The University of Texas at Austin
In-Class Work 6

M339G Predictive Analytics

October 20, 2026

Principal Component Analysis.

Source: SRM Sample Questions.

Problem 6.1. Consider the following statements:

I. Principal Component Analysis (PCA) provide low-dimensional linear surfaces that are closest to the observations.

II. The first principal component is the line in p-dimensional space that is closest to the observations.

III. PCA finds a low dimension representation of a dataset that contains as much variation as possible.

IV. PCA serves as a tool for data visualization.

Determine which of the statements are correct.

a. Statements I, II, and III only
b. Statements I, II, and IV only
c. Statements I, III, and IV only
d. Statements II, III, and IV only
e. Statements I, II, III, and IV are all correct

Solution. (e)
Statement I is correct – Principal components provide low-dimensional linear surfaces that are closest to the observations.
Statement II is correct – The first principal component is the line in p-dimensional space that is closest to the observations.
Statement III is correct – PCA finds a low dimension representation of a dataset that contains as much variation as possible.
Statement IV is correct – PCA serves as a tool for data visualization.

Problem 6.2. Consider the following statements:

I. The proportion of variance explained by the next principal component never decreases as more principal components are added.

II. The cumulative proportion of variance explained never decreases as more principal components are added.

III. Loading vectors of different principal components are orthogonal.

IV. Scree plots help us determine the number of principal components to use.

Which of the statements is **FALSE**?

a. I only
b. II only
c. III only
d. I, II, and III
e. None of the above.

Solution. (a)
By the PCA algorithm, I is **FALSE**. In fact, it's the opposite. Statement II is **TRUE** as are statements III and IV.

Problem 6.3. Principal component analysis is applied to a large data set with four variables. Loadings for the first four principal components are estimated. Determine which of the following statements is/are true with respect to the loadings.

I. The loadings are unique.

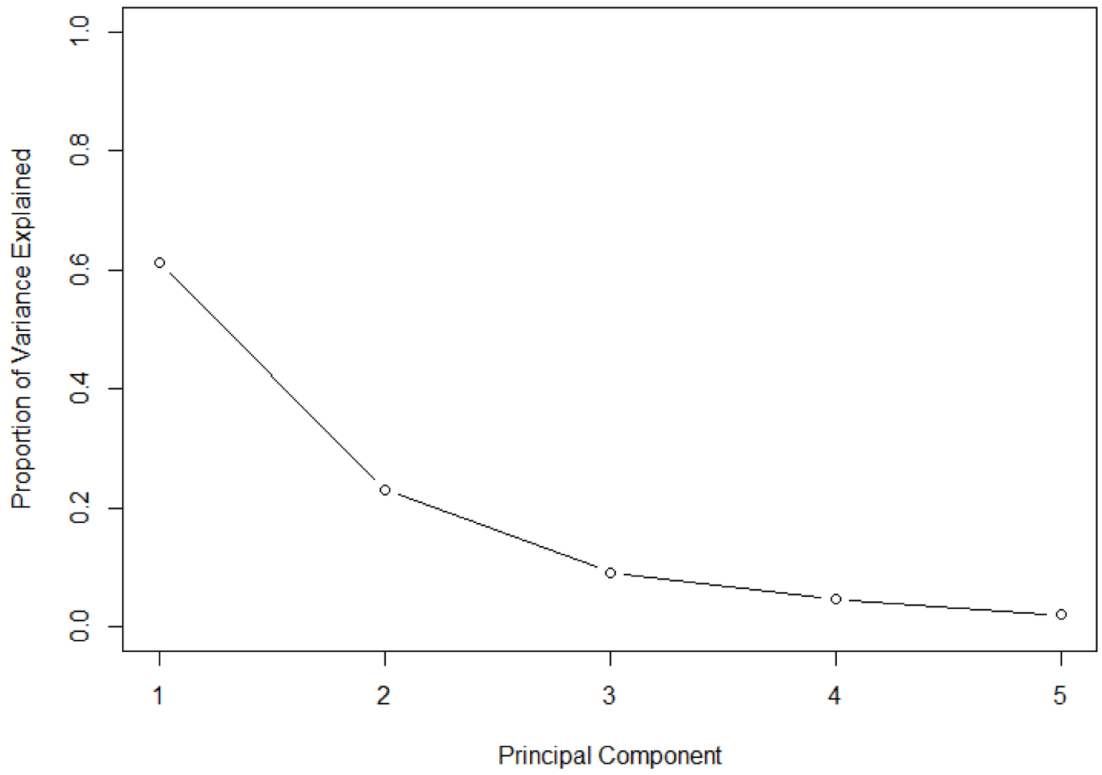
II. For a given principal component, the sum of the squares of the loadings across the four variables is one.

III. Together, the four principal components explain 100% of the variance.

a. None
b. I and II only
c. I and III only
d. II and III only
e. The correct answer is not given by (a), (b), (c), or (d).

Solution. (d)
I is false because the loadings are unique only up to a sign flip.
II is true. Principal components are designed to maximize variance. If there are no constraints on the magnitude of the loadings, the variance can be made arbitrarily large. The PCA algorithm's constraint is that the sum of the squares of the loadings equals 1.
III is true because four components can capture all the variation in four variables, provided there are at least four data points (note that the problem states that the data set is large).

Problem 6.4. Consider the following scree plot.



Determine the minimum number of principal components that are needed to explain at least 80% of the variance of the original dataset.

a. One
b. Two
c. Three
d. Four
e. It cannot be determined from the information given.

Solution. (b)
The first PC explains about 62% of the variance. The second PC explains about 23% of the variance. Combined they explain about 85% of the variance, and hence two PCs are sufficient.

Problem 6.5. Analysts W, X, Y, and Z are each performing Principal Components Analysis on the same data set with three variables. They use different programs with their default settings and discover that they have different factor loadings for the first principal component. Their loadings are:

	Variable 1	Variable 2	Variable 3
W	−0.549	−0.594	0.587
X	−0.549	0.594	0.587
Y	0.549	−0.594	−0.587
Z	0.140	−0.570	−0.809

Determine which of the following is/are plausible explanations for the different loadings.

I. Loadings are unique up to a sign flip and hence X's and Y's programs could make different arbitrary sign choices.

II. Z's program defaults to not scaling the variables while Y's program defaults to scaling them.

III. Loadings are unique up to a sign flip and hence W's and X's programs could make different arbitrary sign choices.

a. None
b. I and II only
c. I and III only
d. II and III only
e. The correct answer is not given by (a), (b), (c), or (d).

Solution. (b)
I is true. Uniqueness up to a sign flip means all three signs must be flipped. This is true for X and Y.
II is true. The presence or absence of scaling can change the loadings.
III is false. Uniqueness up to a sign flip means all three signs must be flipped. For W and X only the second loading is flipped.

Problem 6.6. You have run a principal components analysis on a veterinarian's database of dogs. The result is provided below.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Weight (lb)	0.506	0.017	0.003	−0.002	0.004	−0.002	0.006	−0.072	−0.389	−0.766
Height (in)	0.504	0.009	0.003	−0.004	0.004	−0.003	−0.002	−0.104	−0.575	0.636
Food (cups per day)	0.496	0.013	−0.006	−0.002	0.004	0.010	0.018	−0.603	0.620	0.070
Stride Length (in)	0.493	0.015	0.002	0.001	−0.006	−0.010	−0.005	0.787	0.364	0.067
Age (years)	−0.007	0.275	−0.628	0.066	0.469	−0.493	−0.252	−0.004	0.002	0.000
Vet Visits per Year	−0.002	0.197	−0.648	0.036	−0.680	0.278	0.031	0.001	−0.008	0.001
Miles Walked per Week	0.009	−0.566	−0.239	0.007	−0.080	−0.424	0.661	0.006	−0.004	0.000
Resting HR	−0.014	0.461	0.350	−0.040	−0.471	−0.663	0.039	−0.014	0.002	0.002
Body Fat (%)	−0.021	0.593	0.009	0.027	0.297	0.246	0.705	0.015	−0.007	0.007
Barks per Day	0.004	−0.019	0.081	0.996	−0.033	−0.008	−0.006	−0.003	−0.002	0.001

Determine which of the following statements is true.

a. Loadings close to 0 represent noise in the dataset.
b. PCA is only capturing the linear relationships between the predictors.
c. Principal components determine features contributing the least amount of variance.
d. The loading of PC1 for Age (years) indicates that it is not a strong predictor and can be discarded.
e. The first several principal components are capturing the noise and should not be used in a predictive model.

Solution. (b)
(a) is false; loadings close to zero indicate that the variable makes little contribution to that component.
(b) is true; PCA uses linear relationships.
(c) is false; principal components try to capture the most amount of variance.
(d) is false, because principal components are not predictors and the first component is the best one to retain.
(e) is false; the first components capture the most variance, it is the remaining components that capture noise.

Conceptual Short-Answer Questions.

Problem 6.7. An analyst runs PCA on a dataset of policyholders with two variables: annual income (in dollars, standard deviation $\approx 25,000$) and age (in years, standard deviation ≈ 12). The variables are **not** standardized before PCA.

a. Describe what the first principal component loading vector will look like, and explain why.
b. Would your answer change if income were recorded in thousands of dollars? What does this tell you about unscaled PCA?
c. Describe a situation in which it would be reasonable **not** to standardize the variables.

Solution.
1. The loading vector will be approximately $(1, 0)$, i.e., almost entirely on income. PCA finds the direction of maximum variance, and income's variance ($\approx 6.25 \times 10^8$) dwarfs age's (≈ 144), purely because of units.
2. Yes. In thousands, income's standard deviation is ≈ 25, which is comparable to age's, so both variables would contribute to PC1. Unscaled PCA results depend on the arbitrary choice of units.
3. When all variables are measured in the same units and their relative variances are meaningful, e.g., monthly claim counts for the same line of business across regions.

Problem 6.8.
a. Give two equivalent geometric descriptions of the first principal component loading vector φ_1.
b. Why does PCA impose the constraint $\sum_{j=1}^{p} \varphi_{j1}^2 = 1$?
c. The second principal component is constrained to be uncorrelated with the first. What does this imply geometrically about φ_1 and φ_2?

Solution.
1. (i) φ_1 is the direction in p-dimensional feature space along which the (centered) data have the greatest variance. (ii) The line through the data's mean in the direction φ_1 is the line closest to the observations, in the sense of minimizing the sum of squared perpendicular distances.
2. Without the constraint, the variance of the scores $z_{i1} = \sum_j \varphi_{j1} x_{ij}$ could be made arbitrarily large by scaling up the loadings, so the maximization problem would have no solution.
3. φ_2 is orthogonal to φ_1, i.e., $\varphi_1 \cdot \varphi_2 = 0$.

Problem 6.9. PCA is performed on five **standardized** variables. The variances of the principal component scores are

$$2.6, \quad 1.3, \quad 0.6, \quad 0.3, \quad 0.2.$$

a. Explain why these variances must sum to 5.
b. Compute the proportion of variance explained (PVE) and the cumulative PVE for each component.
c. Sketch the scree plot. Based on the plot, how many components would you keep? Justify your answer.
d. Explain why the PVE of each successive component can never exceed that of the previous one.

Solution.

1. Each standardized variable has variance 1, so the total variance is 5. Using all $p = 5$ components preserves all of the variance.
2. PVE: $0.52, 0.26, 0.12, 0.06, 0.04$.
   Cumulative PVE: $0.52, 0.78, 0.90, 0.96, 1.00$.
3. The plot drops steeply from PC1 to PC2 and from PC2 to PC3, then flattens; the elbow is at PC3. Keeping 2 or 3 components is reasonable (2 explain 78%, 3 explain 90%). There is no single correct answer; the choice is a judgment call.
4. Each component maximizes variance subject to being orthogonal to the previous ones. Adding constraints cannot increase the maximum, so the variance explained is non-increasing.

Problem 6.10. PCA is run on standardized auto-insurance data. The first two loading vectors are:

Variable	PC1	PC2
Claim frequency	0.60	0.04
Claim severity	0.50	−0.31
Premium	0.60	−0.02
Driver age	0.15	0.95

a. Give a practical interpretation of PC1 and PC2.
b. A policyholder has a large positive PC1 score and a large negative PC2 score. Describe this policyholder.
c. On a biplot of PC1 vs. PC2, describe the direction of the arrows for claim frequency and driver age. What does the angle between them suggest?
d. Another program reports PC1 loadings of $(-0.60, -0.50, -0.60, -0.15)$. Is this a contradiction? Explain.

Solution.

1. PC1 is a roughly equal weighting of frequency, severity, and premium: an overall "riskiness/cost" index. PC2 is dominated by driver age.
2. High overall claims and premium, and relatively young (with slightly higher-than-average severity, since severity loads negatively on PC2).
3. Claim frequency points almost horizontally along PC1; driver age points almost vertically along PC2. The near-right angle suggests the two variables are close to uncorrelated.
4. No. Loadings are unique only up to a sign flip of the entire vector; the direction (line) is the same, and the interpretation is unchanged.