1. Minimal sufficiency for correlated normals

Suppose that \((X_i, Y_i), i = 1, \ldots, n\) are sampled i.i.d. from the bivariate normal distribution
\[
\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix} \right).
\]
with \(\theta \in \Theta = (-1, 1)\). (Note: this is an example of a curved exponential family)

(a) Find a two-dimensional minimal sufficient statistic and show it is minimal.
(b) Prove that the minimal sufficient statistic found in (a) is not complete.
(c) Prove that \(Z_1 = \sum_{i=1}^{n} X_i^2\) and \(Z_2 = \sum_{i=1}^{n} Y_i^2\) are both ancillary, but that \((Z_1, Z_2)\) is not ancillary.

**Moral of the story:** To check ancillarity of a random vector, it is not enough to just check that the individual coordinates of the vector are (marginally) ancillary; we need to consider the joint distribution of the vector.

2. Bayesian interpretation of sufficiency

Assume we have a family \(\mathcal{P}\) of densities \(p_\theta(x)\) with respect to a common measure \(\mu\) on \(X\), for \(\theta \in \Theta \subseteq \mathbb{R}^n\). Additionally, assume the parameter \(\theta\) is itself random, following prior density \(q(\theta)\) with respect to the Lebesgue measure on \(\Theta\).

Then, we can write the posterior density (distribution of \(\theta\) given \(X = x\)) as
\[
q_{\text{post}}(\theta \mid x) = \frac{p_\theta(x)q(\theta)}{\int_\Theta p_\zeta(x)q(\zeta)\,d\zeta}.
\]
(Note: this is a slightly careless manipulation of the densities, but one that generally works. Feel free to make similar non-rigorous manipulations yourself in the problem).

In this setting, prove the following claims:

(a) Suppose a statistic \(T(X)\) has the property that, for any prior distribution \(q(\theta)\), the posterior distribution \(q_{\text{post}}(\theta \mid x)\) depends on \(x\) only through \(T(x)\). Show that \(T(X)\) is sufficient for \(\mathcal{P}\).
(b) Conversely, show that, if \(T(X)\) is sufficient for \(\mathcal{P}\) then, for any prior \(q\), the posterior depends on \(x\) only through \(T(x)\).

**Moral:** If we have a prior opinion about \(\theta\) in the form of a distribution, and then we rationally update our opinion after observing \(X\), then we will naturally adhere to the sufficiency principle. This gives an alternative epistemological motivation for the principle.
3. Ancillarity in location-scale families

In a parameterized family where $\theta = (\zeta, \lambda)$, we say a statistic $T$ is ancillary for $\zeta$ if its distribution is independent of $\zeta$; that is, if $T(X)$ is ancillary in the subfamily where $\lambda$ is known, for each possible value of $\lambda$.

Suppose that $X_1, \ldots, X_n \in X = \mathbb{R}$ are an i.i.d. sample from a location-scale family $\mathcal{P} = \{F_{a,b}(x) = F((x-a)/b) : a \in \mathbb{R}, b > 0\}$, where $F(\cdot)$ is a known cumulative distribution function. The real numbers $a$ and $b$ are called the location and scale parameters respectively. (Note: recall it is not enough to prove ancillarity of the coordinates.)

(a) Show that the vector of differences $(X_1 - X_i)_{i=2}^n$ is ancillary for $a$.

(b) Show that the vector of ratios $(X_i - a)_{i=2}^n$ is ancillary for $b$. (Note: this is only a statistic when $a$ is known).

(c) Show that the vector of difference ratios $(X_i - X_j)_{i=3}^n$ is ancillary for $(a, b)$.

(d) Let $X_1, \ldots, X_n$ be mutually independent with $X_i \sim \Gamma(k_i, \theta)$. Show that $X_+ = \sum_{i=1}^n X_i$ is independent of $(X_1, \ldots, X_n)/X_+$.

4. Interpretation of completeness

The concept of completeness for a family of measures was introduced in Lehmann and Scheffé (1950) as a precursor to their definition, in the same paper, of a complete statistic. The definition of a complete family did not stick, and lives on only in the (consequently confusingly named) idea of complete statistic (in particular it has nothing to do with the definition of a complete measure that you can find on Wikipedia).

If $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a family of measures on $X$, we say that $\mathcal{P}$ is complete if

$$\int f(x) dP_\theta(x) = 0, \forall \theta \Rightarrow P_\theta(\{x : f(x) \neq 0\}) = 0, \forall \theta.$$ 

That is, the family is not complete if there is some nonzero function $f$ which is “orthogonal” to every $P_\theta$ in some sense. We will try to gain some intuition for this definition and, thereby, for the definition of a complete statistic.

For the following parts, let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a family of probability measures on $X$, assume $T(X)$ is a statistic, and let $\mathcal{T} = T(X)$ be the range of the statistic $T(X)$. Let $\mathcal{P}^T = \{P_\theta^T : \theta \in \Theta\}$ denote the induced model of push-forward probability measures on $T$ denoting the possible distributions of $T(X)$:

$$P_\theta^T(B) = P_\theta(T^{-1}(B)) = \mathbb{P}_\theta(T(X) \in B).$$

(a) Show that $T(X)$ is a complete statistic for the family $\mathcal{P}$ if and only if $\mathcal{P}^T$ is a complete family.

(b) Assume (for this part only) that $X$ is a finite set, i.e. $X = \{x_1, \ldots, x_n\}$ for some $n < \infty$, and assume without loss of generality that every $x \in X$ has $P_\theta(\{x\}) > 0$ for at least one value of $\theta$ (otherwise we could truncate the sample space).

Let $p_\theta(x) = P_\theta(X = x) \geq 0$, and $v^\theta = (p_\theta(x_1), \ldots, p_\theta(x_n)) \in \mathbb{R}^n$. Show that $\mathcal{P}$ is complete if and only if $\text{Span}\{v^\theta : \theta \in \Theta\} = \mathbb{R}^n$.

(c) Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Pois}(\theta)$ for $\theta \in \Theta = \{\theta_1, \ldots, \theta_m\}$ with $2 \leq m < \infty$. Find a sufficient statistic that is minimal but not complete (prove both properties).

(d) In the same scenario but with $\Theta = \pi \mathbb{Z}_+ = \{0, \pi, 2\pi, \ldots\}$, show that the same statistic is minimal but not complete.
(e) **Optional** (not graded, no extra points). Let \(X_1, \ldots, X_n \sim \text{Pois}(\theta)\) for \(\theta \in \Theta\), and assume that \(\Theta\) has an accumulation point at 0, i.e., \(\Theta\) includes an infinite sequence of positive values \(\theta_1, \theta_2, \ldots \in \Theta\) such that \(\lim_{m \to \infty} \theta_m = 0\). Find a complete sufficient statistic and prove it is complete sufficient.

**Hint:** suppose \(f\) is a counterexample function; what is \(f(0)\)? It may be helpful to recall that \(\int f \, d\mu\) is undefined unless either \(\int \max(0, f(x)) \, d\mu(x)\) or \(\int \max(0, -f(x)) \, d\mu(x)\) is finite; as a result \(\int f \, d\mu = 0 \Rightarrow \int |f| \, d\mu < \infty\).

5. **Uniform location family**

Let \(X_1, \ldots, X_n \sim \text{Unif}[\theta - \frac{1}{2}, \theta + \frac{1}{2}]\), with \(\theta \in \mathbb{R}\) unknown.

(a) Show that \(T(X) = (X_{(1)}, X_{(n)})\) is minimal sufficient but not complete.

(b) Suppose that we wish to estimate \(\theta\) under the squared error loss \(L(\theta, d) = (\theta - d)^2\).

The sample mean \(\bar{X}\) may appear to be a reasonable estimator of \(\theta\), but we might worry about the fact that it is not a function of \(T(X)\).

Guided by the sufficiency principle, we could instead consider the estimator \(\delta(X) = (X_{(1)} + X_{(n)})/2\). Compute the MSE of each estimator as a function of \(n\) and \(\theta\). What happens to the ratio \(\text{MSE}(\theta, \bar{X})/\text{MSE}(\theta, \delta)\) as \(n \to \infty\)?

**Moral:** When we understand the statistical structure of a given model, we may be able to do much better than we could with naive methods.

**References**