BST 764: Applied Statistical Modeling
Breheny

Assignment 5: Principal Components
Due: Tuesday, December 6

Mathematical concepts and derivations

1. Let $\mathbf{A}$ be a symmetric matrix. Derive an expression for $\mathbf{A}^n$, where $\mathbf{A}^n = \prod_{i=1}^n \mathbf{A}$.

2. Let $\mathbf{A}$ be a symmetric matrix. The equation

$$|\mathbf{A} - \lambda \mathbf{I}| = 0$$

is known as the *characteristic equation* of $\mathbf{A}$ because it characterizes the eigenvalues of $\mathbf{A}$. Prove that λ satisfies the above equation if and only if λ is an eigenvalue of $\mathbf{A}$.

Simulation

3. Conduct a simulation study comparing principal components regression with ordinary least squares. Consider the following data generating mechanism:

$$\begin{aligned}
V_{ij} &\overset{\text{iid}}{\sim} N(0,1) \text{ for } i = 1, 2, \ldots, 30 \text{ and } j = 1, 2, 3 \\
X_{ik} &\overset{\text{iid}}{\sim} V_{i1} + N(0,1) \text{ for } i = 1, 2, \ldots, 30 \text{ and } k = 1, 2, 3, 4 \\
X_{ik} &\overset{\text{iid}}{\sim} V_{i2} + N(0,1) \text{ for } i = 1, 2, \ldots, 30 \text{ and } k = 5, 6, 7, 8 \\
X_{ik} &\overset{\text{iid}}{\sim} V_{i3} + N(0,1) \text{ for } i = 1, 2, \ldots, 30 \text{ and } k = 9, 10, 11, 12
\end{aligned}$$

In words, there are 12 observable explanatory variables ($\mathbf{x}$) that are derived from 3 unobservable latent factors ($\mathbf{v}$).

   (a) Suppose $y_i \overset{\text{iid}}{\sim} v_{i1} + v_{i2} + N(0,1)$ for $i = 1, 2, \ldots, 30$. Fit three models: an intercept-only model, a linear regression model with all 12 explanatory variables, and a principal component regression model using only the first three components. For each iteration, compare the accuracy of the approaches by using the model fits to predict 1,000 new observations generated according to the same mechanism. Report the average model error (mean squared prediction error minus irreducible error) for each of the three methods across 100 simulated data sets of sample size 30. Comment on the relative ordering of principal components and linear regression – which method performs the best, and why?

   (b) Repeat part (a), only with $y_i \overset{\text{iid}}{\sim} x_{i1} + x_{i5} + N(0,1)$. Again, comment on the relative ordering of principal components and linear regression – which method performs the best, and why?

Application

4. Drug abuse among adolescents is a major public health concern. Patterns of drug consumption tend to exist among the various licit and illicit substances that are in common use. In an investigation of these patterns, researchers from UCLA collected data on drug usage rates for 1634 students in the seventh to ninth grades in 11 schools in the greater metropolitan area of Los Angeles. Each participant completed a questionnaire about the number of times a particular substance had ever been used, on a five-point scale: (1) never tried, (2) only once, (3) a few times, (4) many times, and (5) regularly, for the following substances:

   - `cigarettes`
   - `beer`
   - `wine`
   - `liquor`
   - `cocaine`
   - `tranquilizers`
   - `medication`: Over the counter medication
   - `heroin`
   - `marijuana`
   - `hashish`
   - `inhalants`
   - `hallucinogenics`
   - `amphetamine`

   The responses are online in the file `druguse.txt`; NOTE: Only the correlation matrix and the sample size were published, not the raw data. The file online is synthetic (*i.e.*, made up), but does reflect the actual sample size and correlation matrix. Any other patterns, such as the marginal distribution of scores for a given substance, are entirely fabricated and do not necessarily bear any resemblance to reality.

   Carry out a principal components analysis on the drug use data. Decide on an appropriate number of components and interpret them in the context of the problem. Briefly state your main conclusions as far as the primary patterns of drug use.

5. The course website contains a data set (`bodyfat.txt`) from a study relating the amount of body fat to several body measurements which serve as predictors. The sample consists of 20 healthy females aged 25-34. The outcome, `BodyFat`, is given as a percentage of total weight and is obtained by a very accurate, yet cumbersome and intrusive process involving underwater submersion. Ideally, one could predict body fat % using simple body measurements. Three such measurements were obtained:

   - `Triceps`: Triceps skinfold thickness (mm)
   - `Thigh`: Thigh circumference (cm)
   - `Midarm`: Midarm circumference (cm)

   (a) Analyze the data using linear regression with all three predictors in the model. Comment on your results.

   (b) Analyze the data using linear regression, but use the three principal components as explanatory variables. Contrast the results of this model with the results of the model in (a). Give particular attention to predictive ability and estimation accuracy. Is one model able to predict more accurately than the other? Why or why not? Is one model able to estimate the regression coefficients more accurately than the other? Why or why not?

   (c) Interpret the principal components and their loadings.

   (d) Decide on a model that you feel is the best way to analyze this data (it may be one of (a) or (b) above, or something different), and interpret the results of the model, providing insight into the scientific questions of interest.