4.1 Sufficiency

Sufficiency provides a model-based data dimension reduction. (Think about some model-free data dimension reduction techniques you learned.)

**Definition 1** Let $\mathcal{F} = \{f(\cdot|\theta), \theta \in \Theta\}$ be a parametric family. $T = T(X)$ is called a sufficient statistic for $\mathcal{F}$ (or simply for $\theta$) if the conditional distribution of $X$ given $T$ does not depend on $\theta$.

**Example 4.1** Let $X_1, ..., X_n$ be iid random variables sharing a common distribution $f(\cdot|\theta)$. Then $S_n = X_1 + \cdots + X_n$ is a sufficient statistic in each of the following cases for $f(\cdot|\theta)$: (a) Bernoulli distribution with parameter $\theta$; (b) Poisson distribution with parameter $\theta$; (c) $N(\theta, \sigma^2)$ where $\sigma > 0$ is known.

Definition 1 only enables us to verify sufficiency of a given statistic, while the following theorem, due to Fisher and Neyman, gives a useful clue for recognition of a sufficient statistic.

**Theorem 1** Let $X$ follow the distribution $f(x|\theta)$. $T = T(X)$ is a sufficient statistic for $\theta$ $\iff$ $f(x|\theta) = g(T(x)|\theta) h(x)$ for some functions $g$ and $h$.

**Proof:** We consider discrete models because proving other cases requires measure theory. Without loss of generality, consider $t$ with $P_\theta(T = t) > 0$ and $x$ with $T(x) = t$.

"$\Leftarrow$"

$$P_\theta(X = x|T = t) = \frac{P_\theta(X = x)}{\sum_{x': T(x') = t} P_\theta(X = x')} = \frac{g(t|\theta) h(x)}{\sum_{x': T(x') = t} g(t|\theta) h(x')}$$

which does not involve $\theta$.

"$\Rightarrow$" Note that

$$P_\theta(X = x) = P_\theta(X = x, T = t) = P_\theta(T = t) P_\theta(X = x|T = t) \triangleq g(t|\theta) h(x),$$

where $g(t|\theta) = P_\theta(T = t)$, and $h(x) = P_\theta(X = x|T = t)$ which does not really depend on $\theta$ due to the sufficiency of $T$. QED.
Example 4.2  Denote by \( U(a, b) \) the uniform distribution over interval \([a, b]\). Let \( X_1, \ldots, X_n \) be iid random variables with \( X_1 \sim U(a, b) \).

(i) Set \( a = -\theta \) and \( b = \theta \). Then \( f(x^n|\theta) = (2\theta)^{-n}I(\theta \geq \max\{x_{(n)}, -x_{(1)}\}) \). Theorem 1 implies that \( T = \max\{X_{(n)}, -X_{(1)}\} \) is a sufficient statistic, where \( X_{(1)} \leq \cdots \leq X_{(n)} \) are order statistics for \( X^n \).

(ii) Set \( a = \theta - 1/2 \) and \( b = \theta + 1/2 \). Then
\[
f(x^n|\theta) = I(\theta-1/2 \leq x_{(1)} \leq x_{(n)} \leq \theta+1/2) = I(x_{(n)}-1/2 \leq \theta \leq x_{(1)}+1/2).
\]
Hence \((X_{(1)}, X_{(n)})\) is a 2D sufficient statistic. Later we will show in case (ii), \( \not\exists \) 1D sufficient statistics for \( \theta \).

Example 4.3  Let \( X_1, \ldots, X_n \) be iid random variables with a common density
\[
f(x|\theta) = \{\pi[1 + (x - \theta)^2]\}^{-1}, \quad x \in \mathbb{R},
\]
i.e. Cauchy distribution with location parameter \( \theta \). Later we will show that \( \not\exists \) sufficient statistics of dimensionality lower than \( n \).

4.2 Minimal sufficiency

Observe that for any bijection \( \gamma \) (a 1-1 onto mapping) and sufficient statistic \( T \), \( \gamma(T) \) is also a sufficient statistic. Hence there are infinitely many “equivalent” sufficient statistics. A natural question: can we find a sufficient statistic \( T \) in a given problem such that \( T \) reduces the data to the lowest dimension without loss of information? First, we introduce a criterion:

**Definition 2**  \( T \) is called a minimal sufficient statistic if it is sufficient and can be expressed as a function of any other sufficient statistic.

Next, we provide a recipe to find a minimal sufficient statistic.

**Recipe:**

- Define a relation "\( \sim \)" using the likelihood \( f(x|\theta) \): \( x \sim x' \) for \( x, x' \in \mathcal{X} \) if and only if \( \frac{f(x|\theta)}{f(x'|\theta)} \triangleq H(x, x') \) (does not depend on \( \theta \) \( \forall \theta \in \Theta \), for some function \( H \).
- Verify that \( \sim \) satisfies the properties:
  \( x \sim x \) (reflexivity);
x ∼ x′ ⇔ x′ ∼ x (symmetry);
If x ∼ x′ and x′ ∼ x′′ then x ∼ x′′ (transitivity).

Hence ∼ is an equivalence relation.

• Partition \( \mathcal{X} \) via ∼ as a disjoint union of equivalence classes \( \mathcal{X} = \bigcup \mathcal{C}_t \) such that any pair \( x \sim x' \) must belong to the same equivalence class \( \mathcal{C}_t \). (Note: If \( x \in \mathcal{C}_t \) and \( x \in \mathcal{C}_{t'} \), then \( \mathcal{C}_t = \mathcal{C}_{t'} \). Why?)

• Define a statistic \( T \) such that in each equivalence class \( \mathcal{C}_t \), \( T(x) = \) constant \( \forall x \in \mathcal{C}_t \).

Then \( T \) is a minimal sufficient statistic.

Here is the justification:

**Theorem 2** A statistic \( T^* \) that remains constant in each equivalence class is a minimal sufficient statistic. (Note that \( T^* \) may take different constant values over different equivalence classes.)

**Proof:** Again, consider only a discrete model. Let \( T^*(x) = t \) and \( \mathcal{C}_t = \{ x' : T^*(x') = t \} \). Then

\[
P_{\theta}(X = x | T^* = t) = \frac{f(x|\theta)}{\sum_{x' \in \mathcal{C}_t} f(x'|\theta)} = \frac{1}{\sum_{x' \in \mathcal{C}_t} H(x', x)},
\]
which does not depend on \( \theta \). Hence \( T^* \) is sufficient.

To show \( T^* \) is minimal sufficient, let \( T \) be an arbitrary sufficient statistic with \( T(x) = T(x') \). Then

\[
\frac{f(x|\theta)}{f(x'|\theta)} \cdot \frac{g(T(x)|\theta)}{g(T(x')|\theta)} \cdot \frac{h(x)}{h(x')} = \frac{h(x)}{h(x')} \triangleq H(x, x'),
\]
which implies \( x \sim x' \). Therefore, \( T^*(x) = T^*(x') \). \( QED \).

**Note:** Theorem 2 indicates that every sufficient statistic \( T \) yields a partition for \( \mathcal{X} \) in terms of distinct values of \( T \), called the sufficient partition. Those values of \( T \) themselves are not essential. They serve only as labels which can be changed. A minimal sufficient statistic gives rise to the unique coarsest sufficient partition for \( \mathcal{X} \), called the minimal sufficient partition. Any non-minimal sufficient partition is a refinement of the minimal sufficient partition.

**Example 4.4** Let \( X_1, ..., X_n \) be iid Bernoulli random variables with parameter \( \theta \). Following the recipe, it can be shown that both \( S_n = \sum_{i=1}^{n} X_i \) and \( \bar{X} = S_n/n \) are minimal sufficient statistics. However, they are different labels of the same (unique) minimal sufficient partition \( \mathcal{X} = \bigcup_{i=0}^{n} \mathcal{C}_i \) where \( \mathcal{C}_i \) contains \( \binom{n}{i} \) vectors \((x_1, ..., x_n)\) satisfying \( x_1 + \cdots + x_n = i \).
**Example 4.5**  \(X_1, \ldots, X_n\) be iid \(N(\mu, \sigma^2)\) random variables with a 2D unknown parameter \(\theta = (\mu, \sigma^2)\). Show that \(T = (T_1, T_2)\) is a 2D minimal sufficient statistic where \(T_1 = \bar{X}\) and \(T_2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2\).

**Example 4.6** Revisit Example 4.2 (ii), show that \(T = (X(1), X(n))\) is a 2D minimal sufficient statistic for the 1D parameter \(\theta\).

**Example 4.7** Revisit Example 4.3. Now we show the order statistics \((X(1), \ldots, X(n))\) is a \(n\)-dimensional minimal sufficient statistic. Let \(x = (x_1, \ldots, x_n), x' = (x'_1, \ldots, x'_n)\) with \(x \sim x'\), thus \(\frac{f(x|\theta)}{f(x'|\theta)} = c = c(x, x') \forall \theta\), i.e.

\[
\prod_{j=1}^{n} \left[1 + (x'_j - \theta)^2\right] = c \prod_{j=1}^{n} \left[1 + (x_j - \theta)^2\right],
\]

hence

\[
\prod_{j=1}^{n} \left[\theta^2 - 2x'_j \theta + (x'_j)^2 + 1\right] = c \prod_{j=1}^{n} \left[\theta^2 - 2x_j \theta + x_j^2 + 1\right],
\]

where both sides are polynomials in \(\theta\) of degree \(2n\), with the same set of zeros, denoted by \(\mathcal{O}_L = \mathcal{O}_R\). We can spell out

\[
\mathcal{O}_L = \{x'_j \pm i, \ j = 1, \ldots, n\} \quad \text{and} \quad \mathcal{O}_R = \{x_j \pm i, \ j = 1, \ldots, n\},
\]

where \(i = \sqrt{-1}\) (the imaginary unit). Therefore, \(\{x_1, \ldots, x_n\} = \{x'_1, \ldots, x'_n\}\), i.e. the two sets are permutations of each other. Which statistic takes the same value over all permutations of \((X_1, \ldots, X_n)\)? Answer: \((X(1), \ldots, X(n))\).

**Sufficiency Principle:** If \(T\) is a sufficient statistic for \(\theta\), then any inference on \(\theta\) should depend on the data \(X\) only through \(T\), i.e. for two samples \(x\) and \(x'\) with \(T(x) = T(x')\), the inference on \(\theta\) should be the same whether \(x\) or \(x'\) is observed. This principle is interpreted as no information is lost when just focusing on sufficient statistics in an inference problem.