The Annals of Statistics
1984, Vol. 12, No. 2, 786–791

BOOK REVIEW

J. PFANZAGL WITH THE ASSISTANCE OF W. WEFELMEYER, *Contributions to a General Asymptotic Statistical Theory. Springer Lecture Notes in Statistics* **13,** 1982, vii + 315 pages, $16.80.

Review by P. J. BICKEL

University of California, Berkeley

In this monograph Pfanzagl has made an important contribution to asymptotic estimation and testing theory in nonparametric models. The main questions he addresses are the following:

Consider models according to which we observe $X_1, \cdots, X_n$ which take values in a sample space $\mathscr{X}$ and are independent and identically distributed according to $P \in \mathscr{P}$.

1) How well can we (asymptotically) estimate a Euclidean parameter $K(P)$?

Here $K : \mathscr{P} \to R^m$ for some m.

2) How well can we test hypotheses of the form $H : K(P) = c$, $P \in \mathscr{P}$?

If $\mathscr{P}$ is "parametric", $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$, Θ Euclidean, $\theta \to P_\theta$ smooth, the answers are standard.

Pfanzagl develops a method introduced by Koshevnik and Levit (1976), itself based on an old idea of Stein's (1956), for obtaining "information bounds" in "nonparametric" or what one might call semiparametric models. Here are a few examples of such models; a wealth of others can be found in Chapters 2, 14–18 of Pfanzagl.

a) The symmetric location model: The parameter of interest $K(P)$ here is the centre of symmetry.
b) The linear regression model with stochastic independent variables and i.i.d. but not necessarily normally distributed errors. The parameters of interest are the regression coefficients other than the constant.
c) The Cox (1972) regression model with time independent covariates. Here stochastic independent variables (covariates) and survival times are observed, the latter possibly with censoring independent of the survival time given the covariates. The hazard rate of a survival time (precensoring) given the covariates c is given by

$$\lambda(t \mid c) = \exp\{c^T\beta\}\lambda_0(t)$$

where λ_0 is an unknown fixed hazard rate. The parameter β is of interest.

As these examples suggest, the models of interest are described through a parametrization $(\theta, G) \to P_{(\theta,G)}$ where θ is Euclidean and G ranges over an abstract space, typically a set of probability distributions on some space. The parameters

Received November 1983.

$\theta(P)$ are defined implicitly through identifiability rather than being given explicitly as a function $\theta : \mathscr{P} \to R^m$. Typically, these models can be viewed as generalizations of classical parametric models $\{P_{(\theta,G_0)}\}$ with G_0 assumed known. In these cases a question appears naturally:

3) Can we do as well not knowing G_0 as knowing it?

If we can we are dealing with the phenomenon of adaptation (see Bickel, 1982, for example). [Pfanzagl dismisses this question as a special case of question 1. This is, of course, true. Before going on to praise the book further, I must note that Pfanzagl's polemical discussion of this question, the issues of robustness, and testing against multidimensional alternatives put at least this reader off.] These unnecessary polemics, Pfanzagl's difficult notation, e.g., $P(X)$ for the expectation of X under P, and a choice of metric on $\mathscr{P}$ which complicates the exposition, makes this excellent work less accessible than it should be.

Pfanzagl's presentation falls into three parts.

1) Development of the concepts leading up to lower bounds on the efficiency of estimation and the probability of type II error of tests in these general models (Chapters 1–9);

2) a largely heuristic discussion of methods of achieving these bounds in estimation and testing (Chapters 10–12);

3) a large number of examples in which the bounds are computed. In essentially all of these cases the bounds are shown to be sharp through existence of well known procedures achieving them. The heuristics of Chapters 10–12 play only a small role here (Chapters 13–18).

The key ideas (introduced by Koshevnik and Levit) used in developing the lower bounds are the tangent cone at P_0 in $\mathscr{P}$ and the canonical gradient of K. The tangent cone $\mathscr{T}$ is a subset of $L_2(P_0)$ defined by: $h \in \mathscr{T} \Leftrightarrow$ there exists a mapping (called a path) from [0, 1] to $L_2(P_0)$, $t \to dP_t/dP_0$ such that $\| (dp_t/dP_0) - 1 - th \|_o = o(t)$ as $t \downarrow 0$ where $(\cdot, \cdot)_o$, $\| \cdot \|_o$ are the Hilbert inner product and norm in $L_2(P_0)$. This definition implicitly supposes that P_t is dominated by P_0 and $dP_t/dP_0 \in L_2(P_0)$. Pfanzagl weakens the latter requirement and broadens his notion of path. However, as Le Cam (1983) shows, the set of all paths obtained using the weaker definition is just the set of all $h = 2g$ where $g\sqrt{dP_0}$ is the (Hellinger) derivative at 0 of the mapping $t \to \sqrt{dP_t}$ from [0, 1] to the Hilbert space of equivalence classes $\xi\sqrt{dQ}$, $\xi \in L_2(Q)$ where

$$\text{a)} \quad \xi\sqrt{dQ} \equiv \eta\sqrt{dR} \Leftrightarrow \xi \sqrt{\frac{dQ}{d(Q+R)}} = \eta \sqrt{\frac{dR}{d(Q+R)}} \quad \text{a.e.} \quad Q + R$$

$$\text{b)} \quad (\xi\sqrt{dQ}, \eta\sqrt{dR}) = \int \xi\eta \sqrt{\frac{dQ}{d(Q+R)}} \sqrt{\frac{dR}{d(Q+R)}}\, d(Q+R).$$

Pfanzagl's choice of local metric $\| \cdot \|_o$ leads to awkwardnesses in definitions in Chapter 7, and long and tedious discussions of approximations by global metrics

such as the Hellinger metric in Chapter 6. Use of the Hellinger metric and/or the local metric $2 \| (dQ/dP)^{1/2} - 1 \|_o$ greatly simplifies the discussion.

If $\mathscr{P}$ is parametric, $\mathscr{P} = \{P_\theta : \theta \in R^k\}$, $\mathscr{T}$ is (under the usual regularity conditions) the linear span in $L_2(P_0)$ of $(\partial/\partial\theta_i)\log p(X_1, \theta^0)$, $i = 1, \cdots, k$, where $p(\cdot, \theta)$ is the density of P_θ with respect to $\mu \gg \mathscr{P}$ and $P_0 = P_{\theta_0}$. The tangent cone in general corresponds essentially to the set of all $(\partial/\partial t)\log p_t(X_1) \mid_{t=0}$ where p_t is the density of P_t and $\{P_t\}$ ranges over smooth 1-dimensional subfamilies of $\mathscr{P}$ with P_0 as an endpoint. Typically the tangent cone is a closed linear space as above and is then referred to as a tangent space. The tangent space is an essential component of the differential geometric structure of $\mathscr{P}$ as studied by Amari (1982), following Efron (1975).

The 1-dimensional parameter K has a gradient $\dot{K}(P_0) \in L_2(P_0) \Leftrightarrow$ for all paths $\{P_t\}$,

$$K(P_t) = K(P_0) + \left(\dot{K}(P_0), \frac{dP_t}{dP_0} - 1\right)_o + o\left(\left\| \frac{dP_t}{dP_0} - 1 \right\|_o\right)$$

$$= K(P_0) + t(\dot{K}(P_0), h_o) + o(t)$$

where

$$\left\| \frac{dP_t}{dP_0} - 1 - th_0 \right\|_o = o(t).$$

The gradient is in general not unique. Evidently if $\dot{K}(P_0)$ is a gradient so is its projection onto the tangent space, which we shall refer to as $\tilde{K}$.

The lower bounds of Chapters 8 and 9 can now be expressed as follows: For estimates $\hat{K}$ of K such that $\Delta_n = \sqrt{n}(\hat{K} - K(P_t))$ converge in law uniformly on paths $\{P_t\}$ the distribution of Δ_n under P_0 is at least as dispersed as $N(0, \| \tilde{K}(P_0) \|_o^2)$. The power of tests of $H : K(P) = K(P_0)$ vs. $K(P) = K(P_0) + (t/\sqrt{n})$ is, for $t \geq 0$, under regularity conditions on $\mathscr{P}$ and H, no larger than $1 - \Phi(z - t \| \tilde{K}(P_0) \|^{-1}) + o(1)$ where the level of the test is $1 - \Phi(z) + o(1)$.

In smooth parametric models and in the presence of identifiability, these bounds are classical and attained. In the general case, construction of procedures achieving these bounds and even their achievement requires considerably more study.

Pfanzagl gives two lines of attack on the problem of estimating $K(P)$ for K having a gradient.

i) Say $\tilde{P}_n \in \mathscr{P}$, a sequence of estimates of P, is asymptotically efficient if, for all f in the tangent space of P,

$$\int f(x)\tilde{P}_n(dx) = n^{-1} \textstyle\sum_{i=1}^n f(X_i) + o_p(n^{-1/2}).$$

Then for K differentiable in a strong sense and $(d\tilde{P}_n)/dP$ sufficiently close to 1 in $\| \cdot \|_o$, $K(\tilde{P}_n)$ is asymptotically efficient, i.e., achieves the lower bound (Theorem 11.2.1).

Unfortunately the construction of $\tilde{P}_n$ in other than smooth parametric models

is unclear. One possibility is to choose $\tilde{P}_n \in \mathscr{P}$ to minimize Hellinger distance bewteen the empirical distribution $\hat{P}_n$ and $\mathscr{P}$. Pfanzagl gives (in Theorem 10.4.8 as corrected below) conditions under which such a method yields an asymptotically efficient $\tilde{P}_n$. Unfortunately even the weakened conditions of Remark 10.4.11 do not hold in a case as simple as $\mathscr{P}$ = {absolutely continuous symmetric distributions}—see the discussion in Section 15.2 for instance.

ii) Given a reasonable estimate $\tilde{P}_n$ of P, $\tilde{P}_n \in \mathscr{P}$, use a 1-step Newton iteration,

$$K(\tilde{P}_n) + (1/n) \sum_{i=1}^{n} \tilde{K}(X_i, \tilde{P}_n)$$

where $\tilde{K}(\cdot, \tilde{P}_n) = \tilde{K}(\tilde{P}_n)(\cdot)$.

This method has been shown to work in special cases under suitable conditions by Levit (1975), Ibragimov-Hasminskii (1979) and Bickel (1982).

The last group of chapters giving examples is in many ways the most novel and interesting. Here are what I view as the most interesting results:

CHAPTER 14. The most important topic here is estimation of parameters in the presence of unknown nuisance parameters which vary stochastically from observation to observations. These "mixture models" which go back to Neyman and Scott (1948) have been treated by E. Andersen (1973) and others, most recently by B. Lindsay (1983). Pfanzagl's main result here (earlier obtained by Godambe, 1976, Theorem 3.2 in a different setting) is that if there is a complete sufficient statistic T for the nuisance parameter with the parameter of interest fixed, then conditional inference (given T) is asymptotically efficient. For the Neyman-Scott model, this was shown by Lindsay and independently by Hammerstrom (1978).

CHAPTER 15. This chapter deals with symmetric probability measure models and contains an application of method (i) of obtaining an asymptotically efficient estimate of P. The treatment is essentially heuristic and rather unsatisfactory in view of the results of Beran (1974) and Stone (1975).

CHAPTER 16. The unsurprising key result here is that the product of asymptotically efficient estimates of the marginal distributions is the asymptotically efficient estimate of P if we know P to be a product measure.

CHAPTER 17. This chapter deals with independence-dependence problems. The most striking result is that, in testing for independence in models such as that of Bhuchongkul (1964), the upper bound to the power is the same whether we assume the distributions of the variables under the hypothesis known up to a change of location and scale or completely unknown. In principle, adaptation is possible here.

CHAPTER 18. This chapter dealing with two sample (or rather one bivariate sample) problems contains one particularly interesting result.

SECTION 18.5. An extension to group models of a result of Stein (1956) that if the second sample is obtained from the first by shift and change of scale, estimation of both changes should be as easy (on the basis of lower bounds) with the population shape of the first sample unknown as with it known up to a change of location and scale.

In this chapter, Pfanzagl also gives the lower bound for estimation of the ratio of hazard rates in the proportional hazards model. This calculation due to Begun and Wellner (1983a) has now been superseded in Begun et al. (1983b) which presents some interesting additional geometry.

Pfanzagl views this book as a basis for a unified asymptotic statistical theory. This is, I think, too ambitious. Such a basis, in a much more general context, is to be found in Le Cam's forthcoming monumental treatise (to appear). However, I believe at least elements of the Pfanzagl (Koshevnik-Levit) approach will prove important in the development of theory, at a level where it can be immediately applied to a wide range of examples. Topics which I see as requiring much greater development are:

1) extension of the geometry to nonidentically distributed as well as dependent observations, e.g., regression models with fixed covariate values, Cox regression models with time varying covariates and censoring;

2) construction of efficient procedures including a thorough examination of "nonparametric" maximum likelihood and the method of sieves and regularization (Grenander, 1981);

3) calculation of the bounds (and of efficient procedures) in further important examples;

4) uniformity of convergence of efficient estimates. This topic is touched on briefly in Section 9.4. Attainment of the bounds at any point of the parameter space is not enough. If convergence is not uniform in reasonable neighborhoods of the point, a sample size guaranteeing reasonable results cannot really be specified.

5) Can efficiency in this sense and robustness be reconciled?

6) How do such procedures behave for moderate size data sets? In some cases, e.g. the Cox estimate in censored regression, some simulation studies are available, in others, e.g. adaptive estimation of location, very little is available.

In a forthcoming monograph based in part on the Mathematical Sciences Lectures I gave at Johns Hopkins University in June, 1983, C. Klaassen, J. Wellner and I hope to carry out some of this program. Our thinking on these questions was greatly stimulated by Pfanzagl's book. I expect the book will similarly influence others.

I conclude with an omission and a misprint:

Page 158 (9.3.3): $\hat{K}^{\cdot}(\underline{x}, P) = n^{-1} \sum_{i=1}^{n} K^{\cdot}(x_i, P)$ appears not to be defined.

Page 186: The theorem is proved for $\Delta(P_n(\underline{x}, \cdot); P) = O_p(n^{-1/2})$. The stated condition $o_p(n^{-1/4})$ requires additional conditions—see Remark 10.4.11.

REFERENCES

AMARI, S. (1982). Differential geometry of curved exponential families—curvature and information loss. *Ann. Statist.* **10** 357–385.

ANDERSEN, E. (1973). Conditional inference and models for measuring. *Mentalhygienisk Vorlag*, Copenhagen.

BEGUN, J. and WELLNER, J. (1983a). Asymptotic efficiency of relative risk estimates. In *Contributions to Statistics: Essays in Honor of Norman L. Johnson* 47–62. (P. K. Sen, ed.), North Holland, Amsterdam.

BEGUN, J., HALL, W. J., HUANG, W. M., and WELLNER, J. (1983b). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11** 432–452.

BERAN, R. (1974). Asymptotically efficient adaptive rank estimates in decision models. *Ann. Statist.* **2** 63–74.

BHUCHONGKUL, S. (1964). A class of nonparametric tests for independence in bivariate populations. *Ann. Math. Statist.* **35** 138–149.

BICKEL, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10** 647–671.

COX, D. R. (1972). Regression models and life tables. *J. Roy. Statist. Soc. B* **34** 187–220.

EFRON, B. (1975). Defining the curvature of a statistical problem. *Ann. Statist.* **3** 1189–1242.

GODAMBE, V. P. (1976). Conditional likelihood and unconditional ôptimum estimation equations. *Biometrika* **63** 277–284.

GRENANDER, U. (1981). *Abstract Inference*. Wiley, New York.

HAMMERSTROM, T. (1978). Thesis. University of California at Berkeley.

HASMINSKII, R., and IBRAGIMOV, I. (1979). On the nonparametric estimation of functions. In *Proceedings of the 2nd Prague Symposium on Asymptotic Statistics* 41–51. (P. Mandl and M. Huskova, eds.). North-Holland, Amsterdam.

KOSHEVNIK, Y. and LEVIT, B. (1976). On a non-parametric analogue of the information matrix. *Theor. Probab. Appl.* **21** 738–753.

LE CAM, L. (1983). Differentiability, tangent spaces and Gaussian auras. Unpublished manuscript.

LE CAM, L. (1985). *Asymptotic Theory*. Forthcoming book.

LEVIT, B. (1975). On the efficiency of a class of non-parametric estimates. *Theor. Probab. Appl.* **20** 723–740.

LINDSAY, B. (1983). Efficiency of the conditional score in a mixture setting. *Ann. Statist.* **11** 486–497.

NEYMAN, J. and SCOTT, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1–32.

STEIN, C. (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 187–195.

STONE, C. (1975). Adaptive maximum likelihood estimators of a location parameter. *Ann. Statist.* **3** 267–284.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720