \partial_j (f \circ \phi) - d_x f \cdot \partial_i \partial_j \phi] uu' + o(\|hu\|^2)$$

$$= h^2 B uu' + o(\|hu\|^2).$$

Because $\beta = [f(x_0) \ d_x f]'$, we can rewrite the expectation of the expression as

$$\mathbb{E}(B_n) = \mathbb{E}\left(h^{p-d} K_H (x - x_0)(f(x) - x_{x_0}\beta)' x_{x_0} \begin{bmatrix} 1 & 0 \\ 0 & 1/h^2 \mathbb{I} \end{bmatrix} \right) H^{1/2}$$

$$= \left(\int_{\mathbb{R}^d} \left\{ K(d_u\phi \cdot u) h^2 u' u M' \begin{bmatrix} 1 & \frac{1}{h} d_u\phi \cdot u + \frac{1}{2} \partial_i \partial_j \phi \cdot uu' \end{bmatrix} \right. \right.$$

$$\left. \times (F(0) + h d_u F(0) \cdot u) \right\} du$$

$$\left. + o(h^2) \right) H^{1/2}$$

$$= \sqrt{nh^d} h^2 \left[F(0) \int_{\mathbb{R}^d} K(d_u\phi \cdot u) uu' \, du M' + o(1) \quad O(\sqrt{h^2}) \right],$$

where the last line follows because of the odd symmetries in the integrand. Since $h = \kappa n^{-1/(d+4)}$, this expectation becomes

$$\mathbb{E}(B_n) = B + o(1)\mathbb{1}_{1 \times (p+1)}.$$

The result follows from application of the weak law of large numbers. $\quad\square$

Let

$$\text{(A.4)} \qquad V = F(0) \int_{\mathbb{R}^d} (K(d_u\phi \cdot u))^2 \begin{bmatrix} 1 & 0 \\ 0 & d_u\phi \cdot uu' \cdot d_u\phi \end{bmatrix} du,$$

then the following lemma describes the asymptotic distribution of the error residuals.

LEMMA A.4. *If $h = \kappa n^{-1/(d+4)}$, then*

$$h^p \varepsilon' W_{x_0} X_x H^{-1/2} \xrightarrow{d} \mathcal{N}(0, \sigma^2 V).$$

PROOF. Since $\mathbb{E}(\varepsilon) = 0$ and ε is independent of x, we have that

$$\mathbb{E}(\sqrt{n} h^p \varepsilon K_H (x - x_0) x_{x_0} H^{-1/2}) = 0.$$

The variance of this quantity is

$$
\begin{aligned}
\mathrm{Var}&\big(h^p \sqrt{n}\varepsilon K_H(x - x_0)x_{x_0}H^{-1/2}\big) \\
&= \mathbb{E}\big((h^p \sqrt{n}\varepsilon K_H(x - x_0)x_{x_0}H^{-1/2})'(h^p \sqrt{n}\varepsilon K_H(x - x_0)x_{x_0}H^{-1/2})\big) \\
&= nh^{2p}\sigma^2 \mathbb{E}\big((K_H(x - x_0))^2[1 \quad (x - x_0)']'H^{-1}[1 \quad (x - x_0)']\big) \\
&= \sigma^2\left\{\int_{\mathbb{R}^d}(K(d_u\phi \cdot u))^2\begin{bmatrix} 1 & \left(d_u\phi \cdot u + \dfrac{h}{2}\partial_i\partial_j\phi \cdot uu'\right)' \\ \cdot & d_u\phi \cdot uu' \cdot d_u\phi \end{bmatrix}\right. \\
&\qquad\qquad \left. \times \big(F(0) + hd_u F(0)\cdot u\big)\,du + o(h^2)\right\} \\
&= \sigma^2\big(V + o(h)\mathbb{I}\big).
\end{aligned}
$$

Thus, the central limit theorem implies that

$$
\frac{h^p \sqrt{n}\varepsilon' W_{x_0}X_x H^{-1/2}}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \sigma^2 V). \qquad \square
$$

PROOF OF THEOREM 4.2. This proof follows the framework of [30, 59] but with significant modifications to deal with our estimator. For notational convenience, we define the indices of β such that: $\beta_0 = f(x_0)$ and $[\beta_1 \ \cdots \ \beta_p] = d_x f$. Let $\tilde{\beta} = \beta + H^{-1/2}u$ and

$$
\begin{aligned}
\Psi_n(u) &= h^p\big\| W_{x_0}^{1/2}\big(Y - X_{x_0}(\beta + H^{-1/2}u)\big)\big\|_2^2 \\
&\quad + \lambda_n\| P_n \cdot (\beta + H^{-1/2}u)\|_2^2.
\end{aligned}
$$

Let $\hat{u}^{(n)} = \arg\min \Psi_n(u)$; then $\hat{\beta}^{(n)} = \beta + H^{-1/2}\hat{u}^{(n)}$. Note that $\Psi_n(u) - \Psi_n(0) = V_4^{(n)}(u)$, where

$$
\begin{aligned}
V_4^{(n)}(u) &= u'H^{-1/2}(h^p X'_{x_0}W_{x_0}X_{x_0} + \lambda_n P_n)H^{-1/2}u \\
&\quad + 2\big(h^p(Y - X_{x_0}\beta)'W_{x_0}X_{x_0} + \lambda_n\beta' P_n\big)H^{-1/2}u.
\end{aligned}
$$

If $\lambda_n/nh^{d+2} \to \infty$ and $h\lambda_n/nh^{d+2} \to 0$, then for every u

$$
\lambda_n\beta' P_n H^{-1/2}u = \lambda_n\beta' Pu/\sqrt{nh^{d+2}}O_p(1) + \lambda_n h/nh^{d+2}O_p(1),
$$

where we have used Lemma A.2. It follows from the definition of β (4.5) and Lemma A.2 that $\beta' P \equiv 0$; thus, $\lambda_n\beta' P_n H^{-1/2}u = \lambda_n h/nh^{d+2}O_p(1) = o_p(1)$. For all $u \in \mathcal{N}(P)$, we have

$$
\lambda_n/nh^{d+2}u' P_n u = \lambda_n/nh^{d+2}O_p(1/nh^d) = o_P(h\lambda_n/nh^{d+2}),
$$

and for all $u \notin \mathcal{N}(P)$, we have

$$
\lambda_n/nh^{d+2}u' P_n u = \lambda_n/nh^{d+2}u' P_n u O_p(1) \to \infty.
$$

Let $W \sim \mathcal{N}(0, \sigma^2 V)$. Then, by Slutsky's theorem we must have that $V_4^{(n)}(u) \xrightarrow{d} V_4(u)$ for every u, where

$$V_4(u) = \begin{cases} u'Cu - 2u'(W + B), & \text{if } u \in \mathcal{N}(P), \\ \infty, & \text{otherwise.} \end{cases}$$

Lemma 5 shows that $V_4^{(n)}(u)$ is convex with high-probability, and Lemma 5 also shows that $V_4(u)$ is convex. Consequently, the unique minimum of $V_4(u)$ is given by $u = C^\dagger(W + B)$, where $C^\dagger$ denotes the Moore–Penrose pseudoinverse of C. Following the epi-convergence results of [20, 30], we have that $\hat{u}^{(n)} \xrightarrow{d} C^\dagger(W + B)$. This proves asymptotic normality of the estimator, as well as convergence in probability.

The proof for the NALEDE estimator comes for free. The proof formulation that we have used for the consistency of nonparametric regression in (4.6) allows us to trivially extend the proof of [59] to prove asymptotic normality and consistency. □

LEMMA A.5.   *Consider $A_n, B_n \in \mathbb{R}^{p_n \times p_n}$ that are symmetric, invertible matrices. If $\|A_n - B_n\|_2 = O_p(\gamma_n)$, $\|A_n^{-1}\|_2 = O_p(1)$ and $\|B_n^{-1}\|_2 = O_p(1)$, then $\|A_n^{-1} - B_n^{-1}\|_2 = O_p(\gamma_n)$.*

PROOF.   Consider the expression

$$\|A_n^{-1} - B_n^{-1}\|_2 = \|A_n^{-1}(B_n - A_n)B_n^{-1}\|_2$$
$$\leq \|A_n^{-1}\|_2 \cdot \|A_n - B_n\|_2 \cdot \|B_n^{-1}\|_2,$$

where the last line follows because the induced, matrix norm $\|\cdot\|_2$ is sub-multiplicative for square matrices.   □

PROOF OF THEOREM 5.3.   Under our set of assumptions, the results from [7] apply:

$$(A.5) \qquad \|T_t(X'X/n) - (\Sigma_\xi + \sigma_v^2 \mathbb{I})\|_2 = O_p\left(c_n\sqrt{\frac{\log p}{n}}\right),$$

$$(A.6) \qquad \|T_t(X'Y/n) - \Sigma_\xi \overline{\beta}\|_2 = O_p\left(c_n\sqrt{\frac{\log p}{n}}\right).$$

An argument similar to that given in Lemma A.2 implies that

$$\|\hat{P}_n - P_n\| = O_p(c_n\sqrt{\log p/n}).$$

Consequently, it holds that

$$(A.7) \qquad \begin{aligned} &\|T_t(X'X/n) - \sigma_v^2 \mathbb{I} + \lambda_n \hat{P}_n - (\Sigma_\xi + \lambda_n P_n)\|_2 \\ &= O_p\left(c_n(\lambda_n + 1)\sqrt{\frac{\log p}{n}}\right). \end{aligned}$$

Next, observe that

$$\Sigma_\xi + \lambda_n P_n = [\, U_R \quad U_N \,]\, \mathrm{diag}(\sigma_1, \ldots, \sigma_d, \lambda_n, \ldots, \lambda_n)\, [\, U_R \quad U_N \,]'.$$

Recall that we only consider the case in which $d < p$. We have that:

(a) $\|(\Sigma_\xi + \lambda_n P_n)^{-1}\|_2 = O(1)$, because of (5.2);
(b) $\|\Sigma_\xi^\dagger - (\Sigma_\xi + \lambda_n P_n)^{-1}\|_2 = O_p(1/\lambda_n)$.

Weyl's theorem [6] and (A.7) imply that

$$\left\|\left(T_t(X'X/n) - \sigma_\nu^2 \mathbb{I} + \lambda_n \hat{P}_n\right)^{-1}\right\|_2 = O_p(1).$$

Additionally, Lemma A.5 implies that

$$\left\|\left(T_t(X'X/n) - \sigma_\nu^2 \mathbb{I} + \lambda_n \hat{P}_n\right)^{-1} - (\Sigma_\xi + \lambda_n P_n)^{-1}\right\| = O_p\left(c_n \lambda_n \sqrt{\frac{\log p}{n}}\right).$$

Note that the solution to the estimator defined in (5.6) is:

$$\hat{\beta} = \left(T_t(X'X/n) - \sigma_\nu^2 \mathbb{I} + \lambda_n \hat{P}_n\right)^{-1} T_t(X'Y/n).$$

Next, we define

$$\beta^{(n)} \triangleq \left(T_t(X'X/n) - \sigma_\nu^2 \mathbb{I} + \lambda_n \hat{P}_n\right)^{-1} \Sigma_\xi \overline{\beta},$$

and note that the projection matrix onto the range space of Σ_ξ is given by $P_{\Sigma_\xi} = \Sigma_\xi^\dagger \Sigma_\xi$. Thus, $\beta = P_{\Sigma_\xi} \overline{\beta} = \Sigma_\xi^\dagger \Sigma_\xi \overline{\beta}$. Consequently, we have that

$$
\begin{aligned}
\|\hat{\beta} - \beta\|_2 &\leq \|\hat{\beta} - \beta^{(n)}\|_2 + \|\beta^{(n)} - \beta\|_2 \\
&\leq \left\|\left(T_t(X'X/n) - \sigma_\nu^2 \mathbb{I} + \lambda_n \hat{P}_n\right)^{-1}\right\|_2 \cdot \|T_t(X'Y/n) - \Sigma_\xi \overline{\beta}\|_2 \\
&\quad + \left\|\left(T_t(X'X/n) - \sigma_\nu^2 \mathbb{I} + \lambda_n \hat{P}_n\right)^{-1} - \Sigma_\xi^\dagger\right\|_2 \cdot \|\Sigma_\xi \overline{\beta}\|_2,
\end{aligned}
$$

(A.8)

where the inequality comes about because $\|\cdot\|_2$ is an induced, matrix norm and the expressions are of the form $\mathbb{R}^{p \times p}(\mathbb{R}^{p \times p}\mathbb{R}^p)$. Recall that for symmetric matrices, $\|A\|_1 = \|A\|_\infty$; ergo, $\|A\|_2 \leq \sqrt{\|A\|_1 \|A\|_\infty} = \|A\|_1$. Because of (5.1), we can use this relationship on the norms to calculate that $\|\Sigma_\xi\|_2 = O(c_n)$ and $\|\overline{\beta}\| = O(c_n)$. Consequently,

$$(A.8) \leq O_p\left(c_n \lambda_n \sqrt{\frac{\log p}{n}}\right) + O_p(c_n^2/\lambda_n).$$

The result follows from the relationship

$$\lambda_n = O\left(\sqrt{c_n}\left(\frac{n}{\log p}\right)^{1/4}\right). \qquad \square$$

We can show that the bias of the terms of the nonparametric exterior derivative estimation goes to zero at a certain rate.

LEMMA A.6. *Under the assumptions of Section 5, we have that*

$$|\mathbb{E}([\hat{C}_n]_{ij}) - [C_n]_{ij}| = O(h^2 c_n^2 (2\Omega)^{2d}),$$

$$|\mathbb{E}([\hat{R}_n]_{ij}) - [R_n]_{ij}| = O(h^2 c_n^2 (2\Omega)^{2d}),$$

where i, j denote the components of the matrices. Similarly, we have that

$$\mathrm{Var}([n\hat{C}_n]_{ij}) = O(1/h^d),$$

$$\mathrm{Var}([n\hat{R}_n]_{ij}) = O(1/h^d).$$

PROOF.   By the triangle inequality and a change of variables,

$$\mathrm{Bias}(\hat{C}_n^{11}) = \left| \int_{\mathcal{B}_{0,\Omega/h}^d} \frac{1}{h^d} K\left(\frac{\phi(z) - \phi(0)}{h} \right) F(z)\, dz - \int_{\mathcal{B}_{0,\Omega}^d} K(d_u\phi \cdot u) F(0)\, du \right|$$

$$\leq \left| \int_{\mathcal{B}_{0,\Omega}^d} \left[K\left(\frac{\phi(hu) - \phi(0)}{h} \right) - K(d_u\phi \cdot u) \right] F(hu)\, du \right|$$

$$+ \left| \int_{\mathcal{B}_{0,\Omega}^d} K(d_u\phi \cdot u)[F(hu) - F(0)]\, du \right| = T_1 + T_2.$$

The Taylor remainder theorem implies that

$$K\left(\frac{\phi(hu) - \phi(0)}{h} \right)$$

$$= K(d_u\phi \cdot u)$$

$$+ \partial_k K(d_u\phi \cdot u) \times (h\, \partial_{ij}\phi^k|_0 u^i u^j/2 + h^2\, \partial_{ijm}\phi^k|_w u^i u^j u^m/6)$$

$$+ \partial_{kl} K(v)/2 \times (h\, \partial_{ij}\phi^k|_0 u^i u^j/2 + h^2\, \partial_{ijm}\phi^k|_w u^i u^j u^m/6)$$

$$\times (h\, \partial_{ij}\phi^l|_0 u^i u^j/2 + h^2\, \partial_{ijm}\phi^l|_w u^i u^j u^m/6),$$

where $w \in \overline{\mathcal{B}_{0,\Omega}^d}$ and $v \in \overline{\mathcal{B}_{d_u\phi \cdot u, h\, \partial_{ij}\phi^k|_0 u^i u^j/2 + h^2\, \partial_{ijm}\phi^k|_w u^i u^j u^m/6}^d}$, and

$$F(hu) = F(0) + h\, \partial_i F|_0 u^i + h^2\, \partial_{ij} F|_v u^i u^j/2,$$

where $v \in (0, hu)$.

The odd-symmetry components of the integrands of T_1 and T_2 will be equal to zero, and so we only need to consider even-symmetry terms of the integrands. Recall that $K(\cdot), \partial_k K(\cdot), \partial_{kl} K(\cdot)$ are, respectively, even, odd and even. By the sparsity assumptions, we have that

$$T_1 = O(h^2 d^6 c_n^2 (2\Omega)^d),$$

$$T_2 = O(h^2 d^2 (2\Omega)^d).$$

Consequently, $T_1 + T_2 = O(h^2 d^6 c_n^2 (2\Omega)^d) = O(h^2 c_n^2 (2\Omega)^{2d})$.

We can compute the variance of $n\hat{C}_n^{11}$ to be

$$
\begin{aligned}
\operatorname{Var}(n\hat{C}_n^{11}) &= \int_{\mathcal{B}_{0,\Omega/h}^d} h^{-2d}\big[K\big((\phi(z)-\phi(0))/h\big)\big]^2 (F(z))^2 \, dz \\
&\quad - (\mathbb{E}(n\hat{C}_n^{11}))^2 \\
&= h^{-d} \int_{\mathcal{B}_{0,\Omega}^d} \big[K\big((\phi(hu)-\phi(0))/h\big)\big]^2 (F(hu))^2 \, dy \\
&\quad - (\mathbb{E}(n\hat{C}_n^{11}))^2 \\
&= O(1/h^d).
\end{aligned}
$$

The remainder of the results follow by similar, lengthy calculations. Note that for the variance of terms involving Y_i, a σ^2 coefficient appears, but this is just a finite-scaling factor which is irrelevant in O-notation. $\square$

PROOF OF THEOREM 5.1.   The key to this proof is to provide an exponential concentration inequality for the terms in $\hat{C}_n$ and $\hat{R}_n$. Having done this, we can then piggyback off of the proof in [7] to immediately get the result. The proofs for $\hat{C}_n$ and $\hat{R}_n$ are identical; so we only do the proof for $\hat{C}_n$.

Using the Bernstein inequality [36] and the union bound,

$$
\begin{aligned}
&\mathbb{P}\Big(\max_{i,j}\|[\hat{C}_n]_{ij} - \mathbb{E}[\hat{C}_n]_{ij}\| > t\Big) \\
&\qquad \le 2p^2 \exp\Big(-\frac{nt^2}{2\operatorname{Var}(n[\hat{C}_n]_{ij}) + \max(|n[\hat{C}_n]_{ij}|)2t/3}\Big).
\end{aligned}
$$

Since the ith component of X obeys: $|[X]_i| \le M$, it follows that

$$
\max(|n[\hat{C}_n]_{ij}|) = 2M/h^\eta,
$$

where $\eta \in \{0, 1, 2\}$ depending on i and j. Using this bound and Lemma A.6 gives

$$
\max_{i,j}|[\hat{C}_n]_{ij} - \mathbb{E}[\hat{C}_n]_{ij}| = O_p(\sqrt{\log p/nh^d}).
$$

Recall that

$$
\max_{i,j}|[\hat{C}_n]_{ij} - [C_n]_{ij}| \le \max_{i,j}|[\hat{C}_n]_{ij} - \mathbb{E}[\hat{C}_n]_{ij}| + \max_{i,j}|\mathbb{E}([\hat{C}_n]_{ij}) - [C_n]_{ij}|.
$$

However, this second term is $o(\sqrt{\log p/nh^d})$. Consequently,

$$
(\text{A.9}) \qquad \max_{i,j}|[\hat{C}_n]_{ij} - [C_n]_{ij}| = O_p(\sqrt{\log p/nh^d}).
$$

Using (A.9), we can follow the proof of Theorem 1 in [7] to prove the result. $\square$

REFERENCES

[1] AITCHISON, P. W. (1982). Generalized inverse matrices and their applications. *Internat. J. Math. Ed. Sci. Tech.* **13** 99–109. MR0646552

[2] ANDERSSON, M. (2009). A comparison of nine PLS1 algorithms. *J. Chemometrics* **23** 518–529.

[3] ASWANI, A., BICKEL, P. and TOMLIN, C. (2009). Statistics for sparse, high-dimensional, and nonparametric system identification. In *IEEE International Conference on Robotics and Automation* 2133–2138. IEEE Press, Piscataway, NJ.

[4] ASWANI, A., KERÄNEN, S., BROWN, J., FOWLKES, C., KNOWLES, D., BIGGIN, M., BICKEL, P. and TOMLIN, C. (2010). Nonparametric identification of regulatory interactions from spatial and temporal gene expression data. *BMC Bioinformatics* **11** 413.

[5] BELKIN, M., NIYOGI, P. and SINDHWANI, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.* **7** 2399–2434. MR2274444

[6] BHATIA, R. (2007). *Perturbation Bounds for Matrix Eigenvalues. Classics in Applied Mathematics* **53**. SIAM, Philadelphia, PA. MR2325304

[7] BICKEL, P. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

[8] BICKEL, P. and LI, B. (2007). Local polynomial regression on unknown manifolds. In *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **54** 177–186. Inst. Math. Statist., Beachwood, OH. MR2459188

[9] BICKEL, P. and FREEDMAN, D. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217. MR0630103

[10] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

[11] CARROLL, R., MACA, J. and RUPPERT, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika* **86** 541–554. MR1723777

[12] COSTA, J. and HERO, A. (2004). Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Trans. Signal Process.* **52** 2210–2221. MR2085582

[13] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[14] FAN, J. and TRUONG, Y. (1993). Nonparametric regression with errors in variables. *Ann. Statist.* **21** 1900–1925. MR1245773

[15] FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and its Applications. Monographs on Statistics and Applied Probability* **66**. Chapman and Hall, London. MR1383587

[16] FRANK, I. E. and FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35** 109–135.

[17] FRIEDMAN, J., HASTIE, T., HÖFLING, H. and TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.* **1** 302–332. MR2415737

[18] FU, W. (2000). Ridge estimator in singular design with application to age-period-cohort analysis of disease rates. *Comm. Statist. Theory Methods* **29** 263–278.

[19] FU, W. (2008). A smooth cohort model in age-period-cohort analysis with applications to homicide arrest rates and lung cancer mortality rates. *Sociol. Methods Res.* **36** 327–361. MR2422770

[20] GEYER, C. (1994). On the asymptotics of constrained m-estimation. *Ann. Statist.* **22** 1993–2010. MR1329179

[21] GOLDBERG, Y., ZAKAI, A., KUSHNIR, D. and RITOV, Y. (2008). Manifold learning: The price of normalization. *J. Mach. Learn. Res.* **9** 1909–1939. MR2438829

[22] GOLUB, G. and VAN LOAN, C. (1996). *Matrix Computations*, 3rd ed. Johns Hopkins Univ. Press, Baltimore, MD. MR1417720

[23] HEIN, M. and AUDIBERT, J.-Y. (2005). Intrinsic dimensionality estimation of submanifolds in $\mathbb{R}^d$. In *International Conference on Machine Learning* 289–296. ACM, New York.

[24] HELLAND, I. (1988). On the structure of partial least squares regression. *Comm. Statist. Simulation Comput.* **17** 581–607. MR0955342

[25] HOCKING, R. (1976). The analysis and selection of variables in linear regression. *Biometrics* **32** 431–453. MR0398008

[26] HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **8** 27–51.

[27] HUFFEL, S. V. and VANDEWALLE, J. (1991). *The Total Least Squares Problem: Computational Aspects and Analysis*. SIAM, Philadelphia, PA. MR1118607

[28] JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693.

[29] KAROUI, N. E. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011

[30] KNIGHT, K. and FU, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

[31] KRITCHMAN, S. and NADLER, B. (2008). Determining the number of components in a factor model from limited noisy data. *Chemometrics and Intelligent Laboratory Systems* **94** 19–32.

[32] LAFFERTY, J. and WASSERMAN, L. (2006). Rodeo: Sparse nonparametric regression in high dimensions. In *Advances in Neural Information Processing Systems (NIPS) 18* 707–714. MIT Press, Cambridge, MA.

[33] LEDOIT, O. and WOLF, M. (2003). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. MR2026339

[34] LEE, J. (2003). *Introduction to Smooth Manifolds*. Springer, New York. MR1930091

[35] LEVINA, E. and BICKEL, P. (2005). Maximum likelihood estimation of intrinsic dimension. In *Advances in NIPS 17* 777–784. MIT Press, Cambridge, MA.

[36] LUGOSI, G. (2006). Concentration-of-measure inequalities. Technical report, Pompeu Fabra Univ.

[37] MASSY, W. F. (1965). Principal components regression in exploratory statistical research. *J. Amer. Statist. Assoc.* **60** 234–246.

[38] MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. MR2488351

[39] MEINSHAUSEN, N. and BUEHLMANN, P. (2010). Stability selection. *J. Roy. Statist. Soc. Ser. B* **72** 417–473.

[40] MISNER, C. W., THORNE, K. S. and WHEELER, J. A. (1973). *Gravitation*. W. H. Freeman and Co., San Francisco, CA. MR0418833

[41] NADLER, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *Ann. Statist.* **36** 2791–2817. MR2485013

[42] NG, A. (1997). Preventing overfitting of cross-validation data. In *14th International Conference on Machine Learning* 245–253. Morgan Kaufmann, San Francisco, CA.

[43] NIYOGI, P. (2008). Manifold regularization and semi-supervised learning: Some theoretical analyses. Technical Report TR-2008-01, Univ. Chicago, Computer Science Dept.

[44] RAO, R., FUNG, G. and ROSALES, R. (2008). On the dangers of cross-valdation: An experimental evaluation. In *SIAM Data Mining*. SIAM, Philadelphia, PA.

[45] REUNANEN, J. (2003). Overfitting in making comparisons between variable selection methods. *J. Mach. Learn. Res.* **3** 1371–1382.

[46] RUPPERT, D. and WAND, M. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22** 1346–1370. MR1311979

[47] SASTRY, S. (1999). *Nonlinear Systems*. Springer, New York. MR1693648

[48] SHAO, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88** 486–494. MR1224373

[49] SHAO, J. (1994). Bootstrap sample size in nonregular cases. *Proc. Amer. Math. Soc.* **122** 1251–1262. MR1227529

[50] SHAO, J. (1996). Bootstrap model selection. *J. Amer. Statist. Assoc.* **91** 655–665. MR1395733

[51] SPIVAK, M. (1965). *Calculus on Manifolds. A Modern Approach to Classical Theorems of Advanced Calculus*. W. A. Benjamin, Inc., New York.

[52] STEWART, G. and SUN, J. (1990). *Matrix Perturbation Theory*. Academic Press, Boston, MA. MR1061154

[53] TENENBAUM, J. B., DE SILVA, V. and LANGFORD, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** 2319–2323.

[54] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[55] WOLD, H. (1975). Soft modeling by latent variables: the nonlinear iterative partial least squares approach. In *Perspectives in Probability and Statistics, Papers in Honour of M. S. Bartlett* (J. Gani, ed.) 117–142. Univ. Sheffield, Sheffield. MR0431486

[56] WRIGHT, J., YANG, A., GANESH, A., SASTRY, S. and MA, Y. (2009). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31** 210–227.

[57] WU, T. T. and LANGE, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Statist.* **2** 224–244. MR2415601

[58] YANG, Y., FU, W. and LAND, K. (2004). A methodological comparison of age-period-cohort models: The intrinsic estimator and conventional generalized linear models. *Sociological Methodology* **34** 75–110.

[59] ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

[60] ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 301–320. MR2137327

[61] ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286. MR2252527

DEPARTMENT OF ELECTRICAL ENGINEERING
  AND COMPUTER SCIENCES
UNIVERSITY OF CALIFORNIA, BERKELEY
253 CORY HALL
BERKELEY, CALIFORNIA 94720-1770
USA
E-MAIL: aaswani@eecs.berkeley.edu
        tomlin@eecs.berkeley.edu

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
367 EVANS HALL
BERKELEY, CALIFORNIA 94720-3860
USA
E-MAIL: bickel@stat.berkeley.edu