IEOR 265 – Lecture 5
Abstract Structure

1 Principal Component Analysis

Recall that the EDE estimator was defined as

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

where Π is a projection matrix that projects onto the $(p-d)$ smallest eigenvectors of the sample covariance matrix $\frac{1}{n}X'X$. Also recall that the PCR estimator was defined as a special case of the EDE estimator. It is worth discussing in greater detail what Π is and what the d largest eigenvectors of $\frac{1}{n}X'X$ are.

Principal component analysis (PCA) is a popular visualization and dimensionality reduction technique. One of its uses is to convert a complex high-dimensional data set into a two- or three-dimensional data set, which can then be visualized to see relationships between different data points. Because the sample covariance matrix is symmetric (and positive semidefinite), it can be diagonalized by an orthogonal matrix:

$$\tfrac{1}{n}X'X = U\mathrm{diag}(s_1, \ldots, s_p)U',$$

where $U \in \mathbb{R}^{p\times p}$ is an orthogonal matrix and $s_1 \geq \ldots \geq s_p \geq 0$. The idea of PCA is to define a coordinate transformation from $\mathbb{R}^p$ to $\mathbb{R}^d$ by the following linear transformation

$$T_d = \begin{bmatrix}\mathbb{I}_d \\ 0\end{bmatrix} U'.$$

Graphically, the idea of PCA is to identify the directions of most variation of the data points x_i.

The reason that PCA is relevant to the EDE estimator is that Π is computed using PCA. The majority of the variation of the data lies within the hyperplane with basis defined by the column-span of T_d. In the collinearity model, the exterior derivative must lie within the this hyperplane, and so we penalize for deviation of the estimates from this hyperplane by specifically penalizing for any component of the estimate that lies in the space orthogonal to this hyperplane. This orthogonal component is given by $\Pi\beta$, where

$$\Pi = U\begin{bmatrix}0 \\ \mathbb{I}_{p-d}\end{bmatrix} U' = \mathbb{I}_p - T_d'T_d.$$

2 Manifold Structure

Imagine a generalization of the collinearity model previously considered. Specifically, assume that the x_i lie on an embedded submanifold $\mathcal{M}$ with dimension $d < p$, where this manifold is unknown *a priori*. Suppose that the system has model $y_i = f(x_i) + \epsilon_i$, where $f(\cdot)$ is an unknown nonlinear function. The reason that this situation is interesting is that the LLR estimate converged at rate $O_p(n^{-2/(p+4)})$, but if we knew the manifold then we could do a coordinate change into a lower-dimensional space and then the LLR estimate would converge at rate $O_p(n^{-2/(d+4)})$.

Even though this manifold is unknown, we could imagine that if we were able to somehow learn this manifold and then incorporate this knowledge into our estimator, then we could achieve the faster convergence rate. This is in fact the idea behind the nonparametric exterior derivative estimator (NEDE), which is defined as

$$\begin{bmatrix}\hat{\beta}_0[x_0]\\ \hat{\beta}[x_0]\end{bmatrix} = \arg\min_{\beta_0,\beta} \left\| W_h^{1/2}\left(Y - \begin{bmatrix}1_n & X_0\end{bmatrix}\begin{bmatrix}\beta_0\\ \beta\end{bmatrix}\right)\right\|_2^2 + \lambda\|\Pi\beta\|_2^2,$$

where $X_0 = X - x_0' 1_n$, Π is a projection matrix that projects onto the $(p-d)$ smallest eigenvectors of the sample local covariance matrix $\frac{1}{nh^{d+2}}X_0' W_h X_0$, and

$$W_h = \operatorname{diag}\left(K(\|x_1 - x_0\|/h), \ldots, K(\|x_n - x_0\|/h)\right).$$

It can be shown that the error in this estimate converges at rate $O_p(n^{-2/(d+4)})$, even though the regression is being computed for coefficients that lie within a p-dimensional space. Furthermore, it can be shown that $\hat{\beta}[x_0]$ is a consistent estimate of the exterior derivative of f at x_0 (i.e., $df|_{x_0}$). This improved convergence rate can be very useful if $d \ll p$.

The idea of lower-dimensional structure either in a hyperplane or manifold context is an important abstract structure. It is important because there are many methods that can exploit such structure to provide improved estimation.

3 Sparsity Structure

Consider a linear model $y_i = x_i'\beta + \epsilon_i$, where $x_i \in \mathbb{R}^p$ and we have n measurements: (x_i, y_i) for $i = 1, \ldots, n$. In the classical setting, we assume that p is fixed and n increases towards infinity. However, an interesting situation to consider is when p is roughly the same size as (or even larger than) n. In this high-dimensional setting, many of the estimators that we have defined are no longer consistent.

In general, this situation is hopeless. But the situation is improved given certain structure. The coefficients β of the model are s-sparse if at most s values are non-zero:

$$\sum_{i=1}^{p} |\beta_i|^0 \leq s.$$

In a situation with exact sparsity, the majority of coefficients are exactly zero. This idea can be relaxed. The coefficients β of the model are approximately-s_q-sparse if

$$\sum_{i=1}^{p} |\beta_i|^q \leq s_q,$$

for $q \in [0, 1]$. Note that in the special case where $q = 0$, this is the same as the condition for an s-sparse model. The idea of approximate sparsity is that even though most of the coefficients are non-zero, the coefficients β can be well-approximated by another set of coefficients $\tilde{\beta}$ that are exactly sparse. Such sparsity is important because consistent estimators can be designed for sparse models.

There are numerous extensions of sparse structure, which can be exploited. One example is group sparsity. Suppose we partition the coefficients into blocks $\beta' = \begin{bmatrix}\beta^{1'} & \dots & \beta^{m'}\end{bmatrix}'$, where the blocks are given by:

$$\begin{aligned}
\beta^{1'} &= \begin{bmatrix}\beta_1 & \dots & \beta_k\end{bmatrix} \\
\beta^{2'} &= \begin{bmatrix}\beta_{k+1} & \dots & \beta_{2k}\end{bmatrix} \\
&\vdots \\
\beta^{m'} &= \begin{bmatrix}\beta_{(m-1)k+1} & \dots & \beta_{mk}\end{bmatrix}.
\end{aligned}$$

Then the idea of group sparsity is that most blocks of coefficients are zero.

4 Lasso Regression

The M-estimator which had the Bayesian interpretation of a linear model with Laplacian prior

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1,$$

has multiple names: Lasso regression and $L1$-penalized regression.

4.1 Computation of Lasso Regression

Computation of this estimator is a complex topic because the objective is not differentiable, but for pedagogy we talk about how the corresponding optimization can be rewritten as a constrained quadratic program (QP). If we use an epigraph formulation, then we can rewrite the optimization as

$$\begin{aligned}
\hat{\beta} = \arg\min_{\beta}\ & \|Y - X\beta\|_2^2 + t \\
\text{s.t. }\ & t \geq \lambda\|\beta\|_1.
\end{aligned}$$

But because $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$ (by definition), we can rewrite the above optimization as a constrained QP

$$\begin{aligned}\hat{\beta} = \arg\min_{\beta} \; & \|Y - X\beta\|_2^2 + t \\ \text{s.t. } & t \geq \lambda \textstyle\sum_{j=1}^p \mu_j \\ & -\mu_j \leq \beta_j \leq \mu_j, \forall j = 1, \ldots, p.\end{aligned}$$

It is worth stressing that this is not an efficient way to compute Lasso regression.