Supervised Principal Components

Marvin Ward Jr.

April 19, 2020

1 Introduction

For better or worse, linear regression models serve as the workhorse of social science. "They are simple and often provide an adequate and interpretable description of how inputs affect the ouput." [2] The parsimony of the linear model allows researchers to not only assign factor loadings to the various phenomena which may affect an outcome of interest, but also to straightforwardly convey the results to a wide variety of audiences. For those who wish to design policy, digestibly understanding the relationship between inputs and output is essential.

A linear model is linear in the **coefficient vector** β, but not necessarily in the **feature matrix** X.

The linear regression model seeks to interpret the joint distribution of a response Y and related variables $X_1, X_2, \ldots, X_p$ that may or may not have a causal relationship with Y. We impose linearity to ease the separability of effect. Each observation is a function of p linearly integrated regressors:

$$y_i = \beta_0 + \Sigma_{j=1}^{p} \beta_j x_{ij} = X_i \beta \tag{1}$$

We estimate the mapping from X to Y because we do not actually know the underlying data generating process and true functional form of *f(X)*. Consequently, we collect many observations and fit the p-dimensional vector $\hat{\beta}$ such that the distance between Y and $X\beta$ is minimized. Assuming we choose the 2-norm as the distance measurement, $||y - X\beta||_2$, linear regression is classic least squares approximation. [7] The challenge addressed in this paper is that the estimation of $\hat{\beta}$ assumes we can observe all of the relevant regressors!

$$\hat{Y} = E(Y|X)E(X) \tag{2}$$

Latent features or factors are inputs to a model we cannot directly observe.

If this assumption is violated, what recourse do we have? Ideally, we would want to find some way to include latent features into the analysis. The **Supervised Principal Components** (SPC) approach combines **Principal Components Analysis** (PCA) and linear regression together in a way that yields estimates of latent factors, which can then be used directly in the estimation of the full model.

1.1 Hypothetical Grounding

Suppose that we seek to study structural changes in consumption activity during the COVID-19 pandemic using a sample of people, and we can only observe attributes of their financial transactions. Some members of the sample have lost their jobs while others have not. We want to understand the impact of job loss, but we cannot directly observe which parties lost their jobs. The ability to maintain spending levels depends on income, but the separability between the employed and unemployed groups is obscured because everyone is reducing their consumption due to the pandemic. If we could observe job loss, we could include it directly in our regression specification. Since we cannot, how should we proceed?

For simplicity, suppose that the consumption response is entirely a function of job loss.

We expect that each person, i, in the sample will have a consumption response to that is a function of *unobserved* job loss, u_i, and a vector of other factors. We must consider a set of *observable* regressors, $X_i = x_{i1} + x_{i2} + \cdots + x_{ip}$, we believe have some relationship to job loss.

$$Y = \beta_0 + \beta_1 U + \epsilon = \begin{bmatrix} 1 & U \end{bmatrix} B + \epsilon \tag{3}$$

$$X_j = \alpha_{0j} + \alpha_{1j} U + \eta_j = \begin{bmatrix} 1 & U \end{bmatrix} A + \eta_j \tag{4}$$

On average, people who keep their jobs are more likely to maintain higher levels of spending, but since everyone is generally changing their spending behavior the distribution of spending change in the newly unemployed portion of the sample is likely to overlap strongly with the distribution of spending change among those who retained their jobs. While administrative data on financial account activity provides a very broad set of indicators, many of them will be well correlated with each other. We could use PCA to identify covariance among features, but there is no guarantee that the features that belong to the dominant components are related to job loss or changes in consumption in general. SPC provides a way of sifting through the principal components of the regressor data to identify the direction of variation that best aligns with changes in the response. In the remainder of this paper we will discuss the mechanics behind PCA, explain how PCA can be used jointly with linear regression, and use that foundation to present the SPC estimation approach.

The **response variable** is the variable to predicted or estimated Y.

2 Principal Components Analysis

The big idea that motivates PCA is the desire to collapse the information contained in n-dimensional data into a m-dimensional space without materially affecting the information content, for some $m < n$. Getting new, information rich features can be accomplished by changing the basis of our matrix to match the orthogonal set of eigenvectors that characterize our data. Once we have identified the set of eigenvectors, we can simply select the subset we feel captures an adequate proportion of overall variation.

For instance, if there are 100 features in our input data, can we reduce to 10?

2.1 Overview

A **transformation** $T : A \to B$ maps values from the **domain** A to values in the **codomain** B.

Collapsing information from n to m dimensions starts with selecting a new n-dimensional basis in the codomain and then selecting a subset of m basis vectors. Once we have restricted the dimensional representation, we then map the reduced data back to the original basis to complete the process.

The input features X define the basis in the domain, while the eigenbasis of the features defines the basis in the codomain.

For now, assume the existence of a known linear transformation $T : \mathbb{C}^n \to \mathbb{C}^n$ that maps values from the domain to a codomain that retains all previous information, but now with a new basis. Assume also that the inverse T^{-1} exists that maps the data back from the codomain to the original basis $T^{-1} : \mathbb{C}^m \to \mathbb{C}^m$. Let f represent the algorithm for selecting the m dimensions from the total set of n: $f : X_n \to X_m$. The following provides a high-level view of the overall PCA dimension reduction algorithm (where the data matrix X is indexed by domain/codomain and n-/m-dimensions):

$$T(X_{d,n}) = X_{c,n} \tag{5}$$

$$f(X_{c,n}) = X_{c,m} \tag{6}$$

$$T^{-1}(X_{c,m}) = X_{d,m} \tag{7}$$

2.2 Choice of Basis

How should we identify the basis that will best capture the information in our data? To the extent that we can consider variation as information, the variance-covariance matrix, $V[X] = X^T X$ is a reasonable choice. To see why, consider the following minimal example in which the matrix X is comprised of two $n \times 1$ column vectors x_1 and x_2.

$$X = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \tag{8}$$

$$X^T X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \begin{bmatrix} x_1 & x_2 \end{bmatrix} \tag{9}$$

$$= \begin{bmatrix} x_1 \cdot x_1 & x_1 \cdot x_2 \\ x_2 \cdot x_1 & x_2 \cdot x_2 \end{bmatrix} \tag{10}$$

We assume that all features are standardized to $\mu = 0$ and $\sigma = 1$.

The absolute value of each dot product in the final matrix is maximized when the input vectors are colinear. The diagonal elements directly return the variance scaled by n, $\Sigma_{j=1}^{n} x_{ij}^2$, for the column vectors (assuming $E[X_i] = 0$). The off-diagonal elements return the scaled covariance $\Sigma_{j=1}^{n} x_{1j} x_{2j}$. The variance-covariance matrix contains all relevant information about how the columns in our matrix are related to each other. [6]

In the case of PCA, our goal is to define a lower-dimensional basis that allows us to retain as much of the original variance as possible. [3] That is, we want to maximize the variance of each coordinate z_i for $z \in \mathbb{R}^n$ for $i = 1, 2, \ldots, n$. Assuming we have identified the eigenbasis A for our input features X, a full

vector in the codomain is given by the following:

$$z_n = A^T x_n = \begin{bmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_n^T \end{bmatrix} x_n \tag{11}$$

The first scalar coordinate z_{1n} of the transformed x_n is given by $z_{1n} = a_1^T x_n$. We can maximize the variance in the first coordinate across all vectors x_i for $i = 1, 2, \ldots, n$ by maximizing the following expression:

$$\begin{aligned} \sigma_1 &= \frac{1}{n}\Sigma_{i=1}^n z_{1n}^2 && (12)\\ &= \frac{1}{n}\Sigma_{i=1}^n (a_1^T x_i)^2 && (13)\\ &= \frac{1}{n}\Sigma_{i=1}^n a_1^T x_i x_i^T a_1 && (14)\\ &= a_1^T (\frac{1}{n}\Sigma_{i=1}^n x_i x_i^T) a_1 && (15)\\ & && (16) \end{aligned}$$

We normalize a_1 to length 1 so that only its direction affects $a_1^T S a_1$ and not its magnitude.

But what is $\frac{1}{n}\Sigma_{i=1}^n x_i x_i^T$? It is the variance of the entire input feature set, X, which we will label S. Therefore, we are seeking some vector a_1 that maximizes $\sigma_1 = a_1^T S a_1$. Now we can picture a normal vector a_1 sweeping about in n-dimensional space, all the while being projected onto the vectors x_i for $i = 1, 2, \ldots, n$. That projection is maximized when a_1 and all x_i are most colinear in aggregate, and correspondingly, $a_1^T S a_1$ is also maximized. We can use the maximization of $a_1^T S a_1$ to *identify the direction of greatest variation*, which now reduces to straightforward constrained optimization problem: max $a_1^T S a_1$ s.t. $||a_1||^2 = 1$. To resolve it, we need the partial derivatives of the following Lagrangian:

$$\mathcal{L}(a_1, \lambda) = a_1^T S a_1 + \lambda_1 (1 - a_1^T a_1) \tag{17}$$

First for a_1 ...

$$\begin{aligned} \frac{\delta \mathcal{L}}{\delta a_1} &= 2a_1 S - 2\lambda a_1 = 0 && (18)\\ \rightarrow a_1^T S &= \lambda a_1^T && (19)\\ (a_1^T S)^T &= (\lambda a_1^T)^T && (20)\\ S a_1 &= \lambda a_1 && (21) \end{aligned}$$

... and then λ:

$$\frac{\delta \mathcal{L}}{\delta \lambda} = 1 - a_1^T a_1 \tag{22}$$

$$\rightarrow a_1^T a_1 = 1 \tag{23}$$

Upon resolution of the Lagrangian, it becomes clear that a_1 is an eigenvector of the variance-covariance matrix, S, of X. The constraint term, λ, is the associated eigenvalue. Furthermore, it can be seen that $\sigma_1 = a_1^T S a_1 = a_1^T \lambda_1 a_1 = \lambda_1 (a_1^T a_1) = \lambda_1$, so the eigenvalue is also the associated variance of X in the direction of the eigenvector a_1, *our first principal component.*

2.3 Singular Value Decomposition

Singular Value Decomposition: Every $p \times q$ matrix X can be decomposed as $X = U\Sigma V^T$ where U is $p \times p$ and unitary, V is $q \times q$ and unitary, and Σ is $p \times q$ and diagonal.

We have established that principal components of a given matrix X are the eigenvectors of the associated variance-covariance matrix, but we still need a reliable method of extracting those eigenvectors: singular value decomposition. Given SVD, the following is implied:

$$S = X^T X = (U\Sigma V^T)^T (U\Sigma V^T) \tag{24}$$

$$= (V\Sigma^T U^T)(U\Sigma V^T) \tag{25}$$

$$= V\Sigma^T \Sigma V^T \tag{26}$$

$$\rightarrow \Sigma^T \Sigma = V^T (X^T X) V \tag{27}$$

$$= V^T S V \tag{28}$$

The diagonalization of S into $D = \Sigma^T \Sigma$, which must contain real, non-negative values along the diagonal, implies that V is comprised of the eigenvectors of S. [5] To be precise, each eigenvector $v_1, v_2, \ldots v_q \in V$ and associated eigenvalue $\sigma_1, \sigma_2, \ldots, \sigma_q \in \Sigma$ represent a principal component of X and the associated variance, respectively. The components are ordered by the magnitude of the variance, with the largest being defined as the first principal component, insofar as it captures the direction of most variance in X.

2.4 Dimension Reduction

Once we have our ordered set of principal components, a choice must be made about tolerable error. The full set contains all of the variance information in the input data matrix X, but the amount of variation in each subsequent component decreases. We can choose a subset of $m < n$ components to approximate the information contained in X, and that process of subset selection is how we achieve dimension reduction. The variance contained in the first m principal components is simply the sum of the associate variances.

$$\sigma_m = \Sigma_{i=1}^{m} \lambda_i \tag{29}$$

The variance lost is that contained in the $n - m$ components:

$$\sigma_{\text{lost}} = \Sigma_{i=m+1}^{n} \lambda_i \tag{30}$$

Once this selection takes place, we transform the data from the codomain back to the domain by multiplying the reduced data by V^{-1}, which is just V^T since V is unitary.

3 PCA and Linear Regression

Given a set of data X and observed outcomes Y one can find the least squares solution β for the system $X\beta = Y$ so long as X is of full rank:

$$X\beta = Y \tag{31}$$

$$X^T X\beta = X^T y \tag{32}$$

$$\beta = (X^T X)^{-1} X^T y \tag{33}$$

The **inner product** $\langle \cdot, \cdot \rangle$ between two real-valued vectors is just the **dot product**: $\langle x_1, x_2 \rangle = \Sigma_{i=1}^{n} x_{1i} x_{2i}$.

This approach to approximation has wide application in social science, but when dealing with real data, ideal conditions are often not met. Among other considerations, one may have a variety of measures that are related to each other, thereby violating the full rank assumption. PCA is one approach to reducing the number of columns needed in the regression specification while also removing any existing collinearity across columns. Specifically, we can find the principal components of the input data X and use them to calculate our new observations in the regression. Each new column z_i would be some linear combination of the input columns $z_i = \gamma_1 a_1 + \gamma_2 a_2 + \cdots + \gamma_q a_q$. Since $\langle z_i, z_j \rangle = 0 \ \forall \ i \neq j$, the regression resolves to the sum of univariate regressions of b on each principal component z_i. [4]

$$b = \Sigma_{i=1}^{m} \frac{\langle z_i, b \rangle}{\langle z_i, z_i \rangle} z_i \tag{34}$$

PCA regression of this sort can be useful for both prediction and estimation of relationships. With respect to the latter, however, we want some way of recovering the importance of the original inputs in X. The score for each input factor x_i can be recovered by projecting it onto the principal components to determine the extent to which it covaries with them. [1]

4 Supervised Principal Components

As discussed above, a primary advantage of PCA is the ability to decide which components are most useful for the estimation in question. Quite frequently, we are concerned with only those components that capture most of the variation in the input matrix X, so we select the top m components sorted by their associated singular values. However, SPC envisions a different scenario in which our

ability to observe all relevant inputs is impaired in some way. The hypothetical grounding in Section 1.1 is one such scenario, insofar as we could not directly observe job loss. Reconsider the initial model:

$$Y = \beta_0 + \beta_1 U + \epsilon = \begin{bmatrix} 1 & U \end{bmatrix} B + \epsilon \tag{35}$$

$$X_j = \alpha_{0j} + \alpha_{1j} U + \eta_j = \begin{bmatrix} 1 & U \end{bmatrix} A + \eta_j, \quad \forall \ j \in \mathcal{P} \tag{36}$$

In this system, all other observed regressors $X_k \ \forall \ k \neq j$ are independent of U. The set of regressors $\mathcal{P}$ can be thought of as indirect observations of U, and can therefore be used to estimate its value so long as we can isolate the set. If we can estimate U, we can use the $\hat{U}$ to estimate Y.

4.1 Why Subset the Feature Set?

In many cases we use PCA regression to avoid collinearity issues and increase the efficiency of our estimation by simply regressing Y on the principal components of the entire input set X. Screening out some of the features prior to PCA is a notable departure. The key consideration is that PCA on input features does not incorporate any information about the relationship between Y and X, which means there is no guarantee that the first principal component will have any relationship to our response variable of interest Y or the underlying latent variable U. Insofar as we seek to explicitly model our latent variable for use in estimating Y, subsetting the feature set allows us to isolate the features that enable estimation of $\hat{U}$. We are also able to identify the principal components that are most related to Y itself.

4.2 The Supervised Principal Components Procedure

The algorithm proceeds as follows [1]:

1. Compute the univariate standard regression coefficients for each feature in the input data matrix A.
2. Identify a threshold value θ for comparison with the regression coefficients from the previous step. The set of regressors in $\hat{\mathcal{P}}$, our estimate of $\mathcal{P}$, is comprised of the features that have regression coefficients greater than θ.
3. Perform SVD on the subset of X in $\hat{\mathcal{P}}$, X_θ. The largest principal component $u_{\theta 1}$ is $\hat{U}$, our estimate of U.
4. Fit the original regression with $\hat{U}$ and estimate Y.
5. Calculate the importance scores for each $x_j \in X_\theta$ as the inner product between each feature and $\hat{U}$. The larger the score, the greater the contribution to the prediction of Y.

Once the features x_j with the highest importance scores are identified, they can be used as inputs to policy design or downstream analysis.

5 Conclusion

SPC can be a powerful approach to estimating a wide range of models. The flexibility to restrict the feature set provides a structured way for researchers to address latent factors that might be otherwise obscured with conventional PCA regression.

That being said, there are a couple notable limitations to consider when one seeks to deploy the approach. First, Bair et al [1] do not offer a theoretical grounding to ensure that the univariate regression screening approach for selecting a subset of features will always select the "correct" features. It cannot, for example, be guaranteed that they are all functions of the latent variable U (even if selection tends to work well in practice). Second, the procedure largely rests on the assumption that marginal dependence of the response on individual features implies marginal dependence on the joint distribution of features, and vice versa. Should this assumption be violated in some way, the procedure would fail.

Despite these limitations, SPC can be a formidable tool. The utility of the approach lies in the flexibility of the general linear model, and the incorporation of unsupervised learning (via PCA) to identify feature relationships that can be used to model latent features that cannot be observed. Just as important, the procedure is quite straightforward, which makes it easy to deploy in practice.

References

[1] Eric Bair, Trevor Hastie, Debashis Paul, Robert Tibshirani. *Prediction by Supervised Principal Components*, 2006, Journal of the American Statistical Association, Vol. 101, No. 473.

[2] Alexandre Belloni, Victor Chernozhukov, Christian Hansen. *High-Dimensional Methods and Inference on Structural and Treatment Effects*, 2014, Journal of Economic Perspectives, Vol. 28, No. 2.

[3] Marc Peter Deisenroth, A.Aldo Faisal, Cheng Soon Ong. *Mathematics for Machine Learning*, 2020, Pre-publication Offering, Cambridge University Press.

[4] Trevor Hastie, Robert Tibshirani, Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2009, Second Edition, Springer Series in Statistics.

[5] Ben Noble, James W. Daniel. *Applied Linear Algebra*, 1988, Third Edition, Prentice-Hall.

[6] Alvin C. Rencher, G. Bruce Schaalje. *Linear Models in Statistics*, 2008, Second Edition, John Wiley & Sons, Inc.

[7] Thomas S. Shores. *Applied Linear Algebra and Matrix Analysis*, 2018, Second Edition, Springer International Publishing.