Sufficiency and the preservation of Fisher Information
David Pollard (http://www.stat.yale.edu/∼pollard)
JSM Utah, July 2007

1

Densities $f_\theta(x) = dP_\theta/d\mu$ for $\theta \in \Theta$, an open subset of $\mathbb{R}$.
Hellinger differentiability/DQM:

$$\xi_{\theta+\delta}(x) := \sqrt{f_{\theta+\delta}(x)} = \sqrt{f_\theta(x)} + \delta\dot{\xi}_\theta(x) + r_\theta(x,\delta) \qquad \text{with } r_\theta = o(|\delta|) \text{ in } \mathcal{L}^2(\mu) \text{ norm as } \delta \to 0.$$

Equivalently,

$$\sqrt{\frac{dP_{\theta+\delta}}{dP_\theta}} = 1 + \tfrac{1}{2}\delta\Delta_\theta(x) + R_\theta(x,\delta) \qquad \text{with } R_\theta = o(|\delta|) \text{ in } \mathcal{L}^2(P_\theta) \text{ norm as } \delta \to 0,$$

where $\Delta_\theta(x) := \{\xi_\theta > 0\} 2\dot{\xi}_\theta(x)/\xi_\theta(x)$ (the **score function**). cf. $\partial\sqrt{f_\theta}/\partial\theta \stackrel{?}{=} \dot{f}_\theta/2\sqrt{f_\theta}$

Fisher information $I(\theta) := 4\mu(\dot{\xi}_\theta^2) = 4P_\theta(\Delta_\theta^2) = 4\int \Delta_\theta(x)^2 P_\theta(dx)$.

2

Statistic T with distribution Q_θ under P_θ.

$$\sqrt{\frac{dQ_{\theta+\delta}}{dQ_\theta}} = 1 + \tfrac{1}{2}\delta\widetilde{\Delta}_\theta(x) + \widetilde{R}_\theta(x,\delta) \qquad \text{with } \widetilde{R}_\theta = o(|\delta|) \text{ in } \mathcal{L}^2(Q_\theta) \text{ norm as } \delta \to 0,$$

where $\widetilde{\Delta}_\theta(t) := P_\theta(\Delta \mid T = t)$, the conditional expectation of Δ_θ given $T = t$.

See: Ibragimov & Has'minskii (1981, Section 1.7??); van der Vaart (1988, page 181); Le Cam & Yang (1988, Section 7); Bickel, Klaassen, Ritov & Wellner (1993, page 461).

Fisher information $\widetilde{I}(\theta) := 4Q_\theta\widetilde{\Delta}_\theta^2 = 4P_\theta\widetilde{\Delta}_\theta(Tx)^2$.

3

Geometry:

$$I(\theta) = 4P_\theta\left(\Delta_\theta(x) - \widetilde{\Delta}_\theta(Tx) + \widetilde{\Delta}_\theta(Tx)\right)^2 = 4P_\theta\left(\Delta_\theta(x) - \widetilde{\Delta}_\theta(Tx)\right)^2 + \widetilde{I}(\theta).$$

No loss of Fisher information if and only if $\Delta_\theta(x) = \widetilde{\Delta}_\theta(Tx)$ a.s. $[P_\theta]$. cf. Pitman (1979, page 19)

4

No loss of Fisher information if T is a sufficient statistic.
Kagan & Shepp (2005): example of zero loss of Fisher information without sufficiency.

5

Example based on K&S. Densities $\{g_\theta(t) : \theta \in \Theta\}$ and $\{h_\theta(t) : \theta \in \Theta\}$ with respect to λ, both DQM:

$$\xi_{\theta+\delta,g}(t) := \sqrt{g_{\theta+\delta}(t)} = \sqrt{g_\theta(t)} + \delta\dot{\xi}_{\theta,g}(t) + o(|\delta|) \text{ in } \mathcal{L}^2(\lambda)$$
$$\xi_{\theta+\delta,h}(t) := \sqrt{h_{\theta+\delta}(t)} = \sqrt{h_\theta(t)} + \delta\dot{\xi}_{\theta,h}(t) + o(|\delta|) \text{ in } \mathcal{L}^2(\lambda)$$

Put $x = (t, y)$ with $y \sim \nu$, for a known (nondegenerate) probability measure ν on $(0, 1)$, mean $\alpha \neq 1/2$. Define $f_\theta(x) = yg_\theta(t) + (1-y)h_\theta(t)$, a density with respect to $\mu := \lambda \otimes \nu$. The corresponding marginal density for the statistic $T(x) = t$ is $\alpha g_\theta(t) + (1-\alpha)h_\theta(t)$.

If we make sure that $G_\theta := \{t : g_\theta(t) > 0\}$ is disjoint from $H_\theta := \{t : g_\theta(t) > 0\}$, for each θ, then

$$\sqrt{f_\theta(x)} = \xi_\theta(x) = \sqrt{y}\ \xi_{\theta,g}(t)\{t \in G_\theta\} + \sqrt{1-y}\ \xi_{\theta,h}(t)\{t \in H_\theta\}$$
$$\dot{\xi}_\theta(x) = \sqrt{y}\ \dot{\xi}_{\theta,g}(t)\{t \in G_\theta\} + \sqrt{1-y}\ \dot{\xi}_{\theta,h}(t)\{t \in H_\theta\}$$
$$\Delta_\theta(x) = 2\frac{\dot{\xi}_\theta(x)}{\xi_\theta(x)}\{t \in G_\theta \cup H_\theta\} = 2\frac{\dot{\xi}_{\theta,g}(t)}{\xi_{\theta,g}(t)}\{t \in G_\theta\} + 2\frac{\dot{\xi}_{\theta,h}(t)}{\xi_{\theta,h}(t)}\{t \in H_\theta\}.$$

The score function does not depend on y. There is no loss of Fisher information but T is not sufficient:

$$y \mid t \sim \nu_g\{t \in G_\theta\} + \nu_h\{t \in H_\theta\} \qquad \text{under } P_\theta(\cdot \mid T = t),$$

where ν_g and ν_h are the probability measures defined by

$$\frac{d\nu_g}{d\nu} = \frac{y}{\alpha} \qquad \text{and} \qquad \frac{d\nu_h}{d\nu} = \frac{1-y}{1-\alpha}.$$

6

Take independent observations $x_1 = (t_1, y_1), \ldots, x_n = (t_n, y_n)$ from P_θ. The statistic $T = (t_1, \ldots, t_n)$ is not sufficient but there is no loss of Fisher information. The y_i's are conditionally independent given T, with

$$(*) \qquad y_i \mid t_i \sim \nu_g\{t_i \in G_\theta\} + \nu_h\{t_i \in H_\theta\}.$$

How much more about θ do we learn from the y_i's once we know T? Not much. For example, the asymptotic distributions for efficient estimators are the same with or without the y_i's.

7

K&S example: $g_\theta(t) = \psi(t-\theta)$ and $h_\theta(t) = \psi(\theta - t)$ where $\psi(t) = \frac{1}{2}t^2 e^{-t}\{t > 0\}$ and λ equal to Lebesgue measure on the real line. Thus $G_\theta = (\theta, \infty)$ and $H_\theta = (-\infty, \theta)$.

Generate independent observations $y_1', \ldots, y_n'$ from ν_g and $y_1'', \ldots, y_n''$ from ν_h.

Given T (and secret knowledge of θ), define $y_i^* = y_i'\{t_i > \theta\} + y_i''\{t_i < \theta\}$ and $x_i^* := (t_i, y_i^*)$.
The variables $x_1^*, \ldots, x_n^*$ have the same joint distribution as the sample $x_1, \ldots, x_n$.

Given a $\sqrt{n}$-consistent estimator $\widehat{\theta}_n$ for θ, define $y_i^{**} = y_i'\{t_i > \widehat{\theta}_n\} + y_i''\{t_i < \widehat{\theta}_n\}$ and $x_i^{**} := (t_i, y_i^{**})$.

Use the fact that $\min_{i \le n} |t_i - \theta|$ is of order $n^{-1/3}$ to show that

$$\mathbb{P}_\theta\{x_i^* = x_i^{**} \text{ for each } i \le n\} = \mathbb{P}_\theta\{\text{ no } t_i\text{'s between } \theta \text{ and } \widehat{\theta}_n\} \to 1 \qquad \text{as } n \to \infty.$$

More concretely, for $\widehat{\theta}_n = \text{sample median} - c_\alpha$, with a suitable constant c_α that depends on α, get

$$\mathbb{P}_\theta\{x_i^* \ne x_i^{**} \text{ for at least one } i \le n\} = O\left(\frac{\log^{3/2} n}{\sqrt{n}}\right)$$

We can almost reproduce the behaviour of any statistic based on $x_1, \ldots, x_n$ by a (randomized) statistic based on $x_1^{**}, \ldots, x_n^{**}$.

Compare with the Le Cam distance between experiments.

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. & Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins University Press, Baltimore.

Ibragimov, I. A. & Has'minskii, R. Z. (1981), *Statistical Estimation: Asymptotic Theory*, Springer, New York.

Kagan, A. & Shepp, L. A. (2005), 'A sufficiency paradox: an insufficient statistic preserving the Fisher information', *The American Statistician* **59**, 54–56.

Le Cam, L. & Yang, G. L. (1988), 'On the preservation of local asymptotic normality under information loss', *Annals of Statistics* **16**, 483–520.

Pitman, E. J. G. (1979), *Some Basic Theory for Statistical Inference*, Chapman and Hall.

van der Vaart, A. (1988), *Statistical estimation in large parameter spaces*, Center for Mathematics and Computer Science. CWI Tract 44.