10 Sufficient Statistics

In many settings, the observed data can have very large dimension. As such, performing inference based on such data directly can be rather cumbersome. As a result, an attractive architecture for inference involves first preprocessing the data to obtain a more compact set of data from which we make inferences, rather than directly from the data itself. Two obvious questions arise when considering such an architecture. First, what kinds of preprocessing are lossless (with respect to the task of interest—inference)? Second, what is a meaningful measure of compactness when designing the preprocessing?

With respect to the first question, we have already seen that at least for binary hypothesis testing, the decision rule used the data through the likelihood ratio only. Similarly, we also observed that for exponential families, the values of natural statistics and the log base distribution fully determined the probability distribution values. Both examples suggest that we ought to be able to more generally develop useful forms of lossless preprocessing, as we now develop.

In such an architecture, the preprocessing of interest consists of computing a set of functions, or *statistics*, of the data. Lossless preprocessing, as described above, is, in turn, captured by the notion of *sufficient* statistics. Finally, in pursuit of preprocessing with good compaction properties, we develop the notion of a *minimal* sufficient statistics.

We start the development by modeling the parameters of interest as nonrandom. We then extend the notion of sufficient statistics to the Bayesian case.

10.1 The NonBayesian Case

We define a statistic $\mathbf{t}(\cdot)$ to be a deterministic function that takes values of $\mathbf{y}$ as an argument. Note that $\mathbf{t} = \mathbf{t}(\mathbf{y})$ is itself a random variable. As some properties of distributions involving $\mathbf{t}$, it is easy to see that[1]

$$p_{\mathbf{t}|\mathbf{y}}(\mathbf{t}|\mathbf{y};\mathbf{x}) = \mathbb{1}_{\mathbf{t}=\mathbf{t}(\mathbf{y})},$$

and that $p_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t})$ is only defined for $\mathbf{t} = \mathbf{t}(\mathbf{y})$. In turn, it follows that the joint distribution for $\mathbf{y}$ and $\mathbf{t}$ is

$$p_{\mathbf{y},\mathbf{t}}(\mathbf{y},\mathbf{t};\mathbf{x}) = p_{\mathbf{y}}(\mathbf{y};\mathbf{x}) \cdot \mathbb{1}_{\mathbf{t}=\mathbf{t}(\mathbf{y})},$$

and that the marginals take the forms

$$p_{\mathbf{t}}(\mathbf{t};\mathbf{x}) = \sum_{\mathbf{y}\in\mathcal{Y}} p_{\mathbf{y},\mathbf{t}}(\mathbf{y},\mathbf{t};\mathbf{x}) = \sum_{\{\mathbf{y}\in\mathcal{Y}\,:\,\mathbf{t}(\mathbf{y})=\mathbf{t}\}} p_{\mathbf{y}}(\mathbf{y};\mathbf{x})$$

[1]Recall that The notation $\mathbb{1}_{\mathcal{E}}$ means that the value is 1 if $\mathcal{E}$ is true, and 0 otherwise.

and

$$p_{\mathbf{y}}(\mathbf{y};\mathbf{x}) = \sum_{\mathbf{t}} p_{\mathbf{y},\mathbf{t}}(\mathbf{y},\mathbf{t};\mathbf{x}) = p_{\mathbf{y},\mathbf{t}}(\mathbf{y},\mathbf{t}(\mathbf{y});\mathbf{x}) = p_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t}(\mathbf{y});\mathbf{x})\, p_{\mathbf{t}}(\mathbf{t};\mathbf{x}).$$

We use these properties extensively in the derivations below.

Definition 1. *A statistic* $\mathbf{t}(\cdot)$ *is* sufficient *with respect to distribution* $p_{\mathbf{y}}(\cdot;\mathbf{x})$ *if* $p_{\mathbf{y}|\mathbf{t}}(\cdot|\cdot;\mathbf{x})$ *is not a function of* $\mathbf{x}$ *for all* $\mathbf{x} \in \mathcal{X}$.

The definition says that once we know the value of $\mathbf{t}$, the remaining uncertainty about $\mathbf{y}$ is not a function of $\mathbf{x}$. Therefore, we cannot infer additional information about $\mathbf{x}$ by getting access to the value of $\mathbf{y}$ from which this value of $\mathbf{t}$ was computed. The following theorem formalizes the usefulness of sufficient statistics for inference.

Theorem 1 (Likelihood Characterization)**.** *A statistic* $\mathbf{t}(\cdot)$ *is sufficient with respect to distribution* $p_{\mathbf{y}}(\cdot;\mathbf{x})$ *if and only if*

$$\frac{p_{\mathbf{y}}(\mathbf{y};\mathbf{x})}{p_{\mathbf{t}}(\mathbf{t}(\mathbf{y});\mathbf{x})} \tag{1}$$

is not a function of $\mathbf{x}$ *for all* $\mathbf{x} \in \mathcal{X}$ *and* $\mathbf{y} \in \mathcal{Y}$.

Proof. We start with the following expression from above

$$p_{\mathbf{y}}(\mathbf{y};\mathbf{x}) = \sum_{\mathbf{t}} p_{\mathbf{y},\mathbf{t}}(\mathbf{y},\mathbf{t};\mathbf{x}) = p_{\mathbf{y},\mathbf{t}}(\mathbf{y},\mathbf{t}(\mathbf{y});\mathbf{x}) = p_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t}(\mathbf{y});\mathbf{x})\, p_{\mathbf{t}}(\mathbf{t}(\mathbf{y});\mathbf{x}), \tag{2}$$

from which we obtain

$$\frac{p_{\mathbf{y}}(\mathbf{y};\mathbf{x})}{p_{\mathbf{t}}(\mathbf{t}(\mathbf{y});\mathbf{x})} = p_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t}(\mathbf{y});\mathbf{x}). \tag{3}$$

It is easy to see that the ratio is not a function of $\mathbf{x}$ if and only if $p_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t}(\mathbf{y});\mathbf{x})$ is not a function of $\mathbf{x}$, which is exactly the definition of a sufficient statistic. □

This theorem states that a sufficient statistic has the property that inference based on $\mathbf{t}(\mathbf{y})$ is equivalent to inference based on $\mathbf{y}$. More specifically, if we use $L_{\mathbf{y}}(\mathbf{x}) = p_{\mathbf{y}}(\mathbf{y};\mathbf{x})$ and $L_{\mathbf{t}}(\mathbf{x}) = p_{\mathbf{t}}(\mathbf{t}(\mathbf{y});\mathbf{x})$ to denote the likelihood functions for $\mathbf{x}$ based on the data $\mathbf{y}$ and $\mathbf{t}(\mathbf{y})$, respectively, Theorem 1 asserts that $\mathbf{t}(\mathbf{y})$ is a sufficient statistic if and only if $L_{\mathbf{y}}(\mathbf{x}) \propto L_{\mathbf{t}}(\mathbf{x})$. Thus, any inference about $\mathbf{x}$ based the use of likelihoods, such as estimation by the method of maximum likelihood, must yield the same result whether $\mathbf{y}$ or $\mathbf{t}(\mathbf{y})$ is used as the data.

Unfortunately, the definition of sufficiency does not provide a way to construct sufficient statistics. It might also be hard to check sufficiency of a proposed statistic. The following theorem identifies the structure in the distribution that leads to existence of sufficient statistics, bringing us closer to being able to propose and test candidate sufficient statistics.

Theorem 2 (Neyman Factorization Theorem)**.** *A statistic* $\mathbf{t}(\cdot)$ *is sufficient with respect to distribution* $p_{\mathbf{y}}(\cdot;\mathbf{x})$ *if and only if there exist functions* $a(\cdot,\cdot)$ *and* $b(\cdot)$ *such that*

$$p_{\mathbf{y}}(\mathbf{y};\mathbf{x}) = a(\mathbf{t}(\mathbf{y}),\mathbf{x})\, b(\mathbf{y}) \tag{4}$$

for all $\mathbf{x} \in \mathcal{X}$ *and* $\mathbf{y} \in \mathcal{Y}$.

Proof. We first assume that the distribution is factored as defined in (4). Then for $\mathbf{t} = \mathbf{t}(\mathbf{y})$,

$$p_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t};\mathbf{x}) = \frac{p_{\mathbf{y},\mathbf{t}}(\mathbf{y},\mathbf{t};\mathbf{x})}{p_{\mathbf{t}}(\mathbf{t};\mathbf{x})} = \frac{p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\, p_{\mathbf{t}|\mathbf{y}}(\mathbf{t}|\mathbf{y};\mathbf{x})}{\sum\limits_{\mathbf{y}':\mathbf{t}(\mathbf{y}')=\mathbf{t}} p_{\mathbf{y}}(\mathbf{y}';\mathbf{x})} = \frac{p_{\mathbf{y}}(\mathbf{y};\mathbf{x})}{\sum\limits_{\mathbf{y}':\mathbf{t}(\mathbf{y}')=\mathbf{t}} p_{\mathbf{y}}(\mathbf{y}';\mathbf{x})} \tag{5}$$

$$= \frac{a(\mathbf{t},\mathbf{x})\, b(\mathbf{y})}{a(\mathbf{t},\mathbf{x}) \sum\limits_{\mathbf{y}':\mathbf{t}(\mathbf{y}')=\mathbf{t}} b(\mathbf{y}')} = \frac{b(\mathbf{y})}{\sum\limits_{\mathbf{y}':\mathbf{t}(\mathbf{y}')=\mathbf{t}} b(\mathbf{y}')}, \tag{6}$$

which clearly is not a function of $\mathbf{x}$. Therefore $\mathbf{t}(\cdot)$ satisfies the definition of a sufficient statistic.

Now let's assume that $\mathbf{t}(\cdot)$ is a sufficient statistic. Then $p_{\mathbf{y}|\mathbf{t}}(\cdot|\cdot;\mathbf{x})$ is not a function of $\mathbf{x}$, allowing us to define $b(\mathbf{y}) = p_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t}(\mathbf{y}))$. By setting $a(\mathbf{t},\mathbf{x}) = p_{\mathbf{t}}(\mathbf{t};\mathbf{x})$, we construct the required factorization as can be seen from (2). □

Example 1. Let $\mathbf{y} = \begin{bmatrix} y_1 & y_2 \end{bmatrix}^{\mathrm{T}}$ be a two-dimensional vector whose components are i.i.d. Gaussian random variables with mean x and unit variance:

$$p_{\mathbf{y}}(\mathbf{y};\mathbf{x}) = (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{(y_1-x)^2 + (y_2-x)^2}{2}\right\}$$

$$= (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{y_1^2+y_2^2}{2}\right\} \exp\left\{x(y_1+y_2) - x^2\right\}. \tag{7}$$

It is easy to see that $t(\mathbf{y}) = y_1 + y_2$ is a sufficient statistic since $p_{\mathbf{y}}(\mathbf{y};\mathbf{x})$ satisfies the Neyman Factorization Theorem with

$$a(t,x) = e^{xt-x^2}, \quad b(\mathbf{y}) = (2\pi)^{-\frac{1}{2}} e^{-\frac{y_1^2+y_2^2}{2}}. \tag{8}$$

Example 2. Let $p_{\mathbf{y}}(\cdot;\mathbf{x})$ be a member of an exponential family $\mathtt{E}(x;\boldsymbol{\lambda}(\cdot),\mathbf{t}(\cdot),\beta(\cdot))$. Then the (potentially vector-valued) natural statistic of the family $\mathbf{t}(\cdot)$ is a sufficient statistic since $p_{\mathbf{y}}(\cdot;\mathbf{x})$ satisfies the Neyman Factorization Theorem with

$$a(\mathbf{t},\mathbf{x}) = \exp\{\boldsymbol{\lambda}^T(\mathbf{x})\mathbf{t} - \alpha(\mathbf{x})\}, \quad b(\mathbf{y}) = \exp\{\beta(\mathbf{y})\}. \tag{9}$$

Note that $\beta(\mathbf{y})$ is not part of the sufficient statistic. This makes sense, intuitively. Indeed, if we think of $p_{\mathbf{y}}(\mathbf{y};\mathbf{x})$ as a function in the $\mathbf{y}$ and $\mathbf{x}$, then for a fixed $\mathbf{y}$ (i.e.,

a given observation), the relative variations in the likelihood $p_{\mathbf{y}}(\mathbf{y};\mathbf{x})$ with $\mathbf{x}$ is what we have available to exploit in making inferences about the true value of $\mathbf{x}$. However, if two exponential families differ only in the value of $\beta(\cdot)$, then for a fixed $\mathbf{y}$, the relative variations with $\mathbf{x}$ are exactly the same. The likelihoods differ only by the multiplicative factor $e^{\beta(\mathbf{y})}$. Hence, we cannot make better inferences about $\mathbf{x}$ from one over the other, and thus having access to $\beta(\mathbf{y})$ is of no value.

Example 3. For a model $p_{\mathbf{y}}(\cdot;x)$ where the parameter x is from a finite alphabet $\mathcal{X} = \{x_0, x_1, \ldots, x_{L-1}\}$, the vector of likelihoods

$$\mathbf{t}(\mathbf{y}) = \begin{bmatrix} t_0 \\ t_1 \\ \vdots \\ t_{L-1} \end{bmatrix} = \begin{bmatrix} p_{\mathbf{y}}(\mathbf{y};x_0) \\ p_{\mathbf{y}}(\mathbf{y};x_1) \\ \vdots \\ p_{\mathbf{y}}(\mathbf{y};x_{L-1}) \end{bmatrix} \tag{10}$$

is a sufficient statistic. To see this, simply note that we can write $p_{\mathbf{y}}(\mathbf{y};x)$ in the form (4) with[2]

$$a(\mathbf{t}(\mathbf{y}),x) = \sum_{l=0}^{L-1} t_l \cdot \mathbb{1}_{x_l}(x) \quad \text{and} \quad b(\mathbf{y}) = 1.$$

Hence, applying the Neyman Factorization theorem establishes that (10) is a sufficient statistic.

Example 4. For continuous $\mathbf{x}$ whereby $\mathcal{X} \subset \mathbb{R}^k$, the likelihood function $p_{\mathbf{y}}(\mathbf{y};\cdot)$, an infinite-dimensional object, is also generally a sufficient statistic. Not surprisingly, formalizing such a statement requires some care. This result might seem of somewhat limited value, since it is not clear the degree to which this statistic is any more compact than the data $\mathbf{y}$ itself. However, the value of the result lies in the observation that good *approximate* sufficient statistic can be formed from a suitable collection of samples the likelihood function, corresponding to a discretization of $\mathcal{X}$.

More generally, despite these encouraging examples, finding good sufficient statistics is somewhat of an art. Nevertheless, at least in some cases, there are procedures that can help in this pursuit, which we leave to exercises.

10.2 Minimal Sufficient Statistics

Given a sufficient statistic, we can augment it with any additional function of data and obviously still have be sufficient. Indeed, the data itself is a sufficient statistic. In other words, sufficient statistics need not be compact in any way. However, it is natural to seek, among all sufficient statistics, those which summarize those aspects

[2]Recall that as related notation to that introduced earlier, $\mathbb{1}_u(v)$ is 1 if $u = v$, and 0 otherwise.

of the data required for inference most compactly. Such statistics can then be thought of as "minimal."

While it might be tempting to define a minimal sufficient statistic as one having the minimum possible dimension (in the case of continuous-valued parameters), this is not a useful notion. Indeed, we can, for example, encode an tuple of real numbers that corresponds to the data $\mathbf{y}$ into a single real number by simply interleaving the digits in their binary representations, from most- to least-significant. However, this would correspond to everywhere-discontinuous function $t(\cdot)$ of the data that achieves no real compaction.

In contrast, a useful notion of minimality is as follows.

Definition 2. *A sufficient statistic* $\mathbf{t}^*$ *is* minimal *if for any other sufficient statistic* $\mathbf{t}$ *there exists function* $\mathbf{g}(\cdot)$ *such that* $\mathbf{t}^* = \mathbf{g}(\mathbf{t})$.

As a preliminary remark, note that a minimal sufficient statistic is never unique. For example, any one-to-one function of a minimal sufficient statistic must also be minimal by this definition.

Uniqueness issues notwithstanding, this definition of minimality is both physically meaningful and intuitively appealing: a sufficient statistic is minimal if it can be expressed as a function of any other sufficient statistic. However, this definition is not constructive: there is, in general, no systematic test (or computational procedure) that is guaranteed to resolve whether a sufficient statistic is minimal. Nevertheless, while easily-tested necessary and sufficient conditions are not known, simple sufficient conditions are known.

A particularly well-known sufficient condition arises out of the notion of *completeness* of a sufficient statistic.

Definition 3. *A sufficient statistic* $\mathbf{t}^*$ *is* complete *if for any function* $\boldsymbol{\phi}(\cdot)$

$$\left\{\mathbb{E}\left[\boldsymbol{\phi}(\mathbf{t}^*(\mathbf{y}))\right] = 0,\ \forall \mathbf{x} \in \mathcal{X}\right\} \quad \Rightarrow \quad \boldsymbol{\phi}(\cdot) \equiv 0. \tag{11}$$

The resulting sufficient condition is then as follows.

Theorem 3. *A sufficient statistic* $\mathbf{t}$ *is minimal if it is complete.*

Proof. Suppose the sufficient statistic $\mathbf{t}(\mathbf{y})$ is complete and the sufficient statistic $\mathbf{s}(\mathbf{y})$ is minimal. Then by Definition 2, there exists a function $\mathbf{g}(\cdot)$ such that $\mathbf{s} = \mathbf{g}(\mathbf{t})$. Then

$$\mathbb{E}\left[\mathbf{t}\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbf{t}|\mathbf{s}\right]\right] \tag{12}$$

but

$$\mathbb{E}\left[\mathbf{t}|\mathbf{s} = \mathbf{s}\right] = \mathbf{f}(\mathbf{s}) = \mathbf{f}(\mathbf{g}(\mathbf{t})) \triangleq \tilde{\mathbf{f}}(\mathbf{t}). \tag{13}$$

Letting

$$\boldsymbol{\phi}(\mathbf{t}) = \mathbf{t} - \mathbb{E}\left[\mathbf{t}|\mathbf{s} = \mathbf{s}\right] = \mathbf{t} - \tilde{\mathbf{f}}(\mathbf{t}), \tag{14}$$

we see that $\boldsymbol{\phi}(\mathbf{t})$ is zero-mean for all $\mathbf{x} \in \mathcal{X}$. Thus by the completeness of $\mathbf{t}$ we have, via Definition 3, that

$$\mathbf{t} = \mathbb{E}\left[\mathbf{t}|\mathbf{s} = \mathbf{s}\right] = \mathbf{f}(\mathbf{s}), \tag{15}$$

i.e., $\mathbf{t}$ is a function of $\mathbf{s}$. But since $\mathbf{s}$ is minimal, it is a function of every other sufficient statistic. Thus, (15) implies that $\mathbf{t}$ is also, through $\mathbf{f}(\cdot)$, a function of every other sufficient statistic. Thus, by Definition 2, $\mathbf{t}$ is minimal as well. □

We emphasize that completeness is not necessary for minimality; it is merely a sufficient condition that can sometimes be easier to test. In that sense, the notion of completeness is much less fundamental than minimality.

Example 5. If $\mathbf{t}^*(\mathbf{y}) = \begin{bmatrix} t_1(\mathbf{y}) & t_2(\mathbf{y}) \end{bmatrix}^{\mathrm{T}}$ is a minimal sufficient statistic, it is easy to show that both

$$\tilde{\mathbf{t}}^*(\mathbf{y}) = \begin{bmatrix} t_1(\mathbf{y}) + t_2(\mathbf{y}) \\ t_1(\mathbf{y}) - t_2(\mathbf{y}) \end{bmatrix} \quad \text{and} \quad \tilde{\tilde{\mathbf{t}}}^*(\mathbf{y}) = \begin{bmatrix} t_1(\mathbf{y}) \\ t_2(\mathbf{y}) \\ t_1(\mathbf{y}) + t_2(\mathbf{y}) \end{bmatrix}$$

are also minimal sufficient statistics. They are each sufficient because we can reconstruct $\mathbf{t}^*$ from each of them. They are each minimal because they are each functions of a minimal sufficient statistic $\mathbf{t}^*$. These examples illustrate that not only are minimal sufficient statistics not unique, they can even have redundancy in them.

While the set of likelihoods in Example 3 form a sufficient statistic, they are not (quite) minimal. However, they are close to minimal, as we now describe.

Example 6. For a model $p_{\mathbf{y}}(\cdot; x)$ where the parameter x is from a finite alphabet $\mathcal{X} = \{x_0, x_1, \ldots, x_{L-1}\}$, the vector of likelihood ratios

$$\mathbf{t}(\mathbf{y}) = \begin{bmatrix} t_0 \\ t_1 \\ \vdots \\ t_{L-1} \end{bmatrix}, \quad \text{where} \quad t_l(\mathbf{y}) = \frac{p_{\mathbf{y}}(\mathbf{y}; x_l)}{p_{\mathbf{y}}(\mathbf{y}; x_0)}, \tag{16}$$

is a minimal sufficient statistic. Note that this implies, as a special (binary) case, for $\mathcal{X} = \{H_0, H_1\}$, that the familiar likelihood ratio

$$L(\mathbf{y}) = \frac{p_{\mathbf{y}}(\mathbf{y}; H_1)}{p_{\mathbf{y}}(\mathbf{y}; H_0)}$$

is a minimal sufficient statistic. We leave the proofs of sufficiency and minimality as an exercise.

Example 7. Suppose our model is $p_y(y;x) = xe^{-xy}$ where $\mathcal{Y} = \mathbb{R}^+$, i.e., $y > 0$. Now let $t = y$ be a statistic, which is evidently sufficient. To see that y is minimal, we note

$$\mathbb{E}\left[\phi(\mathsf{y})\right] = \int \phi(y)\, xe^{-xy}\, \mathrm{d}y = x \int \phi(y)\, e^{-xy}\, \mathrm{d}y \triangleq \Phi(x) \tag{17}$$

where we note that $\Phi(\cdot)$ is the Laplace transform of $\phi(\cdot)$. Since $\Phi(\cdot) \equiv 0$ if and only if $\phi(\cdot) \equiv 0$ (almost everywhere), it follows that y is complete. Hence y is a minimal sufficient statistic.

It can be shown that under reasonable conditions, the natural statistic $\mathbf{t}(\cdot)$ of an exponential family $\mathtt{E}(\mathbf{x}; \boldsymbol{\lambda}(\cdot), \mathbf{t}(\cdot), \beta(\cdot))$ is minimal, as one might hope. We leave this as an exercise. We further note that at first glance it might seem that the log base distribution $\beta(\cdot)$ is also required, this is not the case. In particular, with respect to calculating likelihoods, upon which inference is based, the based distribution is an irrelevant constant of proportionality.

10.3 The Bayesian Case

If we choose to model $\mathbf{x}$ as a random variable, the conditional distribution $p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})$ replaces $p_{\mathbf{y}}(\mathbf{y};\mathbf{x})$. Definition 1 can be naturally modified for the Bayesian case as follows.

Definition 4. *A statistic* $\mathbf{t}(\cdot)$ *is* sufficient *with respect to distribution* $p_{\mathbf{x},\mathbf{y}}(\cdot,\cdot)$ *if*

$$p_{\mathbf{y}|\mathbf{t},\mathbf{x}}(\mathbf{y}|\mathbf{t}(\mathbf{y}),\mathbf{x}) = p_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t}(\mathbf{y})). \tag{18}$$

for all $\mathbf{x} \in \mathcal{X}$ *and* $\mathbf{y} \in \mathcal{Y}$.

Note that in particular, like Definition 1, Definition 4 states that when $\mathbf{t}$ is sufficient, the distribution $p_{\mathbf{y}|\mathbf{t},\mathbf{x}}(\mathbf{y}|\mathbf{t}(\mathbf{y}),\mathbf{x})$ is not a function of $\mathbf{x}$.

A characterization of sufficiency in the Bayesian case that relates directly to our goal of inference arises from examining the relationship between beliefs based on $\mathbf{y}$ and $\mathbf{t}$, respectively. In particular, we have the following.

Theorem 4 (Belief Characterization)**.** *A statistic* $\mathbf{t}(\cdot)$ *is sufficient with respect to distribution* $p_{\mathbf{x},\mathbf{y}}(\cdot,\cdot)$ *if and only if*

$$p_{\mathbf{x}|\mathbf{y}}(\cdot|\mathbf{y}) = p_{\mathbf{x}|\mathbf{t}}(\cdot|\mathbf{t}(\mathbf{y})). \tag{19}$$

Proof. We first assume that $p_{\mathbf{x}|\mathbf{y}}(\cdot|\mathbf{y}) = p_{\mathbf{x}|\mathbf{t}}(\cdot|\mathbf{t}(\mathbf{y}))$. Then for $\mathbf{t} = \mathbf{t}(\mathbf{y})$,

$$p_{\mathbf{y}|\mathbf{t},\mathbf{x}}(\mathbf{y}|\mathbf{t},\mathbf{x}) = \frac{p_{\mathbf{y},\mathbf{t},\mathbf{x}}(\mathbf{y},\mathbf{t},\mathbf{x})}{p_{\mathbf{t},\mathbf{x}}(\mathbf{t},\mathbf{x})} = \frac{p_{\mathbf{t}}(\mathbf{t})p_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t})p_{\mathbf{x}|\mathbf{t},\mathbf{y}}(\mathbf{x}|\mathbf{t},\mathbf{y})}{p_{\mathbf{t}}(\mathbf{t})p_{\mathbf{x}|\mathbf{t}}(\mathbf{x}|\mathbf{t})} \tag{20}$$

$$= \frac{p_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t})p_{\mathbf{x}|\mathbf{t},\mathbf{y}}(\mathbf{x}|\mathbf{t},\mathbf{y})}{p_{\mathbf{x}|\mathbf{t}}(\mathbf{x}|\mathbf{t})} \tag{21}$$

$$= \frac{p_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t})p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})}{p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})} = p_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t}), \tag{22}$$

where we obtained (22) by recognizing that for a deterministic function $\mathbf{t}(\cdot)$ we have $p_{\mathbf{x}|\mathbf{t},\mathbf{y}}(\mathbf{x}|\mathbf{t}(\mathbf{y}),\mathbf{y}) = p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})$. Thus, $\mathbf{t}(\cdot)$ satisfies the definition of a sufficient statistic.

Now we assume that $\mathbf{t}(\cdot)$ is a sufficient statistic, i.e., $p_{\mathbf{y}|\mathbf{t},\mathbf{x}}(\mathbf{y}|\mathbf{t}(\mathbf{y}),\mathbf{x}) = p_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t}(\mathbf{y}))$ for all $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$. Then for $\mathbf{t} = \mathbf{t}(\mathbf{y})$,

$$p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) = p_{\mathbf{x}|\mathbf{t},\mathbf{y}}(\mathbf{x}|\mathbf{t},\mathbf{y}) = \frac{p_{\mathbf{x},\mathbf{t},\mathbf{y}}(\mathbf{x},\mathbf{t},\mathbf{y})}{p_{\mathbf{t},\mathbf{y}}(\mathbf{t},\mathbf{y})} \tag{23}$$

$$= \frac{p_{\mathbf{t}}(\mathbf{t})p_{\mathbf{x}|\mathbf{t}}(\mathbf{x}|\mathbf{t})p_{\mathbf{y}|\mathbf{t},\mathbf{x}}(\mathbf{y}|\mathbf{t},\mathbf{x})}{p_{\mathbf{t}}(\mathbf{t})p_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t})} = \frac{p_{\mathbf{x}|\mathbf{t}}(\mathbf{x}|\mathbf{t})p_{\mathbf{y}|\mathbf{t},\mathbf{x}}(\mathbf{y}|\mathbf{t},\mathbf{x})}{p_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t})} \tag{24}$$

$$= \frac{p_{\mathbf{x}|\mathbf{t}}(\mathbf{x}|\mathbf{t})p_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t})}{p_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t})} = p_{\mathbf{x}|\mathbf{t}}(\mathbf{x}|\mathbf{t}), \tag{25}$$

which completes the proof. □

An alternative characterization of sufficiency is the following Bayesian version of the Neyman Factorization theorem (Theorem 2).

Theorem 5. *A statistic $\mathbf{t} = \mathbf{t}(\cdot)$ is sufficient with respect to distribution $p_{\mathbf{x},\mathbf{y}}(\cdot,\cdot)$ if and only if*

$$p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) = p_{\mathbf{t}|\mathbf{x}}(\mathbf{t}(\mathbf{y})|\mathbf{x})\, p_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t}(\mathbf{y})) \tag{26}$$

for all $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$.

To relate to the Neyman Factorization, note that the term $p_{\mathbf{t}|\mathbf{x}}(\mathbf{t}(\mathbf{y})|\mathbf{x})$ is a function of $\mathbf{t}$ and $\mathbf{x}$, while the term $p_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t}(\mathbf{y}))$ is a function only of $\mathbf{y}$.

Proof. Letting $\mathbf{t} = \mathbf{t}(\mathbf{y})$ for convenience, it suffices to write

$$p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) = p_{\mathbf{t},\mathbf{y}|\mathbf{x}}(\mathbf{t},\mathbf{y}|\mathbf{x}) = p_{\mathbf{t}|\mathbf{x}}(\mathbf{t}(\mathbf{y})|\mathbf{x})\, p_{\mathbf{y}|\mathbf{t},\mathbf{x}}(\mathbf{y}|\mathbf{t},\mathbf{x}) \tag{27}$$

and note that by Definition 4, the last factor on the right-hand side of (27) is equal to $p_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t})$ if and only if $\mathbf{t}(\cdot)$ is a sufficient statistic. □

10.3.1 Markov Chains

Markov chains provide a useful way to view the relationship between the hidden parameter, the data and the statistic. Although we discussed Markov chains a little in the context of randomized decision rules, we revisit and elaborate on the concept now in this more general scenario.

Definition 5. *Random variables* $\mathbf{a}$*,* $\mathbf{b}$*, and* $\mathbf{c}$ *form a* Markov chain *if*

$$p_{\mathbf{c}|\mathbf{b},\mathbf{a}}(\mathbf{c}|\mathbf{b},\mathbf{a}) = p_{\mathbf{c}|\mathbf{b}}(\mathbf{c}|\mathbf{b}). \tag{28}$$

This definition indicates that $\mathbf{c}$ doesn't depend on $\mathbf{a}$ when $\mathbf{b}$ is given, which suggests the following notation $\mathbf{a} \to \mathbf{b} \to \mathbf{c}$ as a generating mechanism for $\mathbf{c}$.

Such conditional independence is also an equivalent characterization of Markov structure.

Claim 1. *Random variables* $\mathbf{a}$*,* $\mathbf{b}$*, and* $\mathbf{c}$ *form a Markov chain if and only if* $\mathbf{a}$ *and* $\mathbf{c}$ *are conditionally independent given* $\mathbf{b}$*, i.e,*

$$p_{\mathbf{c},\mathbf{a}|\mathbf{b}}(\mathbf{c},\mathbf{a}|\mathbf{b}) = p_{\mathbf{c}|\mathbf{b}}(\mathbf{c}|\mathbf{b})\, p_{\mathbf{a}|\mathbf{b}}(\mathbf{a}|\mathbf{b}). \tag{29}$$

Proof. It suffices to note that

$$\begin{aligned} p_{\mathbf{c},\mathbf{a}|\mathbf{b}}(\mathbf{c},\mathbf{a}|\mathbf{b}) &= p_{\mathbf{c}|\mathbf{a},\mathbf{b}}(\mathbf{c}|\mathbf{a},\mathbf{b})\, p_{\mathbf{a}|\mathbf{b}}(\mathbf{a}|\mathbf{b}) && (30)\\ &= p_{\mathbf{c}|\mathbf{b}}(\mathbf{c}|\mathbf{b})\, p_{\mathbf{a}|\mathbf{b}}(\mathbf{a}|\mathbf{b}), && (31) \end{aligned}$$

where to obtain (31) we have used (28). □

Now Claim 1 also makes clear that Markovianity is direction-free. Indeed, the symmetry of (29) means that $\mathbf{c} \to \mathbf{b} \to \mathbf{a}$ form a Markov chain as well. For this reason, the notation

$$\mathbf{a} \leftrightarrow \mathbf{b} \leftrightarrow \mathbf{c} \tag{32}$$

is common, and what we will use. Note that we can explicitly see that the reversed chain is Markov via

$$p_{\mathbf{a}|\mathbf{b},\mathbf{c}}(\mathbf{a}|\mathbf{b},\mathbf{c}) = \frac{p_{\mathbf{c},\mathbf{a}|\mathbf{b}}(\mathbf{c},\mathbf{a}|\mathbf{b})}{p_{\mathbf{c}|\mathbf{b}}(\mathbf{c}|\mathbf{b})} = \frac{p_{\mathbf{c}|\mathbf{b}}(\mathbf{c}|\mathbf{b})\, p_{\mathbf{a}|\mathbf{b}}(\mathbf{a}|\mathbf{b})}{p_{\mathbf{c}|\mathbf{b}}(\mathbf{c}|\mathbf{b})} = p_{\mathbf{a}|\mathbf{b}}(\mathbf{a}|\mathbf{b}), \tag{33}$$

where to obtain the second to last inequality we have used (29).

More generally, Markovianity implies a particular factorization of the joint distribution. In particular, it follows immediately from the chain rule of probability and (28) that $\mathbf{a} \leftrightarrow \mathbf{b} \leftrightarrow \mathbf{c}$ form a Markov chain if and only if the joint probability distribution of these variables factors as

$$p_{\mathbf{a},\mathbf{b},\mathbf{c}}(\mathbf{a},\mathbf{b},\mathbf{c}) = p_{\mathbf{a}}(\mathbf{a})\, p_{\mathbf{b}|\mathbf{a}}(\mathbf{b}|\mathbf{a})\, p_{\mathbf{c}|\mathbf{b}}(\mathbf{c}|\mathbf{b}). \tag{34}$$

Finally, we note that when $\mathbf{c} = \mathbf{g}(\mathbf{b})$, where $\mathbf{g}(\cdot)$ is a deterministic function, then $\mathsf{a} \leftrightarrow \mathsf{b} \leftrightarrow \mathsf{c}$ form a (trivial) Markov chain. To see this, it suffices to note that

$$p_{\mathsf{c}|\mathsf{b},\mathsf{a}}(\mathbf{c}|\mathbf{b},\mathbf{a}) = \begin{cases} 1 & \mathbf{c} = \mathbf{g}(\mathbf{b}) \\ 0 & \text{otherwise} \end{cases} \tag{35}$$

does not depend on $\mathbf{a}$, and hence $p_{\mathsf{c}|\mathsf{b},\mathsf{a}}(\mathbf{c}|\mathbf{b},\mathbf{a}) = p_{\mathsf{c}|\mathsf{b}}(\mathbf{c}|\mathbf{b})$.

10.3.2 Statistics, Sufficiency, and Markov Chains

When choosing a statistic t, it is obviously important that it not depend directly on the latent variable x. Otherwise, we would not consider the statistic to be *valid*. Constraining the statistic to be a deterministic function of the data, i.e.,

$$\mathbf{t} = \mathbf{t}(\mathbf{y}) \tag{36}$$

for some $\mathbf{t}(\cdot)$ is an obvious way to accomplish this. More generally, however, the validity requirement is naturally expressed in the form of a Markov constraint.

Definition 6. *A statistic* t *is* valid *if the Markov chain* $\mathsf{x} \leftrightarrow \mathsf{y} \leftrightarrow \mathsf{t}$ *is satisfied.*

It is straightforward to verify that any sufficient statistic of the form (36) satisfies Definition 6. At the same time, it is also worth emphasizing that the additional randomness in the construction of t that Definition 6 allows in the construction of statistics generally isn't particularly useful for the purpose of inference. To some degree this isn't surprising. Indeed, in the case of Bayesian hypothesis testing we saw that randomized tests offered no advantage over deterministic tests. Nevertheless, the characterization is convenient.

Finally, now examining (18), we see that Definition 4 says that t is a sufficient statistic if it also satisfies the different Markov constraint $\mathsf{x} \leftrightarrow \mathsf{t} \leftrightarrow \mathsf{y}$. Hence, sufficient statistics satisfy a pair of Markov constraints that make t and y in some sense interchangeable with respect to inference about x.

10.4 Sufficient Statistics as Partitions

In this section, we develop the conceptually useful interpretation of a sufficient statistic as a *partition* of the space $\mathcal{Y}$ of observations. To begin, note that for a model $p_{\mathbf{y}}(\cdot; x)$ over $\mathcal{Y}$, when $L_{\mathbf{y}_1}(x) \propto L_{\mathbf{y}_2}(x)$ it follows that $\mathbf{y}_1$ and $\mathbf{y}_2$ provide exactly the same information about x from the perspective of inference. More generally, the class of all $\mathbf{y}$ such that $L_{\mathbf{y}}(x) \propto L_{\mathbf{y}_1}(x)$ provide the same information about x as $\mathbf{y}_1$, and thus for the purposes of inference it is sufficient to know the class; knowing the member of the class that was observed provides no additional information. This is the key idea underlying the concept of a sufficient statistic, and motivates the process of

partitioning the space $\mathcal{Y}$ of possible observations into classes, each of which has this property.

We now use this specific view to develop alternative characterizations of sufficient and minimaly sufficient statistics. We begin with the former.

Theorem 6 (Sufficient Statistic, Partition Characterization)**.** *A statistic* $\mathbf{t}$ *is sufficient if and only if for all* $\mathbf{y}_1$ *and* $\mathbf{y}_2$ *such that* $\mathbf{t}(\mathbf{y}_1) = \mathbf{t}(\mathbf{y}_2)$ *we have* $L_{\mathbf{y}_1}(x) \propto L_{\mathbf{y}_2}(x)$*, i.e., there exists* $g(\cdot, \cdot)$ *such that* $L_{\mathbf{y}_1}(x) = g(\mathbf{y}_1, \mathbf{y}_2)\, L_{\mathbf{y}_2}(x)$*.*

Proof. For the "only if" part, suppose $\mathbf{t}$ is sufficient. Then by the Neymann factorization we have

$$\frac{L_{\mathbf{y}_1}(x)}{L_{\mathbf{y}_2}(x)} = \frac{p_{\mathbf{y}}(\mathbf{y}_1; x)}{p_{\mathbf{y}}(\mathbf{y}_2; x)} = \frac{a(\mathbf{t}(\mathbf{y}_1); x)\, b(\mathbf{y}_1)}{a(\mathbf{t}(\mathbf{y}_2); x)\, b(\mathbf{y}_2)} = \frac{b(\mathbf{y}_1)}{b(\mathbf{y}_2)} \triangleq g(\mathbf{y}_1, \mathbf{y}_2). \tag{37}$$

For the "if" part, note that given any $\mathbf{t}_0 \in \mathcal{T}$, for all $\mathbf{y}$ such that $\mathbf{t}(\mathbf{y}) = \mathbf{t}_0$ we have[3]

$$\begin{aligned} p_{\mathbf{y}|\mathbf{t}}(\mathbf{y}|\mathbf{t}_0; x) &= \frac{p_{\mathbf{y}}(\mathbf{y}; x)}{\sum_{\{\mathbf{y}' \colon \mathbf{t}(\mathbf{y}') = \mathbf{t}_0\}} p_{\mathbf{y}}(\mathbf{y}'; x)} \\ &= \frac{1}{\sum_{\{\mathbf{y}' \colon \mathbf{t}(\mathbf{y}') = \mathbf{t}_0\}} p_{\mathbf{y}}(\mathbf{y}'; x)/p_{\mathbf{y}}(\mathbf{y}; x)} \\ &= \frac{1}{\sum_{\{\mathbf{y}' \colon \mathbf{t}(\mathbf{y}') = \mathbf{t}_0\}} L_{\mathbf{y}'}(x)/L_{\mathbf{y}}(x)} \\ &= \frac{1}{\sum_{\{\mathbf{y}' \colon \mathbf{t}(\mathbf{y}') = \mathbf{t}_0\}} g(\mathbf{y}', \mathbf{y})}, \end{aligned}$$

which we note does not depend on x. □

Theorem 7 (Minimal Sufficient Statistic, Partition Characterization)**.** *A sufficient statistic* $\mathbf{t}$ *is minimal if and only if for all* $\mathbf{y}_1$ *and* $\mathbf{y}_2$ *such that* $L_{\mathbf{y}_1}(x) \propto L_{\mathbf{y}_2}(x)$ *we have* $\mathbf{t}(\mathbf{y}_1) = \mathbf{t}(\mathbf{y}_2)$*.*

Proof. For the "only if" part, consider $\mathbf{y}_1$ and $\mathbf{y}_2$ such that $L_{\mathbf{y}_1}(x) = g(\mathbf{y}_1, \mathbf{y}_2)\, L_{\mathbf{y}_2}(x)$, and consider the statistic

$$\mathbf{s}(\mathbf{y}) = \begin{cases} \mathbf{y}_2 & \mathbf{y} = \mathbf{y}_1 \\ \mathbf{y} & \mathbf{y} \neq \mathbf{y}_1, \end{cases}$$

for which $\mathbf{s}(\mathbf{y}_1) = \mathbf{s}(\mathbf{y}_2)$. Then

$$p_{\mathbf{y}}(\mathbf{y}; x) = \begin{cases} p_{\mathbf{y}}(\mathbf{y}_2; x)\, g(\mathbf{y}, \mathbf{y}_2) & \mathbf{y} = \mathbf{y}_1 \\ p_{\mathbf{y}}(\mathbf{y}; x) & \mathbf{y} \neq \mathbf{y}_1, \end{cases}$$

[3]For continuous-valued variables $\mathbf{y}$, the summations are replaced with corresponding integrals.

so

$$p_{\mathbf{y}}(\mathbf{y}; x) = a(\mathbf{s}, x)\, b(\mathbf{y})$$

with

$$a(\mathbf{s}, x) = p_{\mathbf{y}}(\mathbf{s}, x) \qquad \text{and} \qquad b(\mathbf{y}) = \begin{cases} g(\mathbf{y}, \mathbf{y}_2) & \mathbf{y} = \mathbf{y}_1 \\ 1 & \mathbf{y} \neq \mathbf{y}_1. \end{cases}$$

Hence, by the Neymann factorization, $\mathbf{s}$ is sufficient.

Next, since $\mathbf{t}$ is a minimal sufficient statistic, there must be a function $\mathbf{f}(\cdot)$ such that $\mathbf{t} = \mathbf{f}(\mathbf{s})$. But then

$$\mathbf{t}(\mathbf{y}_1) = \mathbf{f}(\mathbf{s}(\mathbf{y}_1)) = \mathbf{f}(\mathbf{s}(\mathbf{y}_2)) = \mathbf{t}(\mathbf{y}_2),$$

as claimed.

For the "if" part, let

$$\mathcal{Y}_{\mathbf{t}'} \triangleq \{\mathbf{y} \colon \mathbf{t}(\mathbf{y}) = \mathbf{t}'\}, \quad \mathbf{t}' \in \mathcal{T},$$

and note that each $\mathbf{y} \in \mathcal{Y}$ is in exactly one of these subsets, i.e., the subsets form a partition of $\mathcal{Y}$.

Next, for an arbitrary sufficient statistic $\mathbf{s}(\cdot)$ let

$$\mathcal{S}_{\mathbf{t}'} = \mathbf{s}(\mathcal{Y}_{\mathbf{t}'}) \triangleq \{\mathbf{s}' \in \mathcal{S} \colon \mathbf{s}(\mathbf{y}) = \mathbf{s}' \text{ for some } \mathbf{y} \in \mathcal{Y}_{\mathbf{t}'}\}, \quad \mathbf{t}' \in \mathcal{T}$$

Then each $\mathbf{s} \in \mathcal{S}$ is in exactly one of these subsets, i.e., the subsets form a partition of $\mathcal{S}$. To verify this using a contradiction argument, suppose there exists $\mathbf{t}'$ and $\mathbf{t}''$ such that $\mathcal{S}_{\mathbf{t}'} \cap \mathcal{S}_{\mathbf{t}''} \neq \varnothing$. Then there exists $\mathbf{y}_1 \in \mathcal{Y}_{\mathbf{t}'}$ and $\mathbf{y}_2 \in \mathcal{Y}_{\mathbf{t}''}$ such that $\mathbf{s}(\mathbf{y}_1) = \mathbf{s}(\mathbf{y}_2)$. In turn, since $\mathbf{s}$ is sufficient this means that $L_{\mathbf{y}_1}(x) \propto L_{\mathbf{y}_2}(x)$. But then by the assumption in the hypothesis $\mathbf{t}(\mathbf{y}_1) = \mathbf{t}(\mathbf{y}_2)$, i.e., $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}_{\mathbf{t}'} \cap \mathcal{Y}_{\mathbf{t}''} = \varnothing$, which is a contradiction.

Finally, define $\mathbf{g} \colon \mathcal{S} \mapsto \mathcal{T}$ according to

$$\mathbf{g}(\mathbf{s}') = \mathbf{t}' \quad \text{for the unique } \mathbf{t}' \text{ such that } \mathbf{s}' \in \mathcal{S}_{\mathbf{t}'},$$

for $\mathbf{s}' \in \mathcal{S}$. Hence, for any $\mathbf{y}' \in \mathcal{Y}$ we have

$$\mathbf{t}(\mathbf{y}') = \mathbf{g}(\mathbf{s}(\mathbf{y}')),$$

and thus $\mathbf{t}$ is minimal. □

Summarizing, Theorems 6 and 7 formally express the following characterizations. First, a sufficient statistic corresponds to a partition of $\mathcal{Y}$ such that the likelihood functions of any pair of observation values within each subset are proportional. Second, a minimal sufficient statistic corresponds to the coarsest possible partition among those corresponding to sufficient statistics.

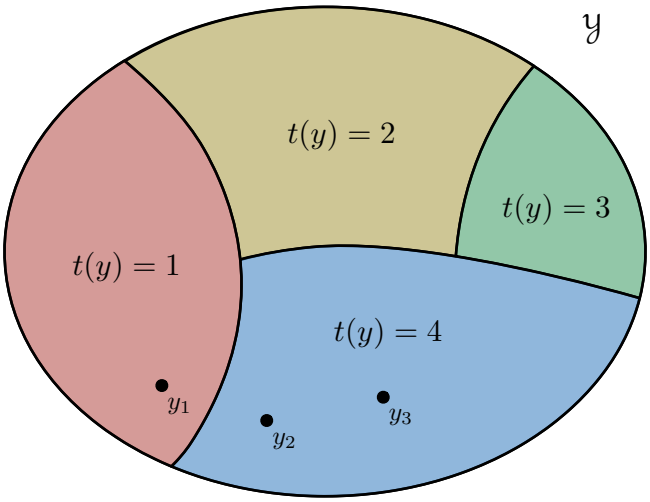

Figure 1: The maximally-coarse partition of the observation space $\mathcal{Y}$ corresponding to a minimal sufficient statistic $t(\cdot)$. Each color corresponds to a different value of t in the alphabet $\mathcal{T} = \{1, \ldots, 4\}$.

Moreover, from the perspective of our development note that invertible transformations of s and t correspond to simply (invertibly) relabeling the elements of $\mathcal{S}$ and $\mathcal{T}$, respectively. Clearly, such relabelings do not change the partition. As such, it is the partition induced by the statistic that governs whether it is sufficient, minimal sufficient, or neither.

Figs. 1, 2, and 3 illustrate these relationships for a given model family $p_{\mathsf{y}}(\cdot; x)$. In Fig. 1, since y_2 and y_3 are in a region of the same color (meaning $t(y_2) = t(y_3)$), we have that $L_{y_2}(x) \propto L_{y_3}(x)$ because t is sufficient. On the other hand, since y_1 and y_2 are in regions that are differently colored (meaning $t(y_1) \neq t(y_2)$), we have that $L_{y_2}(x) \not\propto L_{y_3}(x)$ because t is minimal. In Fig. 2, since the partition is finer, s is not minimal. As a result, for example, although y_2 and y_3 are in regions that are differently colored (meaning $s(y_2) \neq s(y_3)$), we have that $L_{y_2}(x) \propto L_{y_3}(x)$. In Fig. 3, the partition for the statistic r is such that it does not refine that of the minimal sufficient statistic t, so r is not sufficient. In particular, $r(y_1) = r(y_2)$ but $L_{y_1}(x) \not\propto L_{y_2}(x)$.

10.5 Summary

To summarize the developments in this section, the crucial implication of sufficiency is that for problems of inference using this probabilistic model, we can first compute the value of statistic as an initial data processing step, and then use the resulting values in inference computations without any loss of relevant information. Minimality further

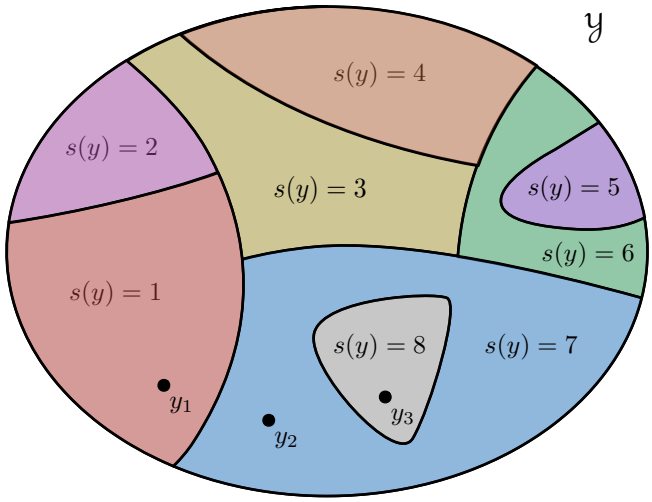

Figure 2: The finer partition of the observation space $\mathcal{Y}$ corresponding to a sufficient statistic $s(\cdot)$ that is not minimal. Each color corresponds to a different value of s in the alphabet $\mathcal{S} = \{1, \ldots, 8\}$.

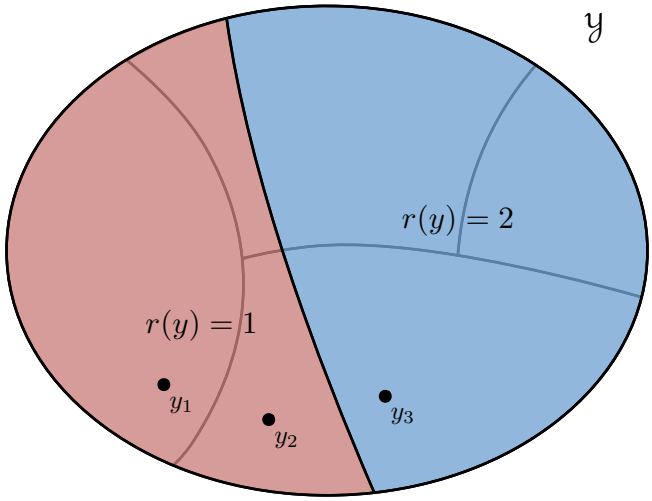

Figure 3: A partition of the observation space $\mathcal{Y}$ corresponding to a statistic $r(\cdot)$ that is not sufficient. Each color corresponds to a different value of r in the alphabet $\mathcal{R} = \{1, 2\}$.

guarantees that the amount of information we retain is as small as the original model allows.

Finding good sufficient statistics is often a central research challenge in many application areas, and is the key to efficient inference.

10.6 Further reading

There are many texts with detailed treatments of sufficient statistics. See, for example:

J. M. Bernardo and A. F. M. Smith, *Bayesian Theory,* Wiley, 1994.

E. L. Lehmann and G. Casella, *Theory of Point Estimation,* Springer, 1998.

Y. Pawitan, *In All Likelihood: Statistical Modelling and Inference Using Likelihood,* Oxford University Press, 2001.