Partial sufficiency and density estimation
April 17, 2009

J. E. Chacón, Departamento de Matemáticas, Universidad de Extremadura.
J. Montanero, Departamento de Matemáticas, Universidad de Extremadura.
A. G. Nogales, Departamento de Matemáticas, Universidad de Extremadura.
P. Pérez, Departamento de Matemáticas, Universidad de Extremadura.

Abstract. In this paper we study the usefulness of the concept of partial sufficiency, in the sense of Fraser, when it is applied to the problem of nonparametric density estimation.

Keywords: nonparametric density estimation, sufficiency, Fraser’s partial sufficiency

2000 AMS Subject Classifications: Primary 62G07, Secondary 62B05

Running head: Partial sufficiency and density estimation
1 Introduction

Several works by Wertz explore the relationships between the density estimation problem and the concepts of sufficiency (Wertz, 1974), unbiasedness (Wertz, 1975), invariance (Wertz, 1976) and quasi-invariance (Wertz, 1992). Here we will extend these studies to cover the relationship with the concept of partial sufficiency, in the sense of Fraser (1956).

Sufficiency is usually recognized as the main contribution of Fisher to the parametric statistical theory. Nevertheless, the appealing intuitive concept of a partially sufficient statistic (a statistic keeping all the relevant information about a subparameter) is more elusive, as Fisher himself pointed out. Since then, the problem of fixing a natural mathematical definition of partial sufficiency in a classical setting has been considered in a good deal of famous papers as Neyman and Pearson (1936), Kolmogorov (1942), Fraser (1956) or Hájek (1965), but a fully satisfactory solution is not available yet.

A good survey paper containing several results on the usefulness of partial sufficiency in point estimation and testing statistical hypotheses, as well as some illustrative examples, is Basu (1977). More recently, but still in the parametric context, Montanero et al. (2005) provide a characterization of the probability families admitting a partially sufficient $\sigma$-field, together with a factorization theorem for the concept partial sufficiency, thus establishing a formal theory from a heuristic approximation given in Basu (1977).

In this paper, we will translate this background to the context of the density estimation problem. In Section 2 we will review all the concepts involved and Wertz’s theorem showing the usefulness of a sufficient $\sigma$-field in for density estimation. Also, we will state and prove our main result, giving an analogous theorem for the concept of partial sufficiency, in the sense of Fraser (1956). Finally, in Section 3 we include three examples illustrating the use of partially sufficient statistics for three specific nonparametric problems.
2 Notation, definitions and main result

First, we will recall briefly some widely used definitions and Wertz’s theorem for sufficiency and density estimation. Once we have settled down this background, we will establish the corresponding theorem for partial sufficiency in the sense of Fraser (1956).

2.1 Sufficiency and density estimation

Let \((\Omega, \mathcal{A}, \mathcal{P})\) be a statistical experiment and \(\mathcal{B}\) a sub-\(\sigma\)-field of \(\mathcal{A}\). For every \(P \in \mathcal{P}\) we will denote \(P(A|\mathcal{B})\) the conditional probability of \(A\) given \(\mathcal{B}\), that is, the set of all the real \(\mathcal{B}\)-measurable functions \(g \equiv g_{P,A}\) such that

\[
P(A \cap B) = \int_B gdP, \quad \forall B \in \mathcal{B}.
\]

(1)

With this notation, \(\mathcal{B}\) is said to be sufficient (for \(P\)) if \(\cap_{P \in \mathcal{P}} P(A|\mathcal{B}) \neq \emptyset\) for every \(A \in \mathcal{A}\). This means that for every \(A \in \mathcal{A}\) there exists a real \(\mathcal{B}\)-measurable function \(g \equiv g_A\) (not depending on \(P\)) such that (1) holds for every \(P \in \mathcal{P}\). As usual, we will also write \(P(A|\mathcal{B})\) for any of the versions of this equivalence class.

Now, suppose that there is a \(\sigma\)-finite measure \(\mu\) dominating \(\mathcal{P}\) in the sense that, for every \(A \in \mathcal{A}\), \(\mu(A) = 0\) implies \(P(A) = 0\) for all \(P \in \mathcal{P}\). Then, we can represent the family of probability measures as \(\mathcal{P} = \{P_f : f \in \Delta\}\), where \(\Delta\) is the set of densities of the probability measures in \(\mathcal{P}\) with respect to \(\mu\).

A density estimator is an \((\mathcal{A} \times \mathcal{A}^n)\)-measurable function \(f_n : \Omega \times \Omega^n \to \mathbb{R}\). Wertz (1974) contains a Rao-Blackwell type theorem showing, in particular, that if we have a sufficient \(\sigma\)-field \(\mathcal{B} \subset \mathcal{A}^n\), then it is enough to consider \((\mathcal{A} \times \mathcal{B})\)-mesurable density estimators \(f_n^* : \Omega \times \Omega^n \to \mathbb{R}\). Let us make this statement precise:

Let \(\lambda\) be another \(\sigma\)-finite measure over \((\Omega, \mathcal{A})\). Given \(p \geq 1\) let us consider the seminorm assigning

\[
\|f_n\|_p := \left\{ \sup_{f \in \Delta} \mathbb{E}_f \int |f_n(\omega, \bar{\omega})|^pd\lambda(\omega) \right\}^{1/p} = \left\{ \sup_{f \in \Delta} \int \int |f_n(\omega, \bar{\omega})|^pd\lambda(\omega)dP^n_f(\bar{\omega}) \right\}^{1/p}
\]
to every \((A \times A^n)\)-measurable density estimator \(f_n\) and let us denote by \(\Lambda_p(A \times A^n)\) the class of all \((A \times A^n)\)-measurable density estimators \(f_n\) for which \(\|f_n\|_p < \infty\). Besides, for such an \(f_n\) let us also denote its mean \(L_p\) error as

\[
ML_p E(f_n, f) := \left\{ \mathbb{E}_f \int |f_n(\omega, \bar{\omega}) - f(\omega)|^p d\lambda(\omega) \right\}^{1/p}.
\]

With respect to these error measures, Theorem 3 in Wertz (1974) states the following:

**Theorem 1** (Wertz, 1974). Suppose that \(\sup_{f \in \Delta} \int f^p d\lambda < \infty\) and let \(\mathcal{B}\) be a countably generated sub-\(\sigma\)-field of \(A^n\). If \(\mathcal{B}\) is sufficient for \(P^n\) then for every \(f_n \in \Lambda_p(A \times A^n)\) there exists \(f^*_n \in \Lambda_p(A \times \mathcal{B})\) such that \(ML_p E(f^*_n, f) \leq ML_p E(f_n, f), \forall f \in \Delta\). Moreover, \(f^*_n\) can be chosen in a way such that \(f^*_n(\omega, \cdot)\) be a version of the conditional expectation \(\mathbb{E}_f[f_n(\omega, \cdot)|\mathcal{B}]\), \(\lambda\)-a.e.; that is, in a way such that the set of all \(\omega \in \Omega\) for which

\[
\int_B f^*_n(\omega, \bar{\omega}) dP^n_\mathcal{B}(\bar{\omega}) = \int_B f_n(\omega, \bar{\omega}) dP^n_\mathcal{B}(\bar{\omega}), \quad \forall B \in \mathcal{B}, \forall f \in \Delta,
\]

does not hold, has null \(\lambda\)-measure.

**Remark 1.** The usual case \(\lambda = \mu\) corresponds to the density estimation problem using a global \((L_p)\) error measure; however, allowing \(\lambda\) to differ from \(\mu\) we also treat other related problems, as density estimation at a fixed point \(\omega_0 \in \Omega\) (setting \(\lambda\) to be the degenerated distribution at \(\omega_0\)) or the possibility of using weight functions in the error measure. If \(\lambda\) is absolutely continuous with respect to \(\mu\), then its density \(d\lambda/d\mu\) would be an adequate weight function. See Wertz (1978), p. 13-14.

**Example 1.** Consider for instance the problem of estimating a univariate density \(f\) symmetric about 0 using a sample \(X_1, \ldots, X_n\) drawn from its corresponding probability measure \(P_f\). Since the statistic \(Z = (|X_1|, \ldots, |X_n|)\) is sufficient for this family of densities, Wertz’s theorem assures that every estimator of the density \(f_n\) is improved by \(f^*_n = \mathbb{E}_f[f_n|Z]\) in the sense that \(\mathbb{E}_f \int |f^*_n - f| d\lambda \leq \mathbb{E}_f \int |f_n - f| d\lambda\). For example, for any kernel \(K\) and bandwidth \(h > 0\), the kernel estimator

\[
\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^n K \left( \frac{x - X_i}{h} \right)
\]
is improved by

\[ f_n^*(x) = \frac{1}{2nh} \sum_{i=1}^{n} \left[ K\left(\frac{x - |X_i|}{h}\right) + K\left(\frac{x + |X_i|}{h}\right) \right]. \]

### 2.2 Partial sufficiency and density estimation

Now, we recall Fraser’s definition of partial sufficiency. Taking into account Theorem 2.4 in Montanero et al. (2005), we set the following structure: consider two \( \sigma \)-finite measure spaces \( (\Omega_i, \mathcal{A}_i, \mu_i), i = 1, 2 \), a family \( \Delta_1 \) of probability densities with respect to \( \mu_1 \) and a family \( \mathcal{L} \) of measurable nonnegative functions \( \ell : (\Omega_1 \times \Omega_2, \mathcal{A}_1 \times \mathcal{A}_2) \rightarrow [0, +\infty) \) such that \( \int_{\Omega_2} \ell(\omega_1, \omega_2) d\mu_2(\omega_2) = 1 \), for every \( \omega_1 \in \Omega_1 \). Denote \( (\Omega, \mathcal{A}) = (\Omega_1 \times \Omega_2, \mathcal{A}_1 \times \mathcal{A}_2) \) and \( \Delta = \{ f \otimes \ell : f \in \Delta_1, \ell \in \mathcal{L} \} \), where

\( (f \otimes \ell)(\omega_1, \omega_2) = f(\omega_1)\ell(\omega_1, \omega_2) \).

Notice that every \( f \otimes \ell \in \Delta \) is a probability density on \( (\Omega, \mathcal{A}) \). Therefore we can consider the family of probability measures on \( (\Omega, \mathcal{A}) \) indexed by \( \Delta \), namely \( \mathcal{P} = \{ P_{f \otimes \ell} : f \in \Delta_1, \ell \in \mathcal{L} \} \).

Our purpose is to estimate the density \( f \in \Delta_1 \) from a sample of size \( n \) taken from \( f \otimes \ell \), leaving \( \ell \) as a nuisance parameter. The next definitions specify what is an estimator of \( f \) in this context, as well as the notions of ancillarity and partial sufficiency in the sense of Fraser (1956).

**Definition 1.** Any measurable function \( f_n : (\Omega_1 \times \Omega^n, \mathcal{A}_1 \times \mathcal{A}^n) \rightarrow \mathbb{R} \) will be called an estimator of \( f \).

**Definition 2.** A sub-\( \sigma \)-field \( \mathcal{B} \subset \mathcal{A}^n \) is said to be ancillary for the parameter \( \ell \) if, for any \( f \in \Delta_1 \) and \( \ell, \ell' \in \mathcal{L} \), we have

\[ P^n_{f \otimes \ell}(B) = P^n_{f \otimes \ell'}(B), \quad \forall B \in \mathcal{B}. \]

In this sense, a statistic \( f_n : \Omega_1 \times \Omega^n \rightarrow \mathbb{R} \) is said to be ancillary for the parameter \( \ell \) if for every \( \omega_1 \in \Omega_1 \) and every \( f \in \Delta_1 \), the \( P^n_{f \otimes \ell} \)-distribution of \( f_n(\omega_1, \cdot) \) does not depend on \( \ell \).
Definition 3. A sub-$\sigma$-field $B \subset A^n$ is said to be partially sufficient for $f$ if it is ancillary for the parameter $\ell \in L$ and, for any $\ell \in L$ and $A \in A^n$,

$$\bigcap_{f \in \Delta_1} P_{f \otimes \ell}^n(A|B) \neq \emptyset.$$ 

Remark 2. It is obvious that, if $B \subset A^n$ is ancillary for the parameter $\ell$ and $f_n(\omega_1, \cdot)$ is $B$-measurable for all $\omega_1 \in \Omega_1$, then $f_n$ is ancillary for the parameter $\ell$.

Remark 3. Under the assumptions of this section, the projection $\pi_1 : (\Omega^n, A^n) \to (\Omega^n_1, A^n_1)$ is clearly partially sufficient for $f$ (in the sense that so is the induced $\sigma$-field $\pi_1^{-1}(A^n_1)$). This result will be useful in the examples of Section 3.

In the following, let $\lambda$ be a $\sigma$-finite measure on $(\Omega_1, A_1)$. In the same spirit as Theorem 1, the next result states that, if a $\sigma$-field $B \subset A^n$ is partially sufficient for $f$ then, among all the density estimators which are ancillary for the parameter $\ell$, it is enough to consider only the $A_1 \times B$-measurable ones.

Theorem 2. Suppose that, for every $\ell \in L$, we have $\sup_{f \in \Delta_1} \int (f \otimes \ell)^d\lambda < \infty$ and let $B$ be a countably generated sub-$\sigma$-field of $A^n$. If $B$ is partially sufficient for $f$ then, for every estimator $f_n \in \Lambda_p(A_1 \times A^n)$ which is ancillary for the parameter $\ell$, there exists $f_n^* \in \Lambda_p(A_1 \times B)$ such that $ML_pE(f_n^*, f) \leq ML_pE(f_n, f)$. Moreover, $f_n^*$ can be chosen in a way such that $f_n^*(\omega_1, \cdot)$ be a version of the conditional expectation $E_{f \otimes \ell}[f_n(\omega_1, \cdot)|B]$, $\lambda$-a.e.; that is, there is a $\lambda$-null set $N_1 \in A_1$ such that

$$\int_B f_n^*(\omega_1, \bar{\omega})dP_{f \otimes \ell}^n(\bar{\omega}) = \int_B f_n(\omega_1, \bar{\omega})dP_{f \otimes \ell}^n(\bar{\omega}), \quad \forall B \in B, \forall f \otimes \ell \in \Delta, \forall \omega_1 \notin N_1.$$

Proof. Let us suppose, in the first place, that $\lambda$ is a probability measure. Fix $\ell \in L$ and let us prove that $A_1 \times B$ is a sufficient $\sigma$-field for the family

$$\lambda \times \mathcal{P}_{\otimes \ell}^n := \{\lambda \times P : P \in \mathcal{P}_\otimes^n\},$$

where $\mathcal{P}_\otimes^n := \{P_{f \otimes \ell}^n : f \in \Delta_1\}$. As the family $\lambda \times \mathcal{P}_{\otimes \ell}^n$ is dominated by the measure $\nu := \lambda \times (\mu_1 \times \mu_2)^n$, by means of Lemma 1.3.18 in Pfanzagl (1994), to prove that $A_1 \times B$ is sufficient for $\lambda \times \mathcal{P}_{\otimes \ell}^n$ it is enough to show that we can choose versions $\phi_{f, \ell}$
of the densities of $\lambda \times P_{f \otimes \ell}^n$ with respect to $\nu$ in a way such that, for any $f, f' \in \Delta_1$, we have that

$$\varphi_{f, f', \ell} := \frac{\phi_{f, \ell}}{\phi_{f, \ell} + \phi_{f', \ell}}$$

is $(A_1 \times B)$-measurable. But that is an immediate consequence of the equality

$$\frac{d}{d\nu}(\lambda \times P_{f \otimes \ell}^n) = \frac{dP_{f \otimes \ell}^n}{d(\mu_1 \times \mu_2)^n}(\omega_{11}, \omega_{12}, ..., \omega_{n1}, \omega_{n2}) = \frac{dP_{f \otimes \ell}^n}{d(\mu_1 \times \mu_2)^n}(\omega_{11}, \omega_{12}, ..., \omega_{n1}, \omega_{n2}).$$

and the fact that, as $B$ is sufficient for $P_{f \otimes \ell}^n$, there are versions $\psi_{f, \ell}$ of the densities of $P_{f \otimes \ell}^n$ with respect to $(\mu_1 \times \mu_2)^n$ in a way such that, for any $f, f' \in \Delta_1$, the function

$$\psi_{f, \ell}/(\psi_{f, \ell} + \psi_{f', \ell})$$

is $B$-measurable.

As $A_1 \times B$ is sufficient for $\lambda \times P_{f \otimes \ell}^n$, given $f_n \in \Lambda_p(A_1 \times A^n)$, there exists

$$f_{n, \ell}^* \in \bigcap_{f \in \Delta_1} E_{\lambda \times P_{f \otimes \ell}^n}[f_n|A_1 \times B].$$

By definition of conditional expectation it holds that, given $A_1 \in A_1$ and $B \in B$,

$$\int_{A_1} \int_B f_{n, \ell}^*(\omega_1, \omega)dP_{f \otimes \ell}^n(\omega)d\lambda(\omega_1) = \int_{A_1} \int_B f_{n}(\omega_1, \omega)dP_{f \otimes \ell}^n(\omega)d\lambda(\omega_1).$$

(2)

By Fubini’s theorem and the fact that $f_n$ is ancillary for the parameter $\ell$ we deduce that the right hand side of the previous equality does not depend on $\ell$. Therefore, the left hand side does not depend on $\ell$ either and, as $f_{n, \ell}^*$ is $A_1 \times B$-measurable, it holds that $f_{n, \ell}^* = f_{n, \ell'}^*$ for any $\ell, \ell' \in L$. We can denote, then, $f_n^*$ instead of $f_{n, \ell}^*$.

Now, let us show that there exists $N_1 \in A_1$ fulfilling $\lambda(N_1) = 0$ and

$$f_n^*(\omega_1, \cdot) \in \bigcap_{f \in \Delta_1} E_{P_{f \otimes \ell}^n}[f_n(\omega_1, \cdot)|B], \quad \forall \omega_1 \notin N_1.$$

Indeed, it follows from (2) that, for every $B \in B$, there exists a $\lambda$-null set $N_B \in A_1$ such that, if $\omega_1 \notin N_B$, then

$$\int_B f_n^*(\omega_1, \cdot)dP_{f \otimes \ell}^n = \int_B f_n(\omega_1, \cdot)dP_{f \otimes \ell}^n.$$

(3)

In particular, that is true for all the sets $B$ in a countable field $B_0$ generating $B$, so if we define $N_1 := \bigcup_{B \in B_0} N_B$ we obtain a $\lambda$-null set such that in its complementary (3) holds, for every $B \in B_0$. 
The class of sets $B \in \mathcal{B}$ satisfying (3) for $\omega_1 \notin N_1$ is monotone and contains $\mathcal{B}_0$; therefore, it coincides with $\mathcal{B}$. This observation, together with standard properties of the conditional expectation (see for instance Billingsley, 1995), yields the proof of the theorem when $\lambda$ is a probability measure.

In the general case, where $\lambda$ is a $\sigma$-finite measure, there exists a countable measurable partition $(A_{1k})_k$ of $\Omega_1$ with $0 < \lambda(A_{1k}) < +\infty$ for all $k$. Let $\mathcal{A}_{1k} = A_{1k} \cap \mathcal{A}_1$ and $\lambda_k$ the restriction of $\lambda$ to $\mathcal{A}_{1k}$ divided by $\lambda(A_{1k})$. By the argument above, we can construct a $\lambda_k$-null set $N_{1k} \in \mathcal{A}_{1k}$ and an $\mathcal{A}_{1k} \times \mathcal{B}$-measurable function $f_{n,k}^*$ such that

$$\int_B f_{n,k}^*(\omega_1, \cdot) dP_{f \otimes \ell} = \int_B f_n(\omega_1, \cdot) I_{A_{1k}}(\omega_1) dP_{f \otimes \ell},$$

for every $\omega_1 \in A_{1k} \setminus N_{1k}$ and every $B \in \mathcal{B}$. Hence $f_n^* := \sum_k f_{n,k}^*$ is $\mathcal{A}_1 \times \mathcal{B}$-measurable and satisfies

$$\int_B f_n^*(\omega_1, \cdot) dP_{f \otimes \ell} = \sum_k \int_B f_n(\omega_1, \cdot) I_{A_{1k}}(\omega_1) dP_{f \otimes \ell} = \int_B f_n(\omega_1, \cdot) dP_{f \otimes \ell},$$

for every $\omega_1 \notin N_1 := \cup_k N_{1k}$ and every $B \in \mathcal{B}$.

\[\square\]

3 \hspace{0.5cm} Examples

In this section we give three illustrative examples on the use of partial sufficiency in density estimation.

**Example 2.** Consider the nonparametric family of real distributions

$$\{G_{f,\phi}: f \text{ a density function on } [0, \infty) \text{ and } 0 < \phi < 1\}$$

where $G_{f,\phi}$ has density $g_{f,\phi}(x) = (1 - \phi)f(-x) + \phi f(x)$. The distribution of the bimeasurable map $x \in \mathbb{R} \mapsto (|x|, I_{[0,\infty)}(x)) \in [0, \infty) \times \{0, 1\}$ with respect to $G_{f,\phi}$ is the product $P_{f \otimes l_\phi}$ of the distribution with density $f$ and the Bernoulli distribution $l_\phi$ with parameter $\phi$. To estimate the density $f$ from a sample $X_1, \ldots, X_n$ drawn from this distribution avoiding the information carried out by $\phi$, we should choose a $\phi$-ancillary estimator; according to Remark 3, the statistic $Z = (|X_1|, \ldots, |X_n|)$ is
partial sufficient for $f$ and then Theorem 2 assures that every $\phi$-ancillary estimator can be improved by another one depending on $Z$. For instance, in these conditions, a kernel estimator of $f$ should also depend on $X_1, \ldots, X_n$ in the way of Example 1.

**Example 3.** Assume that we have $(Y_i, Z_i), i = 1, \ldots, n,$ independent copies of a random bivariate variable $(Y, Z),$ and that there exists an unknown diffeomorphism $g$ such that $Y_i = g(X_i) + Z_i, i = 1, \ldots, n,$ where $X_1, \ldots, X_n$ are independent copies of a real random variable $X$ having a known continuous distribution $P,$ with density $f.$ Suppose that the conditional distribution of $Z$ given $X$ admits a conditional density of the form $\ell(x, z),$ satisfying the conditions given in the previous section. From the density $f$ we can construct, taking into account the Change of Variables Theorem, the density of $g(X),$ namely $f_g.$ If our initial goal is to estimate $f_g$ ignoring any information coming from the distribution of the error $Z,$ we should consider a new statistical experiment whose observations are $(Y_i - Z_i, Z_i), i = 1, \ldots, n.$ These observations are characterized by the family of densities of the form $f_g(\omega)\ell(\omega, z).$ With these conditions we can ensure that the statistic $T_n(Y_1, Z_1, \ldots, Y_n, Z_n) = (Y_1 - Z_1, \ldots, Y_n - Z_n)$ is partially sufficient for the original problem. Therefore, every estimator of the density $f_g$ depending on the observations $(Y_i, Z_i), i = 1, \ldots, n,$ can be improved, in the sense of Theorem 2 above, by another one depending only on $Y_i - Z_i, i = 1, \ldots, n.$ Once $f_g$ has been estimated, using that estimate it would be possible to find an analytic expression for $g.$

**Example 4.** With the notation of the previous example, now assume that we observe $(X_i, Y_i), i = 1, \ldots, n,$ that the diffeomorphism $g$ is fully known, and that $f,$ the density of $X,$ is unknown. Let $f$ and $\ell$ be the densities of $X$ and $Z,$ respectively, and suppose that we are interested in estimating the density $\ell$ ignoring any information coming from the distribution of $X.$ We should consider a new statistical experiment, whose observations are $(X_i, Y_i - g(X_i)), i = 1, \ldots, n.$ These observations are characterized by the family of densities of the form $f(x)\ell(z).$ Therefore, arguing as in the previous example, the statistic $T_n(X_1, Y_1, \ldots, X_n, Y_n) = (Y_1 - g(X_1), \ldots, Y_n - g(X_n)),$ is partially sufficient for the original problem.
Acknowledgments. The authors appreciate the comments and suggestions of an anonymous referee. This research has been supported by the Spanish Ministerio de Ciencia y Tecnología project MTM2006-06172.

References


