Model Selection and Regression Shrinkage Methods

TrungTin Nguyen

STATIFY team, Inria centre at the University Grenoble Alpes, France

Statistical Analysis and Document Mining

Complementary Course, Master of Applied Mathematics in Grenoble

Outline

Outline

Previous episode: linear regression via least squares

- Multiple linear regression takes the form

$$Y = \beta_0 + \sum_{p=1}^{P} \beta_p X_p + \epsilon, \text{ where } \epsilon \text{ is error term,}$$

β_0: intercept coefficient, $\beta_p, p \in [P] \equiv \{1, \ldots, P\}$: slope coefficients.

Previous episode: linear regression via least squares

- Multiple linear regression takes the form

$$Y = \beta_0 + \sum_{p=1}^{P} \beta_p X_p + \epsilon, \text{ where } \epsilon \text{ is error term,}$$

  β_0: intercept coefficient, $\beta_p, p \in [P] \equiv \{1, \ldots, P\}$: slope coefficients.
- We are given a training dataset $\mathcal{D}_N \equiv (\mathbf{x}_{[N]}, y_{[N]}) \equiv \{(\mathbf{x}_n, y_n)\}_{n \in [N]}$ with N independent observations sampled from PDF of $(\mathbf{X}, Y)$.

Previous episode: linear regression via least squares

- Multiple linear regression takes the form

$$Y = \beta_0 + \sum_{p=1}^{P} \beta_p X_p + \epsilon, \text{ where } \epsilon \text{ is error term,}$$

β_0: intercept coefficient, $\beta_p, p \in [P] \equiv \{1, \ldots, P\}$: slope coefficients.

- We are given a training dataset $\mathcal{D}_N \equiv (\mathbf{x}_{[N]}, y_{[N]}) \equiv \{(\mathbf{x}_n, y_n)\}_{n\in[N]}$ with N independent observations sampled from PDF of $(\mathbf{X}, Y)$.
- Using ordinary least squares (LS), we can obtain coefficient estimates $\widehat{\beta}_0$ and $\widehat{\beta}_p, p \in [P]$, such that

$$y_n \approx \widehat{\beta}_0 + \sum_{p=1}^{P} \widehat{\beta}_p x_{np}, \quad n \in [N]. \tag{1}$$

Previous episode: linear regression via least squares

- Multiple linear regression takes the form

$$Y = \beta_0 + \sum_{p=1}^{P} \beta_p X_p + \epsilon, \text{ where } \epsilon \text{ is error term,}$$

β_0: intercept coefficient, $\beta_p, p \in [P] \equiv \{1, \dots, P\}$: slope coefficients.

- We are given a training dataset $\mathcal{D}_N \equiv (\mathbf{x}_{[N]}, y_{[N]}) \equiv \{(\mathbf{x}_n, y_n)\}_{n \in [N]}$ with N independent observations sampled from PDF of $(\mathbf{X}, Y)$.
- Using ordinary least squares (LS), we can obtain coefficient estimates $\widehat{\beta}_0$ and $\widehat{\beta}_p, p \in [P]$, such that

$$y_n \approx \widehat{\beta}_0 + \sum_{p=1}^{P} \widehat{\beta}_p x_{np}, \quad n \in [N]. \tag{1}$$

- For any new observation (x^*, y^*), do we have

$$y^* \approx \widehat{\beta}_0 + \sum_{p=1}^{P} \widehat{\beta}_p x_p^* ? \tag{2}$$

Outline

1. Previous Episode: Multiple Linear Regression
   - Linear regression via least squares
   - How to improve linear models?
2. Methods for Linear Model Selection and Regularization
   - The general model selection paradigm
   - Linear model selection and regularization
3. Shrinkage Methods
   - Ridge regression
   - The Lasso
   - Comparing the Lasso and Ridge Regression
   - An application to the credit data
   - An application to the prostate cancer
4. A Deep Dive into Ridge Regression
   - Solution of the ridge regression in matrix form
   - Singular value decomposition of centered input matrix
   - Principal components and ridge regression

How to improve linear models?

- In spite of its simplicity, the linear model has clear advantages in terms of its interpretability and it often shows a good predictive performance.

How to improve linear models?

- In spite of its simplicity, the linear model has clear advantages in terms of its interpretability and it often shows a good predictive performance.
- How to improve linear models?

How to improve linear models?

- In spite of its simplicity, the linear model has clear advantages in terms of its interpretability and it often shows a good predictive performance.
- How to improve linear models? ☛ Replacing LS fitting with some **alternative fitting procedures** to improve:
  - Model Interpretability: By **dropping irrelevant features**, *i.e.*, setting their coefficient estimates to zero, we can obtain a model that will be easier to interpret ☛ We will now present a few approaches for carrying out the feature selection or variable selection automatically.

How to improve linear models?

- In spite of its simplicity, the linear model has clear advantages in terms of its interpretability and it often shows a good predictive performance.
- How to improve linear models? ☛ Replacing LS fitting with some **alternative fitting procedures** to improve:
  - Model Interpretability: By **dropping irrelevant features**, *i.e.,* setting their coefficient estimates to zero, we can obtain a model that will be easier to interpret ☛ We will now present a few approaches for carrying out the feature selection or variable selection automatically.
  - Prediction Accuracy: By **constraining or shrinking the estimated coefficients** ☛ substantially reduce the variance at the cost of a negligible increase in bias, especially when $P > N$ or $P \gg N$.

How to improve linear models?

- In spite of its simplicity, the linear model has clear advantages in terms of its interpretability and it often shows a good predictive performance.
- How to improve linear models? Replacing LS fitting with some **alternative fitting procedures** to improve:
  - Model Interpretability: By **dropping irrelevant features**, *i.e.,* setting their coefficient estimates to zero, we can obtain a model that will be easier to interpret We will now present a few approaches for carrying out the feature selection or variable selection automatically.
  - Prediction Accuracy: By **constraining or shrinking the estimated coefficients** substantially reduce the variance at the cost of a negligible increase in bias, especially when $P > N$ or $P \gg N$.
    Substantial improvements in the accuracy for predict the response for observations not used in model training (x^*, y^*).

How to improve linear models?

- In spite of its simplicity, the linear model has clear advantages in terms of its interpretability and it often shows a good predictive performance.
- How to improve linear models? Replacing LS fitting with some **alternative fitting procedures** to improve:
  - Model Interpretability: By **dropping irrelevant features**, *i.e.,* setting their coefficient estimates to zero, we can obtain a model that will be easier to interpret We will now present a few approaches for carrying out the feature selection or variable selection automatically.
  - Prediction Accuracy: By **constraining or shrinking the estimated coefficients** substantially reduce the variance at the cost of a negligible increase in bias, especially when $P > N$ or $P \gg N$.
    Substantial improvements in the accuracy for predict the response for observations not used in model training (x^*, y^*).

We will discuss linear model selection and regularization methods in more detail on the next slides!

Outline

Outline

The general model selection paradigm

- **Model selection problem:** let $(S_m)_{m \in \mathcal{M}}$ be a family of models. For every $m \in \mathcal{M}$, let $\widehat{s}_m(\mathcal{D}_N)$ be a minimum contrast estimator, *e.g.*, least-squares contrast in LS, over S_m. **Our goal is to choose the best data-driven model $\widehat{m} \equiv \widehat{m}(\mathcal{D}_N) \in \mathcal{M}$ from data.**

---

[1] Mallows, C. L. (1973). "Some Comments on C_P". Technometrics.

[2] Akaike, H. (1974). "A new look at the statistical model identification". IEEE Transactions on Automatic Control.

[3] Schwarz, G. (1978). "Estimating the dimension of a model". The Annals of Statistics.

[4] Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions". JRSSSB.

[5] Geisser, S. (1975). "The predictive sample reuse method with applications". J. Amer. Statist. Assoc.

[6] Arlot, S., & Celisse, A. (2010). "A survey of cross-validation procedures for model selection". Statist. Surv.

[7] Birgé, L. and Massart, P. (2007). "Minimal penalties for Gaussian model selection". Probability Theory and Related Fields.

[8] **Nguyen, T.**, Nguyen, H. D., Chamroukhi, F., & Forbes, F. (2022). "A non-asymptotic approach for model selection via penalization in high-dimensional mixture of experts models". Electronic Journal of Statistics.

The general model selection paradigm

- **Model selection problem:** let $(S_m)_{m \in \mathcal{M}}$ be a family of models. For every $m \in \mathcal{M}$, let $\widehat{s}_m(\mathcal{D}_N)$ be a minimum contrast estimator, *e.g.,* least-squares contrast in LS, over S_m. **Our goal is to choose the best data-driven model $\widehat{m} \equiv \widehat{m}(\mathcal{D}_N) \in \mathcal{M}$ from data.**
- **Some model selection criteria:**
  1. Asymptotic approach: Mallows's C_p[1],

[1] Mallows, C. L. (1973). "Some Comments on C_P". Technometrics.

[2] Akaike, H. (1974). "A new look at the statistical model identification". IEEE Transactions on Automatic Control.

[3] Schwarz, G. (1978). "Estimating the dimension of a model". The Annals of Statistics.

[4] Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions". JRSSSB.

[5] Geisser, S. (1975). "The predictive sample reuse method with applications". J. Amer. Statist. Assoc.

[6] Arlot, S., & Celisse, A. (2010). "A survey of cross-validation procedures for model selection". Statist. Surv.

[7] Birgé, L. and Massart, P. (2007). "Minimal penalties for Gaussian model selection". Probability Theory and Related Fields.

[8] **Nguyen, T.**, Nguyen, H. D., Chamroukhi, F., & Forbes, F. (2022). "A non-asymptotic approach for model selection via penalization in high-dimensional mixture of experts models". Electronic Journal of Statistics.

The general model selection paradigm

- **Model selection problem:** let $(S_m)_{m \in \mathcal{M}}$ be a family of models. For every $m \in \mathcal{M}$, let $\widehat{s}_m(\mathcal{D}_N)$ be a minimum contrast estimator, *e.g.,* least-squares contrast in LS, over S_m. **Our goal is to choose the best data-driven model $\widehat{m} \equiv \widehat{m}(\mathcal{D}_N) \in \mathcal{M}$ from data.**
- **Some model selection criteria:**
  1. Asymptotic approach: Mallows's C_p[1], Akaike information criterion[2] (AIC),

---

[1] Mallows, C. L. (1973). "Some Comments on C_P". Technometrics.

[2] Akaike, H. (1974). "A new look at the statistical model identification". IEEE Transactions on Automatic Control.

[3] Schwarz, G. (1978). "Estimating the dimension of a model". The Annals of Statistics.

[4] Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions". JRSSSB.

[5] Geisser, S. (1975). "The predictive sample reuse method with applications". J. Amer. Statist. Assoc.

[6] Arlot, S., & Celisse, A. (2010). "A survey of cross-validation procedures for model selection". Statist. Surv.

[7] Birgé, L. and Massart, P. (2007). "Minimal penalties for Gaussian model selection". Probability Theory and Related Fields.

[8] **Nguyen, T.**, Nguyen, H. D., Chamroukhi, F., & Forbes, F. (2022). "A non-asymptotic approach for model selection via penalization in high-dimensional mixture of experts models". Electronic Journal of Statistics.

The general model selection paradigm

- **Model selection problem:** let $(S_m)_{m \in \mathcal{M}}$ be a family of models. For every $m \in \mathcal{M}$, let $\widehat{s}_m(\mathcal{D}_N)$ be a minimum contrast estimator, *e.g.*, least-squares contrast in LS, over S_m. **Our goal is to choose the best data-driven model $\widehat{m} \equiv \widehat{m}(\mathcal{D}_N) \in \mathcal{M}$ from data.**
- **Some model selection criteria:**
  1. Asymptotic approach: Mallows's C_p[1], Akaike information criterion[2] (AIC), Bayesian information criterion[3] (BIC): no finite sample guarantees.

[1] Mallows, C. L. (1973). "Some Comments on C_P". Technometrics.

[2] Akaike, H. (1974). "A new look at the statistical model identification". IEEE Transactions on Automatic Control.

[3] Schwarz, G. (1978). "Estimating the dimension of a model". The Annals of Statistics.

[4] Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions". JRSSSB.

[5] Geisser, S. (1975). "The predictive sample reuse method with applications". J. Amer. Statist. Assoc.

[6] Arlot, S., & Celisse, A. (2010). "A survey of cross-validation procedures for model selection". Statist. Surv.

[7] Birgé, L. and Massart, P. (2007). "Minimal penalties for Gaussian model selection". Probability Theory and Related Fields.

[8] **Nguyen, T.**, Nguyen, H. D., Chamroukhi, F., & Forbes, F. (2022). "A non-asymptotic approach for model selection via penalization in high-dimensional mixture of experts models". Electronic Journal of Statistics.

The general model selection paradigm

- **Model selection problem:** let $(S_m)_{m\in\mathcal{M}}$ be a family of models. For every $m \in \mathcal{M}$, let $\widehat{s}_m(\mathcal{D}_N)$ be a minimum contrast estimator, *e.g.*, least-squares contrast in LS, over S_m. **Our goal is to choose the best data-driven model $\widehat{m} \equiv \widehat{m}(\mathcal{D}_N) \in \mathcal{M}$ from data.**
- **Some model selection criteria:**
  1. Asymptotic approach: Mallows's C_p[1], Akaike information criterion[2] (AIC), Bayesian information criterion[3] (BIC): no finite sample guarantees.
  2. Cross-validation procedures:[4] [5] [6] K-Fold, leave-one-out.

[1] Mallows, C. L. (1973). "Some Comments on C_P". Technometrics.
[2] Akaike, H. (1974). "A new look at the statistical model identification". IEEE Transactions on Automatic Control.
[3] Schwarz, G. (1978). "Estimating the dimension of a model". The Annals of Statistics.
[4] Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions". JRSSSB.
[5] Geisser, S. (1975). "The predictive sample reuse method with applications". J. Amer. Statist. Assoc.
[6] Arlot, S., & Celisse, A. (2010). "A survey of cross-validation procedures for model selection". Statist. Surv.
[7] Birgé, L. and Massart, P. (2007). "Minimal penalties for Gaussian model selection". Probability Theory and Related Fields.
[8] **Nguyen, T.**, Nguyen, H. D., Chamroukhi, F., & Forbes, F. (2022). "A non-asymptotic approach for model selection via penalization in high-dimensional mixture of experts models". Electronic Journal of Statistics.

The general model selection paradigm

- **Model selection problem:** let $(S_m)_{m \in \mathcal{M}}$ be a family of models. For every $m \in \mathcal{M}$, let $\widehat{s}_m(\mathcal{D}_N)$ be a minimum contrast estimator, *e.g.*, least-squares contrast in LS, over S_m. **Our goal is to choose the best data-driven model $\widehat{m} \equiv \widehat{m}(\mathcal{D}_N) \in \mathcal{M}$ from data.**
- **Some model selection criteria:**
  1. Asymptotic approach: Mallows's C_p[1], Akaike information criterion[2] (AIC), Bayesian information criterion[3] (BIC): no finite sample guarantees.
  2. Cross-validation procedures:[4] [5] [6] K-Fold, leave-one-out.
  3. Non-asymptotic approach: **slope heuristic**[7] [8], which is particularly useful for high-dimensional small data sets, *e.g.*, $N \ll P$.

[1] Mallows, C. L. (1973). "Some Comments on C_P". Technometrics.

[2] Akaike, H. (1974). "A new look at the statistical model identification". IEEE Transactions on Automatic Control.

[3] Schwarz, G. (1978). "Estimating the dimension of a model". The Annals of Statistics.

[4] Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions". JRSSSB.

[5] Geisser, S. (1975). "The predictive sample reuse method with applications". J. Amer. Statist. Assoc.

[6] Arlot, S., & Celisse, A. (2010). "A survey of cross-validation procedures for model selection". Statist. Surv.

[7] Birgé, L. and Massart, P. (2007). "Minimal penalties for Gaussian model selection". Probability Theory and Related Fields.

[8] **Nguyen, T.**, Nguyen, H. D., Chamroukhi, F., & Forbes, F. (2022). "A non-asymptotic approach for model selection via penalization in high-dimensional mixture of experts models". Electronic Journal of Statistics.

Outline

Methods for Linear Model Selection and Regularization

1. **Subset Selection:** by best subset or stepwise selection of $\mathcal{P} \equiv$ all possible subset models of P predictors, $\text{card}(\mathcal{P}) = 2^P$.

[9] Pearson, K (1894). "Contributions to the mathematical theory of evolution". Philosophical Transactions of the Royal Society of London.

Methods for Linear Model Selection and Regularization

1. **Subset Selection:** by best subset or stepwise selection of $\mathcal{P} \equiv$ all possible subset models of P predictors, $\text{card}(\mathcal{P}) = 2^P$.
   1. **Define a suitable collection of model** $\mathcal{M} = \{m : m \in \mathcal{P}\} \subset \mathcal{P}$ **and fit models** using LS, maximum likelihood estimation (MLE) or method of moments[9].

[9] Pearson, K (1894). "Contributions to the mathematical theory of evolution". Philosophical Transactions of the Royal Society of London.

1. **Subset Selection:** by best subset or stepwise selection of $\mathcal{P} \equiv$ all possible subset models of P predictors, $\text{card}(\mathcal{P}) = 2^P$.
   1. **Define a suitable collection of model** $\mathcal{M} = \{m : m \in \mathcal{P}\} \subset \mathcal{P}$ **and fit models** using LS, maximum likelihood estimation (MLE) or method of moments[9].
   2. **Identify the best model** $\widehat{m} \in \mathcal{M}$ **via suitable model selection criteria**. (Pedro talked about this in CM3).

[9] Pearson, K (1894). "Contributions to the mathematical theory of evolution". Philosophical Transactions of the Royal Society of London.

Methods for Linear Model Selection and Regularization

1. **Subset Selection:** by best subset or stepwise selection of $\mathcal{P} \equiv$ all possible subset models of P predictors, $\text{card}(\mathcal{P}) = 2^P$.
   1. **Define a suitable collection of model** $\mathcal{M} = \{m : m \in \mathcal{P}\} \subset \mathcal{P}$ **and fit models** using LS, maximum likelihood estimation (MLE) or method of moments[9].
   2. **Identify the best model** $\widehat{m} \in \mathcal{M}$ **via suitable model selection criteria**. (Pedro talked about this in CM3).
2. **Dimension Reduction:** project the P predictors into a lower dimensional subspace, *e.g.*, principal components regression, partial least squares. (We will see how to do this in CM4).

[9] Pearson, K (1894). "Contributions to the mathematical theory of evolution". Philosophical Transactions of the Royal Society of London.

3. **Shrinkage Methods:**
   - **Fit a model involving all P predictors**, using a technique that **constrains** or **regularizes** the coefficient estimates, or equivalently, that the **estimated coefficients are shrunken towards zero**.

[10] Hoerl, A. E., & Kennard, R. W. (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems", Technometrics.

[11] Tibshirani, R. (1996). "Regression Shrinkage and Selection via the lasso". JRSS. Series B.

[12] Santosa, F., & Symes, W. W. (1986). "Linear inversion of band-limited reflection seismograms". SIAM journal on scientific and statistical computing.

3. **Shrinkage Methods:**
   - **Fit a model involving all P predictors**, using a technique that **constrains** or **regularizes** the coefficient estimates, or equivalently, that the **estimated coefficients are shrunken towards zero**.

   Not immediately obvious why such a constraint should improve the fit.

   This shrinkage (also known as **regularization**) has the effect of reducing variance (to be verified soon!) and can also perform **variable selection**.

[10] Hoerl, A. E., & Kennard, R. W. (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems", Technometrics.

[11] Tibshirani, R. (1996). "Regression Shrinkage and Selection via the lasso". JRSS. Series B.

[12] Santosa, F., & Symes, W. W. (1986). "Linear inversion of band-limited reflection seismograms". SIAM journal on scientific and statistical computing.

3. **Shrinkage Methods:**
   - **Fit a model involving all P predictors**, using a technique that **constrains** or **regularizes** the coefficient estimates, or equivalently, that the **estimated coefficients are shrunken towards zero**.

   Not immediately obvious why such a constraint should improve the fit.

   This shrinkage (also known as **regularization**) has the effect of reducing variance (to be verified soon!) and can also perform **variable selection**.

   The two best-known techniques: ridge regression[10] and Lasso [11] [12].

[10] Hoerl, A. E., & Kennard, R. W. (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems", Technometrics.

[11] Tibshirani, R. (1996). "Regression Shrinkage and Selection via the lasso". JRSS. Series B.

[12] Santosa, F., & Symes, W. W. (1986). "Linear inversion of band-limited reflection seismograms". SIAM journal on scientific and statistical computing.

Outline

Outline

Ridge regression vs LS

- Recall that the LS estimates $\widehat{\beta}^{\text{ls}} = (\widehat{\beta}_0^{\text{ls}}, \widehat{\beta}_1^{\text{ls}}, \ldots, \widehat{\beta}_P^{\text{ls}})^\top$ is given by:

$$\widehat{\beta}^{\text{ls}} = \underset{\beta}{\operatorname{argmin}} \sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2 . \quad (3)$$

Ridge regression vs LS

- Recall that the LS estimates $\widehat{\beta}^{\text{ls}} = (\widehat{\beta}_0^{\text{ls}}, \widehat{\beta}_1^{\text{ls}}, \ldots, \widehat{\beta}_P^{\text{ls}})^\top$ is given by:

$$\widehat{\beta}^{\text{ls}} = \underset{\beta}{\operatorname{argmin}} \sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2. \tag{3}$$

- In contrast, the ridge regression coefficient estimates $\widehat{\beta}^{\text{ridge}}$ [James et al., 2021, Section 6.2], [Hastie et al., 2009, Section 3.4] is defined as:

$$\widehat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left[\underbrace{\sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2}_{\text{RSS}} + \underbrace{\lambda \sum_{p=1}^{P} \beta_p^2}_{\text{Shrinkage penalty} \equiv \text{pen}_R} \right]. \tag{4}$$

Here $\lambda \geq 0$ is a tunning parameter, to be determined separately using previous model selection criteria.

How does the ridge regression work?

The ridge regression coefficient estimates $\widehat{\beta}^{\text{ridge}}$ is defined as follows:

$$\widehat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left[\underbrace{\sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2}_{\text{RSS}} + \underbrace{\lambda \sum_{p=1}^{P} \beta_p^2}_{\text{pen}_R = \lambda \|\beta_{[P]}\|_2^2} \right].$$

How does the ridge regression work?

The ridge regression coefficient estimates $\widehat{\beta}^{\text{ridge}}$ is defined as follows:

$$\widehat{\beta}^{\text{ridge}} = \operatorname*{argmin}_{\beta} \left[\underbrace{\sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2}_{\text{RSS}} + \underbrace{\lambda \sum_{p=1}^{P} \beta_p^2}_{\text{pen}_R = \lambda \|\beta_{[P]}\|_2^2} \right].$$

- Like least squares, ridge regression seeks coefficient estimates that fit the data well by making the RSS small.

How does the ridge regression work?

The ridge regression coefficient estimates $\widehat{\beta}^{\text{ridge}}$ is defined as follows:

$$\widehat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left[\underbrace{\sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2}_{\text{RSS}} + \underbrace{\lambda \sum_{p=1}^{P} \beta_p^2}_{\text{pen}_R = \lambda \|\beta_{[P]}\|_2^2} \right].$$

- Like least squares, ridge regression seeks coefficient estimates that fit the data well by making the RSS small.
- However, the second term pen_R is small when $\beta_{[P]}$ are close to zero, and so it has the effect of shrinking the estimates of $\beta_{[P]}$ towards zero.

How does the ridge regression work?

The ridge regression coefficient estimates $\widehat{\beta}^{\text{ridge}}$ is defined as follows:

$$\widehat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left[\underbrace{\sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2}_{\text{RSS}} + \underbrace{\lambda \sum_{p=1}^{P} \beta_p^2}_{\text{pen}_R = \lambda \|\beta_{[P]}\|_2^2} \right].$$

- Like least squares, ridge regression seeks coefficient estimates that fit the data well by making the RSS small.
- However, the second term pen_R is small when $\beta_{[P]}$ are close to zero, and so it has the effect of shrinking the estimates of $\beta_{[P]}$ towards zero.
- The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates.

How does the ridge regression work?

The ridge regression coefficient estimates $\widehat{\beta}^{\text{ridge}}$ is defined as follows:

$$\widehat{\beta}^{\text{ridge}} = \operatorname*{argmin}_{\beta} \left[\underbrace{\sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2}_{\text{RSS}} + \underbrace{\lambda \sum_{p=1}^{P} \beta_p^2}_{\text{pen}_R = \lambda \|\beta_{[P]}\|_2^2} \right].$$

- Like least squares, ridge regression seeks coefficient estimates that fit the data well by making the RSS small.
- However, the second term pen_R is small when $\beta_{[P]}$ are close to zero, and so it has the effect of shrinking the estimates of $\beta_{[P]}$ towards zero.
- The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates.

Selecting a good value for λ is critical using previous model selection criteria.

Ridge regression: scaling of predictors

- The standard LS coefficient estimates $\widehat{\beta}^{\text{ls}}$ are **scale equivariant**: multiplying X_p by a constant C simply leads to a scaling of the least squares coefficient estimates by a factor of $1/C$. In other words, regardless of how the pth predictor is scaled, $X_p\widehat{\beta}^{\text{ls}}$ will remain the same.

Ridge regression: scaling of predictors

- The standard LS coefficient estimates $\widehat{\beta}^{\text{ls}}$ are **scale equivariant**: multiplying X_p by a constant C simply leads to a scaling of the least squares coefficient estimates by a factor of $1/C$. In other words, regardless of how the pth predictor is scaled, $X_p\widehat{\beta}^{\text{ls}}$ will remain the same.
- In contrast, the ridge regression coefficient estimates $\widehat{\beta}^{\text{ridge}}$ can **change substantially** when multiplying a given predictor by a constant, due to the sum of squared coefficients term in the penalty part of the ridge regression objective function.

Ridge regression: scaling of predictors

- The standard LS coefficient estimates $\widehat{\beta}^{\text{ls}}$ are **scale equivariant**: multiplying X_p by a constant C simply leads to a scaling of the least squares coefficient estimates by a factor of $1/C$. In other words, regardless of how the pth predictor is scaled, $X_p\widehat{\beta}^{\text{ls}}$ will remain the same.
- In contrast, the ridge regression coefficient estimates $\widehat{\beta}^{\text{ridge}}$ can **change substantially** when multiplying a given predictor by a constant, due to the sum of squared coefficients term in the penalty part of the ridge regression objective function.
- Final fit will not depend on the predictors' scale

Ridge regression: scaling of predictors

- The standard LS coefficient estimates $\widehat{\beta}^{\text{ls}}$ are **scale equivariant**: multiplying X_p by a constant C simply leads to a scaling of the least squares coefficient estimates by a factor of $1/C$. In other words, regardless of how the pth predictor is scaled, $X_p \widehat{\beta}^{\text{ls}}$ will remain the same.
- In contrast, the ridge regression coefficient estimates $\widehat{\beta}^{\text{ridge}}$ can **change substantially** when multiplying a given predictor by a constant, due to the sum of squared coefficients term in the penalty part of the ridge regression objective function.

Final fit will not depend on the predictors' scale **apply ridge regression after standardizing the predictors**, using the formula

$$\widetilde{x}_{np} = \frac{x_{np}}{\underbrace{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_{np}-\overline{x}_p)^2}}_{s_{x_p} \equiv \text{ estimated standard deviation.}}}. \text{ Check that } s_{\widetilde{x}_p} = 1. \quad (5)$$

Why does ridge regression improve over least squares?

$\boxed{\text{Why does ridge regression improve over least squares?}}$ Bias-variance trade-off!

Why does ridge regression improve over least squares? ☛ Bias-variance trade-off!

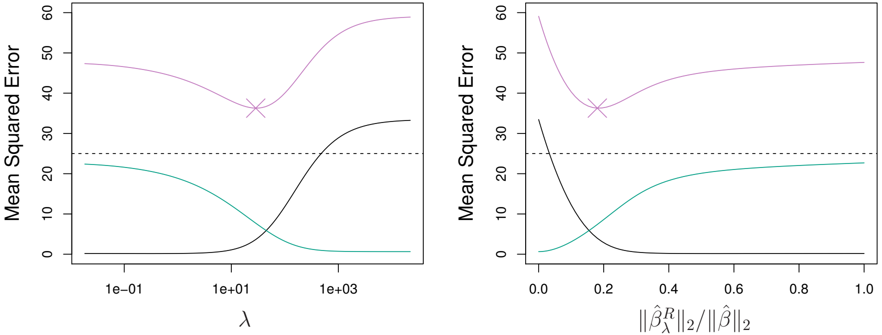


Figure 1: Simulated data with $N = 50$ observations, $P = 45$ predictors, all having nonzero coefficients. Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of λ and $\|\widehat{\beta}^{\text{ridge}}_\lambda\|_2/\|\widehat{\beta}^{\text{ls}}\|_2$. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest [James et al., 2021, Figure 6.5].

Outline

The Lasso vs ridge regression

- **Ridge regression has one obvious drawback**: unlike subset selection, which generally selects models that include only a subset of variables, ridge regression **includes all P predictors in the final model**.

The Lasso vs ridge regression

- **Ridge regression has one obvious drawback**: unlike subset selection, which generally selects models that include only a subset of variables, ridge regression **includes all P predictors in the final model**. ❓ Challenge in model interpretation in settings in which the number of variables P is quite large!
- The Lasso is a relatively recent alternative to ridge regression that overcomes this disadvantage. The lasso coefficients $\widehat{\beta}^{\text{lasso}}$ minimize the quantity

$$\widehat{\beta}^{\text{lasso}} = \operatorname*{argmin}_{\beta} \left[\underbrace{\sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2}_{\text{RSS}} + \underbrace{\lambda \sum_{p=1}^{P} |\beta_p|}_{\text{pen}_L = \lambda \|\beta_{[P]}\|_1} \right].$$

The Lasso vs ridge regression

- **Ridge regression has one obvious drawback**: unlike subset selection, which generally selects models that include only a subset of variables, ridge regression **includes all P predictors in the final model**. ❓ Challenge in model interpretation in settings in which the number of variables P is quite large!
- The Lasso is a relatively recent alternative to ridge regression that overcomes this disadvantage. The lasso coefficients $\widehat{\beta}^{\text{lasso}}$ minimize the quantity

$$\widehat{\beta}^{\text{lasso}} = \operatorname*{argmin}_{\beta} \left[\underbrace{\sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2}_{\text{RSS}} + \underbrace{\lambda \sum_{p=1}^{P} |\beta_p|}_{\text{pen}_L = \lambda \|\beta_{[P]}\|_1} \right].$$

😱 The constraint pen_L makes the solutions nonlinear in the y_n, and there is no closed form expression for $\widehat{\beta}^{\text{lasso}}$ as in ridge regression!

The Lasso vs best subset selection

- As with ridge regression, the Lasso **shrinks the coefficient estimates towards zero.**
- However, in the case of the Lasso, the $\text{pen}_L \equiv l_1$ penalty has the effect of **forcing some of the coefficient estimates to be exactly equal to zero** when the tuning parameter λ is sufficiently large.
- Hence, much like best subset selection, the Lasso performs variable selection. We say that the Lasso yields sparse models, that is, models that involve only a subset of the variables.
- As in ridge regression, selecting a good value of λ for the Lasso is critical. Model selection criteria such as cross-validation or slope heuristic are again the methods of choices.

Equivalent best subset selection (bss), Lasso and ridge

There is a one-to-one correspondence between the parameters λ and t!

$$\widehat{\beta}^{\text{bss}} = \underset{\beta}{\operatorname{argmin}} \sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2 \text{ subject to } \sum_{p=1}^{P} \mathbb{1}_{\beta_p \neq 0} \leq t, \tag{6}$$

$$\widehat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2 \text{ subject to } \sum_{p=1}^{P} \beta_p^2 \leq t, \tag{7}$$

$$\widehat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2 \text{ subject to } \sum_{p=1}^{P} |\beta_p| \leq t. \tag{8}$$

What is the most difficult problem?

Equivalent best subset selection (bss), Lasso and ridge

There is a one-to-one correspondence between the parameters λ and t!

$$\widehat{\beta}^{\text{bss}} = \underset{\beta}{\operatorname{argmin}} \sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2 \text{ subject to } \sum_{p=1}^{P} \mathbb{1}_{\beta_p \neq 0} \leq t, \tag{6}$$

$$\widehat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2 \text{ subject to } \sum_{p=1}^{P} \beta_p^2 \leq t, \tag{7}$$

$$\widehat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2 \text{ subject to } \sum_{p=1}^{P} |\beta_p| \leq t. \tag{8}$$

What is the most difficult problem? Solving (6) is computationally infeasible when P is large.

Equivalent best subset selection (bss), Lasso and ridge

There is a one-to-one correspondence between the parameters λ and t!

$$\widehat{\beta}^{\text{bss}} = \underset{\beta}{\operatorname{argmin}} \sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2 \text{ subject to } \sum_{p=1}^{P} \mathbb{1}_{\beta_p \neq 0} \leq t, \tag{6}$$

$$\widehat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2 \text{ subject to } \sum_{p=1}^{P} \beta_p^2 \leq t, \tag{7}$$

$$\widehat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2 \text{ subject to } \sum_{p=1}^{P} |\beta_p| \leq t. \tag{8}$$

What is the most difficult problem? Solving (6) is computationally infeasible when P is large. Computationally feasible alternatives to bss that replace the intractable form of the budget in (6) with forms that are much easier to solve via (7) and $\boxed{\text{especially (8) with variable selection}}$.

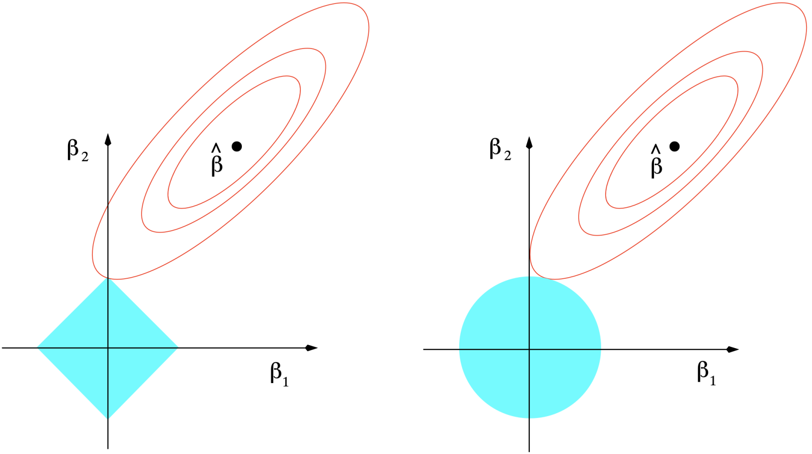


Figure 2: Estimation picture for the Lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions [Hastie et al., 2009, Figure 3.11]. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function. **When $P > 2$, the diamond becomes a rhomboid, and has many corners, flat edges and faces; there are many more opportunities for the estimated parameters to be zero.**

Outline

1. Previous Episode: Multiple Linear Regression
   - Linear regression via least squares
   - How to improve linear models?
2. Methods for Linear Model Selection and Regularization
   - The general model selection paradigm
   - Linear model selection and regularization
3. Shrinkage Methods
   - Ridge regression
   - The Lasso
   - Comparing the Lasso and Ridge Regression
   - An application to the credit data
   - An application to the prostate cancer
4. A Deep Dive into Ridge Regression
   - Solution of the ridge regression in matrix form
   - Singular value decomposition of centered input matrix
   - Principal components and ridge regression

Ridge outperforms the Lasso in terms of prediction error

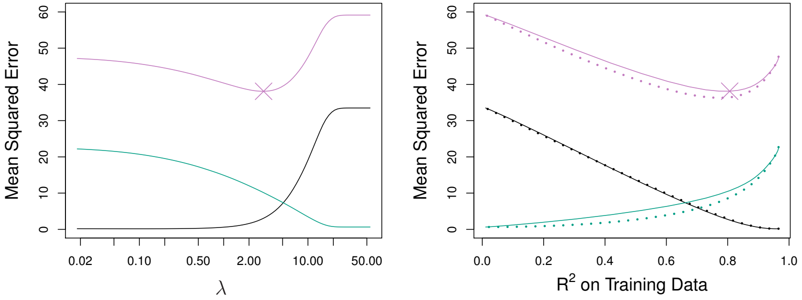


Figure 3: **Left:** Plots of squared bias (black), variance (green), and test MSE (purple) for the Lasso on simulated data set from Figure 1. **Right:** Comparison of squared bias (black), variance (green), and test MSE (purple) between Lasso (solid) and ridge (dashed). Both are plotted against their R^2 on the training data, as a common form of indexing. The crosses in both plots indicate the Lasso model for which the MSE is smallest [James et al., 2021, Figure 6.8].

The Lasso outperforms Ridge in terms of prediction error

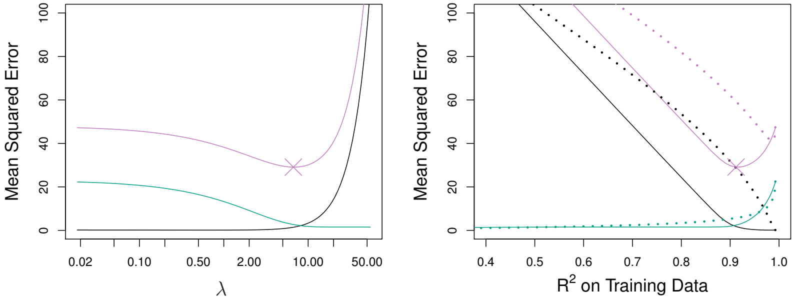


Figure 4: **Left:** Plots of squared bias (black), variance (green), and test MSE (purple) for the Lasso on simulated data set from Figure 1, except that now only two predictors are related to the response. **Right:** Comparison of squared bias (black), variance (green), and test MSE (purple) between Lasso (solid) and ridge (dashed). Both are plotted against their R^2 on the training data, as a common form of indexing. The crosses in both plots indicate the Lasso model for which the MSE is smallest [James et al., 2021, Figure 6.9].

Outline

An application to the credit data

- **Description:** the response is balance (average credit card debt for 400 individuals) and there are 6 quantitative predictors: income (in thousands of dollars), limit (credit limit), rating (credit rating), cards (number of credit cards), age, education (years of education), and 4 qualitative variables: own (house ownership), student (student status), married (Yes or No), and region (East, West or South). [James et al., 2021, Section 3.3].

An application to the credit data

- **Description:** the response is balance (average credit card debt for 400 individuals) and there are 6 quantitative predictors: income (in thousands of dollars), limit (credit limit), rating (credit rating), cards (number of credit cards), age, education (years of education), and 4 qualitative variables: own (house ownership), student (student status), married (Yes or No), and region (East, West or South). [James et al., 2021, Section 3.3].
- **Goal:** develop an accurate model that can be used to **predict balance** on the basis of 10 predictors ← Using glmnet package in $\boldsymbol{R}$.

```
> head(Credit)                                                          Y
   Income Limit Rating Cards Age Education Own Student Married Region Balance
1  14.891  3606    283     2  34        11  No      No     Yes  South     333
2 106.025  6645    483     3  82        15 Yes     Yes     Yes   West     903
3 104.593  7075    514     4  71        11  No      No      No   West     580
4 148.924  9504    681     3  36        11 Yes      No      No   West     964
5  55.882  4897    357     2  68        16  No      No     Yes  South     331
6  80.180  8047    569     4  77        10  No      No      No  South    1151
```

In the left-hand panel, each curve corresponds to the ridge regression coefficient estimate for one of the ten variables, plotted as a function of λ.

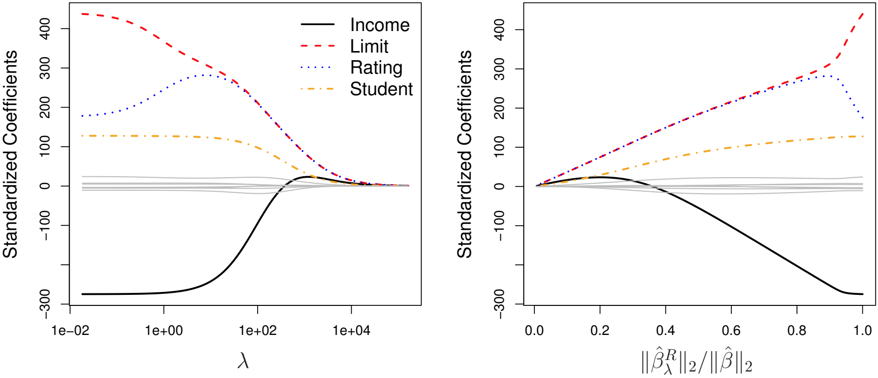


Figure 5: The standardized ridge regression coefficients are displayed for the Credit data set, as a function of λ and $\|\widehat{\beta}^{\text{ridge}}_\lambda\|_2/\|\widehat{\beta}^{\text{ls}}\|_2$ [James et al., 2021, Figure 6.4].

In the right-hand panel, a small value of the x-axis indicates that the ridge regression coefficient estimates have been shrunken very close to zero.

In the left-hand panel, each curve corresponds to the Lasso regression coefficient estimate for one of the ten variables, plotted as a function of λ.

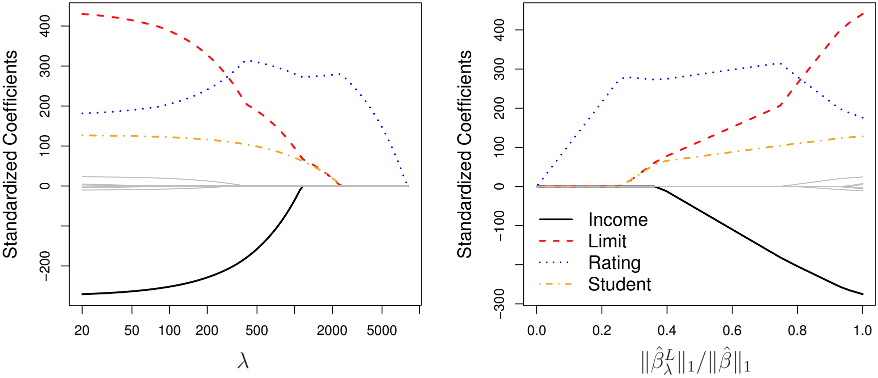


Figure 6: The standardized Lasso coefficients are displayed for the Credit data set, as a function of λ and $\|\widehat{\beta}^{\text{lasso}}_\lambda\|_1/\|\widehat{\beta}^{\text{ls}}\|_1$ [James et al., 2021, Figure 6.6].

In the right-hand panel, a small value of the x-axis indicates that the Lasso regression coefficient estimates have been shrunken to zero.

Outline

An application to the prostate cancer

- **Description:** represent the correlation between the **level of prostate specific antigen** (PSA) and a number of **clinical measures**, in 97 men who were about to receive a radical prostatectomy[13].

[13] Stamey et al. (1989). "Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II radical prostatectomy treated patients", Journal of Urology.

An application to the prostate cancer

- **Description:** represent the correlation between the **level of prostate specific antigen** (PSA) and a number of **clinical measures**, in 97 men who were about to receive a radical prostatectomy[13].
- **Goal:** predict the log of PSA (lpsa) from a number of measurements including log cancer volume (lcavol), log prostate weight (lweight), age, log of benign prostatic hyperplasia amount (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45).

```
> head(df)
      lcavol  lweight age      lbph svi       lcp gleason pgg45       lpsa
1 -0.5798185 2.769459  50 -1.386294   0 -1.386294       6     0 -0.4307829
2 -0.9942523 3.319626  58 -1.386294   0 -1.386294       6     0 -0.1625189
3 -0.5108256 2.691243  74 -1.386294   0 -1.386294       7    20 -0.1625189
4 -1.2039728 3.282789  58 -1.386294   0 -1.386294       6     0 -0.1625189
5  0.7514161 3.432373  62 -1.386294   0 -1.386294       6     0  0.3715636
6 -1.0498221 3.228826  50 -1.386294   0 -1.386294       6     0  0.7654678
```

[13] Stamey et al. (1989). "Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II radical prostatectomy treated patients", Journal of Urology.

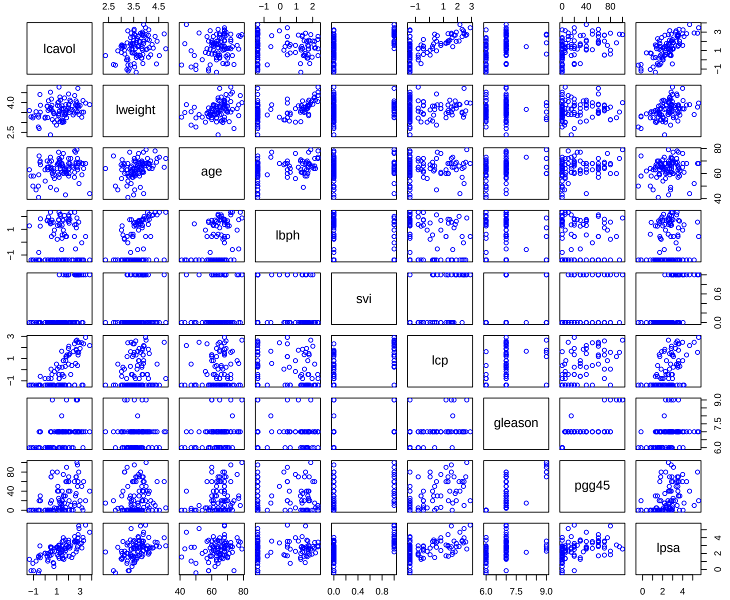


Figure 7: Pairwise scatterplots between predictors of prostate cancer sample.

Mathematical equation for linear regression model in prostate cancer

Goal: predict the log of PSA (lpsa) from a number of measurements including log cancer volume (lcavol), log prostate weight (lweight), age, log of benign prostatic hyperplasia amount (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45).

Mathematical equation for linear regression model in prostate cancer

Goal: predict the log of PSA (lpsa) from a number of measurements including log cancer volume (lcavol), log prostate weight (lweight), age, log of benign prostatic hyperplasia amount (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45).

$$\begin{aligned}\texttt{lpsa} = \beta_0 &+ \beta_1\ \texttt{lcavol} + \beta_2\ \texttt{lweight} + \beta_3\ \texttt{age} \\ &+ \beta_4\ \texttt{lbph} + \beta_5\ \texttt{svi1} + \beta_6\ \texttt{lcp} \\ &+ \beta_7\ \texttt{gleason7} + \beta_8\ \texttt{gleason8} + \beta_9\ \texttt{gleason9} \\ &+ \beta_{10}\ \texttt{pgg45} + \varepsilon.\end{aligned}$$

$$\texttt{svi1} = \begin{cases} 1 & \text{if } \texttt{svi} \text{ is } 1 \\ 0 & \text{if } \texttt{svi} \text{ is not } 1 \end{cases} \qquad \texttt{gleason7} = \begin{cases} 1 & \text{if } \texttt{gleason} \text{ is } 7 \\ 0 & \text{if } \texttt{gleason} \text{ is not } 7 \end{cases}$$

$$\texttt{gleason8} = \begin{cases} 1 & \text{if } \texttt{gleason} \text{ is } 8 \\ 0 & \text{if } \texttt{gleason} \text{ is not } 8 \end{cases} \qquad \texttt{gleason9} = \begin{cases} 1 & \text{if } \texttt{gleason} \text{ is } 9 \\ 0 & \text{if } \texttt{gleason} \text{ is not } 9 \end{cases}$$

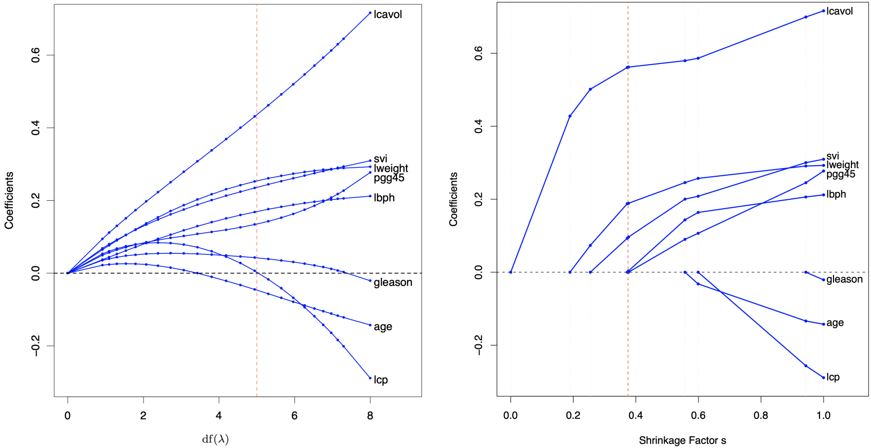


Figure 8: Profiles of ridge (left-hand panel, plotted versus $df(\lambda)$, the effective degrees of freedom) and Lasso (right-hand panel, plotted versus the standardized tunning parameter $s = t/\|\widehat{\beta}^{\text{ls}}\|_1$, role of t is similar to λ) coefficients for the prostate cancer example. The vertical lines are drawn at $df = 5.0$ and at $s = 0.36$, the values chosen by cross-validation. [Hastie et al., 2009, Figures 3.8 and 3.10].

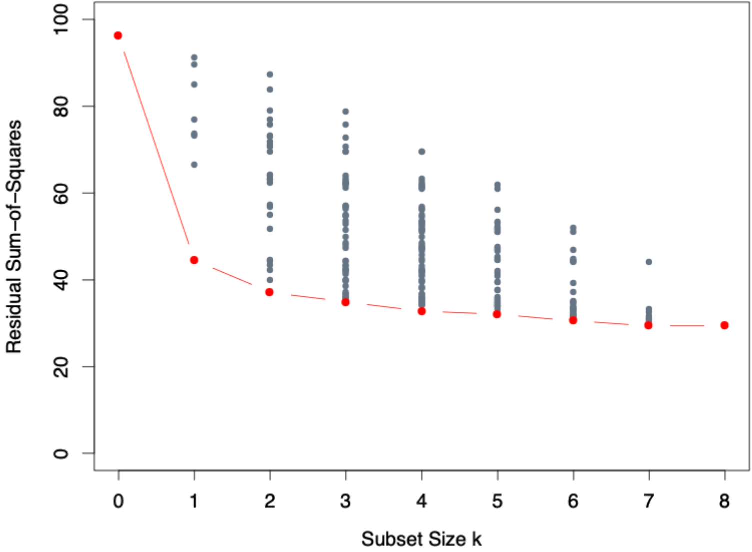


Figure 9: Best subset selection for the prostate cancer example. At each subset size is shown the RSS for each model of that size [Hastie et al., 2009, Figure 3.5].

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
`Intercept`	2.465	2.477	2.452	2.468	2.497	2.452
`lcavol`	0.680	0.740	0.420	0.533	0.543	0.419
`lweight`	0.263	0.316	0.238	0.169	0.289	0.344
`age`	−0.141		−0.046		−0.152	−0.026
`lbph`	0.210		0.162	0.002	0.214	0.220
`svi`	0.305		0.227	0.094	0.315	0.243
`lcp`	−0.288		0.000		−0.051	0.079
`gleason`	−0.021		0.040		0.232	0.011
`pgg45`	0.267		0.133		−0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152

Figure 10: Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data. The blank entries correspond to variables omitted [Hastie et al., 2009, Table 3.3].

Each method has a **complexity parameter**, and this was chosen to minimize an estimate of prediction error based on **model selection criteria, *e.g.,* tenfold cross-validation.**
Original data (97) = training set (67, cross-validation) + test set (30, juge performance of the selected model).

Outline

Outline

Solution of the ridge regression in matrix form

- Recall that the LS estimates $\widehat{\beta}^{\text{ls}} = (\widehat{\beta}_0^{\text{ls}}, \widehat{\beta}_1^{\text{ls}}, \ldots, \widehat{\beta}_P^{\text{ls}})^\top = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ minimizes

$$\sum_{n=1}^{N}\left(y_n - \beta_0 - \sum_{p=1}^{P}\beta_p x_{np}\right)^2 \equiv (\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta)\text{: quadratic function of } \beta.$$

Here, $\mathbf{X} = (x_{np})_{N\times(P+1)} \equiv N \times (P+1)$ matrix, $\mathbf{y} = (y_n)_{N\times 1} \equiv N$ vector of outputs.
If $\text{rank}(\mathbf{X}) = P + 1$ then $\mathbf{X}^\top\mathbf{X}$ is positive definite and $\widehat{\beta}^{\text{ls}}$ is the unique solution.

Solution of the ridge regression in matrix form

- Recall that the LS estimates $\widehat{\beta}^{\mathrm{ls}} = (\widehat{\beta}_0^{\mathrm{ls}}, \widehat{\beta}_1^{\mathrm{ls}}, \ldots, \widehat{\beta}_P^{\mathrm{ls}})^\top = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ minimizes

$$\sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2 \equiv (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta)\text{: quadratic function of } \beta.$$

  Here, $\mathbf{X} = (x_{np})_{N \times (P+1)} \equiv N \times (P+1)$ matrix, $\mathbf{y} = (y_n)_{N \times 1} \equiv N$ vector of outputs. If $\mathrm{rank}(\mathbf{X}) = P + 1$ then $\mathbf{X}^\top \mathbf{X}$ is positive definite and $\widehat{\beta}^{\mathrm{ls}}$ is the unique solution.
- Recall $\widehat{\beta}^{\mathrm{ridge}}$ [Hastie et al., 2009, Section 3.4] minimizes

$$\sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2 + \lambda \sum_{p=1}^{P} \beta_p^2 \text{ ? Matrix form } \ldots. \tag{9}$$

Solution of the ridge regression in matrix form

- Recall that the LS estimates $\widehat{\beta}^{\text{ls}} = (\widehat{\beta}_0^{\text{ls}}, \widehat{\beta}_1^{\text{ls}}, \ldots, \widehat{\beta}_P^{\text{ls}})^\top = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ minimizes

$$\sum_{n=1}^{N}\left(y_n - \beta_0 - \sum_{p=1}^{P}\beta_p x_{np}\right)^2 \equiv (\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta)\text{: quadratic function of } \beta.$$

Here, $\mathbf{X} = (x_{np})_{N\times(P+1)} \equiv N \times (P+1)$ matrix, $\mathbf{y} = (y_n)_{N\times 1} \equiv N$ vector of outputs. If $\text{rank}(\mathbf{X}) = P+1$ then $\mathbf{X}^\top\mathbf{X}$ is positive definite and $\widehat{\beta}^{\text{ls}}$ is the unique solution.

- Recall $\widehat{\beta}^{\text{ridge}}$ [Hastie et al., 2009, Section 3.4] minimizes

$$\sum_{n=1}^{N}\left(y_n - \beta_0 - \sum_{p=1}^{P}\beta_p x_{np}\right)^2 + \lambda\sum_{p=1}^{P}\beta_p^2 \quad \text{? Matrix form} \ldots. \tag{9}$$

After reparametrization using centered inputs, *i.e.*, replace $x_{np} \equiv x_{np} - \overline{x}_p$, ? then minimizing (9) w.r.t. $\widehat{\beta}^{\text{ridge}}$ is equivalent to minimize (10) w.r.t. centered coefficient $\widehat{\beta}^{\text{ridge}\ c}$,

$$\sum_{n=1}^{N}\left(y_n - \beta_0^c - \sum_{p=1}^{P}(x_{np} - \overline{x}_p)\beta_p^c\right)^2 + \lambda\sum_{p=1}^{P}\beta_p^{c2}. \tag{10}$$

Explanation of equivalent centered ridge regression problem

1. By inserting zero as $\overline{x}_p - \overline{x}_p$, $\overline{x}_p = \frac{1}{N}\sum_{n=1}^{N} \overline{x}_{np}$, we obtain

$$\mathrm{RSS}_R(\beta^c) \equiv \sum_{n=1}^{N} \left(y_n - \underbrace{\left(\beta_0 + \sum_{p=1}^{P} \overline{x}_p \beta_p \right)}_{\beta_0^c} - \sum_{p=1}^{P} (x_{np} - \overline{x}_p) \underbrace{\beta_p}_{\beta_p^c} \right)^2 + \lambda \sum_{p=1}^{P} \underbrace{\beta_p^2}_{\beta_p^{c2}}.$$

The equivalence of the minimization results from the fact that if $\widehat{\beta}^{\text{ridge}}$ minimize its respective functional, the $\widehat{\beta}^{\text{ridge}\,c}$'s will do the same.

Explanation of equivalent centered ridge regression problem

1. By inserting zero as $\overline{x}_p - \overline{x}_p$, $\overline{x}_p = \frac{1}{N}\sum_{n=1}^{N} \overline{x}_{np}$, we obtain

$$\mathrm{RSS}_R(\beta^c) \equiv \sum_{n=1}^{N}\left(y_n - \underbrace{\left(\beta_0 + \sum_{p=1}^{P} \overline{x}_p \beta_p\right)}_{\beta_0^c} - \sum_{p=1}^{P}(x_{np} - \overline{x}_p)\underbrace{\beta_p}_{\beta_p^c}\right)^2 + \lambda \sum_{p=1}^{P} \underbrace{\beta_p^2}_{\beta_p^{c2}}.$$

The equivalence of the minimization results from the fact that if $\widehat{\beta}^{\text{ridge}}$ minimize its respective functional, the $\widehat{\beta}^{\text{ridge }c}$'s will do the same.

2. We **compute the value of $\widehat{\beta}_0^{\text{ridge }c}$** in the above expression by setting the derivative of $\mathrm{RSS}_R(\beta^c)$ w.r.t. β_0^c equal to zero, we have

$$\sum_{n=1}^{N}\left(y_n - \beta_0^c - \sum_{p=1}^{P}(x_{np} - \overline{x}_p)\beta_p\right) = 0 \Leftrightarrow \beta_0^c = \frac{1}{N}\sum_{n=1}^{N} y_n \equiv \overline{y}. \tag{11}$$

Explanation of equivalent centered ridge regression problem

1. By inserting zero as $\overline{x}_p - \overline{x}_p$, $\overline{x}_p = \frac{1}{N}\sum_{n=1}^{N} \overline{x}_{np}$, we obtain

$$\mathrm{RSS}_R(\beta^c) \equiv \sum_{n=1}^{N}\left(y_n - \underbrace{\left(\beta_0 + \sum_{p=1}^{P} \overline{x}_p \beta_p\right)}_{\beta_0^c} - \sum_{p=1}^{P} (x_{np} - \overline{x}_p) \underbrace{\beta_p}_{\beta_p^c} \right)^2 + \lambda \sum_{p=1}^{P} \underbrace{\beta_p^2}_{\beta_p^{c2}}.$$

The equivalence of the minimization results from the fact that if $\widehat{\beta}^{\text{ridge}}$ minimize its respective functional, the $\widehat{\beta}^{\text{ridge } c}$'s will do the same.

2. We **compute the value of** $\widehat{\beta}_0^{\text{ridge } c}$ in the above expression by setting the derivative of $\mathrm{RSS}_R(\beta^c)$ w.r.t. β_0^c equal to zero, we have

$$\sum_{n=1}^{N}\left(y_n - \beta_0^c - \sum_{p=1}^{P} (x_{np} - \overline{x}_p)\beta_p \right) = 0 \Leftrightarrow \beta_0^c = \frac{1}{N}\sum_{n=1}^{N} y_n \equiv \overline{y}. \tag{11}$$

The remaining coefficients get estimated by a ridge regression without intercept using the centered x_{np}. Henceforth we assume that this centering has been done, so that the input matrix $\mathbf{X}$ has P (rather than $P+1$) columns and $\beta^c \equiv \beta = (\beta_p)_{P\times 1}$.

Solution of the ridge regression in matrix form

- Recall that the LS estimates $\widehat{\beta}^{\text{ls}} = (\widehat{\beta}_0^{\text{ls}}, \widehat{\beta}_1^{\text{ls}}, \ldots, \widehat{\beta}_P^{\text{ls}})^\top = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ minimizes

$$\sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2 \equiv (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta): \text{ quadratic function of } \beta.$$

Here, $\mathbf{X} = (x_{np})_{N \times (P+1)} \equiv N \times (P+1)$ matrix, $\mathbf{y} = (y_n)_{N \times 1} \equiv N$ vector of outputs.
If $\text{rank}(\mathbf{X}) = P + 1$ then $\mathbf{X}^\top \mathbf{X}$ is positive definite and $\widehat{\beta}^{\text{ls}}$ is the unique solution.

Solution of the ridge regression in matrix form

- Recall that the LS estimates $\widehat{\beta}^{\text{ls}} = (\widehat{\beta}_0^{\text{ls}}, \widehat{\beta}_1^{\text{ls}}, \ldots, \widehat{\beta}_P^{\text{ls}})^\top = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ minimizes

$$\sum_{n=1}^{N} \left(y_n - \beta_0 - \sum_{p=1}^{P} \beta_p x_{np} \right)^2 \equiv (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta): \text{ quadratic function of } \beta.$$

Here, $\mathbf{X} = (x_{np})_{N \times (P+1)} \equiv N \times (P+1)$ matrix, $\mathbf{y} = (y_n)_{N \times 1} \equiv N$ vector of outputs.
If $\text{rank}(\mathbf{X}) = P + 1$ then $\mathbf{X}^\top \mathbf{X}$ is positive definite and $\widehat{\beta}^{\text{ls}}$ is the unique solution.

- After reparametrization using centered inputs $\mathbf{X}$ and output $\mathbf{y}$, we proved that $\widehat{\beta}_0^{\text{ridge}} = \overline{y}$ and $\widehat{\beta}^{\text{ridge}} = (\widehat{\beta}_p^{\text{ridge}})_{P \times 1} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$ minimizes

$$\sum_{n=1}^{N} \left(y_n - \sum_{p=1}^{P} \beta_p x_{np} \right)^2 + \lambda \sum_{p=1}^{P} \beta_p^2 \equiv (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) + \lambda \beta^\top \beta: \text{ quadratic of } \beta.$$

Here, $\mathbf{X} = (x_{np})_{N \times (P)}$ centered input matrix, $\mathbf{y} = (y_n)_{N \times 1}$ centered output vector.

Solution of the ridge regression in matrix form

- Recall that the LS estimates $\widehat{\beta}^{\text{ls}} = (\widehat{\beta}_0^{\text{ls}}, \widehat{\beta}_1^{\text{ls}}, \dots, \widehat{\beta}_P^{\text{ls}})^\top = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ minimizes

$$\sum_{n=1}^{N}\left(y_n - \beta_0 - \sum_{p=1}^{P}\beta_p x_{np}\right)^2 \equiv (\mathbf{y}-\mathbf{X}\beta)^\top(\mathbf{y}-\mathbf{X}\beta)\text{: quadratic function of } \beta.$$

Here, $\mathbf{X} = (x_{np})_{N\times(P+1)} \equiv N \times (P+1)$ matrix, $\mathbf{y} = (y_n)_{N\times 1} \equiv N$ vector of outputs.
If $\text{rank}(\mathbf{X}) = P+1$ then $\mathbf{X}^\top\mathbf{X}$ is positive definite and $\widehat{\beta}^{\text{ls}}$ is the unique solution.

- After reparametrization using centered inputs $\mathbf{X}$ and output $\mathbf{y}$, we proved that $\widehat{\beta}_0^{\text{ridge}} = \overline{y}$ and $\widehat{\beta}^{\text{ridge}} = (\widehat{\beta}_p^{\text{ridge}})_{P\times 1} = (\mathbf{X}^\top\mathbf{X} + \lambda\,\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$ minimizes

$$\sum_{n=1}^{N}\left(y_n - \sum_{p=1}^{P}\beta_p x_{np}\right)^2 + \lambda\sum_{p=1}^{P}\beta_p^2 \equiv (\mathbf{y}-\mathbf{X}\beta)^\top(\mathbf{y}-\mathbf{X}\beta) + \lambda\beta^\top\beta\text{: quadratic of } \beta.$$

Here, $\mathbf{X} = (x_{np})_{N\times(P)}$ centered input matrix, $\mathbf{y} = (y_n)_{N\times 1}$ centered output vector.
Even if $\text{rank}(\mathbf{X}) < P$ then $(\mathbf{X}^\top\mathbf{X} + \lambda\,\mathbf{I})^{-1}$, $\lambda > 0$, is always positive definite and $\widehat{\beta}^{\text{ridge}}$ is the unique solution Main motivation for ridge regression!

Outline

Singular value decomposition of centered input matrix

- Singular value decomposition (SVD) of $N \times P$ matrix $\mathbf{X}$ has the form

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top. \tag{12}$$

Here $\mathbf{U}$ and $\mathbf{V}$ are $N \times P$ and $P \times P$ orthogonal matrices,*i.e.*, $\mathbf{U}^\top\mathbf{U} = \mathbf{U}^\top\mathbf{U} = \mathbf{I}$, with the columns of $\mathbf{U}$ spanning the column space of $\mathbf{X}$, and the columns of $\mathbf{V}$ spanning the row space. $\mathbf{D}$ is a $P \times P$ diagonal matrix, with diagonal entries $d_1 \geq d_2 \geq \ldots \geq d_P \geq 0$ called the singular values of $\mathbf{X}$. If one or more values $d_p = 0$, $\mathbf{X}$ is singular.

Singular value decomposition of centered input matrix

- Singular value decomposition (SVD) of $N \times P$ matrix $\mathbf{X}$ has the form

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top. \tag{12}$$

  Here $\mathbf{U}$ and $\mathbf{V}$ are $N \times P$ and $P \times P$ orthogonal matrices,*i.e.*, $\mathbf{U}^\top\mathbf{U} = \mathbf{U}^\top\mathbf{U} = \mathbf{I}$, with the columns of $\mathbf{U}$ spanning the column space of $\mathbf{X}$, and the columns of $\mathbf{V}$ spanning the row space. $\mathbf{D}$ is a $P \times P$ diagonal matrix, with diagonal entries $d_1 \geq d_2 \geq \ldots \geq d_P \geq 0$ called the singular values of $\mathbf{X}$. If one or more values $d_p = 0$, $\mathbf{X}$ is singular.
- Using the SVD we can write the LS fitted vector as

$$\mathbf{X}\widehat{\beta}^{\mathrm{ls}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \ldots = \mathbf{U}\mathbf{U}^\top\mathbf{y} \text{ ? Verify this using orthogonal matrices...} \tag{13}$$

Singular value decomposition of centered input matrix

- Singular value decomposition (SVD) of $N \times P$ matrix $\mathbf{X}$ has the form

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top = [\mathbf{u}_1 \ldots \mathbf{u}_P] \operatorname{diag}(d_1, \ldots, d_P)[\mathbf{v}_1 \ldots \mathbf{v}_P]^\top = \sum_{p=1}^{P} d_p \mathbf{u}_p \mathbf{v}_p^\top. \tag{14}$$

  Here $\mathbf{U}$ and $\mathbf{V}$ are $N \times P$ and $P \times P$ orthogonal matrices, *i.e.*, $\mathbf{U}^\top\mathbf{U} = \mathbf{U}^\top\mathbf{U} = \mathbf{I}$, with the columns of $\mathbf{U}$ spanning the column space of $\mathbf{X}$, and the columns of $\mathbf{V}$ spanning the row space. $\mathbf{D}$ is a $P \times P$ diagonal matrix, with diagonal entries $d_1 \geq d_2 \geq \ldots \geq d_P \geq 0$ called the singular values of $\mathbf{X}$. If one or more values $d_p = 0$, $\mathbf{X}$ is singular.
- Using the SVD we can write the LS fitted vector as

$$\begin{aligned}
\mathbf{X}\widehat{\beta}^{\text{ls}} &= \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{U}\mathbf{D}\mathbf{V}^\top[(\mathbf{U}\mathbf{D}\mathbf{V}^\top)^\top\mathbf{U}\mathbf{D}\mathbf{V}^\top]^{-1}(\mathbf{U}\mathbf{D}\mathbf{V}^\top)^\top\mathbf{y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}^\top(\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{U}\mathbf{D}\mathbf{V}^\top)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{y} = \mathbf{U}\mathbf{D}\mathbf{V}^\top(\mathbf{V}\mathbf{D}^2\mathbf{V}^\top)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}^\top(\mathbf{V}^\top)^{-1}\mathbf{D}^{-2}\mathbf{V}^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{y} = \mathbf{U}\mathbf{U}^\top\mathbf{y}.
\end{aligned} \tag{15}$$

  Note that $\mathbf{U}^\top\mathbf{y}$ are the coordinates of $\mathbf{y}$ with respect to the orthonormal basis $\mathbf{U}$.

Singular value decomposition of centered input matrix

- Singular value decomposition (SVD) of $N \times P$ matrix $\mathbf{X}$ has the form

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top = [\mathbf{u}_1 \ldots \mathbf{u}_P] \operatorname{diag}(d_1, \ldots, d_P)[\mathbf{v}_1 \ldots \mathbf{v}_P]^\top = \sum_{p=1}^{P} d_p \mathbf{u}_p \mathbf{v}_p^\top. \quad (14)$$

  Here $\mathbf{U}$ and $\mathbf{V}$ are $N \times P$ and $P \times P$ orthogonal matrices,*i.e.*, $\mathbf{U}^\top\mathbf{U} = \mathbf{U}^\top\mathbf{U} = \mathrm{I}$, with the columns of $\mathbf{U}$ spanning the column space of $\mathbf{X}$, and the columns of $\mathbf{V}$ spanning the row space. $\mathbf{D}$ is a $P \times P$ diagonal matrix, with diagonal entries $d_1 \geq d_2 \geq \ldots \geq d_P \geq 0$ called the singular values of $\mathbf{X}$. If one or more values $d_p = 0$, $\mathbf{X}$ is singular.
- Using the SVD we can write the LS fitted vector as

$$\begin{aligned}
\mathbf{X}\widehat{\beta}^{\mathrm{ls}} &= \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{U}\mathbf{D}\mathbf{V}^\top[(\mathbf{U}\mathbf{D}\mathbf{V}^\top)^\top\mathbf{U}\mathbf{D}\mathbf{V}^\top]^{-1}(\mathbf{U}\mathbf{D}\mathbf{V}^\top)^\top\mathbf{y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}^\top(\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{U}\mathbf{D}\mathbf{V}^\top)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{y} = \mathbf{U}\mathbf{D}\mathbf{V}^\top(\mathbf{V}\mathbf{D}^2\mathbf{V}^\top)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}^\top(\mathbf{V}^\top)^{-1}\mathbf{D}^{-2}\mathbf{V}^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{y} = \mathbf{U}\mathbf{U}^\top\mathbf{y}. \quad (15)
\end{aligned}$$

  Note that $\mathbf{U}^\top\mathbf{y}$ are the coordinates of $\mathbf{y}$ with respect to the orthonormal basis $\mathbf{U}$.
- Similarly, the ridge solutions are $\mathbf{X}\widehat{\beta}^{\mathrm{ridge}} = \ldots$?

Singular value decomposition of centered input matrix

- Singular value decomposition (SVD) of $N \times P$ matrix $\mathbf{X}$ has the form

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top = [\mathbf{u}_1 \dots \mathbf{u}_P] \operatorname{diag}(d_1, \dots, d_P)[\mathbf{v}_1 \dots \mathbf{v}_P]^\top = \sum_{p=1}^{P} d_p \mathbf{u}_p \mathbf{v}_p^\top. \tag{14}$$

  Here $\mathbf{U}$ and $\mathbf{V}$ are $N \times P$ and $P \times P$ orthogonal matrices,*i.e.*, $\mathbf{U}^\top\mathbf{U} = \mathbf{U}^\top\mathbf{U} = \mathrm{I}$, with the columns of $\mathbf{U}$ spanning the column space of $\mathbf{X}$, and the columns of $\mathbf{V}$ spanning the row space. $\mathbf{D}$ is a $P \times P$ diagonal matrix, with diagonal entries $d_1 \geq d_2 \geq \ldots \geq d_P \geq 0$ called the singular values of $\mathbf{X}$. If one or more values $d_p = 0$, $\mathbf{X}$ is singular.
- Using the SVD we can write the LS fitted vector as

$$\begin{aligned}
\mathbf{X}\widehat{\beta}^{\mathrm{ls}} &= \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{U}\mathbf{D}\mathbf{V}^\top[(\mathbf{U}\mathbf{D}\mathbf{V}^\top)^\top\mathbf{U}\mathbf{D}\mathbf{V}^\top]^{-1}(\mathbf{U}\mathbf{D}\mathbf{V}^\top)^\top\mathbf{y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}^\top(\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{U}\mathbf{D}\mathbf{V}^\top)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{y} = \mathbf{U}\mathbf{D}\mathbf{V}^\top(\mathbf{V}\mathbf{D}^2\mathbf{V}^\top)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}^\top(\mathbf{V}^\top)^{-1}\mathbf{D}^{-2}\mathbf{V}^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{y} = \mathbf{U}\mathbf{U}^\top\mathbf{y}.
\end{aligned} \tag{15}$$

  Note that $\mathbf{U}^\top\mathbf{y}$ are the coordinates of $\mathbf{y}$ with respect to the orthonormal basis $\mathbf{U}$.
- Similarly, the ridge solutions are $\mathbf{X}\widehat{\beta}^{\mathrm{ridge}} = \ldots$?

$$\begin{aligned}
\mathbf{X}\widehat{\beta}^{\mathrm{ridge}} &= \mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\,\mathrm{I})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{U}\mathbf{D}\mathbf{V}^\top[(\mathbf{U}\mathbf{D}\mathbf{V}^\top)^\top\mathbf{U}\mathbf{D}\mathbf{V}^\top + \lambda\,\mathrm{I}]^{-1}(\mathbf{U}\mathbf{D}\mathbf{V}^\top)^\top\mathbf{y} \\
&= \mathbf{U}\mathbf{D}\mathbf{V}^\top[\mathbf{V}(\mathbf{D}^2 + \lambda\,\mathrm{I})\mathbf{V}^\top]^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{y} = \mathbf{U}\mathbf{D}(\mathbf{D}^2+\lambda\,\mathrm{I})^{-1}\mathbf{D}\mathbf{U}^\top\mathbf{y} \\
&= \sum_{p=1}^{P} \mathbf{u}_p \underbrace{\frac{d_p^2}{d_p^2 + \lambda}}_{\leq 1,\ \text{shrinkage}} \mathbf{u}_p^\top\mathbf{y}, \text{ where } \mathbf{u}_p \text{ are columns of } \mathbf{U}.
\end{aligned} \tag{16}$$

Outline

Principal components and ridge regression

$\boxed{\text{What does a small value of } d_p^2 \text{ mean?}}$

$\boxed{\text{What does a small value of } d_p^2 \text{ mean?}}$

Using SVD $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, we obtain

$$\mathbf{X}^\top\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^\top: \text{ eigen decomposition.} \tag{18}$$

Principal components and ridge regression

$\boxed{\text{What does a small value of } d_p^2 \text{ mean?}}$

Using SVD $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, we obtain

$$\mathbf{X}^\top\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^\top: \text{ eigen decomposition.} \tag{18}$$

- The eigenvectors $\mathbf{v}_p$ (columns of $\mathbf{V}$) are also called the principal components (or Karhunen–Loeve) directions of $\mathbf{X}$.
- The principal components of $\mathbf{X}$ is defined as $\mathbf{z}_p = \mathbf{X}\mathbf{v}_p = \mathbf{u}_p d_p$ and $\operatorname{var}(\mathbf{z}_p) = \operatorname{var}(\mathbf{u}_p d_p) = d_p^2/N$.

Principal components and ridge regression

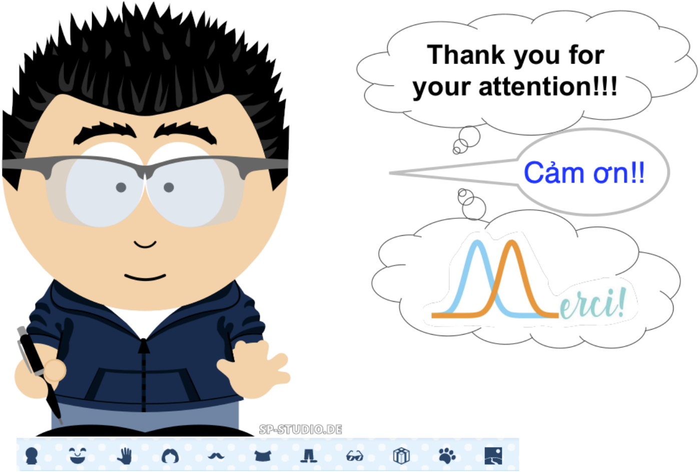
$\boxed{\text{What does a small value of } d_p^2 \text{ mean?}}$

Using SVD $\mathbf{X} = \mathbf{UDV}^\top$, we obtain

$$\mathbf{X}^\top \mathbf{X} = \mathbf{VD}^2\mathbf{V}^\top: \text{ eigen decomposition.} \tag{18}$$

- The eigenvectors $\mathbf{v}_p$ (columns of $\mathbf{V}$) are also called the principal components (or Karhunen–Loeve) directions of $\mathbf{X}$.
- The principal components of $\mathbf{X}$ is defined as $\mathbf{z}_p = \mathbf{X}\mathbf{v}_p = \mathbf{u}_p d_p$ and $\text{var}(\mathbf{z}_p) = \text{var}(\mathbf{u}_p d_p) = d_p^2/N$.
- Recall the fact that from SVD, $d_1 \geq d_2 \geq \ldots \geq d_P \geq 0$. Therefore, the first principal component $\mathbf{z}_1$ of $\mathbf{X}$ has the largest sample variance amongst all normalized linear combinations of the columns of $\mathbf{X}$.

“Essentially, all models are wrong, but some are useful”.[14]



⬆ This is my best data-driven model to approximate myself.

[14] Box, G. E.P. (1979). “Robustness in the strategy of scientific model building”. In Robustness in Statistics (pp. 201-236). Academic Press.

References I

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009).
The elements of statistical learning: data mining, inference, and prediction, volume 2.
Springer.

(Cited on pages 33, 34, 55, 70, 71, 72, 75, 76, and 77.)

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021).
An Introduction to Statistical Learning: with Applications in R.
Springer Texts in Statistics. Springer US.

(Cited on pages 33, 34, 44, 45, 46, 57, 58, 60, 61, 62, and 63.)