

A paradox concerning nuisance parameters and projected estimating functions.

Biometrika (2004) 91(4): 929-941

M. Henmi and S. Eguchi

d.stephens@math.mcgill.ca
www.math.mcgill.ca/dstephens/

Problem addressed

When, in (semi-)parametric inference is it advantageous to use estimated nuisance parameters ?

Abstract:

*This paper is concerned with a paradox associated with parameter estimation in the presence of nuisance parameters. In a statistical model with unknown nuisance parameters, **the efficiency of an estimator of a parameter usually increases when the nuisance parameters are known.***

*However the **opposite phenomenon** can sometimes occur.*

Abstract:

*In this paper, we elucidate the occurrence of this paradox by examining estimating functions. A sufficient condition for the paradox to occur is the **orthogonality** of the two components of the projected estimating functions corresponding to parameters of interest and nuisance parameters.*

Illustrative setting

Robins et al. (1992) model

$$y = \beta s + h(x) + \epsilon \quad (1)$$

$$\text{pr}(S = 1|x) = \text{expit}\{\alpha'x\} = e(x; \alpha) \quad (2)$$

where h is an unknown function, and

$$\mathbb{E}[\epsilon|s, x] = 0.$$

- β : parameter of interest
- α : nuisance parameters

This model is **semiparametric** as the probability distribution of ϵ is **not specified**, there is merely a moment constraint.

Illustrative setting

Estimating equation for β is

$$\sum_{i=1}^n (s_i - e(x_i; \hat{\alpha}))(y_i - \beta s_i) = 0 \quad (3)$$

where $\hat{\alpha}$ is the ML estimator derived from the data $(s_i, x_i), i = 1, \dots, n$. Denote the solution to this equation $\hat{\beta}$.

If α is **known**, it is tempting to use the estimating equation

$$\sum_{i=1}^n (s_i - e(x_i; \alpha))(y_i - \beta s_i) = 0$$

Denote the solution to this equation $\tilde{\beta}$.

Illustrative setting

Under standard regularity conditions

$$\hat{\beta}, \tilde{\beta} \xrightarrow{p} \beta_0$$

with both estimators $\sqrt{n}$ consistent and asymptotically normal.

Key Result:

$$\text{avar}(\tilde{\beta}) \geq \text{avar}(\hat{\beta}) \tag{4}$$

where avar denotes the asymptotic variance, that is

$$\text{avar}(\tilde{\beta}) - \text{avar}(\hat{\beta})$$

is non-negative definite.

This is a well-known result ...

...whose origin is not well understood. Given in various forms by

- ▶ Chamberlain (1987)
- ▶ Newey (1990)
- ▶ BKRW (1992)
- ▶ references in paper

The aim of the Henmi & Eguchi paper is to clarify when the ‘paradox’ is observed to hold, and when it is not. Tools used are the standard definitions and results of semiparametric inference.

Formulation

- ▶ $\mathcal{M} = \{p(x; \theta, \kappa)\}$: family of distribution functions specified by
 - ▶ d -dimensional parameter of interest θ
 - ▶ $\theta = (\beta', \alpha')'$ where β is a parameter of interest, and α is a nuisance parameter
 - ▶ further nuisance parameter κ , possibly infinite dimensional (eg the distribution function of x .)

Let

$$s(x; \theta, \kappa) = \begin{bmatrix} s_{\beta}(x; \theta, \kappa) \\ s_{\alpha}(x; \theta, \kappa) \end{bmatrix} = \frac{\partial}{\partial \theta} \log p(x; \theta, \kappa)$$

be the score for θ , assumed to be bounded in mean square.

Estimating function

Let $u(x, \theta)$ be an estimating function such that

$$\mathbb{E}[u(x, \theta)] = 0 \quad \mathbb{E}[\|u(x, \theta)\|^2] < \infty \quad (5)$$

and that if

$$m(\theta_1, \theta_2, \kappa) = \int u(x, \theta_1) p(x; \theta_2, \kappa) dx$$

then

$$\det \left\{ \frac{\partial}{\partial \theta_1} m(\theta_1, \theta_2, \kappa) \right\}_{\theta_1 = \theta, \theta_2 = \theta} \neq 0 \quad (6)$$

Estimating function

The estimator $\hat{\theta}$ solves

$$\sum_{i=1}^n u(x_i, \theta) = 0$$

and has standard properties (consistency and asymptotic normality).

The asymptotic variance of $\hat{\theta}$ is

$$\text{avar}(\hat{\theta}) = W(\theta, \kappa)^{-1} V(\theta, \kappa) \{W(\theta, \kappa)^{-1}\}' \quad (7)$$

where

$$V(\theta, \kappa) = \text{var}\{u(x, \theta)\} = \mathbb{E}\{u(x, \theta)u(x, \theta)'\}$$

and

$$W(\theta, \kappa) = \mathbb{E}\{u(x, \theta)s(x, \theta, \kappa)'\}$$

Projected Estimating function

Construct projected estimating function u_* by

$$u_*(x, \theta) = W(\theta, \kappa)V(\theta, \kappa)^{-1}u(x, \theta) \quad (8)$$

that is,

$$u_*(x, \theta) = A(\theta, \kappa)u(x, \theta)$$

for the $d \times d$ matrix A .

Clearly $\hat{\theta}$ solves

$$\sum_{i=1}^n u_*(x_i, \theta) = 0.$$

Projected Estimating function

$u_*(x, \theta)$ is the ‘best’ estimating function that we can obtain from $u(x, \theta)$ by linear transformation, that is

$$u_*(x, \theta) = \arg \min_{v=Au} \mathbb{E}[\|s(x, \theta, \kappa) - v(x, \theta)\|^2]$$

That is, if we could pick

$$u_*(x, \theta) = s(x, \theta, \kappa)$$

that would be ideal (theory tells us that this would yield the most efficient estimator), but we are limited to use linear transformations of $u(x, \theta)$ to find the ‘best approximation’ to $s(x, \theta, \kappa)$

Projected Estimating function

By standard theory (geometry of Hilbert spaces), we know that the optimal $u_*(x, \theta)$ is given by (8).

Essentially this is an easy result to prove, by the familiar theory of least-squares estimation.

- ▶ $s(x, \theta, \kappa)$ exists in a (Hilbert) space $H_{\theta, \kappa}$;
- ▶ $u(x, \theta)$ defines a d -dimensional linear subspace $H_{\theta, \kappa}^u$ contained within $H_{\theta, \kappa}$;
- ▶ the ‘closest’ point to $s(x, \theta, \kappa)$ in $H_{\theta, \kappa}^u$ is some linear combination $Au(x, \theta)$, for $d \times d$ matrix A ;
- ▶ this point is at the foot of the perpendicular obtained by projecting $s(x, \theta, \kappa)$ ‘down’ onto $H_{\theta, \kappa}^u$.

Projected Estimating function

We require that the vector

$$u_*(x, \theta) = Au(x, \theta)$$

is perpendicular to the ‘residual’

$$s(x, \theta, \kappa) - u_*(x, \theta) = s(x, \theta, \kappa) - Au(x, \theta)$$

that is

$$\mathbb{E} [Au \{s - Au\}'] = 0$$

a $(d \times d)$ matrix equation, or equivalently

$$\mathbb{E} [u \{s - u_*\}'] = 0.$$

Projected Estimating function

We must have

$$\mathbb{E}[us'] = \mathbb{E}[uu']A'$$

so that

$$A' = \{\mathbb{E}[uu']\}^{-1} \mathbb{E}[us'] = V(\theta, \kappa)^{-1}W(\theta, \kappa).$$

Therefore

$$u_*(x, \theta) = W(\theta, \kappa)'V(\theta, \kappa)^{-1}u(x, \theta)$$

as required.

Lemma 1

As

$$\text{var}\{u_*(x, \theta)\} = A \text{var}\{u(x, \theta)\} A'$$

and

$$\text{var}\{u(x, \theta)\} = V(\theta, \kappa)$$

and

$$A V A' = W' V^{-1} V V^{-1} W = W' V^{-1} W$$

we have

$$\text{avar}(\hat{\theta}) = [\text{var}\{u_*(x, \theta)\}]^{-1} \tag{9}$$

Recall the standard parametric theory:

$$\text{avar}(\hat{\theta}) = \mathcal{I}(\theta)^{-1} = [\text{var}\{s(x, \theta)\}]^{-1} = [\mathbb{E}\{s(x, \theta)s(x, \theta)'\}]^{-1}$$

with $\hat{\theta}$ **efficient**.

Structure of the Paradox

Suppose that

$$u(x; \theta) = \begin{bmatrix} u_\beta(x; \theta) \\ u_\alpha(x; \theta) \end{bmatrix}$$

with u_β satisfying the same non-singularity constraint as u .

Consider a similar partition to the projected estimating function u_*

$$u_*(x; \theta) = \begin{bmatrix} u_{*\beta}(x; \theta) \\ u_{*\alpha}(x; \theta) \end{bmatrix}$$

comprising the projections of s_β and s_α respectively.

Structure of the Paradox

Let

- ▶ $\hat{\theta} = (\hat{\beta}', \hat{\alpha}')'$ be the estimator derived from $u(x, \theta)$ (and $u_*(x, \theta)$).
- ▶ $\tilde{\beta} \equiv \tilde{\beta}(\alpha)$ be the estimator derived from $u_\beta(x, \beta, \alpha)$ with α known.

Finally, let $\Pi \{u(x)|v(x)\}$ denote the orthogonal projection of $u(x)$ onto the subspace spanned by $v(x)$, that is

$$\Pi \{u(x)|v(x)\} = Mv(x)$$

for some optimally chosen matrix M .

Lemma 2

$$\text{avar}(\hat{\beta}) = (\text{var}[u_{*\beta}(x, \theta) - \Pi\{u_{*\beta}(x, \theta)|u_{*\alpha}(x, \theta)\}])^{-1} \quad (11)$$

$$\geq [\text{var}\{u_{*\beta}(x, \theta)\}]^{-1} \quad (12)$$

with equality if and only if

$$\mathbb{E}\{u_{*\beta}(x, \theta)u_{*\alpha}(x, \theta)'\} = 0 \quad (13)$$

or equivalently

$$\mathbb{E}\{u_{*\beta}(x, \theta)s_{\alpha}(x, \theta, \kappa)'\} = 0$$

$$\mathbb{E}\{s_{\beta}(x, \theta, \kappa)u_{*\alpha}(x, \theta)'\} = 0$$

Lemma 2

Note: if $u(x, \theta)$ and $u_*(x, \theta)$ coincide, then

$$\text{avar}(\hat{\beta}) \geq [\text{var}\{u_{*\beta}(x, \theta)\}]^{-1} = \text{avar}(\tilde{\beta}) \quad (14)$$

and there is no efficiency gain in using the estimated α .

This is what happens in the standard parametric setting.

Lemma 3

$$\text{avar}(\tilde{\beta}) = (\text{var} [\Pi \{u_{*\beta}(x, \theta)|u_{\beta}(x, \theta)\}])^{-1} \quad (15)$$

so that

$$\text{avar}(\tilde{\beta}) \geq [\text{var}\{u_{*\beta}(x, \theta)\}]^{-1} \quad (16)$$

with equality if and only if

$$u_{\beta*}(x, \theta) = u_{*\beta}(x, \theta) \quad (17)$$

where $u_{\beta*}(x, \theta)$ is the projected estimating function for $u_{\beta}(x, \theta)$, that is

$$u_{\beta*}(x, \theta) = Au_{\beta}(x, \theta)$$

for some optimally chosen matrix A .

Lemma 3

This follows from Lemma 1, applied to the estimating function $u_\beta(x, \theta)$: we have from Lemma 1

$$\text{avar}(\tilde{\beta}) = [\text{var}\{u_{\beta*}(x, \theta)\}]^{-1} \quad (18)$$

However, $u_{*\beta}(x, \theta)$ is the projection onto a **larger** space; recall

$$u_*(x, \theta) = WV^{-1}u(x, \theta)$$

and $u_{*\beta}(x, \theta)$ is the first chunk of this resulting vector, whereas

$$u_{\beta*}(x, \theta) = Au_\beta(x, \theta)$$

some A i.e. we only use the first chunk of $u(x, \theta)$ on the RHS.

Lemma 3

It is easy to see that

$$u_{\beta*}(x, \theta) = \Pi \{u_{*\beta}(x, \theta) | u_{\beta}(x, \theta)\} \quad (19)$$

that is, $u_{\beta*}(x, \theta)$ is the closest point to $u_{*\beta}(x, \theta)$ that you can reach by taking linear combinations of $u_{\beta}(x, \theta)$.

i.e. we find the best approximation to $u_{*\beta}(x, \theta)$ we can if restricted to the space spanned by $u_{\beta}(x, \theta)$

Figure 1

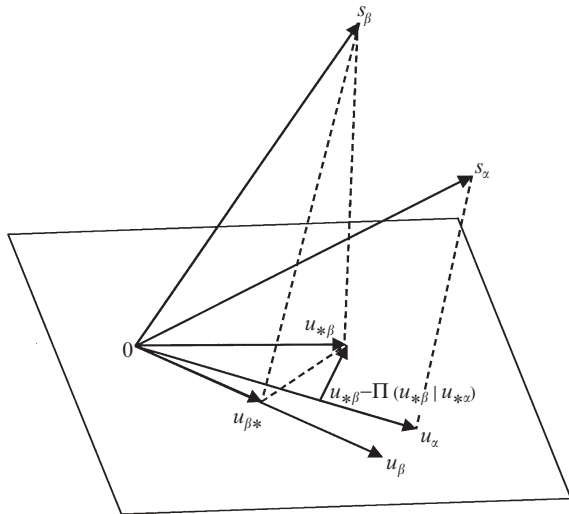


Fig. 1. The relationships between $u_{*\alpha}$, $u_{\beta*}$ and $u_{*\beta}$.

Important Note

p. 934:

In general, there is no relationship between the asymptotic variance matrices of $\hat{\beta}$ and $\tilde{\beta}$.

that is, in general, you cannot say that one is ‘bigger’ or ‘smaller’ than the other, it depends on

$$u_{*\alpha}(x, \theta) \quad u_{*\beta}(x, \theta) \quad u_{\beta*}(x, \theta).$$

Theorem 1

Suppose

$$\mathbb{E}\{u_{*\beta}(x, \theta)u_{*\alpha}(x, \theta)'\} = 0. \quad (20)$$

Then

$$\text{avar}(\tilde{\beta}) \geq \text{avar}(\hat{\beta}) \quad (21)$$

with equality if and only if

$$\mathbb{E}\{u_{\beta}(x, \theta)s_{\alpha}(x, \theta, \kappa)'\} = 0. \quad (22)$$

Theorem 1

In fact

$$\text{avar}(\tilde{\beta}) - \text{avar}(\hat{\beta}) = D \text{avar}(\hat{\alpha}) D' \quad (23)$$

where

$$D = \mathbb{E}\{u_{\beta}(x, \theta)s_{\beta}(x, \theta, \kappa)'\}^{-1}\mathbb{E}\{u_{\beta}(x, \theta)s_{\alpha}(x, \theta, \kappa)'\}$$

That is, estimator $\hat{\beta}$ is more efficient than $\tilde{\beta}$ asymptotically unless (22) holds.

Equality holds in (20) if, for example,

$$\mathbb{E}\{s_{\beta}(x, \theta, \kappa)s_{\alpha}(x, \theta)'\} = 0 \quad (24)$$

$$u_{\alpha}(x, \theta) = s_{\alpha}(x, \theta) \quad (25)$$

and in this case, you **do** get variance improvement.

Here we have (under correct specification)

$$u_{\alpha}(x, \theta) = s_{\alpha}(x, \theta)$$

and β is separated completely from α in the estimation (hence the equation on p.935).

We can compute the difference in asymptotic variances as

$$\frac{1}{[\mathbb{E}\{r(x, \alpha)\}]^2} \mathbb{E}\{h(x)r(x, \alpha)x'\} [\mathbb{E}\{r(x; \alpha)xx'\}]^{-1} \mathbb{E}\{h(x)r(x; \alpha)x\}$$

where

$$r(x; \alpha) = e(x; \alpha)\{1 - e(x; \alpha)\}.$$

Hence, as by assumption $0 < e(x; \alpha) < 1$, the difference is only zero when $h(x)$ is identically zero.

We have the likelihood (conceptually) as

$$L(\theta, \kappa) = L_1(\beta, \kappa)L_2(\alpha)$$

i.e. the likelihood factorizes.

The Robins et al. model is typical of most models we utilize in causal adjustment. Note that it relies on the estimation of α being separated totally from the other inference, that is, in a two-step procedure.

Note also that this is an **asymptotic** result only; however it does seem to hold uniformly in β (that is, there are no pathological cases that would be problematic for finite n).

Proposition 1

Note that if

$$u_{\beta*}(x, \theta) = u_{*\beta}(x, \theta) \quad (26)$$

then

$$\text{avar}(\tilde{\beta}) \leq \text{avar}(\hat{\beta}) \quad (27)$$

This holds when, for example,

$$u_{\beta}(x, \theta) = s_{\beta}(x, \theta)$$

as might occur in the parametric case. Thus

(i) If $\hat{\beta}$ and $\hat{\alpha}$ are asymptotically independent

$$\text{avar}(\tilde{\beta}) \geq \text{avar}(\hat{\beta})$$

(ii) If $\tilde{\beta}$ and $\hat{\alpha}$ are asymptotically independent

$$\text{avar}(\tilde{\beta}) \leq \text{avar}(\hat{\beta})$$

Example

Missing data example: $(y_i, x_i, r_i), i = 1, \dots, n$

- ▶ $r_i = 1$ indicates missing, $r_i = 0$ not missing

Estimating equation

$$\sum_{i=1}^n \frac{1 - r_i}{\text{pr}(R_i = 0|x_i)} \frac{\partial}{\partial \beta} \log f(y_i; \beta) = 0 \quad (28)$$

Missingness mechanism unknown, use model

$$\text{pr}(R_i = 0|x_i) = e(x; \alpha) = \text{expit}(\alpha' x_i)$$

Here

$$\text{avar}(\tilde{\beta}) \geq \text{avar}(\hat{\beta})$$

with the difference quantified by eqn. (29), p. 937.

Example

Difference is

$$I(\beta)^{-1}J(\beta, \alpha)[\mathbb{E}\{r(x; \alpha)x_i x_i'\}]^{-1}J(\beta, \alpha)'I(\beta)^{-1} \quad (29)$$

where

$$\begin{aligned} I(\beta) &= \mathbb{E}\{s_\beta(y_i, \beta)s_\beta(y_i, \beta)'\} \\ J(\beta, \alpha) &= \mathbb{E}[\{1 - e(x_i; \alpha)\}s_\beta(y_i, \beta)x_i'] \\ r(x; \alpha) &= e(x; \alpha)\{1 - e(x; \alpha)\} \end{aligned}$$

Note: (28) is **not** the efficient (AIPWCC) estimating equation, for which the variance bound is achieved.

Simulation

- ▶ (Y_i, X_i)

$$Y_i, X_i \sim \mathcal{N} \left(\begin{pmatrix} \mu \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma\rho \\ \sigma\rho & 1 \end{pmatrix} \right)$$

- ▶ R_i

$$R_i \sim \text{Bernoulli}(1 - e(x; \alpha))$$

Estimator $\tilde{\beta}$ is $(\hat{\mu}, \hat{\sigma}^2)$ where

$$\hat{\mu} = \frac{\sum_{i=1}^n w_i(\alpha) Y_i}{\sum_{i=1}^n w_i(\alpha)} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^n w_i(\alpha) (Y_i - \hat{\mu})^2}{\sum_{i=1}^n w_i(\alpha)}$$

where

$$w_i(\alpha) = \frac{1 - R_i}{e(X_i; \alpha)}$$

For $\hat{\beta}$, replace α by $\hat{\alpha}$.

Simulation

- $\beta = (\mu, \sigma^2) = (0, 1)$
- $\alpha = (\alpha_1, \alpha_2) = (1, 1)$ (or do they mean $(-1, -1)$?)

ρ ranges from -0.9 to 0.9 .

Note: I ‘Monte Carlo’ed the estimators and took the large sample variance. In the paper, they probably did numerical integration for the bits they couldn’t evaluate.

Can compute via (7), p. 931.

Recall: asymptotic variance of $\hat{\theta}$ is

$$\text{avar}(\hat{\theta}) = W(\theta, \kappa)^{-1} V(\theta, \kappa) \{W(\theta, \kappa)^{-1}\}' \quad (30)$$

where

$$V(\theta, \kappa) = \text{var}\{u(x, \theta)\} = \mathbb{E}\{u(x, \theta)u(x, \theta)'\}$$

and

$$W(\theta, \kappa) = \mathbb{E}\{u(x, \theta)s(x, \theta, \kappa)'\}$$

Here

$$u(x, \theta) = \begin{pmatrix} \frac{(1-r)(1-e(x; \alpha))}{e(x; \alpha)^2} \left(-\frac{1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (y - \mu)^2\right) \\ \frac{x(1-r)(1-e(x; \alpha))}{e(x; \alpha)^2} \left(-\frac{1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (y - \mu)^2\right) \\ \frac{1-r}{e(x; \alpha)} (y - \mu) / \sigma^2 \\ \frac{1-r}{e(x; \alpha)} ((y - \mu)^2 / \sigma^2 - 1) / 2\sigma^2 \end{pmatrix}$$

Simulation

Here

$$f(y, x, r|\theta) = f(x, \theta)f(r|x, \theta)f(y|x, r, \theta)$$

so

$$\log f(y, x, r|\theta) = \log f(x, \theta) + \log f(r|x, \theta) + \log f(y|x, r, \theta)$$

and

$$s(x, \theta) = \begin{pmatrix} s_{\alpha_1} \\ s_{\alpha_2} \\ s_{\mu} \\ s_{\sigma}^2 \end{pmatrix} = \begin{pmatrix} r - (1 - e(x, \alpha)) \\ x(r - (1 - e(x, \alpha))) \\ \frac{1 - r}{e(x; \alpha)}(y - \mu)/\sigma^2 \\ \frac{1 - r}{e(x; \alpha)}((y - \mu)^2/\sigma^2 - 1)/2\sigma^2 \end{pmatrix}$$

Thus

$$V = \mathbb{E}\{u(x, \theta)u(x, \theta)'\}$$

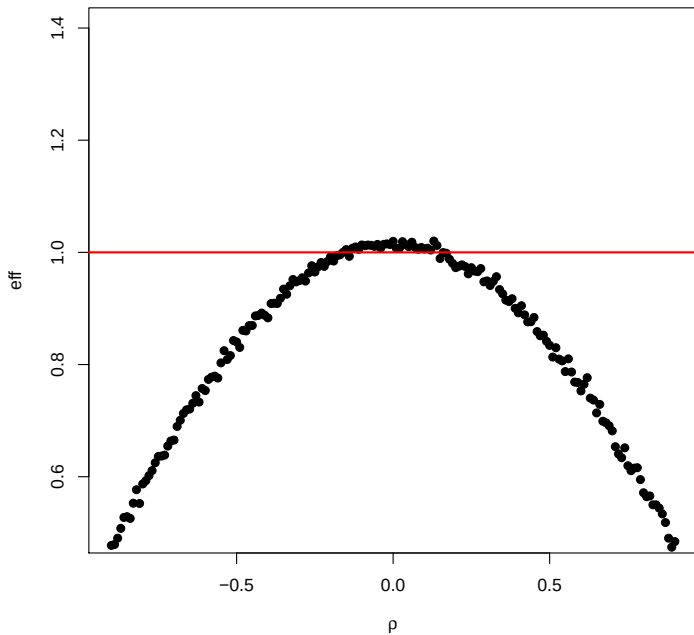
and

$$W = \mathbb{E}\{u(x, \theta)s(x, \theta)'\}$$

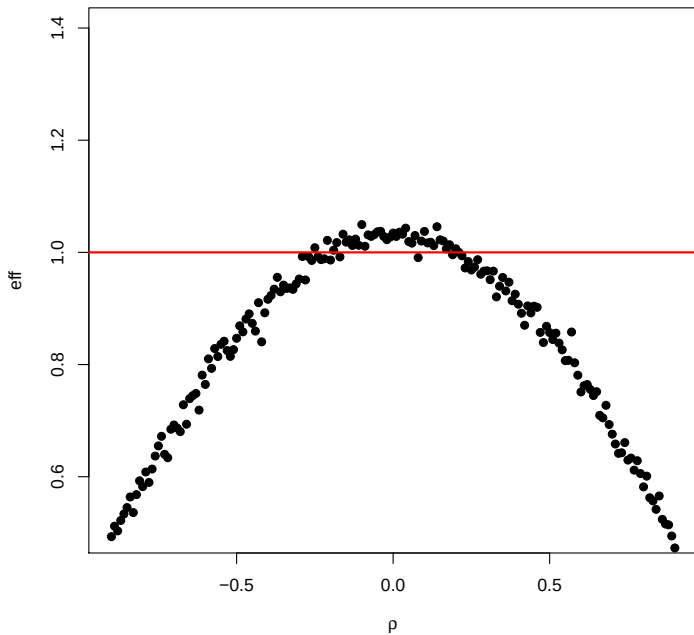
can be estimated using Monte Carlo, yielding

$$W^{-1}V \{W^{-1}\}'.$$

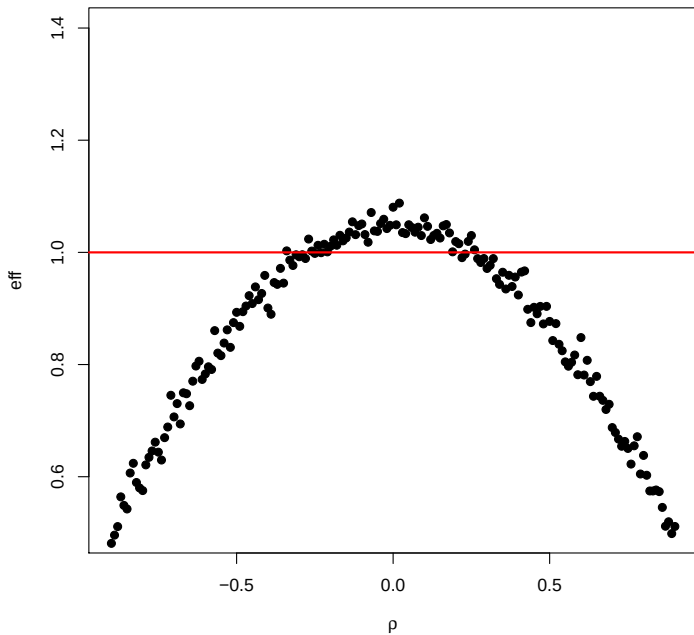
Simulation: $n = 10000$, with $(1, 1)$



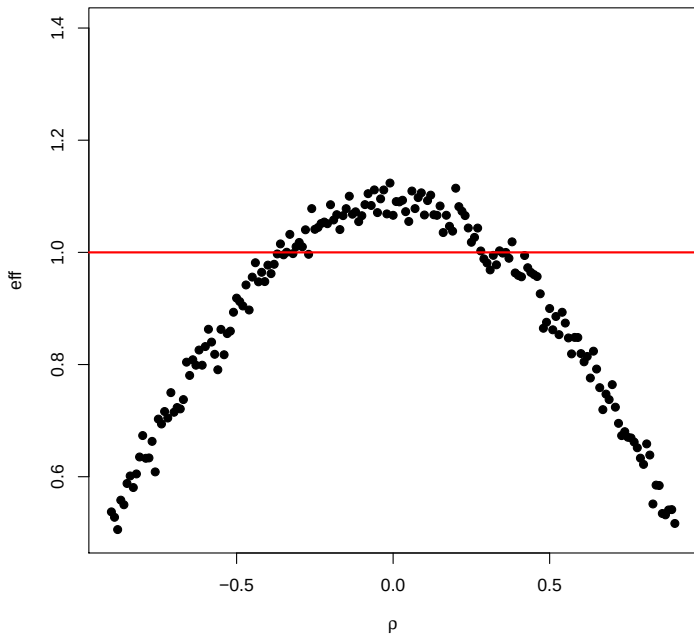
Simulation: $n = 1000$, with $(1, 1)$



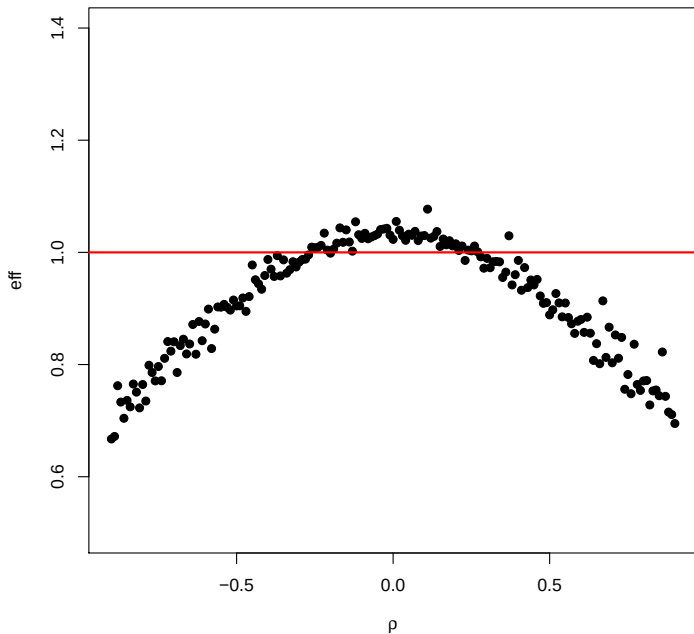
Simulation: $n = 500$, with $(1, 1)$



Simulation: $n = 200$, with $(1, 1)$



Simulation: $n = 200$, with $(-1, -1)$.



Concluding Remarks

- ▶ Some comments on influence functions $u^*(x, \theta)$
- ▶ Paradox does not occur when $u(x, \theta)$ is the efficient estimating function.