ST5215 Advanced Statistical Theory, Lecture 10

HUANG Dongming

National University of Singapore

10 Sep 2020

Overview

Last time

- Point Estimation
  - Method of Moments Estimators (MM estimator)
  - Maximum Likelihood Estimators (MLE)

Today

- More on MLE
- Statistical Decision Theory
- Statistical Inference

Recap: Point Estimators

Suppose $X_1, \ldots, X_n \sim P_\theta \in \mathcal{P}$, where $\theta = (\theta_1, \ldots, \theta_k) \in \Theta$.

- An estimator for estimating θ

$$\widehat{\theta} = w\left(X_1, \ldots, X_n\right)$$

  is a function of the data (it is a statistic)
- The parameter is a fixed, unknown constant, while the estimator is a random variable (a realization of an estimator is called an *estimate*)
- The Method of Moments
  - Express the first k moments as functions of θ: $\mu_j = h_j(\theta), \quad j = 1, \ldots, k$
  - Substitute μ_j's by the sample moments $\hat{\mu}_j$'s: $\hat{\mu}_j = h_j(\hat{\theta})$, $j = 1, \ldots, k$
  - Solve $\hat{\theta} = h^{-1}(\hat{\mu})$
- Maximum likelihood estimator
  - Likelihood function: $\ell(\theta) = f_\theta(X)$
  - An MLE of θ: $\hat{\theta} = \underset{\theta \in \Theta}{\arg\max}\ \ell(\theta)$

Numerical methods

In applications, MLE's typically do not have analytic forms and some numerical methods have to be used to compute MLE's.

- The Newton-Raphson method for solving $\frac{\partial \log \ell(\theta)}{\partial \theta} = \mathbf{0}$:
  1. Start with an initial value $\hat{\theta}^{(0)}$
  2. Repeatedly compute

$$\hat{\theta}^{(t+1)} := \hat{\theta}^{(t)} - \left[\left. \frac{\partial^2 \log \ell(\theta)}{\partial \theta \partial \theta^\top} \right|_{\theta = \hat{\theta}^{(t)}} \right]^{-1} \left. \frac{\partial \log \ell(\theta)}{\partial \theta} \right|_{\theta = \hat{\theta}^{(t)}}, \quad t = 0, 1, \ldots$$

- The Hessian matrix $\partial^2 \log \ell(\theta) / \partial \theta \partial \theta^\top$ is assumed of full rank for every $\theta \in \Theta$
- The rationale: at each time t, we update the current value by expanding $\frac{\partial \log \ell(\theta)}{\partial \theta}$ around $\hat{\theta}^{(t)}$:

$$\mathbf{0} = \frac{\partial \log \ell(\theta)}{\partial \theta} \approx \left. \frac{\partial \log \ell(\theta)}{\partial \theta} \right|_{\theta = \hat{\theta}^{(t)}} + \left[\left. \frac{\partial^2 \log \ell(\theta)}{\partial \theta \partial \theta^\top} \right|_{\theta = \hat{\theta}^{(t)}} \right] (\theta - \hat{\theta}^{(t)})$$

- If the iteration converges, then the limit or $\hat{\theta}^{(t)}$ with a sufficiently large t is a numerical approximation to a solution

Numerical methods (Cont.)

- If $\left[\frac{\partial^2 \log \ell(\theta)}{\partial\theta\partial\theta^\top}\right]\Bigg|_{\theta=\hat{\theta}^{(t)}}$ is replaced by by $\left\{E_\theta\left(\frac{\partial^2 \log \ell(\theta)}{\partial\theta\partial\theta^\top}\right)\right\}\Bigg|_{\theta=\hat{\theta}^{(t)}}$, then the method is known as the *Fisher-scoring method*
- In some applications, ideal observations lead to closed-form MLE but a part of such ideal observations is missing. For such problems, the *Expectation-Maximization (EM) algorithm* will iteratively
  - compute the **expectation** of the log-likelihood w.r.t. the missing data under the population given by the current $\hat{\theta}^{(t)}$, and
  - compute a new $\hat{\theta}^{(t+1)}$ as the **maxima** of this expectation
- In some modern applications, the sample size n is so large that optimizing the likelihood function is intractable. In this case, it is popular to use the *stochastic gradient ascent*, which use a random sub-sample from the data to compute the gradient

MLE in Exponential Families

Suppose that X has a distribution from a natural exponential family so that the likelihood function is

$$\ell(\eta) = \exp\{\eta^\top T(x) - \zeta(\eta)\} h(x), \tag{1}$$

where $\eta \in \Xi$ is a vector of unknown parameters. We observed $X = x$.

- The likelihood equation is then

$$\frac{\partial \log \ell(\eta)}{\partial \eta} = T(x) - \frac{\partial \zeta(\eta)}{\partial \eta} = 0, \tag{2}$$

  which has a unique solution $T(x) = \partial\zeta(\eta)/\partial\eta$, assuming that $T(x)$ is in the range of $\partial\zeta(\eta)/\partial\eta$.

- Note that

$$\frac{\partial^2 \log \ell(\eta)}{\partial\eta\partial\eta^\top} = -\frac{\partial^2 \zeta(\eta)}{\partial\eta\partial\eta^\top} = -\mathrm{Var}(T) \tag{3}$$

- Since $\mathrm{Var}(T)$ is positive definite, $-\log \ell(\eta)$ is convex in η. Thus, $T(x)$ is the unique MLE of the parameter $\mu(\eta) = \partial\zeta(\eta)/\partial\eta$
- By the inverse function theorem, the function $\mu(\eta)$ is one-to-one so that μ^{-1} exists. By the definition, the MLE of η is $\hat{\eta} = \mu^{-1}(T(x))$.

MLE in Exponential Families (Cont.)

- If the distribution of X is in a general exponential family and the likelihood function is

$$\ell(\theta) = \exp\{[\eta(\theta)]^\top T(x) - \xi(\theta)\} h(x)$$

- If the inverse of $\eta(\theta)$ exists and $\hat{\eta}$ is in the range of $\eta(\theta)$, then the MLE of θ is $\hat{\theta} = \eta^{-1}(\hat{\eta})$
- $\hat{\theta}$ is also the solution of the likelihood equation

$$\frac{\partial \log \ell(\theta)}{\partial \theta} = \frac{\partial \eta(\theta)}{\partial \theta} T(x) - \frac{\partial \xi(\theta)}{\partial \theta} = 0, \tag{4}$$

because $\xi(\theta) = \zeta(\eta(\theta))$ and by chain rules $\frac{\partial \xi(\theta)}{\partial \theta} = \frac{\partial \zeta(\eta)}{\partial \eta} \frac{\partial \eta(\theta)}{\partial \theta}$

Decision rules

In the previous estimation problem, we estimate θ by $\hat{\theta}(X)$. This means, after we observe $X = x$, we take an action: estimate θ by $\hat{\theta}(x)$.
More generally,

- Let X be a sample from a population $P \in \mathcal{P}$
- A statistical decision is an action that we take after we observe X e.g.:
  - gives an estimate for a parameter
  - choose between hypotheses
  - make a statement about the range of the parameter
- The set of allowable actions is Θ: *action space*, denoted by $\mathbb{A}$ and endowed with a σ-field $\mathcal{F}_{\mathbb{A}}$
- A *decision rule*: a measurable function from the range of X , say, $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ to $(\mathbb{A}, \mathcal{F}_{\mathbb{A}})$
- If a decision rule T is chosen, then we take the action $T(X) \in \mathbb{A}$ after X is observed

Loss functions, Risks

How to measure the performance of decision rules?

- **Loss function**: a function $L : \mathcal{P} \times \mathbb{A} \to [0, \infty)$ that is Borel for each fixed $P \in \mathcal{P}$
  - When $\mathcal{P}$ is parametric and θ is the parameter, we may also use $L : \Theta \times \mathbb{A} \to [0, \infty)$
  - For a decision rule, $L(P, T(x))$ is the "loss" if we take the action $T(x)$ after observing $X = x$
- The **risk** for a rule is defined by

$$R_T(P) = \mathbb{E}_P L(P, T(X)) = \int L(P, T(X)) \, \mathrm{d}P \tag{5}$$

- Risk is the "average" loss under population P
- Note that the risk depends on T, P, and also the choice of loss function (often predetermined and fixed)

Comparing decision rules

Now we can compare the performance of any two decision rules

- T_1 **is as good as** T_2 if

$$R_{T_1}(P) \leq R_{T_2}(P), \qquad \forall P \in \mathcal{P} \tag{6}$$

- We say T_1 is **better than** T_2, if T_1 is as good as T_2 and $R_{T_1}(P) < R_{T_2}(P)$ for some $P \in \mathcal{P}$
  - We also say T_2 is **dominated by** T_1
- T_1 and T_2 are **equivalent (equivalently good)** if and only $R_{T_1}(P) = R_{T_2}(P)$ for all $P \in \mathcal{P}$

Let $\mathfrak{I}$ be the collection of decision rules under consideration

- T_* is called an $\mathfrak{I}$-**optimal** rule if T_* is as good as any other rule in $\mathfrak{I}$
- T_* is **optimal** if $\mathfrak{I}$ contains all possible rules

Example: Measurement problem revisited

Recall that we are to measure a quantity θ of an object. We take multiple measurements of the object and record the results $X_1, \ldots, X_n$

- Action space $\mathbb{A} = \Theta$ the set of all possible values of θ
- $(\mathbb{A}, \mathcal{F}_{\mathbb{A}}) = (\Theta, \mathcal{B}_{\Theta})$
- A simple decision rule: $T(X) = \overline{X}$
- A commonly used loss function: squared error loss $L(P_\theta, a) = (\theta - a)^2$ for $\theta \in \Theta$ and $a \in \mathbb{A}$
    - The risk function with the squared error loss is called the mean squared error (MSE)
- Suppose $X_1, \ldots, X_n$ are i.i.d. with mean θ and variance σ^2
- The risk of T is

$$
\begin{aligned}
R_T(\theta) &= \mathbb{E}_\theta(\theta - \overline{X})^2 \\
&= (\theta - \mathbb{E}_\theta \overline{X})^2 + \mathbb{E}_\theta(\mathbb{E}_\theta \overline{X} - \overline{X})^2 \\
&= (\theta - \theta)^2 + \mathrm{Var}(\overline{X})^2 \\
&= \sigma^2 / n
\end{aligned}
$$

Example (Cont.)

- In the previous derivation, we have used a well-known decomposition

$$E(X-a)^2 = E(X-EX)^2 + (EX-a)^2$$

- The result is known as MSE $=$ bias^2 $+$ Var

Suppose θ_P is a parameter related to a population P.

- An estimator $T(X)$ for θ_P is said **unbiased** if

$$E_P[T(X)] = \theta_P, \forall P \in \mathcal{P}$$

- If $T(X)$ is not unbiased, the bias of $T(X)$ is defined as

$$b_T(P) = E_P[T(X)] - \theta_P$$

  - Also denoted by $b_T(\theta)$ for a parametric family indexed by θ
- The variance of a biased estimator may be smaller than that of an unbiased estimator
- To get a rule that is optimal in MSE, we may need to *trade-off the bias and variance*

Exercise

Suppose $X_1, \dots, X_n$ are i.i.d. with mean θ and variance σ^2.
Consider $\mathfrak{J}$ the class of estimators $\hat{\theta}_c = c\bar{X}$ for $c \in (0, 1]$.

1. If it is known a-priori that $\theta^2 < \sigma^2$, determine a range of values for c such that $\hat{\theta}_c$ has a smaller MSE than $\bar{X}$.
2. Generally, is there an $\mathfrak{J}$-optimal estimator?

Solution:

- Using the decomposition of MSE,

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}_c) &= (\theta - cE\bar{X})^2 + c^2\mathrm{Var}(\bar{X}) \\ &= (1-c)^2\theta^2 + c^2\sigma^2/n\end{aligned}$$

- $\mathrm{MSE}(\hat{\theta}_c) < \mathrm{MSE}(\hat{\theta}_1) \Leftrightarrow (1-c)^2\theta^2 < (1-c^2)\sigma^2/n$
$\Leftrightarrow 1 - \frac{2}{1+n\theta^2/\sigma^2} < c$
- Since $\theta^2/\sigma^2 < 1$, we have $1 - \frac{2}{1+n\theta^2/\sigma^2} < 1 - 2/(1+n)$
- So for any $c \in (1 - 2/(1+n), 1)$, it holds that $\mathrm{MSE}(\hat{\theta}_c) < \mathrm{MSE}(\bar{X})$

Hypothesis tests

Let $\mathcal{P}$ be a family of distributions, $\mathcal{P}_0 \subset \mathcal{P}$, and $\mathcal{P}_1 = \mathcal{P} \setminus \mathcal{P}_0$.
A general hypothesis testing problem can be formulated as deciding which of the following statements is true:

$$H_0 : P \in \mathcal{P}_0 \qquad \text{versus} \qquad H_1 : P \in \mathcal{P}_1. \tag{7}$$

- Call H_0 *the null hypothesis*, H_1 *the alternative hypothesis*
- The action space $\mathbb{A} = \{0, 1\}$
- A decision rule in this case is called a *test*
- $T : \mathcal{X} \to \{0, 1\}$, so mut be in the form $I_C(X)$ for some $C \subset \mathcal{X}$
- C: the *rejection region* or *critical region* for testing H_0 versus H_1
- A common loss function: 0-1 loss, $L(P, j) = 0$ for $P \in \mathcal{P}_j$ and $L(P, j) = 1$ otherwise, $j = 0, 1$
- The risk is

$$R_T(P) = \begin{cases} P(T(X) = 1) = P(X \in C) & \text{when } P \in \mathcal{P}_0, \\ P(T(X) = 0) = P(X \notin C) & \text{when } P \in \mathcal{P}_1, \end{cases}$$

Type I and Type II errors

- When H_0 is rejected but H_0 is indeed true, the error is called the type I error
- When H_0 is accepted but H_0 is in fact wrong, the error is called the type II error
- Probabilities of making two types of errors
  - Type I error rate:

$$\alpha_T(P) = P(T(X) = 1), \quad P \in \mathcal{P}_0$$

    Type II error rate:

$$1 - \alpha_T(P) = 1 - P(T(X) = 1), \quad P \in \mathcal{P}_1$$

- $\alpha_T(P)$, as a function of P, is called the power function of T
  - If $\mathcal{P}$ is parametric, we may also use $\alpha_T(\theta)$
- Type I and Type II error rates cannot be minimized simultaneously

Significance level and size of the test

- Under the Neyman-Pearson framework, we assign a pre-specified bound α to the Type I error rate:

$$\sup_{P \in \mathcal{P}_0} P(T(X) = 1) \leq \alpha$$

  This number α is called the *significance level* of the test.
    - The choice of significance level is somewhat subjective. Standard values, 0.10 , 0.05, and 0.01, are often used for convenience
- If

$$\sup_{P \in \mathcal{P}_0} P(T(X) = 1) = \alpha'$$

  then α' is called the *size* of the test

Remark.

- The NP framework is slightly different from the decision theory
- Under the NP framework, the two hypotheses should be formulated in a way such that the Type I error is more serious than the Type II error from a practical point-of-view

p-values

- When constructing a test, we usually find a class of tests T_α with varying significance levels α
- Usually, a small significance level leads to a "small" rejection region
- The smallest possible level of significance α at which H_0 would be rejected for the computed $T_\alpha(x)$,

$$\hat{\alpha}(x) = \inf \left\{ \alpha \in (0,1) : T_\alpha(x) = 1 \right\},$$

  is called the *p-value* for the test T_α if $X = x$ is observed
- $\hat{\alpha}(X)$ is a r.v. The *p*-value is the realization of this random variable
- The *p*-value depends on both X and the chosen test
- The *p*-value provides additional information for a test, so using *p*-values is more appropriate than using fixed-level tests in a scientific problem

Confidence sets

Let $\theta \in \Theta \subset \mathcal{R}^k$ be related to the unknown population $P \in \mathcal{P}$

- If $C(X)$ a Borel subset of Θ depending only on the sample X such that
$$\inf_{P \in \mathcal{P}} P(\theta \in C(X)) \geq 1 - \alpha$$
where α is a fixed constant in $(0, 1)$, then $C(X)$ is called a *confidence set* for θ with significance level $1 - \alpha$
- The highest possible level of significance for $C(X)$ is called the *confidence coefficient* of $C(X)$
- A confidence set is a random element that covers the unknown θ with certain probability
- The *coverage probability* of $C(X)$ is at least $1 - \alpha$, although $C(x)$ either covers or does not cover θ whence we observe $X = x$
- In the special case when $k = 1$
  - If $C(X)$ is an interval, it is called a *confidence interval*
  - If $C(X) = (-\infty, \bar{\theta}(X)]$ or $[\underline{\theta}(X), \infty)$, it is called a *confidence bound*

Admissibility

As we have seen in the exercise, optimal rules may not exist. We need a more general notion to describe the performance of decision rules.

Definition

Let $\mathfrak{J}$ be a class of decision rules. A decision rule $T \in \mathfrak{J}$ is called $\mathfrak{J}$-admissible if no $S \in \mathfrak{J}$ is better than T (in terms of the risk)

- If $\mathfrak{J}$ contains all possible rules, simply write "admissible" for "$\mathfrak{J}$-admissible"
- In principle, inadmissible rules shall not be used
- Relationship between admissibility and optimality
  1. If T_* is $\mathfrak{J}$-optimal, then it is $\mathfrak{J}$-admissible
  2. If T_* is $\mathfrak{J}$-optimal and S is $\mathfrak{J}$-admissible, then S is also $\mathfrak{J}$-optimal and is equivalent to T_*
  3. If there are two $\mathfrak{J}$-admissible rules that are not equivalent, then there does not exist any $\mathfrak{J}$-optimal rule

Convex loss functions and sufficient statistics

Theorem (Rao-Blackwell)

Let T be a sufficient statistic for $P \in \mathcal{P}$.
Suppose the action space $\mathbb{A} \subset \mathcal{R}^k$ and is convex, and S_0 is a decision rule satisfying $\mathbb{E}_P\|S_0\| < \infty$ for all $P \in \mathcal{P}$.
Let $S_1 = \mathbb{E}\{S_0(X) \mid T\}$.

1. *If the loss function $L(P, a)$ is convex in a, then $R_{S_1}(P) \leq R_{S_0}(P)$;*
2. *If $L(P, a)$ is strictly convex in a and S_0 is not a function of T, then S_0 is inadmissible and dominated by S_1.*

- Proved by Jensen's inequality (conditional expectation version)
- For a convex loss and any decision rule, we can construct a new rule that may be better than before by taking conditional expectation given a sufficient statistic
- For strictly convex loss function, admissible rules are functions of sufficient statistics

Example: Poisson process

Phone calls arrive at a switchboard according to a Poisson process at an average rate of λ per minute. λ is unknown, but the numbers $X_1, \ldots, X_n$ of phone calls that arrived during n successive one-minute periods are observed.

We want to estimate the probability $\theta = e^{-\lambda}$ that the next one-minute period passes with no phone calls.

- Consider squared error loss $L(\theta, a) = (\theta - a)^2$; strictly convex in a
- Start with the following naive estimator

$$S_0 = \begin{cases} 1 & \text{if } X_1 = 0 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

- Poisson distributions form an exponential family. By Factorization theorem, $T = \sum_{i=1}^n X_i$ is a sufficient statistic
- Since S_0 is not a function of T, it is inadmissible
- Define $S_1(t) = \mathbb{E}\{S_0 \mid T = t\}$. By Rao-Blackwell's theorem, $S_1(T)$ is better than S_0

- In fact, $\sum_{i=1}^n X_i \sim \text{Poisson}(n\lambda)$ (using m.g.f. and Table 1.1)

$$\begin{aligned}
S_1(t) &= \mathbb{E}\{I_{X_1=0} \mid T = t\} \\
&= P\left(X_1 = 0 \mid \sum_{i=1}^n X_i = t\right) \\
&= P\left(X_1 = 0, \sum_{i=2}^n X_i = t\right) / P\left(\sum_{i=1}^n X_i = t\right) \\
&= P\left(X_1 = 0\right) P\left(\sum_{i=2}^n X_i = t\right) / P\left(\sum_{i=1}^n X_i = t\right) \\
&= e^{-\lambda} \frac{((n-1)\lambda)^t e^{-(n-1)\lambda}}{t!} \frac{t!}{(n\lambda)^t e^{-n\lambda}} \\
&= \left(1 - \frac{1}{n}\right)^t .
\end{aligned}$$

- For large n, T/n concentrate around λ, and thus $S_1(T)$ concentrate around $\left(1 - \frac{1}{n}\right)^{n\lambda} \approx e^{-\lambda}$

Tutorial

1. Let $X_1, \ldots, X_n$ be i.i.d. random variables having the Lebesgue p.d.f. $\theta^{-1} e^{-(x-\theta)/\theta} I_{(\theta,\infty)}(x)$, where $\theta > 0$ is an unknown parameter.
   (a) Find a statistic that is minimal sufficient for θ.
   (b) Show whether the minimal sufficient statistic in (a) is complete.
2. Let $X_1, \ldots, X_n$ be i.i.d. from the $N\left(\theta, \theta^2\right)$ distribution, where $\theta > 0$ is a parameter. Find a minimal sufficient statistic for θ and show whether it is complete.
3. Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. random 2–vectors having the normal distribution with
   $EX_1 = EY_1 = 0, \mathrm{Var}\,(X_1) = \mathrm{Var}\,(Y_1) = 1$, and
   $\mathrm{Cov}\,(X_1, Y_1) = \theta \in (-1, 1)$
   (a) Find a minimal sufficient statistic for θ.
   (b) Show whether the minimal sufficient statistic in (a) is complete or not.
   (c) Prove that $T_1 = \sum_{i=1}^n X_i^2$ and $T_2 = \sum_{i=1}^n Y_i^2$ are both ancillary but (T_1, T_2) is not ancillary.

Exercise 1

Let $X_1, \ldots, X_n$ be i.i.d. random variables having the Lebesgue p.d.f. $\theta^{-1} e^{-(x-\theta)/\theta} I_{(\theta,\infty)}(x)$, where $\theta > 0$ is an unknown parameter.
(a) Find a statistic that is minimal sufficient for θ.
(b) Show whether the minimal sufficient statistic in (a) is complete.

Solution: Part (a)

- Let $T(x) = \sum_{i=1}^n x_i$ and $W(x) = \min_{1\leq i\leq n} x_i$, where$x = (x_1, \ldots, x_n)$
- The joint density of $X = (X_1, \ldots, X_n)$ is

$$f_\theta(x) = \frac{e^n}{\theta^n} e^{-T(x)/\theta} I_{(\theta,\infty)}(W(x))$$

- For any two sample points x and y, the density ratio is

$$\frac{f_\theta(x)}{f_\theta(y)} = e^{[T(y)-T(x)]/\theta} \frac{I_{(\theta,\infty)}(W(x))}{I_{(\theta,\infty)}(W(y))}$$

- This ratio is free of θ if and only if $T(x) = T(y)$ and $W(x) = W(y)$
- Hence, $(T(X), W(X))$ is minimal sufficient for θ

Part(b)

- For any $\theta > 0$, $E_\theta[T(X)] = 2n\theta$ and $E_\theta[W(X)] = \left(1 + n^{-1}\right)\theta$
- Hence $E_\theta\left[(2n)^{-1}T - \left(1 + n^{-1}\right)^{-1}W(X)\right] = 0$ for any θ
- $(2n)^{-1}T(x) - \left(1 + n^{-1}\right)^{-1}W(x)$ is not a constant
- Thus, (T, W) is not complete

Exercise 2

Let $X_1, \dots, X_n$ be i.i.d. from the $N\left(\theta, \theta^2\right)$ distribution, where $\theta > 0$ is a parameter. Find a minimal sufficient statistic for θ and show whether it is complete.

Solution:

- The joint Lebesgue density of $X_1, \dots, X_n$ is

$$\frac{1}{\left(2\pi\theta^2\right)^n} \exp\left\{-\frac{1}{2\theta^2}\sum_{i=1}^n x_i^2 + \frac{1}{\theta}\sum_{i=1}^n x_i - \frac{1}{2}\right\}$$

- Let $\eta(\theta) = \left(-\frac{1}{2\theta^2}, \frac{1}{\theta}\right)$, for $\theta > 0$
- Then vectors $\eta\left(\frac{1}{2}\right) - \eta(1) = \left(-\frac{3}{2}, 1\right)$ and $\eta\left(\frac{1}{\sqrt{2}}\right) - \eta(1) = \left(-\frac{1}{2}, \sqrt{2}\right)$ are linearly independent in $\mathcal{R}^2$
- Hence $T = \left(\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i\right)$ is minimal sufficient for θ by the properties of exponential families (Example 2.14 in JS)
- $E_\theta\left(\sum_{i=1}^n X_i^2\right) = nE_\theta X_1^2 = 2n\theta^2$ and $E_\theta\left(\sum_{i=1}^n X_i\right)^2 = n\theta^2 + (n\theta)^2 = \left(n + n^2\right)\theta^2$
- So $\frac{1}{2n}T_1 - \frac{1}{n(n+1)}T_2^2$ has mean 0 but is not a constant
- Hence, T is not complete

Exercise 3

Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. random 2–vectors having the normal distribution with $EX_1 = EY_1 = 0, \text{Var}(X_1) = \text{Var}(Y_1) = 1$, and $\text{Cov}(X_1, Y_1) = \theta \in (-1, 1)$

(a) Find a minimal sufficient statistic for θ.

(b) Show whether the minimal sufficient statistic in (a) is complete or not.

(c) Prove that $T_1 = \sum_{i=1}^n X_i^2$ and $T_2 = \sum_{i=1}^n Y_i^2$ are both ancillary but (T_1, T_2) is not ancillary.

Solution: Part (a)

- The joint Lebesgue density of $(X_1, Y_1), \ldots, (X_n, Y_n)$ is

$$\left(\frac{1}{2\pi\sqrt{1-\theta^2}}\right)^n \exp\left\{-\frac{1}{1-\theta^2}\sum_{i=1}^n \left(x_i^2 + y_i^2\right) + \frac{2\theta}{1-\theta^2}\sum_{i=1}^n x_i y_i\right\}$$

- Let $\eta(\theta) = \left(-\frac{1}{1-\theta^2}, \frac{2\theta}{1-\theta^2}\right)$
- $\eta(0) = (-1, 0)$, $\eta(1/2) = (-4/3, 4/3)$, $\eta(-1/2) = (-4/3, -4/3)$
- Since $\eta(1/2) - \eta(0) = (-1/3, 4/3)$ and $\eta(-1/2) - \eta(0) = (-1/3, -4/3)$ are linearly independent
- By the properties of exponential families, $\left(\sum_{i=1}^n \left(X_i^2 + Y_i^2\right), \sum_{i=1}^n X_i Y_i\right)$ is minimal sufficient

Exercise 3 (Cont.)

Part (b)

- Note that $E_\theta\left[\sum_{i=1}^n\left(X_i^2+Y_i^2\right)-2n\right]=0$, for any θ
- Since $\sum_{i=1}^n\left(X_i^2+Y_i^2\right)-2n$ is not a constant, the statistic we found in (a) is not complete

Part (c)

- Both T_1 and T_2 have the chi-square distribution χ_n^2, which does not depend on θ. Hence both T_1 and T_2 are ancillary
- (T_1, T_2) is not ancillary because $E_\theta\left(T_1T_2\right)$ depend on θ:

$$
\begin{aligned}
E_\theta\left(T_1T_2\right) &= E_\theta\left[\left(\sum_{i=1}^n X_i^2\right)\left(\sum_{j=1}^n Y_j^2\right)\right] \\
&= E\left(\sum_{i=1}^n X_i^2Y_i^2\right)+E\left(\sum_{i\neq j} X_i^2Y_j^2\right) \\
&= nE\left(X_1^2Y_1^2\right)+n(n-1)E\left(X_1^2\right)E\left(Y_1^2\right) \\
&= n\left(1+2\theta^2\right)+n(n-1)
\end{aligned}
$$