45. **Principal Components Regression and Partial Least Squares in Bioinformatics**

In my study of dimension reduction methods in bioinformatics, I've discovered that one of the most famous techniques is Principal Components Regression (PCR). PCR involves a two-step procedure. In the first step, I identify the principal components of the data matrix X. Although I won't delve into the full details of principal components here, the essential idea is straightforward. In step one, I compute the principal components. Then, in step two, I perform a least squares regression using these principal components as predictors.

Principal components are an interesting concept. The first principal component is simply the linear combination of the variables that has the highest variance. The second principal component is the linear combination with the largest variance among all linear combinations that are uncorrelated with the first, and so on. This process provides me with linear combinations or dimensions of the data that exhibit high variance and are uncorrelated with the previous ones. If I'm working with a dataset containing 35 variables, I might compute a few principal components that capture most of the variation in the data in a succinct way, using just a few new variables, such as ​ through ​ or ​ through .

For instance, I performed Principal Components Analysis (PCA) on a simple dataset because within the plot, the x-axis represents population, and the y-axis represents ad spending for a hundred different cities. These data points are shown as purple dots. This dataset contains just two variables (population and ad spending), so I can ask, "What is the linear combination of these variables that has the most variance?" Or equivalently, "What is the direction along which this data varies the most?" I can see that this direction, where the data varies most, aligns with a green line—this is the first principal component direction. If I then ask, "What is the direction with the most variance among all directions uncorrelated with the first?" I get the blue dashed line, which is the second principal component in this dataset. In this example, since there are only two variables (P-2), there are only two principal components.

When I zoom in on what is happening, the green line is the first principal component because it is the direction along which the data varies the most. It’s also the direction along which the distances from the data points (purple points) to the green line (shown in red) are minimized. On a rotated plot, this principal component line becomes horizontal for easier visualization. To better understand these principal components, I can plot each principal component against the original variables, such as population and ad spending. The first principal component is highly correlated with both population and ad spending, suggesting that I am summarizing the data effectively. Rather than using the original two variables to predict sales, for example, I could use just the first principal component, treat it as a predictor in a model, and fit the model using least squares.

I found that when plotting the second principal component against population and ad spending, there is very little relationship between them. This observation indicates that the first principal component summarizes the data effectively, especially when the original variables are correlated, as in this case with population and ad spending. The idea is to compute the principal components, use them as predictors in a regression model, and fit the model using least squares, which can yield good results in many settings.

To illustrate this further, I worked on a simulated dataset with multiple observations and performed PCR with varying numbers of principal components. For example, I used one principal component up to around 45 principal components in this example. I plotted the bias (in black), variance (in green), and mean squared error (MSE, in purple). As the number of components in the model increased, the bias decreased due to the more complex model, but the variance increased. The MSE, which is the sum of squared bias and variance, typically followed a U-shaped curve. I observed that the MSE was minimized with around 18 principal components, suggesting that PCR with 18 predictors was highly effective. In another example, where the MSE remained flat after a certain number of components, I opted for a simpler model with around 25 components, preferring simplicity when models have comparable MSEs.

When using PCR, I need to choose the number of components M. To do this, I often rely on cross-validation to estimate the test mean squared error. For instance, when I performed PCR on a credit dataset, I found that using 10 or 11 components minimized the cross-validated mean squared error. However, when M-11, PCR essentially becomes regular least squares on the original data, which was somewhat disappointing because PCR didn't offer any advantage over standard least squares in this case.

If I don't want to rely on the assumption that the directions of maximum variance in X also relate to the response Y, I turn to Partial Least Squares (PLS). PLS, like PCR, also creates new predictors, ​ through but it does so in a supervised manner. PLS selects these new predictors by considering both the variance in the original features and their relationship with the response Y. For example, I obtain the first PLS direction by performing a regression of Y onto each predictor ​, and I continue this process to determine subsequent directions. This means that the new directions are not only high in variance but are also more closely related to Y.

While it might seem that PLS should outperform PCR because it explicitly considers the response Y, I have often found that it does not provide significant gains over PCR. The results tend to be similar to those from ridge regression and PCR. Mathematically, these methods are closely related; for example, PCR can be seen as a discrete version of ridge regression, where ridge regression continuously shrinks variables while PCR does so more discretely.

Ultimately, my exploration into these methods shows that, in settings where I have many variables, simpler models derived through dimension reduction techniques like PCR and PLS can help avoid overfitting, which is common with least squares when dealing with high-dimensional data. While each method has its strengths and limitations, understanding their nuances helps me better decide which approach to take for improved prediction in bioinformatics research.

Top of Form

Bottom of Form