1 Factorizations, Error sensitivity, Matrices

If $\mathbf{X}$ denotes a matrix, $\kappa(\mathbf{X})$ denotes its condition number: the ratio between the largest singular value and the smallest singular value.

1. If $\mathbf{Q} \times \mathbf{R} = \mathbf{X}$ is the QR décomposition of $\mathbf{X}$, check that $\mathbf{Q} \times \mathbf{Q}^T$ is the orthogonal projection on the subspace generated by the columns of $\mathbf{X}$.

2. Check the little perturbation bound for the OLS solution $\widehat{\theta}$ of $\arg\min_\theta \|\mathbf{X}\theta - y\|_2$ when y is perturbed by $\delta(y)$:
$$\|\delta(\widehat{\theta})\|_2 \leq \kappa(\mathbf{X}) \frac{\|\delta(y)\|_2}{\|\mathbf{X}\|_2} .$$
Hint: use SVD of $\mathbf{X}$.

3. Check the complete perturbation bound for the OLS solution $\widehat{\theta}$ of $\arg\min_\theta \|\mathbf{X}\theta - y\|_2$ when y is perturbed by $\delta(y)$:
$$\|\delta(\widehat{\theta})\|_2 \leq \kappa(\mathbf{X}) \frac{\|\delta(y)\|_2}{\|\mathbf{X}\|_2} .$$
Hint: use SVD of $\mathbf{X}$.

4. Solving least square problems (minimize $\|\mathbf{X}\theta - Y\|_2^2$) by computing $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ may lead to numerical problems. Take
$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ \epsilon & 0 \\ 0 & \epsilon \end{pmatrix} .$$
Assume that ϵ is larger than machine precision while ϵ^2 is not
   (a) Compute $\mathbf{X}^T\mathbf{X}$. Is the machine able to invert $\mathbf{X}^T\mathbf{X}$?
   (b) What are the singular values of $\mathbf{X}$, of $\mathbf{X}^T\mathbf{X}$?
   (c) What is the condition number of $\mathbf{X}$, of $\mathbf{X}^T\mathbf{X}$?
   (d) What is the QR decomposition of $\mathbf{X}$? Does this QR decomposition suffer from the same precision problem as $\mathbf{X}^T\mathbf{X}$?

5. Let $\mathbf{X}$ be an $n \times p$ matrix with rank p. Let $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ denote the Hat matrix.
   (a) Check that $\mathbf{H}$ represents the orthogonal projection of on the linear subspace generated by the columns of $\mathbf{X}$.
   (b) Check that the diagonal coefficients of $\mathbf{H}$ are non-negative and that the trace of $\mathbf{H}$ equals p.
   (c) Check that $\sum_{j=1}^n H[i,j] = 1$.
   (d) Check that the diagonal coefficients of $\mathbf{H}$ are not smaller than $1/n$ and not larger than 1.
   (e) Hat matrix and QR decomposition ...

   In linear regression, the ith diagonal coefficient of the Hat matrix $H[i,i]$ is called the *leverage* of the ith case.

6. The Shermann-Morrison formula. Let A be an invertible matrix and u, v two column vectors such that $1 + v^t A^{-1} u \neq 0$. Prove that
$$(A + uv^t)^{-1} = A^{-1} - \frac{A^{-1}uv^tA^{-1}}{1 + v^tA^{-1}u} .$$
(If a rank-one perturbation of an invertible matrix is invertible, then the inverse of the perturbation is a rank-one perturbation of the inverse).

2 Stochastic inequalities

1. Chebychev association inequality. Let X be a real valued random variable. Let f and g be two increasing functions. Show that

$$\mathbb{E}\left[f(X)g(X)\right] \geq \mathbb{E}\left[f(X)\right] \times \mathbb{E}\left[g(X)\right] .$$

2. Harriss inequality Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be nondecreasing functions with respect to each argument. Let $X_1, \ldots, X_n$ be independent real-valued random variables and define the random vector $X = (X_1, \ldots, X_n)$ taking values in $\mathbb{R}^n$. Prove that

$$\mathbb{E}[f(X)g(X)] \geq \mathbb{E}[f(X)]\mathbb{E}[g(X)] .$$

3. Maximum spacing in a uniform sample. Provide an upper bound on the expected length of the maximum spacing in an n-sample of the uniform distribution over $[0, 1]$. A spacing is the length of the interval between two adjacent sample points.

   You may attempt to prove

$$\mathbb{E}\left[\max_{i \leq n+1} U_{(i)} - U_{(i-1)}\right] \leq \frac{\sum_{i=1}^{n} \frac{1}{i}}{n}$$

   where we agree on $U_{(0)} = 0$ and $U_{(n+1)} = 1$.

   Hint: you may use the fact that order statistics of a uniform sample are distributed like

$$\left(\frac{\sum_{j=1}^{i} E_j}{\sum_{j=1}^{n+1} E_j}\right)_{i \leq n}$$

   where $E_1, \ldots, E_{n+1}$ are independent exponential random variables.

   You may also upper bound the logarithmic moment generating function of spacings and tailor a maximal inequality.

4. Kolmogorov-Smirnov statistics. The KS-statistics is defined as

$$Z_n := \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

   where F_n is the empirical distribution function defined by a sample of n points collected independently from a probability distribution definef by distribution function F.

   Establish an upper bound on $\mathbb{E}Z_n$, prove that there exists a universal constant C such that for all F and n, $\mathbb{E}Z_n \leq C$.

5. McDiarmid inequality (bounded-differences inequality).

Let $(\mathcal{X}, \mathcal{G}, P)$ be a probability space and $\Omega = \mathcal{X}^n, \mathcal{F} = \mathcal{G}^{\otimes n}, \mathbb{P} = P^{\otimes n}$, be the associated product space. Let f be a function from $\mathcal{X}^n \to \mathbb{R}$ such that there exists a sequence of constants $c_1, \ldots, c_n$ satisfying

$$|f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i \mathbb{I}_{x_i \neq x_i'}$$

for all $x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n, x_i'$.

Let $X_1, \ldots, X_n$ denote the random variables that map $\omega = (x_1, \ldots, x_n) \in \Omega = \mathcal{X}^n$, on $X_i(\omega) = x_i$.

Let $Z = f(X_1, \ldots, X_n)$.

Let $\mathcal{F}_i = \sigma(X_1, \ldots, X_i)$.

Let also $M_i = \mathbb{E}[Z \mid \mathcal{F}_i]$ for i from 0 to n.

(a) Why is $(\mathcal{F}_i)_{i \leq n}$ a filtration?

(b) Show that $(M_i)_{i \leq n}$ is an $(\mathcal{F}_i)_{i \leq n}$-adapted martingale.

(c) Show that $\mathrm{var}[Z] \leq v := \sum_{i=1}^{n} c_i^2/4$.

(d) Show that conditionally on $\mathcal{F}_{i-1}$, the distribution of M_i is supported by an interval of length not larger than c_i.

(e) Show that for all $\lambda \geq 0$,

$$\mathbb{E}\left[\mathrm{e}^{\lambda(Z-\mathbb{E}Z)}\right] \leq \mathrm{e}^{\frac{\lambda^2 v}{2}} .$$

(f) Show that for $t \geq 0$

$$\mathbb{P}\left\{Z - \mathbb{E}Z \geq t\right\} \leq \mathrm{e}^{-\frac{t^2}{2v}} .$$

6. Using McDiarmid inequality. Let $\mathcal{F}$ be a (finite) class of functions from $\mathcal{X}$ to $[0,1]$. Let $X_1, \ldots, X_n$ be i.i.d. according to some probability distribution on $\mathcal{X}$. Let

$$Z_n := \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \frac{1}{n} \left(f(X_i) - \mathbb{E}f(X_i)\right)$$

(Z_n is a supremum of a bounded centered empirical process).

Check that Z_n satisfies the conditions of the bounded-differences inequality and that

$$\mathbb{P}\left\{Z_n - \mathbb{E}Z_n \geq t\right\} \leq \exp\left(-2nt^2\right) .$$

7. Efron-Stein-Steele inequalities

Let $X_1, \ldots, X_n$ be independent random variables and let $Z = f(X)$ be a square-integrable function of $X = (X_1, \ldots, X_n)$.

Let $\mathbb{E}^{(i)}$ be a shorthand for $\mathbb{E}\left[\cdot \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\right]$. And let $\mathbb{E}_i$ be a shorthand for $\mathbb{E}\left[\cdot \mid X_1, \ldots, X_i\right]$

(a) Check that $\mathbb{E}_i\left[\mathbb{E}^{(i)}[Z]\right] = \mathbb{E}_{i-1}[Z]$.

(b) Let $\Delta_i := \mathbb{E}_i Z - \mathbb{E}_{i-1} Z$ for $1 \leq i \leq n$. Prove that $\mathsf{Var}\left[Z\right] = \sum_{i=1}^{n} \mathbb{E}\Delta_i^2$.

(c) Prove that $\Delta_i^2 \leq \mathbb{E}_i\left[\left(Z - \mathbb{E}^{(i)} Z\right)^2\right]$

(d) Prove

$$\mathrm{var}\left(Z\right) \leq \sum_{i=1}^{n} \mathbb{E}\left[\left(Z - \mathbb{E}^{(i)} Z\right)^2\right] =: v .$$

Hint :

(e) If $X_1', \ldots, X_n'$ are independent copies of $X_1, \ldots, X_n$ and if we define, for every $i = 1, \ldots, n$,

$$Z_i' := f\left(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n\right) ,$$

prove

$$v = \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\left[\left(Z - Z_i'\right)^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[\left(Z - Z_i'\right)_+^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[\left(Z - Z_i'\right)_-^2\right]$$

where $x_+ = \max(x, 0)$ and $x_- = \max(-x, 0)$ denote the positive and negative parts of a real number x.

(f) Let

$$v = \inf_{Z_i} \sum_{i=1}^{n} \mathbb{E}\left[\left(Z - Z_i\right)^2\right] ,$$

where the infimum is taken over the class of all $X^{(i)}$-measurable and square-integrable variables Z_i, $i = 1, \ldots, n$.

Prove

$$\mathrm{var}(Z) \leq v$$

(a)

(b)

(c)

3 Empirical processes

1. Peeling
2. Normalized empirical process and chaining
3. Chaining for mixed distances

4 Quadratic risk minimization

1. Gaussian linear models.
   (a) Estimation of σ^2. Check that
   $$\widehat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \,.$$
   is an unbiased estimator of homoschedastic noise variance.
   (b) Assuming that noise is Gaussian homoschedastic with known variance, build a confidence region for θ.
   (c) Assuming that noise is Gaussian homoschedastic with unknown variance, build a confidence region for θ.
   (d) Assuming that noise is Gaussian homoschedastic with known variance, build a test for a linear hypothesis $\langle \lambda, \theta \rangle = 0$ where $\lambda \in \mathbb{R}^p, \lambda \neq 0$.
   (e) How would you test the hypothesis : noise is homoschedastic Gaussian ? (this is called *model assessment*)
2. Diagnosing linear model.
   (a) Prove that
   $$\mathrm{var}(\widehat{Y}_i - Y_i) = \sigma^2(1 - H[i,i]) \,.$$
   (b) What is the covariance matrix of $Y - \widehat{Y}$? (fixed design)
   (c) Why do we call $(\widehat{Y}_i - Y_i)/\sqrt{\widehat{\sigma}^2(1 - H[i,i])}$ studentized residuals?
   (d) Cook's distance
   (e) Diagnosing heteroschedasticity
3. Outliers
4. Forward stepwise regression
5. Ridge regression and SVD

5 Lasso

1. Let $\alpha \in \mathbb{R}^p$, $\lambda > 0$, find the minimizer $\widehat{\beta}$ of

$$\|\beta - \alpha\|_2^2 + 2\lambda\|\beta\|_1$$

2.

3. LASSO and orthogonal designs. Assume columns of $\mathbf{X}$ are orthogonal and have unit norms, show that the minimizer of

$$\|\mathbf{X}\beta - Y\|_2^2 + 2\lambda\|\beta\|_1$$

is $\widehat{\beta}(\lambda)$ given by

$$\widehat{\beta}_j(\lambda) = \mathbf{X}_j^T Y \left(1 - \frac{\lambda}{|\mathbf{X}_j^T Y|}\right)_+$$

with convention $\mathbf{X}_j^T Y = \sum_{i=1}^n \mathbf{X}[i,j]Y_i$.

4.

5. Ridge regression and LASSO for Gaussian sequence models

6. For any $\lambda \geq 0$, let $\widehat{\beta}(\lambda)$ denote a minimizer of

$$\|\mathbf{X}\beta - Y\|_n^2 + 2\lambda \|\beta\|_1 \ .$$

   i. Prove that for sufficiently large λ, $\widehat{\beta}(\lambda) = 0$ (give a bound on λ).

   ii. Let $(\lambda_n)_n$ decrease to 0. Does the sequence $(\widehat{\beta}(\lambda_n))_n$ have accumulation points ?

   iii. If yes, prove that any accumulation point minimizes $\|\mathbf{X}\beta - Y\|_n^2$.

   iv. Check that the map $\lambda \mapsto \widehat{\beta}_\lambda$ is piecewise linear.

6 Resampling

6.1 Jackknife

1. Jackknife upper bound for the variance of order statistics. Use Efron-Stein inequality to upper bound the variance of the k^{th} order statistics of n-sample using spacings.
2. Confidence intervals for quantile estimation. For some $p \in (0,1)$ we want to estimate
$$F^{\leftarrow}(p) := \inf\{x : F(x) \geq p\}$$
from an i.i.d. sample $X_1, \ldots, X_n$ from unknown distribution F. We assume that F is absolutely continuous with positive density f on its support (so that $F^{\leftarrow}$ is an inverse function). The estimator is $F_n^{\leftarrow}(p)$ that is the order statistics $X_{(\lfloor np \rfloor, n)}$
   (a) Show that
$$\sqrt{n}\left(X_{(\lfloor np \rfloor, n)} - F^{\leftarrow}(p)\right) \rightsquigarrow \mathcal{N}\left(0, \frac{p(1-p)}{(f(F^{\leftarrow}(p)))^2}\right).$$
   (b) Use the Jackknife estimate of variance for order statistics and normal limit theory to build a confidence interval for $F^{\leftarrow}(p)$.
   (c) Design an estimator of the unknown density at $F^{\leftarrow}(p)$. Check its consistency.
   (d) Use the estimator of $f(F^{\leftarrow}(p))$ to build a confidence interval for $F^{\leftarrow}(p)$ with guaranteed asymptotic coverage.
3. Bias of estimators of smooth functions of moments Let $\theta(P) = f\left(\mathbb{E}_P[X]\right)$ where f is a twice differentiable function. Let $\widehat{\theta}_n := f(\overline{X}_n)$.
   (a) Prove that the bias of this estimator $\mathbb{E}_{P^{\otimes n}}\widehat{\theta}_n - \theta(P)$ scales like $1/n$.
   (b) Try to identify the constant.
4. Estimation of the correlation coefficient between random variables X and Y:
$$\rho(X,Y) = \mathrm{cov}(X,Y)/\sqrt{\mathsf{Var}\left[X\right]\mathsf{Var}\left[Y\right]}.$$
   (a) Define an estimator $\widehat{\theta}_n$ of $\rho := \rho(X,Y)$ as a function of empirical moments.
   (b) Show that its bias scales like $1/n$.
   (c) Compute the limiting distribution of $\sqrt{n}(\widehat{\rho}_n - \rho)$.

6.2 Bootstrap

1. Using the Lindeberg-Feller Central Limit Theorem to establish the consistency of the Bootstrap for the empirical mean.
2. Check that the Kolmogorov-Smirnov distance is a distance
3. Check that Kolmogorov-Smirnov distance metrizes convergence in distribution
4. Delta method and consistency of bootstrap of the mean
5. Use the t-percentile method (Bootstrap) Confidence Ineterval for correlation coefficient.
6. Variance stabilizing transformations for estimating correlation coefficients

6.3 Leave-one out cross-validation

1. loocv and linear regression

We consider a least-square regression problem $Y = \mathbf{X}\theta + \epsilon$ where $\mathbf{X}$ is an $n \times p$ matrix with rank $p < n$, and ϵ an homoschedastic symmetric random noise. $\widehat{\theta}$ is the ols estimate of θ. The ith row of $\mathbf{X}$ is the ith input, it defines a p-dimensional column vector X_i, prediction $\widehat{Y}_i$ is defined by $\widehat{Y}_i = \langle X_i, \widehat{\theta}\rangle$. The matrix $\mathbf{X}^{(i)}$ is obtained from $\mathbf{X}$ by removing the ith row. Vector $\widehat{\theta}^{(i)}$ is the ols solution obtained by removing the ith row from $\mathbf{X}$ and the ith coefficient from Y. Vector $Y^{(i)}$ is obtained by removing the ith coefficient from vector Y. Vector $\widehat{Y}^{(i)}$ is defined by $\widehat{Y}^{(i)} = \mathbf{X}\widehat{\theta}^{(i)}$. The leave-one out cross validation error is defined as:

$$\widehat{\mathrm{CV}}_n := \sum_{i=1}^{n} \frac{1}{n}\left(Y_i - Y_i^{(i)}\right)^2.$$

The aim of this exercise is to show that in a random design setting, $\widehat{\text{CV}}_n$ is an upper bound on the generalization error of $\widehat{\theta}$ and that $\widehat{\text{CV}}_n$ can be computed without much effort from the solution of the OLS problem.

(a) Express $\widehat{\theta}^{(i)}$ as a function of $\mathbf{X}^{(i)}$

(b) Check that for all i

$$\mathbf{X}^{(i)T}\mathbf{X}^{(i)} = \mathbf{X}^T\mathbf{X} - X_i X_i^T$$

and compute $(\mathbf{X}^{(i)T}\mathbf{X}^{(i)})^{-1}$ using the Morrison-Shermann formula.

(c) Check that

$$\widehat{\theta}^{(i)} = \widehat{\theta} - \left(\mathbf{X}^T\mathbf{X}\right)^{-1} X_i \frac{Y_i - \widehat{Y}_i}{1 - H[i,i]}\,.$$

Why do we say that observation i has high leverage when $H[i,i]$ is large (close to 1)?

(d) Show that

$$\text{var}(Y_i - \widehat{Y}_i^{(i)}) = \sigma^2 \left(1 + \mathbf{x}_i^T \left(\mathbf{X}^{(i)T}\mathbf{X}^{(i)}\right)^{(-1)} \mathbf{x}_i\right)\,.$$

(e) Check that

$$\widehat{Y}_i = H[i,i]Y_i + (1 - H[i,i])\widehat{Y}_i^{(i)}$$

and

$$\widehat{\text{CV}}_n = \sum_{i=1}^{n} \frac{1}{n} \left(\frac{Y_i - \widehat{Y}_i}{1 - H[i,i]}\right)^2$$

(f) How can you compute this cross-validation error from the output of `lm()`?

7 Trees and forests

7.1 Regression trees

(Planar) binary trees may be defined recursively as follows. There is one binary tree with 1 leaf and 0 internal node, one binary tree with 1 internal node and 2 leaves (right and left). There are two binary trees with two internal nodes. Let $\mathcal{T}_k$ denote the set of binary trees with k internal nodes. In a regression tree each internal node is labelled with a split, each leave is labelled by a value. The path between a node and the root is called a branch. In a binary tree each branch is uniquely characterized by a string of 0 and 1.

1. How many leaves has a tree with k internal nodes ?
2. Derive a recursion formula for the number of binary trees with k internal nodes.
3. Show that for each $\alpha > 0$, there exists a unique subtree that minimizes the α-criterion.
4. Show that the optimal subtree $\mathcal{T}_{\alpha_1}$ associated with $\alpha_1 > \alpha_0$ is a subtree of $\mathcal{T}_{\alpha_0}$.
5. Describe a algorithm that given a sample, a regression tree grown by CART, computes the sequence of optimal subtrees associated with different values of α.
6. Describe the pruning problem in CART as a model selection problem.

7.2 Bagging and random forests

1.
2.

7.3 Boosting

1.
2.