PCA

Liangjiayi Wang

November 2022

1 ch6

Principal components analysis (PCA) is a popular approach for deriving a low-dimensional set of features from a large set of variables. PCA is a technique for reducing the dimension of an n × p data matrix X. The first principal component direction of the data is that along which the observations vary the most.

Another interpretation for PCA: the first principal component vector defines the line that is as close as possible to the data.

1.1 Example 1: Graphical

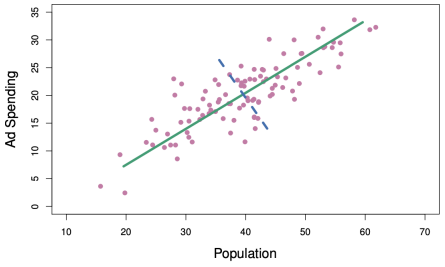

FIGURE 6.14. *The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.*

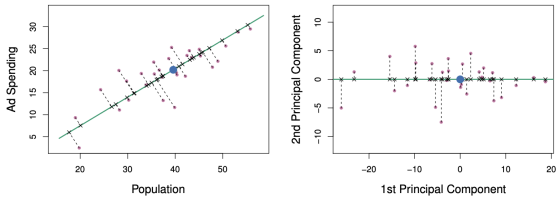

FIGURE 6.15. *A subset of the advertising data. The mean pop and ad budgets are indicated with a blue circle.* Left: *The first principal component direction is shown in green. It is the dimension along which the data vary the most, and it also defines the line that is closest to all n of the observations. The distances from each observation to the principal component are represented using the black dashed line segments. The blue dot represents (pop, ad).* Right: *The left-hand panel has been rotated so that the first principal component direction coincides with the x-axis.*

Figure 6.14 shows population size (pop) in tens of thousands of people, and ad spending for a particular company (ad) in thousands of dollars, for 100 cities. The green solid line represents the first principal component direction of the data. We can see by eye that this is the direction along which there is the greatest variability in the data. That is, if we projected the 100 observations onto this line (as shown in the left-hand panel of Figure 6.15), then the resulting projected observations would have the largest possible variance; projecting the observations onto any other line would yield projected observations with lower variance. Projecting a point onto a line simply involves finding the location on the line which is closest to the point.

For instance, in Figure 6.14, the first principal component line minimizes the sum of the squared perpendicular distances between each point and the line. These distances are plotted as dashed line segments in the left-hand panel of Figure 6.15, in which the crosses represent the projection of each point onto the first principal component line. The first principal component has been chosen so that the projected observations are as close as possible to the original observations.

1.2 Example 1: Math

$$Z_1 = 0.839 * (pop - p\bar{o}p) + 0.544 * (ad - \bar{ad})$$

Here $\phi_{11} = 0.839$ and $\phi_{21} = 0.544$ are the principal component loadings, which define the direction referred to above. The idea is that out of every possible linear combination of pop and ad such that $\phi_{11}^2 + \phi_{21}^2 = 1$. This particular linear combination yields the highest variance: i.e. this is the linear combination for which $Var(\phi_{11}^2 * (pop - p\bar{o}p) + \phi_{21}^2 * (ad - \bar{ad}))$ is maximized.

$$z_{i1} = 0.839 * (pop_i - p\bar{o}p) + 0.544 * (ad_i - \bar{ad})$$

where the values of $z_{11}, ..., z_{n1}$ are known as the principal component scores.

2 Introduction

PCA finds a low-dimensional representation of a data set that contains as much as possible of the variation. The idea is that each of the n observations lives in p-dimensional space, but not all of these dimensions are equally interesting. PCA seeks a small number of dimensions that are as interesting as possible, where the concept of interesting is measured by the amount that the observations vary along each dimension. Each of the dimensions found by PCA is a linear combination of the p features. We now explain the manner in which these dimensions, or principal components, are found.

3 First principle component

The first principal component of a set of features X1, X2, . . . , Xp is the normalized linear combination of the features

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2....\phi_{pi}X_p$$

that has the largest variance. By normalized, we mean that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$ We refer to the elements $\phi_{11}, \phi_{21}...\phi_{p1}$ as the loadings of the first principal loading component; together, the loadings make up the principal component loading vector, $\phi_1 = (\phi_{11}, \phi_{21}...\phi_{p1})^T$.

Given a n × p data set X, we assume that each of the variables in X has been centered to have mean zero (that is, the column means of X are zero).

$$z_{i1} = \phi_{11}X_{i1} + \phi_{21}X_{i2}....\phi_{pi}X_{ip}$$

that has the largest sample variance, subject to $\sum_{j=1}^{p} \phi_{j1}^2 = 1$

We have

$$\underset{\phi_{11},...,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1.$$

which equals $\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2$ since $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$, and avg of z11....zn1 will be zero. Hence the objective that we are maximizing in is just the sample variance of the n values of z_{i1}. We refer to z11, . . . , zn1 as the scores of the first principal component.

The loading vector ϕ_1 with elements $\phi_{11}, \phi_{21}...\phi_{p1}$ defines a direction in feature space along which the data vary the most. If we project the n data points x1,...,xn onto this direction, the projected values are the principal component scores z11, . . . , zn1 themselves.

4 Second principle component

The second principal component is the linear combination of $X_1, ..., X_p$ that has maximal variance out of all linear combinations that are uncorrelated with Z_1. The second principal component scores z12, z22, . . . , zn2 take the form

$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + ... + \phi_{p2}x_{ip}$$

where $\phi 2$ is the second principal component loading vector, with elements $\phi_{12}, \phi_{22}, ..., \phi_{p2}$. It turns out that constraining Z_2 to be uncorrelated with Z_1 is equivalent to constraining the direction ϕ_2 to be orthogonal (perpendicular) to the direction ϕ_1. To find ϕ_2, we solve a problem similar to above with ϕ_2 replacing ϕ_1, and with the additional constraint that ϕ_2 is orthogonal to $\phi 1$

There are at most min(n - 1, p) principal components.

5 Key concept

of PC loading vectors have length $\# \ of \ X + 1(\# \ of \ Y)$

PCA performs after standardizing each variable to have mean 0 and standard deviation 1

6 Example: US Arrests

We illustrate the use of PCA on the USArrests data set. For each of the 50 states in the United States, the data set contains the number of arrests per 100, 000 residents for each of three crimes: Assault, Murder, and Rape. We also record UrbanPop (the percent of the population in each state living in urban areas). The principal component score vectors have length n = 50, and the principal component loading vectors have length p = 4. PCA was performed after standardizing each variable to have mean zero and standard deviation one.

PC scores: Blue state names
PC loading vectors: the orange arrow

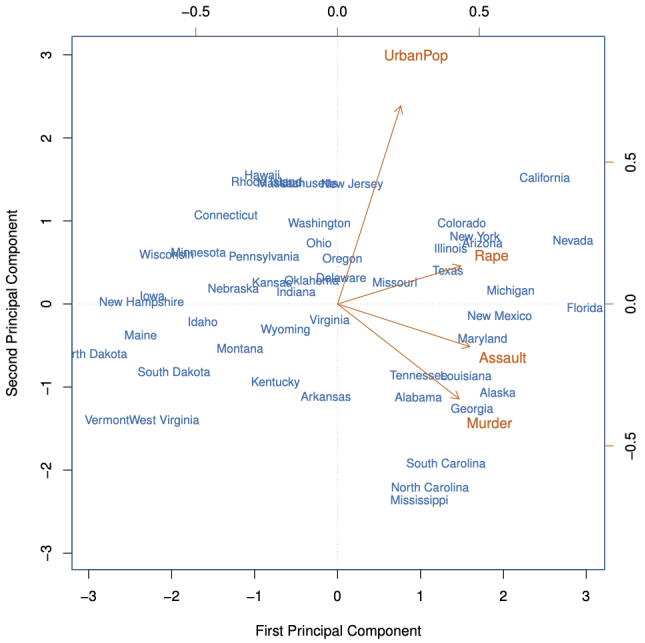

FIGURE 12.1. *The first two principal components for the* `USArrests` *data. The blue state names represent the scores for the first two principal components. The orange arrows indicate the first two principal component loading vectors (with axes on the top and right). For example, the loading for* `Rape` *on the first component is* 0.54*, and its loading on the second principal component* 0.17 *(the word* `Rape` *is centered at the point* $(0.54, 0.17)$*). This figure is known as a biplot, because it displays both the principal component scores and the principal component loadings.*

6.1 Interpretation

1. First loading vector places approximately equal weight on Assault, Murder, and Rape, but with much less weight on UrbanPop. Hence this component roughly corresponds to a measure of overall rates of serious crimes.

2. The second loading vector places most of its weight on UrbanPop and much less weight on the other three features. Hence, this component roughly corresponds to the level of urbanization of the state.

3. Overall, we see that the crime-related variables (Murder, Assault, and Rape) are located close to each other, and that the UrbanPop variable is far from the other three. This indicates that the crime-related variables are correlated with each other—states with high murder rates tend to have high assault and rape rates—and that the UrbanPop variable is less correlated with the other three.

4. Our discussion of the loading vectors suggests that states with large positive scores on the first component, such as California, Nevada and Florida, have high crime rates, while states like North Dakota, with negative scores on the first component, have low crime rates.

5. California also has a high score on the second component, indicating a high level of urbanization, while the opposite is true for states like Mississippi.

6. States close to zero on both components, such as Indiana, have approximately average levels of both crime and urbanization.

7 Another Interpretation of Principal Components

Principal components provide low-dimensional linear surfaces that are closest to the observations. PCA finds hyperplane closest to the observations!!

The first principal component loading vector has a very special property: it is the line in p-dimensional space that is closest to the n observations (using average squared Euclidean distance as a measure of closeness). The appeal of this interpretation is clear: we seek a single dimension of the data that lies as close as possible to all of the data points, since such a line will likely provide a good summary of the data.

The first two principal components of a data set span the plane that is closest to the n observations, in terms of average squared Euclidean distance.

7.1 Example 1

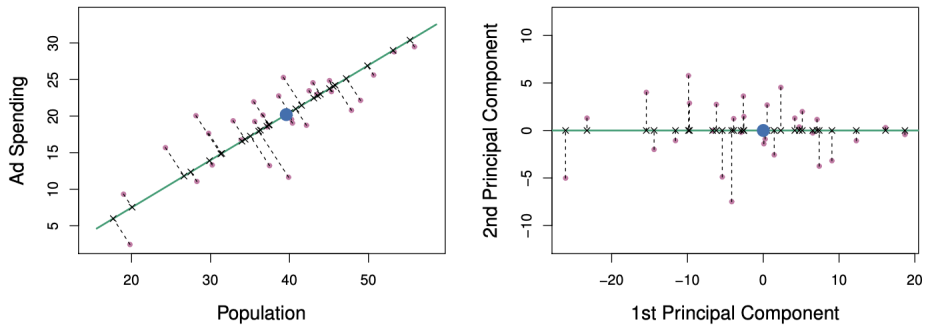

FIGURE 6.15. *A subset of the advertising data. The mean* pop *and* ad *budgets are indicated with a blue circle.* Left: *The first principal component direction is shown in green. It is the dimension along which the data vary the most, and it also defines the line that is closest to all n of the observations. The distances from each observation to the principal component are represented using the black dashed line segments. The blue dot represents* $(\overline{\texttt{pop}}, \overline{\texttt{ad}})$. Right: *The left-hand panel has been rotated so that the first principal component direction coincides with the x-axis.*

This interpretation can be seen in the left-hand panel of Figure 6.15; the dashed lines indicate the distance between each observation and the line defined by the first principal component loading vector.

7.2 Example 2

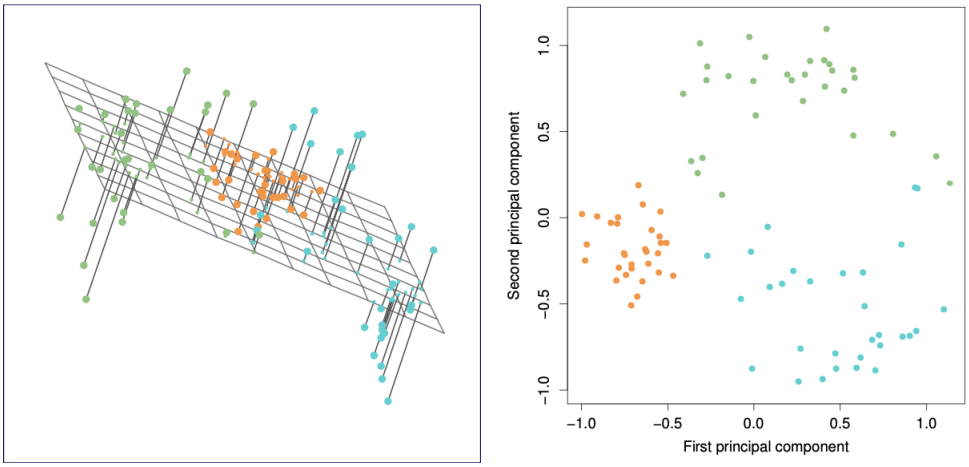

FIGURE 12.2. *Ninety observations simulated in three dimensions. The observations are displayed in color for ease of visualization.* Left: *the first two principal component directions span the plane that best fits the data. The plane is positioned to minimize the sum of squared distances to each point.* Right: *the first two principal component score vectors give the coordinates of the projection of the 90 observations onto the plane.*

The first three principal components of a data set span the three-dimensional hyperplane that is closest to the n observations, and so forth.

8 Formulas

The first M principal component score vectors and the first M principal component loading vectors provide the best M-dimensional approximation (in terms of Euclidean distance) to the ith observation x_{ij}. This representation can be written as

$$x_{ij} \approx \sum_{m=1}^{M} z_{im}\phi_{jm}$$

Suppose the data matrix X is column-centered. Out of all approximations of the form $x_{ij} \approx \sum_{m=1}^{M} a_{im}b_{jm}$, we could ask for the one with the smallest residual sum of squares(RSS) or MSE=1/n*RSS:

$$\underset{\mathbf{A}\in\mathbb{R}^{n\times M},\mathbf{B}\in\mathbb{R}^{p\times M}}{\text{minimize}} \left\{ \sum_{j=1}^{p}\sum_{i=1}^{n} \left(x_{ij} - \sum_{m=1}^{M} a_{im}b_{jm} \right)^2 \right\}. \qquad (12.6)$$

Here, A is a n×M matrix whose (i,m) element is a_{im} (pc score), and B is a p×M element whose (j, m) element is b_{jm} (pc loading vectors).

It can be shown that for any value of M, the columns of the matrices $\hat{A}$ and $\hat{B}$ that solve (12.6) are in fact the first M principal components score and loading vectors. In other words, if $\hat{A}$ and $\hat{B}$ solve (12.6), then $\hat{a}_{im} = z_{im}$ and $\hat{b}_{jm} = \phi_{jm}$. This means that the smallest possible value of the objective in (12.6) is

$$\sum_{j=1}^{p}\sum_{i=1}^{n}\left(x_{ij}-\sum_{m=1}^{M}z_{im}\phi_{jm}\right)^2.$$

In summary, together the M principal component score vectors and M principal component loading vectors can give a good approximation to the data when M is sufficiently large. When $M = min(n-1, p)$, then the representation is exact: $x_{ij} = \sum_{m=1}^{M} a_{im}b_{jm}$

9 The Proportion of Variance Explained

How much of the information in a given data set is lost by projecting the observations onto the first few principal components? That is, how much of the variance in the data is not contained in the first few principal components? More generally, we are interested in knowing the proportion of variance explained (PVE) by each principal component/PVE indicates the strength of each PC.

The total variance present in a data set (assuming that the variables have been centered to have mean zero) is defined as

$$\sum_{j=1}^{p}\mathrm{Var}(X_j)=\sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n}x_{ij}^2, \tag{12.8}$$

and the variance explained by the mth principal component is

$$\frac{1}{n}\sum_{i=1}^{n}z_{im}^2=\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jm}x_{ij}\right)^2. \tag{12.9}$$

Therefore, the PVE of the mth principal component is given by

$$\frac{\sum_{i=1}^{n}z_{im}^2}{\sum_{j=1}^{p}\sum_{i=1}^{n}x_{ij}^2}=\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jm}x_{ij}\right)^2}{\sum_{j=1}^{p}\sum_{i=1}^{n}x_{ij}^2}. \tag{12.10}$$

The PVE of each principal component is a positive quantity. In order to compute the cumulative PVE of the first M principal components, we can simply sum (12.10) over each of the first M PVEs. In total, there are min(n - 1, p) principal components, and their PVEs sum to one.

We showed that the first M principal component load- ing and score vectors can be interpreted as the best M-dimensional approx- imation to the data, in terms of residual sum of squares. It turns out that the variance of the data can be decomposed into the variance of the first M principal components plus the mean squared error of this M-dimensional approximation, as follows:

$$\underbrace{\sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n}x_{ij}^2}_{\text{Var. of data}}=\underbrace{\sum_{m=1}^{M}\frac{1}{n}\sum_{i=1}^{n}z_{im}^2}_{\text{Var. of first } M \text{ PCs}}+\underbrace{\frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n}\left(x_{ij}-\sum_{m=1}^{M}z_{im}\phi_{jm}\right)^2}_{\text{MSE of } M\text{-dimensional approximation}} \tag{12.11}$$

Since the first term is fixed, we see that by maximizing the variance of the first M principal components, we minimize the mean squared error of the M -dimensional approximation, and vice versa. This explains why principal components can be equivalently viewed as minimizing the approximation error (as in Section 12.2.2) or maximizing the variance (as in Section 12.2.1).

Moreover, we can use (12.11) to see that the PVE defined in (12.10) equals

$$1 - \frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\left(x_{ij} - \sum_{m=1}^{M} z_{im}\phi_{jm}\right)^2}{\sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij}^2} = 1 - \frac{\text{RSS}}{\text{TSS}},$$

where TSS represents the total sum of squared elements of X, and RSS represents the residual sum of squares of the M-dimensional approximation given by the principal components. Recalling the definition of R2 from (3.17), this means that we can interpret the PVE as the R^2 of the approximation for X given by the first M principal components.

The PVE of each principal component, as well as the cumulative PVE, is shown in Figure 12.3.

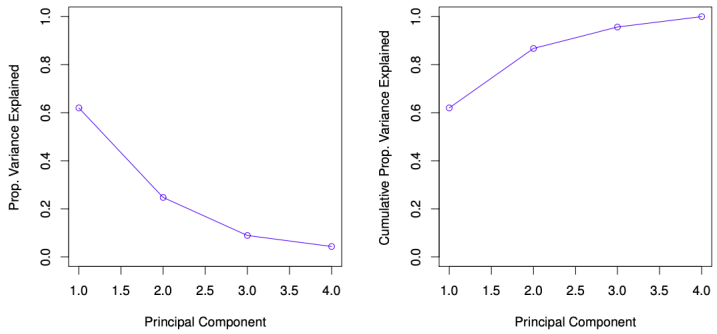

FIGURE 12.3. Left: *a scree plot depicting the proportion of variance explained by each of the four principal components in the* `USArrests` *data.* Right: *the cumulative proportion of variance explained by the four principal components in the* `USArrests` *data.*

9.1 Example: US Arrests

In the USArrests data, the first principal component explains 62.0 % of the variance in the data, and the next principal component explains 24.7 % of the variance. Together, the first two principal components explain almost 87% of the variance in the data, and the last two principal components explain only 13 % of the variance. This means that Figure 12.1 provides a pretty accurate summary of the data using just two dimensions.

10 Uniqueness of the Principal Components

1. Each principal component loading vector is unique, up to a sign flip.
2. The signs may differ because each principal component loading vector specifies a direction in p-dimensional space: flipping the sign has no effect as the direction does not change.
3. The score vectors are unique up to a sign flip, since the variance of Z is the same as the variance of -Z.

11 Deciding How Many Principal Components to Use

1. We typically decide on the number of principal components required to visualize the data by examining a scree plot
2. We choose the smallest number of principal components that are required in order to explain a sizable amount of the variation in the data.
3. This is done by eyeballing the scree plot, and looking for a point at which the proportion of variance explained by each subsequent principal component drops off.
4. For instance, by inspection of Figure 12.3, one might conclude that a fair amount of variance is explained by the first two principal components, and that there is an elbow after the second component. After all, the third principal component explains less than ten percent of the variance in the data, and the fourth principal component explains less than half that and so is essentially worthless.

12 Other Uses for Principal Components

1. Data visualization
2. EDA
3. Clustering
4. Dimension Reduction for regression/classification
5. Missing data imputation

In fact, many statistical techniques, such as regression, classification, and clustering, can be easily adapted to use the n $\times$ M matrix whose columns are the first $M \ll p$ principal component score vectors, rather than using the full n $\times$ p data matrix. This can lead to less noisy results, since it is often the case that the signal (as opposed to the noise) in a data set is concentrated in its first few principal components.

13 PLS

PLS is a dimension reduction method, which first identifies a new set of features Z1,...,ZM that are linear combinations of the original features, and then fits a linear model via least squares using these M new features. PLS identifies these new features in a supervised way—that is, it makes use of the response Y in order to identify new features that not only approximate the old features well, but also that are related to the response. Roughly speaking, the PLS approach attempts to find directions that help explain both the response and the predictors.

After standardizing the p predictors, PLS computes the first direction Z1 by setting each ϕ_{j1} equal to the coefficient from the simple linear regression of Y onto Xj. One can show that this coefficient is proportional to the correlation between Y and Xj. Hence, in computing $Z_1 = \sum_{j=1}^{p} \phi_{j1} X_j$, PLS places the highest weight on the variables that are most strongly related to the response.

14 Principal Components with Missing Values

In Section 12.2.2, we showed that the first M principal component score and loading vectors provide the "best" approximation to the data matrix $\mathbf{X}$, in the sense of (12.6). Suppose that some of the observations x_{ij} are missing. We now show how one can both impute the missing values and solve the principal component problem at the same time. We return to a modified form of the optimization problem (12.6),

$$\underset{\mathbf{A}\in\mathbb{R}^{n\times M},\mathbf{B}\in\mathbb{R}^{p\times M}}{\text{minimize}} \left\{ \sum_{(i,j)\in\mathcal{O}} \left(x_{ij} - \sum_{m=1}^{M} a_{im} b_{jm} \right)^2 \right\}, \tag{12.12}$$

where $\mathcal{O}$ is the set of all *observed* pairs of indices (i, j), a subset of the possible $n \times p$ pairs.

Once we solve this problem:

- we can estimate a missing observation x_{ij} using $\hat{x}_{ij} = \sum_{m=1}^{M} \hat{a}_{im}\hat{b}_{jm}$, where $\hat{a}_{im}$ and $\hat{b}_{jm}$ are the (i, m) and (j, m) elements, respectively, of the matrices $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ that solve (12.12); and

- we can (approximately) recover the M principal component scores and loadings, as we did when the data were complete.

Algorithm 12.1 *Iterative Algorithm for Matrix Completion*

---

1. Create a complete data matrix $\tilde{\mathbf{X}}$ of dimension $n \times p$ of which the (i, j) element equals

$$\tilde{x}_{ij} = \begin{cases} x_{ij} & \text{if } (i,j) \in \mathcal{O} \\ \bar{x}_j & \text{if } (i,j) \notin \mathcal{O}, \end{cases}$$

   where $\bar{x}_j$ is the average of the observed values for the jth variable in the incomplete data matrix $\mathbf{X}$. Here, $\mathcal{O}$ indexes the observations that are observed in $\mathbf{X}$.

2. Repeat steps (a)–(c) until the objective (12.14) fails to decrease:

   (a) Solve

$$\underset{\mathbf{A}\in\mathbb{R}^{n\times M},\mathbf{B}\in\mathbb{R}^{p\times M}}{\text{minimize}} \left\{ \sum_{j=1}^{p}\sum_{i=1}^{n} \left(\tilde{x}_{ij} - \sum_{m=1}^{M} a_{im} b_{jm} \right)^2 \right\} \tag{12.13}$$

   by computing the principal components of $\tilde{\mathbf{X}}$.

   (b) For each element $(i, j) \notin \mathcal{O}$, set $\tilde{x}_{ij} \leftarrow \sum_{m=1}^{M} \hat{a}_{im}\hat{b}_{jm}$.

   (c) Compute the objective

$$\sum_{(i,j)\in\mathcal{O}} \left(x_{ij} - \sum_{m=1}^{M} \hat{a}_{im}\hat{b}_{jm} \right)^2. \tag{12.14}$$

3. Return the estimated missing entries $\tilde{x}_{ij}$, $(i, j) \notin \mathcal{O}$.

In Netflix Problem, $\hat{a}_{im}$ represents the strength of the ith user belonging to the group of customers who enjoys movies of the mth genre; $\hat{b}_{jm}$ represents the strength of the jth movie belonging to the mth genre.