Chapter 6. Principles of Data Reduction
Lecture 22: Sufficiency

Data reduction

We consider a sample $X = (X_1, ..., X_n)$, $n > 1$, from a population of interest (each X_i may be a vector and X may not be a random sample, although most of the time we consider a random sample).

Assume the population is indexed by θ, an unknown parameter vector.

Let $\mathscr{X}$ be the range of X

Let x be an observed data set, a realization of X.

- We want to use the information about θ contained in x.
- The whole x may be hard to interpret, and hence we summarize the information by using a few key features (statistics).
  For example, the sample mean, sample variance, the largest and smallest order statistics.
- Let $T(X)$ be a statistic. For T, if $x \neq y$ but $T(x) = T(y)$, then x and y provides the same information and can be treated as the same.

- T partitions $\mathscr{X}$ into sets

$$A_t = \{x : T(x) = t\}, \quad t \in \mathscr{T} \text{ (the range of } T\text{)}$$

  All points in A_t are treated the same if we are interested in T only.
- Thus, T provides a data reduction.
- We wish to reduce data as much as we can, but not lose any information about θ (or at least important information).

Sufficiency

A sufficient statistic for θ is a statistic that captures all the information about θ contained in the sample.
Formally we have the following definition.

Definition 6.2.1 (sufficiency)

A statistic $T(X)$ is sufficient for θ if the conditional distribution of X given $T(X) = T(x)$ does not depend on θ.

- The sufficiency depends on the parameter of interest.

- If X is discrete, then so is $T(X)$ and sufficiency means that $P(X = x | T(X) = T(x))$ is known, i.e., it does not depend on any unknown quantity.
- Once we observe x and compute a sufficient statistic $T(x)$, the original data x do not contain any further information concerning θ and can be discarded, i.e., $T(x)$ is all we need regarding θ.
- If we do need x, we can simulate a sample y from $P(X = y | T(X) = T(x))$ since it is known; the observed y may not be the same as x, but $T(x) = T(y)$.

Example 6.2.3 (binomial sufficient statistic)

Suppose that $X_1, ..., X_n$ are iid Bernoullie variables with probability θ. The joint pmf is

$$f_\theta(x_1, ..., x_n) = \begin{cases} \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} & x_i = 0, 1, i = 1, ..., n \\ 0 & \text{otherwise} \end{cases}$$

Consider the statistic $T(X) = \sum_{i=1}^{n} X_i$, which is the number of ones in X.

To show T is sufficient for θ, we compute the conditional probability $P(X = x|T = t)$.
For $t = 0, 1, ..., n$, let

$$B_t = \left\{ x = (x_1, ..., x_n) : \; x_i = 0, 1, \; \sum_{i=1}^{n} x_i = t \right\}.$$

If $x \notin B_t$, then $P(X = x|T = t) = 0$.
If $x \in B_t$, then

$$P(X = x, T = t) = P(X = x) = f_\theta(x) = \theta^t (1 - \theta)^{n-t}.$$

Also, since $T \sim binomial(n, p)$,

$$P(T = t) = \binom{n}{t} \theta^t (1 - \theta)^{n-t}$$

Then, for $t = 0, 1, ..., n$,

$$P(X = x|T = t) = \frac{P(X = x, T = t)}{P(T = t)} = \frac{1}{\binom{n}{t}} \quad x \in B_t$$

is a known pmf (does not depend on θ).

Hence $T(X)$ is sufficient for θ.

For any realization x of X, x is a sequence of n ones and zeros.

Since θ is the probability of a one and T is the frequency of ones in x, it has all the information about θ.

Given $T = t$, what is left in the data set x is the redundant information about the positions of t ones, and we can reproduce the data set x if we want by using $T = t$.

How to find sufficient statistics?

To verify that a statistic T is a sufficient statistic for θ by definition, we must verify that for any fixed values of x, the conditional distribution $X|T(X) = T(x)$ does not depend on θ.

This may not be easy but at least we can try.

But how do we find the form of T? By guessing a statistic T that might be sufficient and computing the conditional distribution of $X|T = t$?

For families of populations having pdfs or pmfs, a simple way of finding sufficient statistics is to use the following factorization theorem.

Theorem 6.2.6 (the Factorization Theorem)

Let $f_\theta(x)$ be the joint pdf or pmf of the sample X. A statistic $T(X)$ is sufficient for θ iff there are functions h (which does not depend on θ) and g_θ (which depends on θ) on the range of T such that

$$f_\theta(x) = g_\theta\big(T(x)\big)h(x).$$

In the binomial example, $f_\theta(x) = g_\theta\big(T(x)\big)h(x)$ if we set

$$g_\theta(t) = \theta^t(1-\theta)^{n-t} \text{ and } h(x) = \begin{cases} 1 & x_i = 0, 1, i = 1, ..., n \\ 0 & \text{otherwise} \end{cases}$$

Proof of Theorem 6.2.6 for the discrete case.

Suppose that $T(X)$ is sufficient.
Let $g_\theta(t) = P_\theta(T(X) = t)$ and $h(x) = P(X = x | T(X) = T(x))$.
Then

$$\begin{aligned} f_\theta(x) &= P_\theta(X = x) = P_\theta(X = x, T(X) = T(x)) \\ &= P_\theta(T(X) = T(x))P(X = x | T(X) = T(x)) \\ &= g_\theta(T(x))h(x) \end{aligned}$$

Suppose now that $f_\theta(x) = g_\theta\big(T(x)\big)h(x)$ for $x \in \mathscr{X}$.
Let $q_\theta(t)$ be the pmf of $T(X)$ and $A_x = \{y : T(y) = T(x)\}$.
Then, for any $x \in \mathscr{X}$,

$$
\begin{aligned}
\frac{f_\theta(x)}{q_\theta(T(x))} &= \frac{g_\theta(T(x))h(x)}{q_\theta(T(x))} = \frac{g_\theta(T(x))h(x)}{P_\theta(T(X) = T(x))} \\
&= \frac{g_\theta(T(x))h(x)}{\sum_{y \in A_x} f_\theta(y)} = \frac{g_\theta(T(x))h(x)}{\sum_{y \in A_x} g_\theta(T(y))h(y)} \\
&= \frac{g_\theta(T(x))h(x)}{g_\theta(T(x))\sum_{y \in A_x} h(y)} = \frac{h(x)}{\sum_{y \in A_x} h(y)}
\end{aligned}
$$

which does not depend on θ, i.e., T is sufficient for θ.

Example 6.2.4 (normal sufficient statistic)

Let $X_1, ..., X_n$ be iid $N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$; the joint pdf is

$$
\begin{aligned}
f_\theta(x) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/2\sigma^2} = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right)
\end{aligned}
$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{(n-1)s^2}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right)$$

where $s^2 = (n-1)^{-1}\sum_{i=1}^n (x_i - \bar{x})^2$, the realization of the sample variance $S^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$.

Hence, by Theorem 6.2.6, $(\bar{X}, S^2)$ is a two-dimensional sufficient statistic for $\theta = (\mu, \sigma^2)$.

- If σ^2 is known, then $\bar{X}$ is sufficient for μ.
- If μ is known, then S^2 is sufficient for σ^2.
- If both μ and σ^2 are unknown, we cannot say that $\bar{X}$ is sufficient for μ (or S^2 is sufficient for σ^2); the correct statement is that $\bar{X}$ and S^2 together is sufficient for μ and σ^2.
- We can also say that $(\bar{X}, S^2)$ is sufficient for μ (or σ^2).

Sufficiency for a sub-family

Let θ be a parameter and η be a subset of components of θ.
If T is sufficient for θ, then it is also sufficient for η.

Example 6.2.5 (sufficient order statistics)

Let $X_1,...,X_n$ be iid with a pdf f_θ and $X_{(1)},...,X_{(n)}$ be the order statistics. The joint pdf of $X=(X_1,...,X_n)$ is

$$\prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} f_\theta(x_{(i)})$$

where $x_{(1)},...,x_{(n)}$ are the ordered values of $x_1,...,x_n$.
Then, by the factorization theorem, $(X_{(1)},...,X_{(n)})$ is sufficient for θ.

Intuitively, given the order statistics, what is left in the original data set is the information regarding the positions of $x_1,...,x_n$ and, hence, the set of order statistics is sufficient whenever positions of x_i's are not of interest.

One-to-one transformations of a sufficient statistic

It follows from the factorization theorem that, if T is sufficient and U is a one-to-one function of T, then U is also sufficient.

But this is also true in general by the definition of sufficiency.

In the order statistics problem, $U = (U_1, ..., U_n)$ is a one-to-one function of $(X_{(1)}, ..., X_{(n)})$, where $U_k = \sum_{i=1}^n X_i^k$, $k = 1, ..., n$.
Hence, U is also sufficient for θ.

Example 6.2.8 (uniform sufficient statistic)

Let $X_1, ..., X_n$ be iid from *uniform*$(0, \theta)$, where $\theta > 0$ is the unknown parameter.
The joint pdf of $X_1, ..., X_n$ is

$$\prod_{i=1}^n f_\theta(x_i) = \prod_{i=1}^n \left[\frac{1}{\theta} I(\{0 < x_i < \theta\}) \right] = \frac{1}{\theta^n} I(\{0 < x_{(n)} < \theta\})$$

with $x_{(n)}$ being the largest value of $x_1, ..., x_n$.

Thus, the largest order statistic $X_{(n)}$ is sufficient for θ.

Intuitively, because $X_i \leq \theta$ for all i, if we observe $X_{(n)}$, then we know that $\theta \geq X_{(n)}$ and the values of other X_i's do not provide any additional information about θ.

The same result holds when $X_1, ..., X_n$ are iid from the discrete uniform distribution on $1, 2, ..., \theta$.

Theorem 6.2.10 (exponential families)

Let $X_1,...,X_n$ be iid from a pdf or pmf $f_\theta(x)$ that belongs to an exponential family:

$$f_\theta(x) = h(x)c(\theta)\exp\left(\sum_{j=1}^{k} w_j(\theta)t_j(x)\right)$$

The joint pdf or pmf of $X = (X_1,...,X_n)$ is

$$\prod_{i=1}^{n} f_\theta(x_i) = \left[\prod_{i=1}^{n} h(x_i)\right][c(\theta)]^n \exp\left(\sum_{j=1}^{k} w_j(\theta)\sum_{i=1}^{n} t_j(x_i)\right)$$

It follows from the factorization theorem that the k-dimensional statistic

$$T(x) = \left(\sum_{i=1}^{n} t_1(X_i),...,\sum_{i=1}^{n} t_k(X_i)\right)$$

is sufficient for θ.

Sufficiency Principle

Let X be a sample from a population indexed by $\theta \in \Theta$.
If $T(X)$ is sufficient for θ, then any inference about θ should depend on the sample only through the value $T(X)$.

- Another way to state the sufficiency principle is that, if x and y are two data points (realizations of X), then our decision or inference about θ should be the same when $T(x) = T(y)$.
- The sufficiency principle says that in any inference procedure we should consider functions of a sufficient statistic only.
- In what sense we can be assured that using functions of a sufficient statistic is enough?
- First we should have a criterion to evaluate the performance of inference procedures.
- As an example, we consider here the problem of estimating a function $\vartheta = \psi(\theta)$, where ψ is a known function on the parameter space Θ, but ϑ is unknown.

- Let $U(X)$ be a statistic used to estimate the unknown ϑ.
  A common criterion for the performance of $U(X)$ is the so-called mean squared error (mse) defined as
  $$E_\theta[U(X) - \vartheta]^2 = E_\theta[U(X) - \psi(\theta)]^2, \qquad \theta \in \Theta$$
  where E_θ is the expectation with respect to the population indexed by θ.
- We view $U(X) - \vartheta$ to be the estimation error, which is random since X is random. The mse is simply the average of squared estimation error under the population indexed by θ, and we want to choose a statistic such that the mse is as small as possible.

Rao-Blackwell theorem

Let X be a sample from a population indexed by $\theta \in \Theta$ and $T(X)$ be a sufficient statistic for θ. If $U(X)$ is a statistic used to estimate $\vartheta = \psi(\theta)$ and $E_\theta[U(X) - \vartheta]^2 < \infty$, then the statistic $h(T) = E[U(X)|T]$ satisfies
$$E_\theta[h(T) - \vartheta]^2 < E_\theta[U(X) - \vartheta]^2 \qquad \theta \in \Theta$$
unless $P_\theta(U(X) = h(T(X))) = 1$, $\theta \in \Theta$.

- The Rao-Blackwell theorem says that if $U(X)$ is not a function of the sufficient statistic T, then the new statistic $h(T) = E[U(X)|T]$ is better than $U(X)$ in terms of the mean squared error criterion.
- The theorem is meaningful if a T other than the original data X can be found (such as the minimal sufficient statistic).
- Because $E_\theta[U(X) - \vartheta]^2 < \infty$, $E[U(X)|T]$ is well defined; in fact, we only need $E_\theta|U(X)| < \infty$ for every $\theta \in \Theta$.
- Because T is sufficient, $E[U(X)|T]$ does not depend on θ and is a statistic.
- The Rao-Blackwell theorem actually has a more general form considering a criterion other than the mean squared error.

Proof.

For every $\theta \in \Theta$,

$$
\begin{aligned}
E_\theta[U(X) - \vartheta]^2 &= E_\theta\{[U(X) - h(T)] + [h(T) - \vartheta]\}^2 \\
&= E_\theta[U(X) - h(T)]^2 + E_\theta[h(T) - \vartheta]^2 \\
&\quad + 2E_\theta[U(X) - h(T)][h(T) - \vartheta]
\end{aligned}
$$

Using the properties of conditional expectations, we obtain that

$$\begin{aligned} E_\theta[U(X)-h(T)][h(T)-\vartheta] &= E_\theta\big(E_\theta\{[U(X)-h(T)][h(T)-\vartheta]|T\}\big) \\ &= E_\theta\big([h(T)-\vartheta]E_\theta\{[U(X)-h(T)]|T\}\big) \\ &= E_\theta\big([h(T)-\vartheta]E_\theta[U(X)|T]-h(T)\big) \\ &= 0 \end{aligned}$$

Hence,

$$E_\theta[U(X)-\vartheta]^2 = E_\theta[U(X)-h(T)]^2 + E_\theta[h(T)-\vartheta]^2 > E_\theta[h(T)-\vartheta]^2$$

unless $E_\theta[U(X)-h(T)]^2 = 0$, which implies $P_\theta(U(X)=h(T)) = 1$ from our previous discussion.

The Rao-Blackwell theorem tells us that we should consider functions of a sufficient statistic (if one simpler than X is available).

However, we still need to choose a function such that it provides the best procedure among all functions of the given sufficient statistic.

This will be treated in later chapters.