Statistical Learning, Fall 2024-5

Class notes 5

Principal component analysis (PCA) and PCA regression

Recall the PCA problem: $\mathbb{X}$ is now considered a collection of n points in $\mathbb{R}^p$, and we are looking for directions $v \in \mathbb{R}^p$ with large spread of the data. Finding the first principal component:

$$v_1 = \arg \max_{\|v\|_2=1} \|\mathbb{X}v\|_2^2.$$

The second principal component is an orthogonal direction with the most spread:

$$v_2 = \arg \max_{\|v\|_2=1, v \perp v_1} \|\mathbb{X}v\|_2^2.$$

and so on.

Finding the principal components is closely related to the SVD we discussed above: Recall $\mathbb{X} = UDV^T$, with $d_1 \geq d_2 \geq ... \geq d_p \geq 0$. Since the columns of V are an orthonormal basis of $\mathbb{R}^p$, for any vector v with $\|v\|_2 = 1$, write it as:

$$v = \sum_{j=1}^{p} a_j V_j \ , \ \sum_j a_j^2 = \|v\|^2 = 1 \ (V_j \text{ are the columns of } V).$$

Then we can see:

$$\|\mathbb{X}v\|^2 = v^T V D U^T U D V^T v = (\sum_{j=1}^{p} a_j V_j^T V) D^2 (V^T \sum_{j=1}^{p} a_j V_j) = \left(\sum_{j=1}^{p} a_j e_j^T\right) D^2 \left(\sum_{j=1}^{p} a_j e_j\right) = \sum_{j=1}^{p} a_j^2 d_j^2 \leq d_1^2,$$

however we get equality if $a_1 = 1$ or $v = V_1$. Conclusion: $v_1 = V_1$ the first column of V. Similarly $v_2 = V_2$ and so on.

This is a very important approach for dimensionality reduction and finding "interesting" direction in high dimensional space (p is large) given data X. In our context we can use it for regularization by choosing only the first k principal components to do the regression on: Take the first k principal components (columns of V) and define:

$$Z = \mathbb{X}V_{1..k} = U_{1..k} D_{1..k,1..k},$$

and now perform the regression of $\mathbb{Y}$ on Z instead of $\mathbb{X}$.

In fact, when writing in terms of the SVD, we can relate all three approaches (LS, ridge, PCA regression) through similar operations:

$$\begin{aligned}
\hat{Y}_{LS} &= \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y} = UU^T\mathbb{Y} \\
\hat{Y}_{ridge}(\lambda) &= \mathbb{X}(\mathbb{X}^T\mathbb{X}+\lambda I)^{-1}\mathbb{X}^T\mathbb{Y} = U diag\left(\frac{d_j^2}{d_j^2+\lambda}\right)U^T\mathbb{Y} \\
\hat{Y}_{PCA}(k) &= Z(Z^TZ)^{-1}Z^T\mathbb{Y} = U_{1..k}U_{1..k}^T\mathbb{Y}.
\end{aligned}$$

Lasso

The lasso formulations:

$$\hat{\beta}^{pen}(\lambda) = \arg\min_{\beta} RSS(\beta) + \lambda\sum_j |\beta_j| \ , \ \ \hat{\beta}^{con}(s) = \arg\min_{\beta:\sum_j|\beta_j|\le s} RSS(\beta).$$

Unlike ridge it does not have an algebraic solution (for example, the penalized Lagrange version is not differentiable). A key observation is that this is now a *quadratic programming* (QP) problem, with quadratic objective and linear constraints. This can be seen from the constrained version, which can equivalently be written as:

$$\begin{aligned}
\min \quad & RSS(\beta^+ - \beta^-) \\
\text{s.t.} \quad & \sum_{j=1}^p \beta_j^+ + \beta_j^- \le s \\
& \beta_j^-, \beta_j^+ \ge 0 \ \forall j,
\end{aligned}$$

(the problems are equivalent since in the optimal solution, it is guaranteed that either $\beta_j^+ = 0 \Rightarrow \beta_j = -\beta_j^-$ or $\beta_j^- = 0 \Rightarrow \beta_j = \beta_j^+$.)
Since QP is a standard problem in convex optimization, standard solvers can be used for Lasso (and in fact the original paper by Tibshirani(1996) proposes a special QP variant that fits the structure of this problem).

However, in the early 2000's several groups realized that the problem can be solved with linear algebra tools, by following the set of solutions to the penalized problems $\hat{\beta}^{pen}(\lambda), 0 \le \lambda < \infty$. We will not discuss this approach, best known by the name Least Angle Regression (LARS), but it is interesting both for computation and statistical interpretation.

A key property of Lasso solutions is *sparsity*. This has several expressions:

- In the high dimensional regime $p > n$, Lasso solutions always have at most n non-zero coefficients: $\|\hat{\beta}(\lambda)\|_0 \le n$. This is in contrast to Ridge regression, where all coefficients are always non-zero.
- In the low dimensional regime $p < n$, it is still true that for $\lambda >> 0$ heavy Lasso regularization, we will have $\|\hat{\beta}(\lambda)\|_0 << p$, many zero coefficients (again, in contrast to Ridge and other methods).

- The area of Compressed Sensing, which was extremely widely studied around 2005-2010, shows that under some assumptions:
  - The true model is linear $\mathbb{E}(Y|X) = X^t\beta$, and it is sparse $\|\beta\|_0 = k << p$.
  - The explanatory variables (columns of $\mathbb{X}$) have low correlation between them.
  - The amount of data we have is $n = O(k \log p)$ (which is still $<< p$ when the model is very sparse).

  then using Lasso we can find *with high probability* the true sparsity pattern (that is, $\hat{\beta}(\lambda)_j \neq 0 \Leftrightarrow \beta_j \neq 0$). However, due to the strong assumptions this intriguing area is not necessarily relevant to our settings of interest.

To demonstrate and explain sparsity, we can take two relatively simply views. The first is geometrical and relies on the constrained version, showing that the constraint region $\sum_j |\beta_j| \leq s$ is "diamond shaped" with corners "sticking out". Although we can only draw it in low dimension, we can for example show that in high dimension almost all the volume of the cube is close to the corners. Hence the constrained solution is likely to fall on a corner of this space, which is a point where some of the coefficients are zero.

A more rigorous view can be taken by an algebraic analysis of a simplified version of Lasso, where we assume the matrix $\mathbb{X}$ is orthogonal, i.e., $\mathbb{X}^T\mathbb{X} = I_p$. Then we can write for the penalized version:

$$\begin{aligned} PRSS(\beta) &= RSS(\beta) + \lambda \sum_j |\beta_j| = \mathbb{Y}^T\mathbb{Y} - 2\mathbb{Y}^T\mathbb{X}\beta + \beta^T(\mathbb{X}^T\mathbb{X})\beta + \lambda \sum_j |\beta_j| = \\ &= \sum_j \left(\beta_j^2 + \lambda\beta_j \cdot \text{sign}(\beta_j) - 2\mathbb{X}_j^T\mathbb{Y}\right) + \mathbb{Y}^T\mathbb{Y}, \end{aligned}$$

Now, if we *assume* $\beta_j > 0$ we can differentiate and compare to zero:

$$\frac{\partial PRSS(\beta)}{\partial \beta_j} 2\beta_j + \lambda - 2\mathbb{X}_j^T = 0 \;\Rightarrow\; \beta_j = \mathbb{X}_j^T\mathbb{Y} - \frac{\lambda}{2}.$$

Thus, if $\mathbb{X}_j^T\mathbb{Y} > \lambda/2$ we conclude that we indeed get a positive solution:

$$\hat{\beta}(\lambda)_j = \mathbb{X}_j^T\mathbb{Y} - \frac{\lambda}{2}.$$

Since this is a quadratic function, we know this will yield a minimum.

Similarly, if $\mathbb{X}_j^T\mathbb{Y} < -\lambda/2$ a similar calculation will yield:

$$\hat{\beta}(\lambda)_j = \mathbb{X}_j^T\mathbb{Y} + \frac{\lambda}{2}.$$

With a little more work we can confirm that the optimal solution is:

$$\hat{\beta}(\lambda)_j = \begin{cases} \mathbb{X}_j^T\mathbb{Y} - \frac{\lambda}{2} & \text{if } \mathbb{X}_j^T\mathbb{Y} > \lambda/2 \\ 0 & \text{if } -\lambda/2 \leq \mathbb{X}_j^T\mathbb{Y} \leq \lambda/2 \\ \mathbb{X}_j^T\mathbb{Y} + \frac{\lambda}{2} & \text{if } \mathbb{X}_j^T\mathbb{Y} < -\lambda/2 \end{cases}.$$

This is called the *soft thresholding* solution, and it shows that when λ is big, most of the coefficients $\hat{\beta}(\lambda)_j$ are zero.

We note that the LARS solution without assuming $\mathbb{X}^T\mathbb{X} = I$ is a generalization of similar ideas, with substantially more complex algebra, and leads to similar conclusion that when λ is big, most coefficients are zero.

Summarizing Lasso important properties:

- Statistical: sparsity, compressed sensing
- Computational: convex problem (QP), efficient specialized methods like LARS

Extensions

There are many other regularization methods for linear regression, many of which extend Ridge and Lasso in interesting ways, two of the most important ones:

- Elastic net (Zou and Hastie 2005) combines the ridge and lasso penalties:

  $$\hat{\beta}(\lambda) = \arg\min_{\beta} RSS(\beta) + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2.$$

  Geometrically, this is a "diamond" with curved sides, and they argue that it mitigates the over-sparsity that Lasso tends to have with highly correlated variables. The popularity of this method stems from the very good predictive modeling results it gives, often better than both Lasso and Ridge.

- Group lasso (Yuan and Lin 2006): assume our coefficients are divided into k groups $\mathcal{G}_1, ..., \mathcal{G}_k$, where we assume that within each group the coefficients are similarly important, or should appear together, but we want sparsity between groups, then the following formulation assures that:

  $$\hat{\beta}(\lambda) = \arg\min_{\beta} RSS(\beta) + \lambda \sum_{l=1}^{k} \|\beta_{\mathcal{G}_l}\|_2,$$

  this can be thought of as having Ridge penalty within groups and Lasso between groups.