THEORY AND TECHNIQUES OF DIMENSIONALITY REDUCTION

W. RYAN LEE
CS109/AC209/STAT121 ADVANCED SECTION
INSTRUCTORS: P. PROTOPAPAS, K. RADER
FALL 2017, HARVARD UNIVERSITY

As we saw in the last chapter, one method of dealing with high-dimensional data (that is, when $p \approx n$ or even $p > n$) is regularization. In particular, by using the LASSO estimator, one can simultaneously conduct variable selection and regression by choosing appropriate values for the regularization parameter λ. The elastic net method allows us to combine these strengths with the shrinkage properties of the ridge regressor as well, yielding a fairly robust technique for high-dimensional inference.

While this does allow us to analyze and perform regressions on high-dimensional data, it seems somewhat *naive* in the following sense. Suppose that p is large, whether or not relative to n. Then, the LASSO estimator, for example, would select some $p' < p$ predictors with an appropriate choice of λ. However, it is not at all clear that the chosen p' predictors are the "appropriate" variables to consider in the problem. This may be clearer in light of an example.

Example 0.1. Consider the spring system depicted in **Figure 1**, where, for simplicity, we assume no mass, friction, or air resistance. By understanding the physics of the problem, it is clear that there is only one degree of freedom in the system, which is indicated by the x-axis. However, *a priori*, we may not know this is this case, and therefore measure the position of the ball attached to the spring from, say, three arbitrary angles. This is depicted by the three cameras A, B, C; denote these measured variables as x_A, x_B, x_C respectively. Let us also measure the pressure on the spring, which can be obtained by considering the weight of the spring against the wall. Denote this value as Y.

Suppose that we conduct LASSO regression on this problem, namely

$$Y = \beta_A x_A + \beta_B x_B + \beta_C x_C$$

By sheer luck, it turns out that the values x_A measured by camera A are closest to the true underlying degree of freedom (along the x-axis), and so the LASSO estimator selects x_A and sets $\hat{\beta}_B = \hat{\beta}_C = 0$. Yet scientifically, this is an unsatisfactory conclusion; we would like to be able to discern the true degree of freedom as the predictor, not simply select one of the arbitrary directions we decided to take measurements in. □

In a similar vein, when faced with a dataset with a number or dimensions of predictors, we may suspect that the data actually lie on a lower-dimensional manifold; in the same sense that three measurements were necessary to situate the ball on a spring in the example above, but the data only had one true degree of freedom.

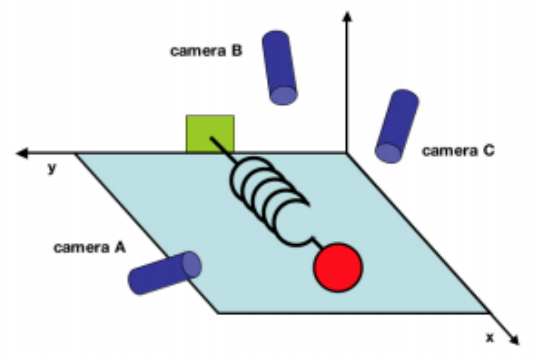

FIGURE 1. Toy example of experimenting on a spring system, taken from Shlens (2003) [3].

Thus, rather than variable selection methods such as LASSO, we may want to consider more sophisticated techniques for learning the intrinsic dimensionality of the data, a field known as **dimensionality reduction** or **manifold learning**.

1. Preliminaries in Linear Algebra and Statistics

The above example and discussion serve to motivate the introduction of **principal component analysis**, or **PCA**. Before we discuss PCA, however, we present some necessary preliminary results from linear algebra and statistics.[1]

1.1. Linear Algebra.

For this section, let X denote an arbitrary $n \times p$ matrix of real numbers, $X \in \mathbb{R}^{n \times p}$. We assume that the reader is familiar with the basic matrix computations, such as matrix multiplication, transpose, row reduction, and eigenvalue/eigenvector determination.

Proposition 1.1. *For any such matrix X, the matrices $X^T X$ and XX^T are symmetric.*

Proof. To show symmetry of a matrix A, it suffices to show that $A^T = A$. Clearly, this holds in our case, since

$$(X^T X)^T = X^T (X^T)^T = X^T X$$

and similarly for XX^T. □

The above proposition, while simple, will prove very useful due to an attractive property of real, symmetric matrices as given in the following theorem. Indeed, the following is often considered the fundamental theorem of linear algebra, and is known as the **spectral theorem**.

Theorem 1.2. *If A is a real, symmetric matrix, then there exists an orthonormal basis of eigenvectors of A.*

[1] Much of the presentation follows that of Jauregui (2012) [2].

In order words, for any such matrix $A \in \mathbb{R}^{m \times m}$, we can find a basis $\{v_1, \dots, v_n\}$ such that the basis is **orthonormal**, which means that the basis vectors are orthogonal ($v_i \perp v_j$, so that $v_i^T v_j = 0$) and normalized to unity ($\|v_i\|_2 = 1$). Moreover, this basis consists of **eigenvectors** of A, so that $Av_i = \lambda_i v_i$ for $\lambda_i \in \mathbb{R}$.

Alternatively, if we stack the eigenvectors as rows, we obtain the matrix U^T, and we can express the **eigendecomposition** of A as

$$A = U \Lambda U^T$$

where $\Lambda = \text{diag}(\lambda_i)$ is the diagonal matrix of eigenvalues, and U is an orthogonal matrix (so that $U^T = U^{-1}$).

The proof of the theorem is quite technical, and we state the theorem here without proof. Moreover, there is a considerable amount of theory involving the set of eigenvalues of A, which is called its **spectrum**. The spectrum of a matrix reveals much about its properties, and though we do not delve into it here, we encourage the reader to refer to the bibliography for further details.

We can, however, discuss one important property of the spectrum for the Gram matrices $X^T X$ and related XX^T; namely, that the eigenvalues are nonnegative.

Proposition 1.3. *The eigenvalues of $X^T X$ and XX^T are nonnegative reals.*

Proof. Suppose λ is an eigenvalue of $X^T X$ with associated eigenvector v. Then,

$$\|Xv\|_2^2 = (Xv)^T (Xv) = v^T (X^T X v) = \lambda v^T v = \lambda \|v\|_2^2$$

Since both $\|Xv\|_2^2, \|v\|_2^2 \geq 0$, we must have $\lambda \geq 0$. The result for XX^T follows from a similar proof using X^T instead of X. □

In fact, it turns out that the nonzero eigenvalues of these matrices are identical, as the following Proposition shows.

Proposition 1.4. *The matrices XX^T and $X^T X$ share the same nonzero eigenvalues.*

Proof. Suppose that λ is a nonzero eigenvalue of XX^T with associated eigenvector v. Then

$$XX^T v = \lambda v \Rightarrow (X^T X) X^T v = \lambda X^T v$$

Thus, λ is also an eigenvalue of $X^T X$, with associated eigenvector $X^T v$ rather than v. Moreover, $X^T v \neq 0$, since otherwise $XX^T v = 0$, which would imply that $\lambda = 0$, which contradicts our assumption. A similar proof concludes the proof for $X^T X$. □

1.2. **Statistics.** In this section, we return to considering $X \in \mathbb{R}^{n \times p}$ as the model matrix. From this point on, we assume that the predictors are all *centered*, which means that for each column X_j of X, we subtract the sample column mean

$$\hat{\mu}_j = n^{-1} \sum_{i=1}^{n} x_{ij}$$

so that we are considering the centered model matrix

$$\tilde{X} = \begin{pmatrix} X_1 - \hat{\mu}_1 & X_2 - \hat{\mu}_2 & \cdots & X_p - \hat{\mu}_p \end{pmatrix}$$

Note that each column now has expectation zero, so that we can consider the **sample covariance matrix**

$$S \equiv (n-1)^{-1} \tilde{X}^T \tilde{X}$$

This is a modified Gram matrix, using the centered columns (or predictors) and scaling by $n-1$. One way to understand the origin of the name is to consider each of the terms in the matrix. The diagonal terms all have the form

$$S_{jj} = (n-1)^{-1} \sum_{i=1}^{n} (x_{ij} - \hat{\mu}_j)^2$$

whereas the off-diagonal terms have the form

$$S_{jk} = (n-1)^{-1} \sum_{i=1}^{n} (x_{ij} - \hat{\mu}_j)(x_{ik} - \hat{\mu}_k)$$

Thus, it is clear that the diagonal terms yield the sample variances of each of the predictors, whereas the off-diagonal terms yield the sample covariances.

2. Principal Component Analysis

With the above preliminaries, the actual methodology of PCA is now quite simple. The main idea is that in order to conduct dimensionality reduction and obtain the irreducible degrees of freedom inherent in the problem, we would like to remove as much redundancy in our predictors as possible. The way that PCA defines such redundancy is by using the correlation (or covariance) between the predictors. For instance, if predictors x_j and x_k are highly correlated, it is likely that one holistic predictor may suffice instead.

Proceeding to the mathematics, we first use **Proposition 1.1** to note that the sample covariance matrix S is a symmetric matrix, and thus we can apply **Theorem 1.2** to obtain an orthonormal basis of eigenvectors of S, such that the eigenvalues are ordered $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, with corresponding eigenvectors $u_1, \ldots, u_p$.

The vector u_i is called the i^{th} **principal component** of S, and λ_i is a measure of the "variance explained" by that principal component. This is because the trace of S,

$$\text{trace}(S) \equiv \sum_{j=1}^{p} S_{jj} = (n-1)^{-1} \sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - \hat{\mu}_j)^2$$

can be considered the "total sample variance" of the predictors, as it sums up the sample variances of each of the p predictor variables. But the trace of S also equals the sum of its eigenvalues

$$\text{trace}(S) = \sum_{j=1}^{p} \lambda_j$$

Moreover, as demonstrated by **Proposition 1.3**, all of the eigenvalues of S are nonnegative. Thus, $\lambda_i / \sum_j \lambda_j = \lambda_i / \text{trace}(S)$ represents, in a heuristic sense, the fraction of the "total sample variance" accounted for by the eigenvector or principal component u_i.

In general, it will often be the case that the largest eigenvalues are orders of magnitude greater than the others, because the data may indeed have fewer degrees of freedom than the number of predictors may indicate. In practice, one keeps only the principal components with the largest eigenvalues, and discards the rest, thereby reducing the dimension of the problem, as shown in **Figure 2**. Thus, a smaller subset of the eigenvalues being significantly larger than the others indicates the possibility of dimensionality reduction. How many components to keep is left to the

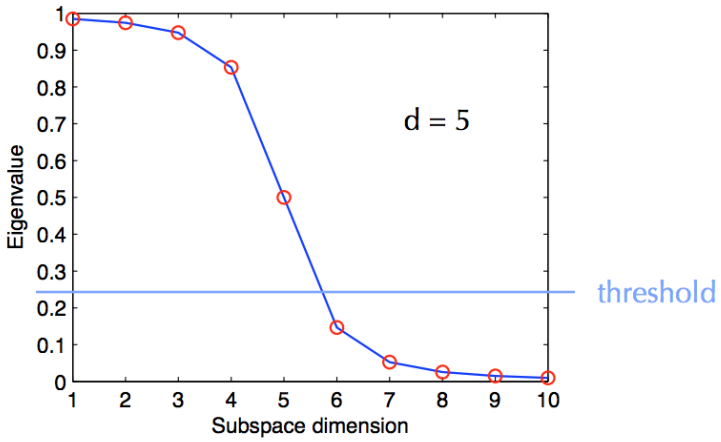

FIGURE 2. An example of dimensionality reduction by PCA, thresholding the eigenvectors to keep based on examination of the eigenvalue magnitudes.

data analyst's discretion, but it is generally clear (when dimensionality reduction is possible).

Intuitively, the principal components u_i denote directions in $\mathbb{R}^n$ that are "natural" for the problem at hand, and are linear combinations of the original coordinates. For example, in the spring system example, we may have $u_1 = (0.9, 0.2, 0.4)$ as the first principal component, which may have $\lambda_1 / \sum_j \lambda_j \approx 1$, as it represents the x-axis. Consequently, the possibility of dimensionality reduction also indicates that there may be fewer but more interpretable variables, represented by the principal components, that are responsible for the variability of a response.

3. Assumptions of Principal Component Analysis

There are a number of assumptions that were both implicitly and explicitly made in order to motivate and justify the PCA method described in the previous section.

A. *Linear change of basis.* Note that all of the operations above are linear operations; indeed, PCA consists essentially of a change of basis, from the Euclidean basis (in which we measure our predictors) to an orthonormal basis of eigenvectors of $X^T X$. Thus, PCA assumes that such a linear change of basis is sufficient for identifying degrees of freedom and conducting dimensionality reduction.

B. *Mean/variance is sufficient.* In applying the PCA technique to our data, we only using the means (for standardizing) and covariance matrix associated with our predictors. Thus, the method assumes that such statistics are *sufficient* for describing the distributions of the predictor variables. This is, in fact, only the case if the predictors are drawn jointly from a multivariable Normal distribution, but may be approximately true in other situations. However, when the predictor distributions heavily violate this sufficiency assumption, one can still conduct PCA, but the resulting components may not be as informative.

C. *High variance indicates importance.* Another fundamental assumption we made when describing the PCA procedure is that the eigenvalues λ_i, which represent the variability in the data "explained by" or associated with with the i^{th} principal component, measure the importance of that component. This is intuitively reasonable, since components corresponding to low variability likely say little about the data, it need not always be true.
D. *Principal components are orthogonal.* When conducting PCA, we explicitly sought orthonormal eigenvectors as our principal components. We did not need to make such an assumption, and it may not be true that the "intrinsic dimensions" are orthogonal. However, this allowed us to use techniques from linear algebra such as the spectral decomposition, and thereby simplify our calculations.

Thus, while most of the assumptions appear plausible, they must be checked in practice before drawing any strong conclusions from PCA. Let us assess which assumptions are fundamental and which are technical. Assumption A is inherent in PCA, as a matrix-based method. Unfortunately, it is also one of the most limiting aspects of PCA. If the data are confined to a subspace, then linear methods will suffice. However, if the data are on some (nonlinear) manifold in the space, as put forth by the **manifold hypothesis**, then linear methods are doomed to fail in general, and we must turn to nonlinear methods (as we do in Section 5).

Assumption B can be problematic, but unlike Assumption A, it can be more easily verified. For example, if any of the predictors appear to be heavily skewed, then the first two moments (mean and variance) are likely insufficient to describe the distribution, and thus PCA may not be very informative. In such a case, a transformation of certain problematic predictor variables (for example, by taking the logarithm) can be an adequate solution. Of course, one should ideally examine the *joint* distribution of the predictors, but this can be difficult in high-dimensional situations.

Finally, Assumptions C and D are not necessarily data-dependent, but rather method-dependent: that is, we make these assumptions as a way to understand the data, and they are not intrinsic to the data itself. Using metrics other than variability and allowing non-orthogonal components are not inherently nonsensical or antithetic to PCA; they will simply yield different methods and solutions to the problem of dimensionality reduction.

4. Multidimensional Scaling and Other Linear Dimensionality Reduction Methods

As noted above, PCA is a *linear* dimensionality reduction method that is based on a certain objective (maximizing variances and minimizing covariances), and substituting other metrics to be optimized yield different methods.[2] Rather than maximizing variances, one may want to instead find lower-dimensional representations of X that preserve the *pairwise distances* between the observations. This leads to the method of **multidimensional scaling**, or **MDS**.

As usual, suppose that we have n observations $\{x_1, \ldots, x_n\| \subset \mathbb{R}^p$, each of which are p-dimensional. Also, define a distance function between observations $d_{ij} = d(x_i, x_j)$, such that it is a *metric*. Namely, it is symmetric ($d_{ij} = d_{ji}$), and has the

[2]For more information regarding methods in this section and the next, we refer the reader to Saul *et al.* (2005) [1].

property that $d_{ii} = 0$ and $d_{ij} > 0$ if $i \neq j$. Often, we will consider the Euclidean distance as our metric, so that $d(x_i, x_j) = \|x_i - x_j\|_2^2$. One can verify that the Euclidean distance satisfies all properties necessary to be a metric.

We can the construct the *distance matrix* D, defined as

$$D \equiv \begin{pmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{pmatrix}$$

where the diagonal terms are zero by definition of the metric. In addition, we can consider a lower-dimensional representation $\{y_1, \dots, y_n\} = \{g(x_1), \dots, g(x_n)\} \subset \mathbb{R}^d$ for $d < p$, and the associated distance matrix. We refer to the original distance matrix as D^X and the distance matrix associated with the lower-dimensional representation as D^Y; note that both matrices are of dimension $n \times n$.

One criterion for ensuring that the lower-dimensional representation is faithful to the original data is to preserve the distances between the observations. Thus, in MDS, one seeks to find a representation such that

$$\min_g \sum_{i,j=1}^{n} (d_{ij}^X - d_{ij}^Y)^2$$

where g is the transformation that yields y.

There are a number of ways one can use this framework for dimensionality reduction, but here we focus on the Euclidean case. In this situation, the following lemma connects the distance matrix to the Gram matrix.

Lemma 4.1. *The distance matrix D for observations $\{x_1, \dots, x_n\}$ and Euclidean metric $d(x_i, x_j) = \|x_i - x_j\|_2^2$ satisfies*

$$XX^T = -\frac{1}{2} HDH$$

where $H = I_n - n^{-1}\mathbf{1}\mathbf{1}^T$ and $\mathbf{1}$ is the vector of all ones.

With this lemma, we can express the above minimization problem in terms of inner products as follows

$$\min_g \sum_{i,j=1}^{n} (x_i^T x_j - y_i^T y_j)^2$$

and it can be shown that the solution to this problem is given by $Y = \Lambda^{1/2} V^T$ where V is the matrix of eigenvectors corresponding to the largest d eigenvalues of XX^T, and Λ is the diagonal matrix of those eigenvalues (and zero otherwise).

However, note from **Proposition 1.4** that, in fact, the largest d eigenvalues of XX^T are exactly the largest d eigenvalues of $X^T X$. Thus, despite approaching the problem from a completely different criterion, MDS actually yields the same dimensionality reduction as PCA. Thus, it is also a linear dimensionality reduction technique (if we use Euclidean distance as our metric) and suffers from the same drawbacks and assumptions as PCA.

5. Nonlinear Dimensionality Reduction Techniques

To surmount the linearity assumptions of PCA and MDS, there are, by now, a large number and variety of *nonlinear dimensionality reduction* techniques, which are also called manifold learning methods. We focus on two salient examples of such methods, which are each based on one of the methods we have discussed.

5.1. **Kernel PCA.** One obvious extension to PCA that allows for nonlinear dimensionality reduction is to first apply a nonlinear map Φ, known as a **feature map**, to the data, yielding a nonlinear representation $\Phi(X)$, then applying PCA to this transformed data. We shall use this technique multiple times in the future. Once we transform the data, we must find the Gram matrix in this transformed space, which we define to be the **kernel**.

$$K \equiv \Phi(X)^T \Phi(X)$$

Once we have achieved this, we can conduct PCA on this Gram matrix, just taking care to ensure that the columns have mean zero. This yields the **kernel PCA** method for nonlinear dimensionality reudction.

Note that we cannot simply standardize each column as before, since that does not conform to the transformation above. Instead, we must modify the feature map itself

$$\tilde{\Phi}(X) = \Phi(X) - E_x[\Phi(X)]$$

and compute the modified kernel $\tilde{K} \equiv \tilde{\Phi}(X)^T \tilde{\Phi}(X)$.

5.2. **Isomap.** Similarly to the case of kernel PCA, one can extend MDS to the nonlinear setting by using a non-Euclidean distance metric. One widely-used alternative yields a technique called **Isomap**. The exact same MDS objective is minimized as before (minimizing the difference in pairwise distances between the original points and the transformed representation). However, we employ a different, particular distance metric $d(x_i, x_j)$.

To construct this metric, one first constructs the k-nearest neighbors (KNN) graph of the data. This entails employing the KNN method on the data, and constructing a graph in which the data points are the nodes, and an undirected edge $\{i, j\}$ indicates that x_i, x_j are one of each other's k-nearest neighbors. Then, one can use a shortest-paths algorithm (such as Djikstra's algorithm) to compute the shortest *geodesic distance* between pairs of observations. That is, $d_{ij} = d(x_i, x_j)$ indicates the length of the shortest path between x_i and x_j in this nearest neighbors graph.

Finally, one can use a standard optimization algorithm or an eigendecomposition of the distance matrix D^X to find the representations Y. This step is identical to that of MDS, and it is noted that one can use the number of "large" eigenvalues of D^X to determine the dimensionality of the representation.

References

[1] Saul, L. K., et al. (2006). "Spectral Methods for Dimensionality Reduction." *Semisupervised Learning*: 293-308.
[2] J. Jauregui (2012). "Principal Component Analysis with Linear Algebra."
[3] J. Shlens (2003). "A Tutorial on Principal Component Analysis."