3.7.4 → PCA + regression

→ overview → Recall that a statistic is described as summarizing a random variables by mapping them to one value. Now consider summarizing a dataset w/ variables $x_1, \ldots, x_p$ into new + fewer variables. We can achieve this by performing principal component analysis (PCA). PCA is an unsupervised way to obtain new variables that summarize $x_1, \ldots, x_p$ in a dataset. A more technical definition of PCA is: a statistical tool that finds a low-dimensional representation of a dataset w/o significant loss of information

→ After explaining how PCA operates, we consider these new variables as predictors in a MLR setting — this is called principal component regression (PCR).

→ For this section, we define $x_1, \ldots, x_p$ as the centered variables of the original.

→ PCA

→ Just as a statistic summarizes RVs via a function, PCA summarizes $x_1, \ldots, x_p$ via a function as well. Specifically, a new variable is created by taking a linear combination of the original variables

→ Denote such a new variable as z_m, which we call the m^{th} principal component. Then z_m is defined as

$$z_m = \sum_{j=1}^{p} \phi_{j,m} x_j$$

→ The coefficients of this linear combination $\phi_{1,m}, \ldots, \phi_{p,m}$ are called the loadings of the mth principal components.

→ Then, for each observation in the dataset, we can obtain an mth principal component score. To find the m^{th} principal component score for the i^{th} observation, evaluate the variables x_j at the corresponding values for the i^{th} observation, i.e.

$$z_{i,m} = \sum_{j=1}^{p} \phi_{j,m} x_{i,j}$$

→ First principal component

→ How do we determine the principal component loadings? In PCA, the principal components are created in a sequence, so we start w/ the first principal component. The goal is for z_1 to explain the largest portion of variability in a dataset, compared to subsequent z_m's. To achieve this, we maximize the (biased) sample variance of z_1. In other words, the values of $\phi_{1,1}, \ldots, \phi_{p,1}$ are determined by maximizing

$$\frac{1}{n}\sum_{i=1}^{n}(z_{i,1}-\bar{z}_1)^2 = \frac{1}{n}\sum_{i=1}^{n} z_{i,1}^2 = \frac{1}{n}\sum_{i=1}^{n}(\phi_{1,1}x_{i,1} + \cdots + \phi_{p,1}x_{i,p})^2$$

($\bar{z}_1 = 0$) ($\bar{z}_1 = 0$ b/c variables are centered)

→ Moreover, we maximize while constrained by $\sum_{j=1}^{p}\phi_{j,1}^2 = 1$ in order to obtain non-trivial results.

→ To solve for these loadings, a technique known as eigen decomposition is often used. Learning this technique is not required for this exam. However, note that the loadings are determined using the $x_1, \ldots, x_p$ variables only. In the absence of a response variable, PCA is an unsupervised (learning supervised) method.

→ Second + subsequent principal components

→ Next consider the second principal component. The goal is for z_2 to explain the next largest portion of remaining variability in a dataset that is not explained by z_1. Aside from the analogous first principal component requirements this means:

→ z_2 is further constrained to be uncorrelated w/ z_1. This is equivalent to saying that the vector of loadings for the first principal component is orthogonal or perpendicular to the vector of loadings for the second principal component, i.e

$$\sum_{j=1}^{p}\phi_{j,1}\cdot\phi_{j,2} = 0 \Rightarrow z_1 \perp z_2$$
$$= \phi_1 \cdot \phi_2$$
↳ dot product

→ The variability explained by the second principal component is less than the variability explained by the first principal component

→ We can generalize this for all subsequent principal components

→ All principal components are uncorrelated w/ one another. In other words,

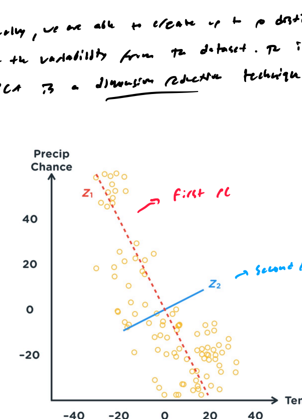
$$\sum_{j=1}^{p}\phi_{j,m}\cdot\phi_{j,u} = 0$$

for all pairs of m + u, $m \neq u$.

→ The variability explained by each subsequent principal component is always less than the variability explained by its previous principal component

→ Usually, we are able to create up to p distinct principal components — the same # of original variables. All p principal components would collectively capture all of the variability from the dataset. The ideal situation is for the first few principal components to explain much of the dataset's variability. Thus, PCA is a dimension reduction technique as all p original variables are being summarized/compressed into fewer variables

→ Example



→ First PC captures the largest variability present in the dataset of two variables. Visually, it is the direction where the variability of the points is the most.

→ Notice that the second PC is ⊥ (uncorrelated/perpendicular) w/ the first PC. b/c p=2, there is only 2 PCs + the second PC captures all remaining variability after the first PC.

→ The first PC loadings are $\phi_{1,1} = -0.4528$ + $\phi_{2,1} = 0.8916$. Therefore

$z_1 = -0.4528x_1 + 0.8916x_2$

→ The second PC loadings are $\phi_{1,2} = -0.8916$ + $\phi_{2,2} = -0.4528$. Hence,

$z_2 = -0.8916x_1 - 0.4528x_2$

→ In addition, note that

→ The PCs intersect the origin. This is a consequence of centering both variables before performing PCA.

→ The sum of squares of the (unrounded) first PC is 1 (likewise for the second).

$\phi_{1,1}^2 + \phi_{2,1}^2 = (-0.4528)^2 + (0.8916)^2 = 1$

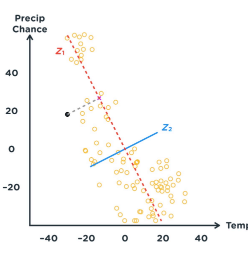
$\phi_{1,2}^2 + \phi_{2,2}^2 = \langle\ldots\rangle = 1$

→ The PCs are uncorrelated. This means the dot product of the loading vectors is 0, i.e:

$$\sum_{j=1}^{2}\phi_{j,1}\cdot\phi_{j,2} = \phi_{1,1}\cdot\phi_{1,2} + \phi_{2,1}\cdot\phi_{2,2}$$
$= \langle\ldots\rangle$
$= 0$

→ To better understand the PC scores, let's study one observation in the dataset: the 12^{th} observation.

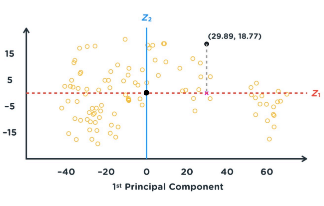
$(x_{12,1}, x_{12,2}) = (-20.279, 18.154)$



→ The observation has a first PC score of $z_{12,1} = -0.4528(-20.279) + 0.8916(18.154)$
↓ $= 29.89$

+ a second PC score of $z_{12,2} = -0.8916(-20.279) - 0.4528(18.154)$
↓ $= 18.77$

→ To grasp the meaning of these scores, we rotate the scatterplot clockwise until the first + second PCs coincide w/ the horizontal + vertical axis, respectively. This results in the following scatterplot where the second PC scores are plotted against the first PCs

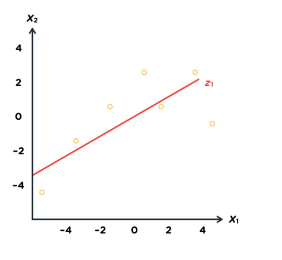


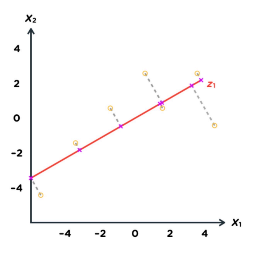
→ We can see that a first PC score is the z_1-direction distance from an observation to the origin. On the other hand, a second PC score is the z_2-direction distance from an observation to the origin.

→ Alternative Interpretation

→ We have viewed PCs as the directions where the data vary the most. However, as the observations live in p-dimensional space, we can also interpret PCs as the lower-dimension surface that is closest to the observations.

→ To demonstrate, imagine a dataset w/ 7 obs + two centered variables. So, the obs live in 2-D space. In this space, a "lower-dimension surface" can only be a line. That line that is closest to the 7 obs is the one given by the first PC.





→ For any observation, the shortest distance to the red line is the ⊥ distance. This aggregate distance is minimized (not the regression line, which is vertical distance) ⇒ best 1-D summary of the dataset

→ Standardized variables

→ PCA is usually performed on a dataset w/ centered variables. However, we may wish to further scale the variables, which leads to a data set of standardized variables.

→ Since the loadings are found by maximizing the sample variance of a PC, an x_j w/ a large sample variance would receive a large loading in absolute value. This may not be desirable, since a large sample variance could be due to the scale that x_j is recorded on, as opposed to the "true" variability in x_j.

→ On the other hand, performing PCA on standardized variables produces loadings that are not swayed by varying scales among the original variables

→ Principal components regression

→ Using PCs as predictors in a MLR is known as principal components regression (PCR) approach. Imagine using the first k PCs $z_1, \ldots, z_k$ in PCR. Then the model equation is

$$Y = \theta_0 + \theta_1 z_1 + \cdots + \theta_k z_k + \varepsilon$$

→ where $\theta_0, \ldots, \theta_k$ are the regression coefficients. These coefficients are estimated using OLS as previously discussed.

→ If $k < p$ + the PCs $z_1, \ldots, z_k$ do explain most of the dataset's variability, then there is a good chance for PCR to outperform the MLR w/ the original p variables as predictors — the loss of information, + fewer coefficients would need to be estimated. However if all PCs are used, i.e $k=p$, then the dimension reduction would not occur. In other words, the results would be equivalent to performing the MLR w/ the p original variables as predictors.

→ The ideal value of k can be determined through CV; the # of PCs used in PCR is a flexibility measure.

→ Examples

→ PCR w/ $k=2$ → $\hat{y} = 76.3875 + 1.7107z_1 + 2.7434z_2$
— — — — — — 4 → $\hat{y} =$ — — — — — — — — — — $+ 3.4333z_3 - 1.8602z_4$ } $p=4$

⇒ Several important things worth mentioning

→ $\hat{\theta}_0 = \bar{y}$, which is a consequence of the PCs each having a sample mean of 0

→ Notice $\hat{\theta}_1$ + $\hat{\theta}_2$ are unchanged, whether 2 or 4 PCs are used in PCR. This holds true for $\hat{\theta}_1, \ldots, \hat{\theta}_k$ in general, since PCs are orthogonal variables

→ The second fitted equation must be equivalent to the original MLR w/ $p=4$ predictors b/c no dimension reduction occurred

→ Given the following info:

→ OLS regression on dataset w/ $p=3$ results:

Variable	Estimate
Intercept	2.439
x_1	0.046
x_2	0.189
x_3	-0.001

→ one of the obs is

Y	x_1	x_2	x_3
14.7	44.5	383	41.1

→ A PCA is performed on the 3 predictors, + all 3 are used in PCR.

→ Predict the response for the obs using PCR model.

→ Since $k=p$ → $\hat{y}$ under PCR = $\hat{y}$ under OLS ⇒ ⟨…⟩

→ Assignment

→ Q11 PCA on the following data:

obs	Height	Weight
1	64	149
2	61	170
3	77	161

→ Given: $\phi_{2,H} = -0.9315$
$\phi_{2,W} > 0$

$z_{2,2} = ??$

→ $\phi_{2,H}^2 + \phi_{2,W}^2 = 1$
$(-0.9315)^2 + ↓ = 1$
⇒ $\phi_{2,W} = \sqrt{0.1323}$

⇒ $z_{2,2} = \sum_{j=1}^{2}\phi_{2,j}x_{2,j}$ ↳ assuming centered
$= -0.9315(61 - \overline{height}) + \sqrt{0.1323}(170 - \overline{weight})$
↳ $= \frac{202}{3}$ ↳ $\frac{480}{3}$
$= 0.1436$

3.7.5 → Partial least squares

→ overview → Similar to PCs, the method of partial least squares (PLS) creates new predictors by taking linear combinations of the recorded variables. These new predictors are called directions.

→ Denote the mth partial least squares direction as z_m. Upon finding k of these directions, we perform a MLR w/ $z_1, \ldots, z_k$ as predictors of the response.

→ The PLS directions differ from PCs in two main ways:

1) PLS directions are created using the response variable. They summarize the original predictors in a supervised way

2) PLS directions are solely used as predictors; they have no other purpose. However PCs are predictors only in the context of PCR; they can also be used in non-regression contexts

→ In this subsection, we define $x_1, \ldots, x_p$ as the standardized variables of the original

→ First PLS direction

→ The first partial least squares direction is calculated as

$$z_1 = \sum_{j=1}^{p}\phi_j x_j$$

→ where ϕ_j is set to equal the OLS slope estimate of y regressed on x_j. Intuitively, the original predictors that are more strongly correlated to the response will contribute more to calculating z_1

→ Second + subsequent PLS directions

→ Similar to the z_1 formula,

$$z_2 = \sum_{j=1}^{p}\phi_j^* x_j^*$$

→ where x_j^* is the residual of the linear regression
→ w/o an intercept
→ w/ x_j as the response variable
→ w/ z_1 as the explanatory variable, +
→ using OLS

+ ϕ_j^* is set to equal the OLS slope estimate of y regressed on x_j^*.

→ The $x_1^*, \ldots, x_p^*$ can be interpreted as the portion of $x_1, \ldots, x_p$ that was not captured by z_1.

→ In generalization, calculating a PLS direction involves using the previous direction to "update the previous predictors"? all subsequent PLS directions are computed iteratively by

(might be incorrect according to author)

→ fitting models w/ the "previous predictors" explained by the previous direction, then
→ using the residuals of those fits as the "current predictors" to calculate the current direction

→ Understanding PLS

→ Besides the aforementioned differences between PCs + PLS directions, the two concepts share many similarities. The PLS directions form a set of orthogonal variables. In addition, PLS performs dimension reduction as well; each direction summarizes all p original predictors. When $k<p$, no dimension reduction occurs, which leads to the same results as the regression of y on $x_1, \ldots, x_p$. Furthermore, k is a flexibility measure whose ideal value can be found via CV.

→ As seen in the linear combination coefficients, the PLS directions are constructed based on information from the response.
↓
⇒ supervised, this is the main difference w/ PCA/PCR.

Hence a drawback to PCR is that the PCs have no concrete reason to be good predictors of the response, even if they summarize the dataset well.