4. Sufficient Statistics

ECE 830, Spring 2014

Parameters and distributions

Suppose that X is an n-dimensional random vector with density determined by a p-dimensional ($p < n$) parameter θ:

$$X \sim p(x|\theta).$$

The parameter θ determines the distribution of X.

In many signal processing applications we need to make some decision about θ from observations of X, where the density of X can be one of many in a family of distributions, $\{p(x|\theta)\}_{\theta \in \Theta}$, indexed by different choices of the parameter θ.

Example: Univariate Gaussian

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, x_i \sim \mathcal{N}(\theta, 1)$$

$$p(x|\theta) =$$

$$=$$

Statistically speaking, any information we are interested in gleaning from x depends only on the p parameters

$$\theta = [\theta_1, ..., \theta_p]^\top.$$

So, it is natural to ask:

Question

Can we compress the n raw measurements into a lower dimensional statistic that carries all the useful information in x?

In particular, we want to use this lower dimensional statistic to estimate θ with the same quality as if we kept all of x.
If so, then to study θ we could discard x and retain only the compressed statistic.

Goal

Given a family of distributions $\{p(x|\theta)\}_{\theta\in\Theta}$ and one or more observations from a particular distribution $p(x|\theta^*)$ in this family, find a data compression strategy that preserves all information pertaining to θ^*. The function identified by such strategies called a *sufficient statistic*.

Definition: Sufficient statistic

Let X be an n-dimensional random vector and let θ denote a p-dimensional parameter of the distribution of X. The statistic $t := T(x)$ is a *sufficient statistic* for θ if and only if the conditional distribution of X given $T(X)$ is independent of θ.

Implications

Let's see what this implies. Let $p(x, t|\theta)$ by the joint density of $(X, T(X))$ conditioned on θ, so that

$$p(x, t|\theta) = \begin{cases} p(x|\theta), & t = T(x) \\ 0 & \text{otherwise} \end{cases}$$

Then

This has the following implications:

1. Parametrization of the probability law of X is manifested completely in $p(t|\theta)$.
2. Any inference strategy based on x can be replaced by an algorithm based on t without loss in accuracy.
3. Information in x regarding θ is contained in t.

Example: Binary Information Source (Scharf p. 78)

Sequence of independent, identically distributed Bernoulli r.v. used to model a communication signal.

$$x = [x_1, x_2, ..., x_N]^\top$$

Each bit $x_n \in \{0, 1\}$;

$$\begin{aligned} p(x_n = 1) &= \theta \\ p(x_n = 0) &= 1 - \theta \end{aligned}$$

Example: (cont.)

Joint probability mass function of X is:

The probability mass function of the number of 1's occurring in N independent Bernoulli trials is binomial:

Example: (cont.)

The joint probability (or mass function) of x and k is

From this we obtain

Example: (cont.)

- Conditional pmf $p(x|k,\theta)$ does not depend on θ!
- This reveals that given k, the full observation x brings no additional discriminating information about θ.
- We say that the conditional distribution of x given k is independent of θ.
- Dependence of a random sample on parameter θ is completely carried on k
- To make inferences about θ, we only need to save k, due to the number of "1"s in x

Fisher-Neyman Factorization Theorem

Working out the above conditional distributions can be challenging in practice, making it difficult to find sufficient statistics directly. The following theorem helps us verify sufficient statistics more readily.

Fisher-Neyman Factorization Theorem

Let X be a discrete random vector with pmf $p(x|\theta)$. The statistic $t = T(x)$ is sufficient for θ *if and only if*

$$p(x|\theta) = a(x)b_\theta(t)$$

for some functions a which depends on x but not θ and b which depends on θ but not x except through t.

Lots of examples

Example: Data itself

x is a sufficient statistic

$$p(x|\theta) = \underbrace{1}_{a(x)} \cdot \underbrace{p(x|\theta)}_{b_\theta(x)}$$

Example: Bernoulli trials revisited

$$p(x|\theta) =$$

$\Rightarrow$

Example: Poisson

Suppose $X_1, ..., X_N$ are iid Poisson random variables:

$$p(x_n|\lambda) = e^{-\lambda}\frac{\lambda^{x_n}}{x_n!}, \lambda \text{is parameter}$$

and we measure $x = [x_1, ..., x_N]^\top$. Then

$$\begin{aligned} p(x|\lambda) &= \prod_{n=1}^{N} e^{-\lambda}\frac{\lambda^{x_n}}{x_n!} \\ &= \end{aligned}$$

$\Rightarrow$

Example: Uniform

Suppose $x = [x_1, ..., x_N]^\top$ are samples from a uniform density on the interval $[a, b]$. What is a sufficient statistic for $\theta = [a\ b]^\top$?

Proof of Fisher-Neyman factorization theorem

(If / Sufficiency)

(Only if / Necessity)

Example: Gaussian observations

Let $x_1, ..., x_N$ be independent observations from a $\mathcal{N}(\mu, \sigma^2)$ distribution and define $x = [x_1, ..., x_N]^\top, \theta = [\mu, \sigma^2]^\top$.

$$\begin{aligned} p(x|\theta) &= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/2\sigma^2} \\ &= (2\pi\sigma^2)^{-N/2} e^{-\sum_{n=1}^{N} -(x_n-\mu)^2/2\sigma^2} \end{aligned}$$

Define the statistics:

Example: (cont.)

$$
\begin{aligned}
p(x|\theta) =& (2\pi\sigma^2)^{-N/2} \exp\left\{\sum_{n=1}^{N} -\frac{1}{2\sigma^2}(x_n - \widehat{\mu} + \widehat{\mu} - \mu)^2\right\} \\
=& (2\pi\sigma^2)^{-N/2} \\
& \cdot \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \widehat{\mu})^2 + 2(x_n - \widehat{\mu})(\widehat{\mu} - \mu) + (\widehat{\mu} - \mu)^2\right\}
\end{aligned}
$$

and note that:

$$
\begin{aligned}
\sum_{n=1}^{N}(x_n - \widehat{\mu})(\widehat{\mu} - \mu) &= (\widehat{\mu} - \mu)\sum_{n=1}^{N}(x_n - \widehat{\mu}) \\
&= (\widehat{\mu} - \mu)\left(\sum_{n=1}^{N} x_n - \sum_{n=1}^{N}\widehat{\mu}\right) \\
&=
\end{aligned}
$$

Example: (cont.)

$$
\begin{aligned}
p(x|\theta) &= (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\widehat{\mu})^2\right\} \\
&\quad \times \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(\widehat{\mu}-\mu)^2\right\} \\
&= \\
&=
\end{aligned}
$$

Therefore $\begin{bmatrix} \widehat{\mu} \\ \widehat{\sigma^2} \end{bmatrix}$ is a sufficient statistic for $\begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$.

Example: Gaussian Vector Observations (Scharf, p. 83)

If $X_1, ..., X_M$ are iid $\mathcal{N}(\mu, R)$, then

$$\mu = \frac{1}{M} \sum_{n=1}^{M} x_n = \text{“Sample Mean”}$$

and

$$\widehat{R} = \frac{1}{M} \sum_{n=1}^{M} (x_n - \widehat{m})(x_n - \widehat{m})^{\top} = \text{“Sample Covariance”}$$

are sufficient statistics for (μ, R).

Sufficient Statistics and Signal Subspace Models

For $H \in \mathbb{R}^{n \times p}, \theta \in \mathbb{R}^{p \times 1}$, suppose we observe a signal

$$s = H\theta$$

in noise

$$w \sim \mathcal{N}(\mathbf{0}, R):$$

$$x = H\theta + w$$

What is the minimal (lowest dimensional) sufficient statistic for θ?

The Rao-Blackwell Theorem

Rao-Blackwell Theorem

Let X be a random variable with pdf $p(x|\theta)$ and let $t(X)$ be a sufficient statistic. Let $f(x)$ be an estimator of θ and define the mean-squared error

$$\mathrm{MSE}(f) := \mathbb{E}\left[\|f(X) - \theta\|^2\right].$$

Next define $g(t) := \mathbb{E}\left[f(X)|t(X)\right]$ and

$$\mathrm{MSE}(g) := \mathbb{E}\left[\|g(t(X)) - \theta\|^2\right].$$

Then

$$\mathrm{MSE}(g) \leq \mathrm{MSE}(f)$$

with equality iff $f(X) \equiv g(t(X))$ with probability one (almost surely).

Observations and Interpretation

1. g is a function of the sufficient statistic (and otherwise independent of x).
2. Given any estimator f that is not a function of a sufficient statistic, there exists a better estimator (with respect to MSE).
3. We may restrict our search for estimators to functions of a sufficient statistic.
4. The conditional expectation

$$\mathbb{E}\left[f(X)|T(X)\right]$$

averages out (or removes) non-informative components in f. We can view this as a filter that eliminates unnecessary components of the data.

Proof of Rao-Blackwell Theorem

(for 1-dimensional θ)

$$
\begin{aligned}
\mathsf{MSE}(f) &= \mathbb{E}[(f(X)-\theta)^2] && \\
&= && \text{smoothing} \\
&\geq && \text{Jensen's} \\
&= && \text{defn of } g \\
&= \mathsf{MSE}(g) &&
\end{aligned}
$$