Asymptotic efficiency of estimates for models with incidental nuisance parameters

Helmut Strasser[*]

November, 1992

I would like to speak about the estimation of a parameter θ if a second parameter η is a nuisance parameter changing from observation to observation.

The starting point is a paper by Pfanzagl which will appear in the Annals of Statistics. In this paper Pfanzagl considers bounds for the asymptotic risk of estimators which are valid under mixtures of probability measures. The question is discussed whether those bounds remain valid if individual sequences of nuisance parameters are considered.

Let us begin with some notation.

$(X, \mathcal{A})$ denotes a sample space and $(P_{\theta,\eta} : \theta \in \Theta, \eta \in H)$ is a family of probability measures on $\mathcal{A}$. θ is the parameter to be estimated and η is the nuisance parameter.

Let $\Gamma|\mathcal{B}(H)$ be a prior distribution for the nuisance parameter. We denote the mixtures by

$$Q_{\theta,\Gamma}(A \times B) = \int_B P_{\theta,\eta}(A)\,\Gamma(d\eta),\ A \in \mathcal{A},\ B \in \mathcal{B}(H),$$

and marginals of the mixtures by

$$Q'_{\theta,\Gamma} = Q_{\theta,\Gamma}|\mathcal{A}.$$

The paper by Pfanzagl contains two main results. The first result deals with the asymptotic behaviour of estimators under individual sequences of random nuisance parameters, and the second result is a counterexample.

Let me give you an impression of the first result by Pfanzagl. It deals with asymptotic linear estimator sequences.

Let $K(x, \theta)$ be a kernel satisfying

$$\int K(.,\theta)\,dQ_{\theta,\Gamma} = 0,\ \int K(.,\theta)^2\,dQ_{\theta,\Gamma} =: \sigma_1^2(\theta) < \infty.$$

An estimator sequence (T_n) is asymptotically linear with influence function K if

$$\sqrt{n}(T_n - \theta) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} K(x_i, \theta) \to 0 \quad \left\{ \begin{array}{ll} (Q'^{n}_{\theta,\Gamma}) & (*) \\ Q'^{\mathbb{N}}_{\theta,\Gamma} - \text{a.e.} & (**) \end{array} \right.$$

[*] Scientific talk given at the meeting on Asymptotic Methods in Statistics in Oberwolfach, 1992

There are two versions of this property depending on the type of convergence of the residual term.

What is clear is the following assertion: The distributions

$$\mathcal{L}\Big(\sqrt{n}(T_n - \theta)\Big|(Q'^n_{\theta,\Gamma})\Big) \to \nu_{0,\sigma_1^2(\theta)} \text{ weakly}$$

of the estimator sequence converge weekly to a certain normal distribution.

Now, let (Y_i) be a sequence of random nuisance parameters which is i.i.d. according to Γ. The question arises what happens with the distributions

$$\mathcal{L}\Big(\sqrt{n}(T_n - \theta)\Big|\bigotimes_{i=1}^{n} P_{\theta,Y_i}\Big) \ ?$$

An answer is given by the result of Pfanzagl. Roughly speaking, the assertion is as follows:

0.1 THEOREM. *(Pfanzagl, 1992) Let*

$$\mu_n(\theta) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int K(.,\theta)\, dP_{\theta,Y_i},\ \sigma_2^2(\theta) := \int [E_{\theta,\eta}(K(.,\theta))]^2\Gamma(d\eta).$$

Then, if (T_n) *satisfies* $(**)$,

$$\mathcal{L}\Big(\sqrt{n}(T_n - \theta)\Big|\bigotimes_{i=1}^{n} P_{\theta,Y_i}\Big) \sim \nu_{\mu_n(\theta),\sigma_1^2(\theta)-\sigma_2^2(\theta)}$$

The tilde can be made precise in many ways. One possibility is that the difference between the distribution functions converges to zero $\Gamma^{\mathbb{N}}$-a.e.

The message is that the asymptotic variance of (T_n) is strictly smaller than σ_1^2 but there are random fluctuations of the mean. Pfanzagl's result is valid for an abstract space of nuisance parameters $\eta \in H$.

I would like to show you an approach to Pfanzagl's theorem which is very different from Pfanzagl's own method. My approach has advantages and disadvantages.

Its main advantage is that the whole problem can be embedded into the usual framework of the local asymptotic theory of statistics. Thus, as we will see later, the machinery of obtaining asymptotic risk bounds can be applied.

The disadvantage is that at least so far the method I'm going to present can only be applied to one-dimensional nuisance parameters which a continuous distribution function. You will see soon, why.

Thus, my approach is not able to replace Pfanzagls's proofs which can be applied to more general situations. But it shows some light into the deeper structure of the problem.

Let us indicate what we want to find.

Given the probability measures $\otimes_{i=1}^{n} P_{\theta+s/\sqrt{n},\eta_i}$ for an individual sequence (η_i) of nuisance parameters we would like to have probability measures $\otimes_{i=1}^{n} R_{n,i}$, say, such that the likelihood ratios

$$\frac{d\bigotimes_{i=1}^{n} P_{\theta+s/\sqrt{n},\eta_i}}{d\bigotimes_{i=1}^{n} R_{n,i}}$$

can be expanded in some sense, for example leading to some LAN-property.

If we take in the denominator a product of i.i.d. components then an LAN-property could be proved but for a class of sequences (η_i) which is too small to cover cases of random nuisance parameters. In order to cover the case of random nuisance parameters we have to choose that the denominator more subtle.

The main idea is as follows.

Let (η_i) be a sequence of nuisance parameters. We assume that the nuisance parameters take their values in the unit interval $\eta_i \in [0,1]$ and that they are uniformly distributed. The case of random nuisance parameters with a continuous distribution function can be reduced to that case.

Let $\eta_{n:1}, \eta_{n:2}, \ldots, \eta_{n:n}$ be the order statistics of $\eta_1, \eta_2, \ldots, \eta_n$. These order statistics can be decomposed in the following way:

$$\eta_{n:i} = \frac{i}{n+1} + \frac{1}{\sqrt{n}}\Big(\sqrt{n}\big(\eta_{n:i} - \frac{i}{n+1}\big)\Big).$$

We center the order statistics at their expectation and rescale the residuals by $\sqrt{n}$. For notational convenience let us write

$$\eta_{n:i} = \tau_{ni} + \frac{1}{\sqrt{n}} t_{ni}.$$

This decomposition is applied in the following way.

Let (T_n) be a sequence of permutation invariant estimators. Then we may write its distribution as

$$\begin{aligned}
\mathcal{L}\Big(\sqrt{n}(T_n-\theta)\Big|\bigotimes_{i=1}^{n} P_{\theta+s/\sqrt{n},\eta_i}\Big) \\
&= \mathcal{L}\Big(\sqrt{n}(T_n-\theta)\Big|\bigotimes_{i=1}^{n} P_{\theta+s/\sqrt{n},\eta_{n:i}}\Big) \\
&= \mathcal{L}\Big(\sqrt{n}(T_n-\theta)\Big|\bigotimes_{i=1}^{n} P_{\theta+s/\sqrt{n},\tau_{ni}+t_{ni}/\sqrt{n}}\Big)
\end{aligned}$$

What we need is an LAN-property for the likelihood ratios

$$\frac{d\bigotimes_{i=1}^{n} P_{\theta+s/\sqrt{n},\tau_{ni}+t_{ni}/\sqrt{n}}}{d\bigotimes_{i=1}^{n} P_{\theta,\tau_{ni}}}$$

Such an LAN-property can in fact be obtained, at least for triangular arrays (t_{ni}) satisfying a certain compactness condition. We will see later that in case of random nuisance parameters the arrays satisfy this compactness condition with large probability.

Let (t_{ni}) be an arbitrary triangular area and define step functions

$$t_n(\eta) := \sum_{i=1}^{n} t_{ni} 1_{((i-1)/n, i/n]}.$$

We define

0.2 DEFINITION. The array (t_{ni}) is relatively compact if the sequence of step functions (t_n) is relatively compact in $L^2([0,1])$.

In order to state the LAN-property we need additional notations.

Let $P_{\theta,\eta} \ll \mu$ and $dP_{\theta,\eta}/d\mu =: p(x,\theta,\eta)$. Then we denote

$$\ell^{(1)}(x,\theta,\eta) = \frac{\partial}{\partial\theta}\log p(x,\theta,\eta),\ \ell^{(2)}(x,\theta,\eta) = \frac{\partial}{\partial\eta}\log p(x,\theta,\eta)$$

and

$$g_\theta(x,\eta,s,t) := s\cdot\ell^{(1)}(x,\theta,\eta) + t\cdot\ell^{(2)}(x,\theta,\eta).$$

Of course, we suppose the validity of some regularity conditions which are not specified in detail.

Now, we are in the position to state the LAN-property.

0.3 THEOREM. *Let (t_{ni}) be relatively compact. Then*

$$\frac{d\bigotimes_{i=1}^{n} P_{\theta+s/\sqrt{n},\tau_{ni}+t_{ni}/\sqrt{n}}}{d\bigotimes_{i=1}^{n} P_{\theta,\tau_{ni}}}$$
$$= \exp\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_\theta(x_i,\tau_{ni},s,t_{ni}) - \frac{1}{2}\int\int g_\theta^2(x,y,s,t_n(\eta))\,dP_{\theta,\eta}(x,y)\,d\eta + r_n(x,\theta,s,t_n)\Big)$$

where

$$\sup_{|s|\le a}|r_n(.,\theta,s,t_n)| \to 0\ \Big(\bigotimes_{i=1}^{n} P_{\theta,\tau_{ni}}\Big)$$

Now, let us show how to apply the LAN-property to the asymptotic behaviour of estimator sequences.

Let (T_n) be a permutation invariant estimator sequence being asymptotically linear with residuals satisfying $(*)$. We do not require the stronger condition $(**)$ used by Pfanzagl, but we have to assume that (T_n) consists of permutation invariant estimators. The reason is that for permutation invariant functions (r_n) the LAN property implies

$$r_n \to 0\ (Q'^{n}_{\theta,\lambda})\ \Rightarrow\ (r_n)\to 0\ \Big(\bigotimes_{i=1}^{n} P_{\theta,\tau_{ni}}\Big).$$

(the measures are contiguous on the permutation invariant sigma fields).

From asymptotic linearity it follows that

$$\mathcal{L}\Big(\sqrt{n}(T_n-\theta)\Big|\bigotimes_{i=1}^{n} P_{\theta,\tau_{ni}}\Big) \to \nu_{0,\sigma_1^2(\theta)-\sigma_2^2(\theta)}$$

The asymptotic variance term is the same as in Pfanzagl's theorem.

And now we apply the LAN-property together with LeCam's third lemma and obtain

0.4 THEOREM. *Let* $(t_{ni}$ *be relatively compact. Then*

$$\mathcal{L}\Big(\sqrt{n}(T_n-\theta)\Big|\bigotimes_{i=1}^{n} P_{\theta+s/\sqrt{n},\tau_{ni}+t_{ni}/\sqrt{n}}\Big) \to \nu_{\mu_n(\theta),\sigma_1^2(\theta)-\sigma_2^2(\theta)}$$

where

$$\mu_n(\theta) = s\int (J_1(\theta,\eta)\,d\eta + \int t_n(\eta)J_2(\theta,\eta)\,d\eta$$

with

$$J_i(\theta,\eta) = \int (K(.,\theta)\ell^{(i)}(.,\theta,\eta)\,dP_{\theta,\eta},\ i=1,2,$$

being the inner products between the influence function of (T_n) *and the partial derivatives of the log likelihood functions.*

This is the version of Pfanzagl's theorem stated for deterministic arrays of nuisance parameters. Of course, such a result is of particular interest if it is possible to obtain the case of stochastic nuisance parameters as a special case.

Let us show how to do this.

Let (Y_i) be a sequence of random nuisance parameters and note by (F_n) the empirical distribution functions. Assume that the standardised empirically processes

$$\sqrt{n}(F_n(\eta)-\eta)$$

converge weakly to the Brownian Bridge (X_η) with paths in $D([0,1])$. (E.g.: Let let (Y_i) be i.i.d and uniformly distributed.)

The order statistics $(Y_{n:i})$ can be decomposed as

$$Y_{n:i} = \frac{i}{n+1} + \frac{1}{\sqrt{n}}\underbrace{\Big(\sqrt{n}(Y_{n:i} - \frac{i}{n+1})\Big)}_{t_{ni}},$$

and what we have to do is to check whether the arrays (t_{ni}) are relatively compact.

For this, we consider the step functions

$$\begin{aligned} t_n(\eta) &= \sum_{i=1}^{n} t_{ni}1_{((i-1)/n,i/n]}(\eta) \\ &= \sum_{i=1}^{n}\Big(\sqrt{n}(Y_{n:i} - \frac{i}{n+1})\Big)1_{((i-1)/n,i/n]}(\eta) \\ &= \sqrt{n}\Big(Y_{n:[n\eta]+1} - \frac{[n\eta]+1}{n+1}\Big) \end{aligned}$$

and observe that these are not nothing else than the quantile processes of (Y_i). It is known that the quantile processes converge weakly to a Brownian Bridge (X_η) in $D([0,1])$ and therefore are concentrated on compact sets with high probability.

Thus, we obtain the following result for random nuisance parameters.

0.5 THEOREM. *Let (T_n) be a sequence of permutation invariant estimatros satisfying the asymptotic linearity condition. Suppose the (Y_i) is a sequence of random nuisance parameters who's quantile processes converge weekly to a Browning Bridge (X_η). Then for every bounded loss function W the distribution of*

$$\int W\big(\sqrt{n}(T_n - \theta)\big)\, d \bigotimes_{i=1}^{n} P_{\theta+s/\sqrt{n},Y_i}$$

converge weakly to the distributions of

$$\int W\Big(x + s\int J_1(\theta,\eta)\,d\eta - \int X_\eta J_2(\theta,\eta),\, d\eta\Big)\,\nu_{0,\sigma_1^2(\theta)-\sigma_2^2(\theta)}(dx).$$

This is nothing else then another version of Pfanzagl's theorem. It states that for individual sequences of stochastic nuisance parameters the asymptotic variance is smaller than for mixtures of probability measures but the distributions of the estimators are centerd around random fluctuations.

The second topic of Pfanzagl's paper is concerned with asymptotic bounds for the risk of estimators.

The following is well known.

Consider the families of probability measures which are called the full mixture model

$$\Big(Q_{\theta+s/\sqrt{n},(1+k/\sqrt{n})\lambda} : s \in \mathbb{R},\; k \in L^2([0,1])\Big).$$

Then there exists an influence function $\hat{K}(.,\theta)$ (which is called the canonical gradient for the estimation of θ) such that under the mixture model the estimator sequence (T_n) with influence function $\hat{K}$ is minimax in the Hajek-LeCam sense.

Let us state result on the minimax bound as an admissibility assertion. For this we have to restrict ourselves to the case $\theta \in \mathbb{R}$.

0.6 THEOREM. *Let (T_n) be an estimator sequence which satisfies the asymptotic linearity condition with the optimal influence function $\hat{K}$:*

$$\sqrt{n}(T_n - \theta) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{K}(x_i,\theta) \to 0 \;(Q'^n_{\theta,\lambda})$$

Let W be a loss function satisfying the usual conditions and let

$$\beta_W := \int W(x)\,\nu_{0,\sigma_1^2(\theta)}(dx).$$

Then the following assertions are valid:

(1) If an estimator sequence (S_n) is not worse than β_W, i.e.

$$\limsup_{n\to\infty} \int W(\sqrt{n}(S_n - \theta) - s)\, dQ'^n_{\theta+s/\sqrt{n},(1+k/\sqrt{n})\lambda} \le \beta_W \text{ for all } s, k,$$

then (S_n) is not better than β_W, i.e.

$$\liminf_{n\to\infty} \int W(\sqrt{n}(S_n - \theta) - s)\, dQ'^{n}_{\theta+s/\sqrt{n},(1+k/\sqrt{n})\lambda} \geq \beta_W \text{ for all } s, k.$$

(2) The sequence (T_n) attains the bound β_W:

$$\lim_{n\to\infty} \int W(\sqrt{n}(T_n - \theta) - s)\, dQ'^{n}_{\theta+s/\sqrt{n},(1+k/\sqrt{n})\lambda} = \beta_W \text{ for all } s, k.$$

This is well known.

In his paper Pfanzagl considers the problem whether the bound β_W remains valid if we consider individual sequences of stochastic nuisance parameters.

At first sight things look promising.

Let (T_n) be an estimator sequence which satisfies the asymptotic linearity condition with the influence function $\hat{K}$ which is optimal for the full mixture model. Then it follows from Pfanzagl's theorem that

$$\lim_{n\to\infty} \int W(\sqrt{n}(T_n - \theta) - s)\, d\bigotimes_{i=1}^{n} P^n_{\theta+s/\sqrt{n},Y_i} = \beta_W\ (\lambda^n).$$

The reason is that for the full mixture model the optimal influence function $\hat{K}$ is not correlated with $\ell^{(2)}(.,\theta,\eta)$. Hence the term $\sigma_2^2(\theta)$ is zero and the stochastic fluctuation of the mean does not exist.

But, although the bound for the mixture model is attained even for individual sequences of nuisance parameters, a counter example by Pfanzagl shows that the bound β_W is not a strict bound for the risks if individual sequences of nuisance parameters are considered.

The counter example by Pfanzagl is as follows.

There exists a sequence of estimators (T_n) which satisfies the asymptotic linearity condition with $\hat{K}$ and hence satisfies

$$\lim_{n\to\infty} \int W(\sqrt{n}(T_n - \theta) - s)\, d\bigotimes_{i=1}^{n} P^n_{\theta+s/\sqrt{n},Y_i} = \beta_W\ \lambda^{\mathbb{N}} - \text{a.e..}$$

but

$$\lim_{n\to\infty} \int W(\sqrt{n}(T_n - \theta) - s)\, d\bigotimes_{i=1}^{n} P^n_{\theta+s/\sqrt{n},\eta_i} < \beta_W$$

for countably many individual sequences (η_i) of nuisance parameters. The sequence (η_i) can even be chosen such that the empirical distribution functions converge to the distribution function of λ.

In view of this counter example only the hope remains to show that the set of sequences (η_i) where such a superefficiency can happen is small in some sense. I would like to show that this is indeed the case.

The basic result is that the LAN-property for individual sequences can be stated if the probability measures are restricted to the symmetric sigma fields.

For any function $f : X \times \Theta \times [0,1] \to \mathbb{R}$ we abbreviate

$$\bar{f}(x,\theta) := E_{Q_{\theta,\lambda}}(f(.,\theta,.)|\mathcal{A}) = \frac{\int f(x,\theta,\eta)p(x,\theta,\eta)\,d\eta}{\int p(x,\theta,\eta)\,d\eta}.$$

Let

$$\bar{g}_\theta(x,s,t_n) = s \cdot \bar{\ell}^{(1)}(x,\theta) + \overline{(t_n \cdot \ell^{(2)})}(x,\theta)$$

where $t_n \in L^2([0,1])$. Moreover, let $\mathcal{B}_n \subseteq \mathcal{A}^n$ be the sub-sigma field of permutation invariant sets. Then we have the following theorem.

0.7 THEOREM. *Let (t_{ni}) be relatively compact. Then*

$$\frac{d\bigotimes_{i=1}^n P_{\theta+s/\sqrt{n},\tau_{ni}+t_{ni}/\sqrt{n}}\Big|\mathcal{B}_n}{d\bigotimes_{i=1}^n P_{\theta,\tau_{ni}}\Big|\mathcal{B}_n}$$
$$= \exp\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n \bar{g}_\theta(x_i,s,t_n) - \frac{1}{2}\int \bar{g}_\theta^2(.,s,t_n)\,dQ_{\theta,\lambda} + r_n(x,\theta,s,t_n)\Big)$$

where

$$\sup_{|s|\le a}|r_n(.,\theta,s,t_n)| \to 0 \,\Big(\bigotimes_{i=1}^n P_{\theta,\tau_{ni}}\Big)$$

This result gives us the possibility to apply the results of the local asymptotic decision theory to permutation invariant estimators.

In general, we obtain canonical gradients for the estimation of θ which are different from the gradients obtained in the full mixture model. But we obtain the same canonical gradients if we are dealing with models of a special structure.

Let us assume that

$$p(x,\theta,\eta) = q(x,\theta)p_0(T(x),\theta,\eta),$$

where T is sufficient for η and $(P_{\theta,\eta} : \eta \in H)$ is complete for T. This structure is valid for example for the case considered in the famous paper by Neyman and Scott. In such a case the canonical gradients for the estimation of θ coincide for the full mixture model and the symmetric model with individual nuisance parameters.

Thus, we arrive at the following assertion.

0.8 THEOREM. *Let S_n be a permutation invariant estimator sequence. If for all relatively compact arrays (t_{ni}) and every s*

$$\limsup_{n\to\infty}\int W(\sqrt{n}(S_n-\theta)-s)\,d\bigotimes P_{\theta+s/\sqrt{n},\tau_{ni}+_{ni}/\sqrt{n}} \le \beta_W$$

then equality holds, that is

$$\liminf_{n\to\infty}\int W(\sqrt{n}(S_n-\theta)-s)\,d\bigotimes P_{\theta+s/\sqrt{n},\tau_{ni}+_{ni}/\sqrt{n}} \ge \beta_W$$

for all relatively compact arrays (t_{ni}) *and all* s.

Now, what is the relation of this assertion to Pfanzagl@s counter example ?

The point is that Pfanzagl's exceptional sequences (η_i) do not give rise to relatively compact arrays (t_{ni}). But we know that for i.i.d. random nuisance parameters the arrays (t_{ni}) are a relatively compact with great probability.

This gives the following final result.

0.9 THEOREM. *Let* (Y_i) *be a sequence of random nuisance parameters such that the empirical processes* $\sqrt{n}(F_n(\eta - \eta)$ *converge to a Brownian Bridge. Let* (S_n) *be a permutation invariant estimator sequence.*

If for all s *and all* $\epsilon > 0$

$$P\Big(\int W(\sqrt{n}(T_n - \theta) - s)\, d\bigotimes_{i=1}^{n} P^n_{\theta+s/\sqrt{n},Y_i} \geq \beta_W + \epsilon\Big) \to 0$$

then for all s *and all* $\epsilon > 0$

$$P\Big(\int W(\sqrt{n}(T_n - \theta) - s)\, d\bigotimes_{i=1}^{n} P^n_{\theta+s/\sqrt{n},Y_i} \leq \beta_W - \epsilon\Big) \to 0$$

This assertion shows that superefficiency is seldom but it does not exclude counter examples of Pfanzahl's type.

To finish, I would like to stress that the connection between random nuisance parameters and triangular arrys can only be treated in the way I did, if the random nuisance parameter is of dimension one. I'm afraid that the idea can not be extended to greater dimension.