Why is ridge regression a shrinkage method? Suppose the design matrix $\mathbf{X}$ has ON [a] columns, $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$. Then the ridge estimate/prediction is a shrinkage version of the LS estimate/prediction.

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{\mathsf{LS}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{y} \\
\hat{\boldsymbol{\beta}}^{\mathsf{ridge}} &= (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y} \\
&= \frac{1}{1+\lambda}\mathbf{X}^T\mathbf{y} = \frac{1}{1+\lambda}\hat{\boldsymbol{\beta}}^{\mathsf{LS}}
\end{aligned}$$

$$\begin{aligned}
\hat{\mathbf{y}}_{\mathsf{LS}} &= \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathsf{LS}} \\
\hat{\mathbf{y}}_{\mathsf{ridge}} &= \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathsf{ridge}} = \frac{1}{1+\lambda}\mathbf{y}_{\mathsf{LS}}
\end{aligned}$$

[a] Orthonormal (ON): orthogonal with norm one.

In case the columns of $\mathbf{X}$ are not orthogonal, we can reformulate the regression on an orthogonal version of $\mathbf{X}$, known as the principal components analysis or SVD. Similarly we can see that the ridge estimate/prediction is a shrinkage version of the LS estimate/prediction. (You can skip the next 6 slides.)

Consider a singular value decomposition (SVD) of $\mathbf{X}$:

$$\mathbf{X}_{n\times p} = \mathbf{U}_{n\times p}\mathbf{D}_{p\times p}\mathbf{V}^T_{p\times p},$$

where

- $\mathbf{U}_{n\times p}$: columns $\mathbf{u}_j$'s form an ON basis for $C(\mathbf{X})$, $\mathbf{U}^T\mathbf{U} = \mathbf{I}_p$.
- $\mathbf{V}_{p\times p}$: columns $\mathbf{v}_j$'s form an ON basis for $\mathbb{R}^p$ with $\mathbf{V}^T\mathbf{V} = \mathbf{I}_p$.
- $\mathbf{D}_{p\times p}$: diagonal matrix with diagonal entries $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ being the singular values of $\mathbf{X}$.

  For ease of exposition we assume $n > p$ and rank$(\mathbf{X}) = p$. Therefore $d_p > 0$.

The Geometric interpretation of SVD:

$$\mathbf{X}_{n\times p} = \mathbf{U}_{n\times p}\mathbf{D}_{p\times p}\mathbf{V}^T_{p\times p}$$

Map a unit circle in $\mathbf{R}^p$ to an ellipse in $\mathbf{R}^n$

$$\mathbf{X}_{n\times p}\mathbf{v}_{j\,p\times 1} = \mathbf{U}_{n\times p}\mathbf{D}_{p\times p}\mathbf{V}^T_{p\times p}\mathbf{v}_{j\,p\times 1} = d_j\mathbf{u}_j.$$

Consider a singular value decomposition (SVD) of $\mathbf{X}$:

$$\mathbf{X}_{n\times p} = \mathbf{U}_{n\times p}\mathbf{D}_{p\times p}\mathbf{V}^T_{p\times p},$$

where

- $\mathbf{U}_{n\times p}$: columns $\mathbf{u}_j$'s form an ON basis for $C(\mathbf{X})$, $\mathbf{U}^T\mathbf{U} = \mathbf{I}_p$.
- $\mathbf{V}_{p\times p}$: columns $\mathbf{u}_j$'s form an ON basis for $\mathbb{R}^p$ with $\mathbf{V}^T\mathbf{V} = \mathbf{I}_p$.
- $\mathbf{D}_{p\times p}$: diagonal matrix with diagonal entries $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ being the singular values of $\mathbf{X}$.

  For ease of exposition we assume $n > p$ and $\text{rank}(\mathbf{X}) = p$. Therefore $d_p > 0$.
- PCA: write $\mathbf{X} = \mathbf{F}\mathbf{V}^T$ where each columns of $\mathbf{F}_{n\times p} = \mathbf{U}\mathbf{D}$ is the so-called principal components and each column of $\mathbf{V}$ is the principal component directions of $\mathbf{X}$;

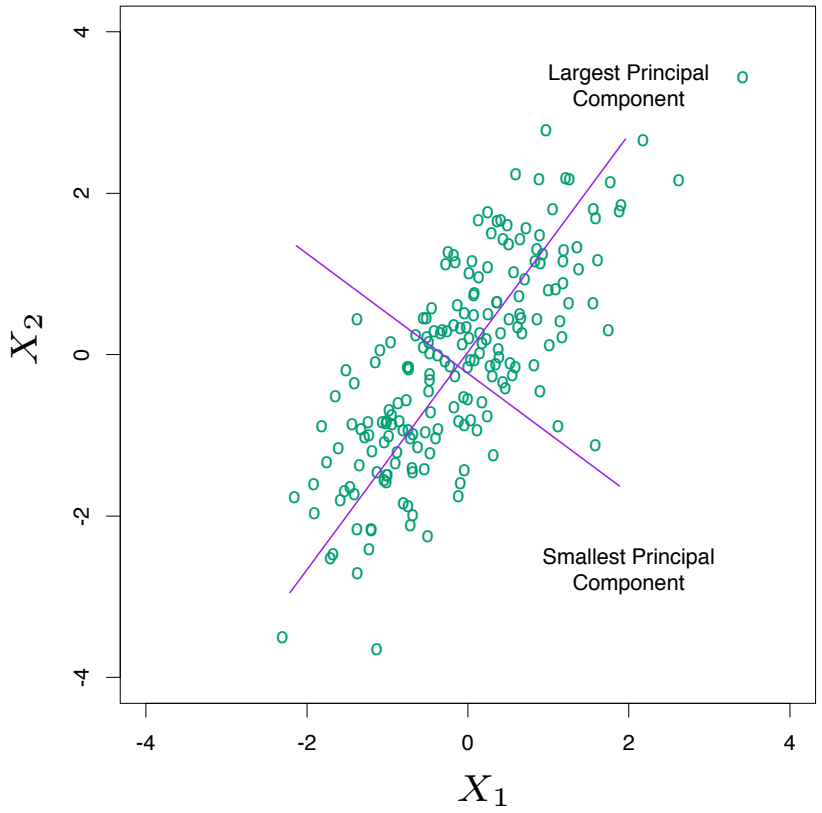

FIGURE 3.9. *Principal components of some input data points. The largest principal component is the direction that maximizes the variance of the projected data, and the smallest principal component minimizes that variance. Ridge regression projects* $\mathbf{y}$ *onto these components, and then shrinks the coefficients of the low-variance components more than the high-variance components.*

Write

$$\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{y} - \mathbf{UDV}\boldsymbol{\beta} = \mathbf{y} - \mathbf{F}\boldsymbol{\alpha}.$$

there is a one-to-one correspondence between $\boldsymbol{\beta}_{p\times 1}$ and $\boldsymbol{\alpha}_{p\times 1}$ and $\|\boldsymbol{\beta}\|^2 = \|\boldsymbol{\alpha}\|^2$. So

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \Longleftrightarrow \min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \|\mathbf{y} - \mathbf{F}\boldsymbol{\alpha}\|^2 + \lambda\|\boldsymbol{\alpha}\|^2.$$

$$\hat{\boldsymbol{\alpha}}^{\text{LS}} = \mathbf{D}^{-1}\mathbf{U}^T\mathbf{y}, \qquad \hat{\boldsymbol{\alpha}}_j^{\text{LS}} = \frac{1}{d_j}\mathbf{u}_j^T\mathbf{y}$$

$$\hat{\boldsymbol{\alpha}}^{\text{ridge}} = \text{diag}\Big(\frac{d_j}{d_j^2+\lambda}\Big)\mathbf{U}^t\mathbf{y}, \qquad \hat{\boldsymbol{\alpha}}_j^{\text{ridge}} = \frac{d_j^2}{d_j^2+\lambda}\hat{\boldsymbol{\alpha}}_j^{\text{LS}}$$

So the ridge estimate $\hat{\boldsymbol{\alpha}}^{\text{ridge}}$ shrinks the LS estimate $\hat{\boldsymbol{\alpha}}^{\text{LS}}$ by the factor $d_j^2/(d_j^2+\lambda)$: directions with smaller eigen values get more shrinkage.

- The LS prediction

$$\mathbf{F}\hat{\boldsymbol{\alpha}}^{\mathsf{LS}} = (\mathbf{UD})\mathbf{D}^{-1}\mathbf{U}^T\mathbf{y} = \mathbf{U}\mathbf{U}^T\mathbf{y} = \sum_{j=1}^{p} \big(\mathbf{u}_j^T\mathbf{y}\big)\mathbf{u}_j.$$

- The ridge prediction

$$\mathbf{F}\hat{\boldsymbol{\alpha}}^{\mathsf{ridge}} = \mathbf{U}\mathrm{diag}\Big(\frac{d_j}{d_j^2+\lambda}\Big)\mathbf{U}^t\mathbf{y} = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2+\lambda}(\mathbf{u}_j^T\mathbf{y})\mathbf{u}_j$$

- So the ridge prediction $\hat{\mathbf{y}}_{\mathsf{ridge}}$ shrinks the LS prediction $\hat{\mathbf{y}}_{\mathsf{LS}}$ by factor $d_j^2/(d_j^2+\lambda)$: directions with smaller eigen values get more shrinkage.