# Applied Statistics Qualifying Exams Coaching

Michael Howes*

Summer 2023

## Contents



*With lots of content credit given to previous applied quals coaches

# 0 Applied 2022[1]

## Problem 1: R-squared and PCA

Key ideas:

- Connections between principal components analysis and the singular value decomposition.
- Orthogonality of principal components and principal component directions.

(a) We are given that $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is standardized so that the columns have mean zero and variance one. We are also given that $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ is the SVD of $\boldsymbol{X}$. Let $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ be the diagonal entries of $\boldsymbol{D}$. The columns of $\boldsymbol{V}$ are thus the principal component directions and $\frac{1}{n}d_j^2 = \frac{1}{n}\|Xv_j\|_2^2$ is the variance of $X$ in the $j$th principal component direction. The cumulative percent variance explained sequence is thus,

$$\rho_k = 100 \times \frac{\sum_{s=1}^k \frac{1}{n}d_s^2}{\sum_{s=1}^p \frac{1}{n}d_s^2} = 100 \times \frac{\sum_{s=1}^k d_s^2}{\sum_{s=1}^p d_s^2},$$

for $k = 1, \dots, p$.

(b) We are given a response $\boldsymbol{y} \in \mathbb{R}^n$ with mean zero and variance one. The fitted values are $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} \in \mathbb{R}^n$. We know that the fitted values $\hat{\boldsymbol{y}}$ are orthogonal to the residuals $\boldsymbol{y} - \hat{\boldsymbol{y}}$. Thus,

$$\|\boldsymbol{y}\|_2^2 = \|\boldsymbol{y} - \hat{\boldsymbol{y}} + \hat{\boldsymbol{y}}\|_2^2 = \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2 + \|\hat{\boldsymbol{y}}\|_2^2.$$

And so

$$\frac{1}{n}\sum_{i=1}^n y_i^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \frac{1}{n}\sum_{i=1}^n \hat{y}_i^2.$$

Since $\boldsymbol{y}$ has mean zero and variance one, this implies that

$$1 = \text{MSE}_0 = \text{MSE} + \text{MSS}.$$

And so,

$$R^2 = 1 - \frac{\text{MSE}}{\text{MSE}_0} = 1 - \text{MSE} = \text{MSS}.$$

(c) We now regress the $j$th column of $\boldsymbol{X}$ on the first $k$ principal components of $\boldsymbol{X}$. Let $\boldsymbol{U}_k = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_k] \in \mathbb{R}^{n\times k}$ be the matrix containing the first $k$ columns of $\boldsymbol{U}$. Since $\boldsymbol{U}_k^\top \boldsymbol{U}_k = \boldsymbol{I}_k$, the fitted values from regressing $\boldsymbol{x}_j$ on $\boldsymbol{U}_k$ are

$$\hat{\boldsymbol{x}}_j = \boldsymbol{U}_k(\boldsymbol{U}_k^\top \boldsymbol{U}_k)^{-1}\boldsymbol{U}_k^\top \boldsymbol{x}_j = \boldsymbol{U}_k\boldsymbol{U}_k^\top \boldsymbol{x}_j.$$

[1]Michael Howes

The average regression sum of squares is thus,

$$\begin{aligned}\mathrm{MSS}_j &= \frac{1}{n}\hat{\boldsymbol{x}}_j^\top \hat{\boldsymbol{x}}_j \\ &= \frac{1}{n}\boldsymbol{x}_j^\top \boldsymbol{U}_k\boldsymbol{U}_k^\top \boldsymbol{U}_k\boldsymbol{U}_k^\top \boldsymbol{x}_j \\ &= \frac{1}{n}\boldsymbol{x}_j^\top \boldsymbol{U}_k\boldsymbol{U}_k^\top \boldsymbol{x}_j.\end{aligned}$$

Since $\boldsymbol{x}_j$ is the $j$th column of $\boldsymbol{X}$ we $\boldsymbol{x}_j = \boldsymbol{X}\boldsymbol{e}_j$ where $\boldsymbol{e}_j \in \mathbb{R}^p$ is the $j$th standard basis vector.

$$\begin{aligned}\boldsymbol{U}_k^\top \boldsymbol{x}_j &= \boldsymbol{U}_k\boldsymbol{X}\boldsymbol{e}_j \\ &= \boldsymbol{U}_k^\top \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{e}_j.\end{aligned}$$

We know that $\boldsymbol{U}_k^\top \boldsymbol{U} = [\boldsymbol{I}_k, \boldsymbol{0}_{k\times(p-k)}] \in \mathbb{R}^{k\times p}$ where $\boldsymbol{0}_{k\times(p-k)}$ is a matrix of all zeros of size $k\times(p-k)$. It follows that

$$\begin{aligned}\boldsymbol{U}_k^\top \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top &= [\boldsymbol{I}_k, \boldsymbol{0}_{k\times(p-k)}]\boldsymbol{D}\boldsymbol{V}^\top \\ &= \boldsymbol{D}_k\boldsymbol{V}_k^\top,\end{aligned}$$

where $\boldsymbol{D}_k = \mathrm{diag}(d_1, \ldots, d_k) \in \mathbb{R}^{k\times k}$ and $\boldsymbol{V}_k \in \mathbb{R}^{p\times k}$ is equal to the first $k$ rows of $\boldsymbol{V}$. Thus,

$$\begin{aligned}\boldsymbol{U}_k^\top \hat{\boldsymbol{x}}_j &= \boldsymbol{D}_k\boldsymbol{V}_k^\top \boldsymbol{e}_j \\ &= \sum_{s=1}^k d_s \left(\boldsymbol{v}_s^\top \boldsymbol{e}_j\right)\boldsymbol{e}_s \\ &= \sum_{s=1}^k d_s v_{sj}\boldsymbol{e}_s\end{aligned}$$

where $v_{sj}$ is the entry of $\boldsymbol{V}$ in row $s$ and column $j$. We thus have

$$\begin{aligned}\mathrm{MSS}_j &= \frac{1}{n}\|\boldsymbol{U}_k^\top \hat{\boldsymbol{x}}_j\|_2^2 \\ &= \qquad = \frac{1}{n}\|\sum_{s=1}^k d_s v_{sj}\boldsymbol{e}_s\|_2^2 \\ &= \frac{1}{n}\sum_{s=1}^k d_s^2 v_{sj}^2.\end{aligned}$$

Since $\boldsymbol{x}_j$ has mean zero and variance one, we are in the setting of part (b) and hence

$$R_j^2 = \mathrm{MSS}_j = \frac{1}{n}\sum_{s=1}^k d_s^2 v_{sj}^2.$$

(d) Note that

$$
\begin{aligned}
\sum_{j=1}^{p} \mathrm{MSS}_j &= \sum_{j=1}^{p} \frac{1}{n} \sum_{s=1}^{k} d_s^2 v_{sj}^2 & &= \sum_{s=1}^{k} \sum_{j=1}^{p} \frac{1}{n} d_s^2 v_{sj}^2 \\
&= \sum_{s=1}^{k} \frac{1}{n} d_s^2 \sum_{j=1}^{p} v_{sj}^2 \\
&= \sum_{s=1}^{k} \frac{1}{n} d_s^2 \|\boldsymbol{v}_s\|_2^2 \\
&= \frac{1}{n} \sum_{s=1}^{k} d_s^2,
\end{aligned}
$$

since all rows of $\boldsymbol{V}$ have norm one. We know that each column of $\boldsymbol{X}$ has variance one and mean zero. Thus,

$$
\begin{aligned}
p &= \sum_{j=1}^{p} \frac{1}{n} \boldsymbol{x}_j^\top \boldsymbol{x}_j \\
&= \frac{1}{n} \sum_{j=1}^{p} \operatorname{tr}\left(\boldsymbol{x}_j^\top \boldsymbol{x}_j\right) \\
&= \frac{1}{n} \sum_{j=1}^{p} \operatorname{tr}\left(\boldsymbol{x}_j \boldsymbol{x}_j^\top\right) \\
&= \frac{1}{n} \operatorname{tr}\left(\sum_{j=1}^{p} \boldsymbol{x}_j \boldsymbol{x}_j^\top\right) \\
&= \frac{1}{n} \operatorname{tr}\left(\boldsymbol{X}\boldsymbol{X}^\top\right) \\
&= \frac{1}{n} \operatorname{tr}\left(\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top\right) \\
&= \frac{1}{n} \operatorname{tr}\left(\boldsymbol{D}^2\right) \\
&= \frac{1}{n} \sum_{s=1}^{p} d_s^2.
\end{aligned}
$$

Thus,

$$
\frac{100}{p} \sum_{j=1}^{p} \mathrm{MSS}_j = \frac{100}{np} \sum_{s=1}^{k} d_s^2 = 100 \times \frac{\sum_{s=1}^{k} d_s^2}{\sum_{s=1}^{p} d_s^2} = \rho_k.
$$

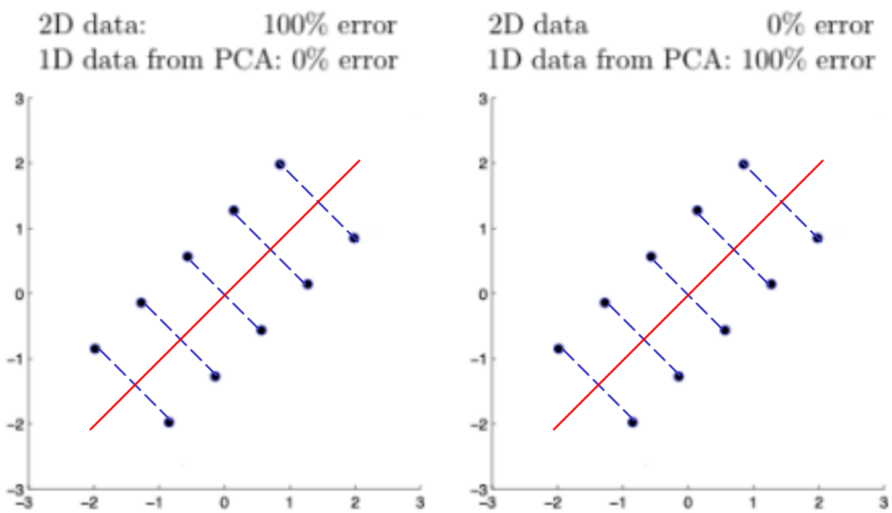
So $\frac{100}{p} \sum_{j=1}^{p} \mathrm{MSS}_j$ is exactly the cumulative percent variance explained sequence.

## Problem 2: LOO CV, PCA and 1NN

Key ideas:

- "Eyeballing" principal component directions.
- Working out nearest neighbor classifiers.

1. We are asked to draw a line corresponding to the first principal component. This does not have to be exact, but we can see that a roughly 45 degree line gives the direction with the most variance. Here the principal component direction is represented as a red solid line. In both plots I have also included dashed lines showing the projection onto the first principal direction.



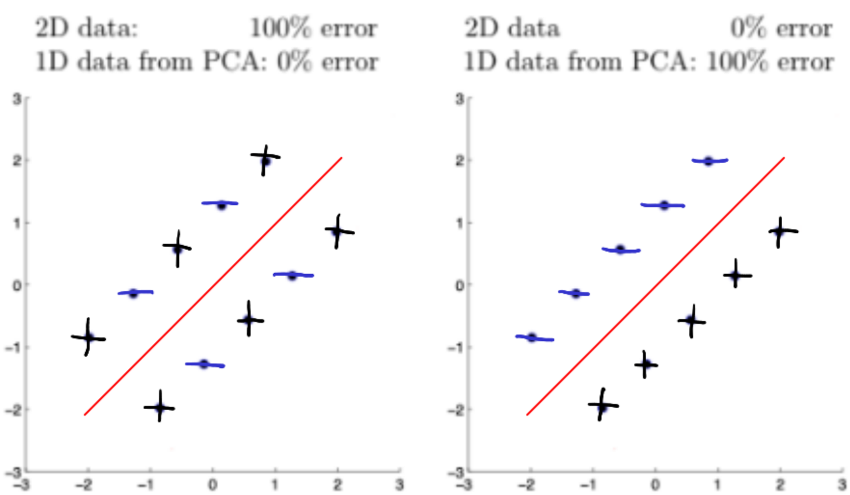

2. Looking at the previous figure we can make two observations.

   - Using the 2D data, the nearest neighbor of a point is one of the adjacent points on the line of points parallel to the PCA direction. For example, for the point at roughly $(-, 1, -2)$, the nearest neighbor is the point at roughly $(0, -1.25)$.
   - Using the projected 1D data from PCA, the nearest neighbor of a point is the point across the PCA direction. This because these points get projected onto the same value when we perform PCA. For example, for the point at roughly $(-1, -2)$, the 1D projected nearest neighbor is the point at roughly $(-2, -1)$.

   To have 100% error on the 2D data we want an alternating sequence of "+"'s and "−"'s along the two lines of points parallel to the PCA direction. If we use the same sequence of "+"'s and "−"'s on both lines of points, then we will also have 0% error when using the 1D data. This is because the 1D-nearest neighbor points will have the same labels.

To have 0% error on the 2D data we want adjacent points on the two lines to all have the same label. If we label only line of points with "+"'s and the other with "−"'s, then the 1D data will also have 100% error. This is because the 1D nearest neighbors will have the opposite labels.

In summary, the labelling below has the specified error rates.



## References