Sankhyā : The Indian Journal of Statistics
2009, Volume 71-A, Part 2, pp. 354-363

Discussion of Michael Kosorok's Article: What's so Special about Semiparametric Methods?

Moulinath Banerjee
University of Michigan, USA

Abstract

I would like to take this opportunity to congratulate Professor Michael Kosorok (MK in the subsequent discussion) on a fine review of the current state of semiparametric methods and inference. MK has dealt with a number of different areas in the article; from publication facts and figures in connection with impact to the scientific utility of semiparametric modeling, theoretical issues of current interest, the closely associated topic of empirical processes, and last but not the least the issue of dissemination of semiparametric methodology among students of statistics through graduate education and to practitioners through the availability of software. I will focus my discussion of MK's paper on a subset of the topics that he has covered, topics that reflect (at some level) my personal interests and those on which I have stronger opinions.

AMS (2000) *subject classification.* Primary 62G05, 62G20; Secondary 62N01, 62N02.
Keywords and phrases. asymptotically pivotal, mixed-case interval-censoring, likelihood-based change point estimation, non-regularity of the parametric component

1 Likelihood Ratios for the Nonparametric Component in Semiparametric Models

I would like to focus on some issues related to likelihood ratio based inference in semiparametric models for inference on the nonparametric component that MK touches upon in Section 3, partly owing to an interest in non-standard likelihood ratios, and partly owing to the lucrative prospects of being able to use asymptotically pivotal likelihood ratio statistics for inference in a number of interesting problems. The discussion will be restricted to shape-restricted nonparametric components for which there is enough hope of being able to construct such asymptotically pivotal statistics.

Nonparametric likelihood ratio statistics for estimating the value of a shape-restricted function at a point were first introduced in the setting of current status

data by Banerjee and Wellner (2001, 2005), and extended to general monotone function settings in Banerjee (2007). These papers basically establish that confidence intervals for the value of a monotone function at a pre-fixed point of interest can be obtained by inverting the likelihood ratio statistic for testing the value of the function at that point, with calibration provided by the quantiles of a new distribution, namely that of a complex functional, $\mathbb{D}$, of Brownian motion with quadratic drift that *does not depend on nuisance parameters*. While these results were obtained purely in the nonparametric context, Banerjee et al. (2006) and Banerjee et al. (2009) extended them to the proper semiparametric context, by studying a class of semiparametric binary regression/binary choice models, where the conditional mean of the binary response variable Δ, given covariates X (vector-valued) and Z (real-valued), could be modeled as a monotone function, a 'so-called' link function of $\beta^T X + \psi(Z)$ with ψ monotone. Special choices of ψ lead to logit and probit models, and also to a binary regression model that can be identified with the Cox PH model with current status data, originally studied in Huang (1996). It was shown that, by appropriately redefining covariates, inference on the conditional probability $P(\Delta = 1|X = x, Z = z)$ is equivalent to inference on the value of ψ at a point, in a slightly altered version of the original model (see, for details, Section 3.3 of Banerjee et al., 2009), and this fact was used critically for constructing confidence intervals for (i) the conditional probability via inversion of appropriately constructed semiparametric LRTs using the quantiles of $\mathbb{D}$, and, (ii) the conditional distribution of the survival time (given covariates) in the Cox PH model with current status data. The great advantage of such LRT based intervals, of course, lies in the fact that nuisance parameters need not be estimated from the data.

As MK points out, the monotonicity requirement is somewhat restrictive. However, as I will demonstrate, the monotonicity condition crops up automatically in survival data contexts, and furthermore, the semiparametric methodology discussed in the previous paragraph extends much further beyond binary response data. A very useful and interesting model in the interval-censored data context that remains unaddressed is the Cox PH model under more general forms of interval-censoring. A very flexible version is provided by the 'mixed-case interval-censoring model' (see Schick and Yu, 2000). In this model, each individual is observed at a random number of observation times, say $C_1, C_2, \ldots, C_N$ (N being random) and one can identify the interval $(C_i, C_{i+1}]$ (define $C_0 = 0$ and $C_{N+1} = \infty$) in which the individual succumbs to disease/infection. A covariate vector X is also observed for each individual, and if S denotes the time to disease/infection, we have $\lambda_S(t|X) = \lambda_0(t)\, e^{\beta^T X}$, where $\lambda_S(\cdot|X)$ is the hazard rate of S given X. In the simplest version of this model, one assumes that given X, S is independent of the entire (random) censoring mechanism. Interest focuses on estimating

$F(t|X = x) = P(S \leq t|X = x)$, i.e. one seeks to make inference on survival probabilities at different covariate profiles. When N is degenerate at a value, say k, we have what is called Case k interval censoring. The effective parameters that drive the likelihood function in this model are β and the integrated baseline hazard, Λ_0, which is *monotone*. As in the Cox PH model with current status data, it is not too difficult to establish that inference on the survival probabilities can be reduced to inference on Λ_0 in a closely-related model. The interesting question, then, is whether the LRT for testing Λ_0 at a point asymptotically behaves like $\mathbb{D}$ as in the current status (Case 1 interval-censoring) case. The answer is unknown, but there are enough reasons to believe that this may be the case. A first attempt to address this question could focus on a study of a highly simplified form of this model: no covariates and N degenerate at 2, i.e. the fully nonparametric Case 2 interval-censoring model (see Groeneboom and Wellner, 1992). How does the LRT for testing the value of the survival distribution F at a pre-specified point of interest behave in this model? Groeneboom (1996) uses some hard analysis to show that under appropriate regularity conditions, the NPMLE of F at the point t, say $\hat{F}(t)$, converges to $F(t)$ at rate $n^{1/3}$ and is asymptotically distributed like Chernoff's distribution (see Groeneboom and Wellner, 2001). Heuristic considerations in this model suggest that, once again, $\mathbb{D}$ should describe the limiting likelihood ratio, though a formalization of such arguments is rendered difficult owing to additional complexities in the likelihood (see Banerjee, 2007, for a detailed discussion of such issues) beyond the current status model. Such complexities become exacerbated in the mixed case context. Conjecturing that $\mathbb{D}$ will continue to describe the limiting LRT for Λ_0 in the Cox PH model with mixed-case interval-censoring is, then, the natural temptation. A rigorous solution to this problem can be expected to be excruciatingly difficult and may take a while, but it may be worthwhile to simulate the LRT in this model, nevertheless, and compare its distribution to that of $\mathbb{D}$. If an extensive simulation study reveals substantial agreement (as this discussant predicts), one could even advocate using the quantiles of $\mathbb{D}$ provisionally (till the hard-core proof comes along) for data-analysis, though I realize that this is a point that many statisticians will want to argue over, on philosophical grounds. There are obvious extensions of the above model to time-varying covariates and more complex dependence structures between the survival time and the censoring mechanism which we avoid owing to space constraints.

Let us now step beyond the 'bread and butter' Cox PH model. Recently, Kosorok et al. (2004) have studied very general extensions of the Cox model in right censored settings. For time independent covariates X, the univariate proportional hazards frailty regression models that they consider posit $\overline{F}(t \mid X) = 1 - F(t \mid X) = H_\gamma(e^{\beta^T X} A(t))$, for a function $A(t)$ which is the anti-derivative of

a non-negative function a; here H_γ (this is called Λ_γ in Kosorok et al., 2004) is a frailty transform and γ may be unknown. Note that A is automatically constrained to be monotone increasing. One such example is the gamma frailty transform with $H_\gamma(u) = (1 + \gamma u)^{-1/\gamma}$. With $\gamma = 1$ in this model, we recover the proportional odds model, while letting γ go to 0 gives the Cox PH model with $A(t) = \Lambda(t)$. One can consider in this context, the possibility of a unified treatment of the likelihood ratio statistic for estimating (the increasing function) $A(t)$, and more generally, the conditional survival distribution, $F(t \mid X = x_0)$, for general frailty models and determine possible limit distributions with *interval censored data*. In particular, to what extent does $\mathbb{D}$ arise in these models? It seems reasonable to postulate that the MLE of A under interval censoring should be $n^{1/3}$ consistent with Chernoff's distribution in the limit (while those of β and γ should be $\sqrt{n}$ consistent), so $\mathbb{D}$ is, yet again, a natural conjecture.

I end this section by pointing out two other directions where asymptotically pivotal LRT's may be obtained. The first is a semiparametric generalization of the monotone response models of Banerjee (2007) that encompass a fairly wide variety of situations. A general formulation runs thus: consider a random vector (Y, X, Z) with Y real-valued, Z real-valued and X vector-valued, and suppose that the conditional distribution of Y given $X = x, Z = z$ is given by $p(y, \beta^T x + \psi(z))$, with $p(x, \theta)$ being a regular one-dimensional parametric model, and ψ monotone increasing. Thinking of Y as the response and (X, Z) as covariates, one can consider making inference on the parameters (β, ψ). In keeping with our discussion, we are interested mainly in ψ, the nonparametric effect on the response. The above formulation is fairly general and closely related to the partially linear regression model $Y = \beta^T X + \psi(Z) + \epsilon$ under a monotone ψ, certain aspects of which have been studied in Huang (2002). Letting $p(x, \theta)$ be a one-parameter exponential family (for example), parametrized naturally, this model encompasses semiparametric logistic regression and Poisson regression among many others and gives a semiparametric version of the GLM with a monotone nonparametric component (overdispersion parameters could have been included in the discussion but are omitted for simplicity). The results in Banerjee (2007), Banerjee et al. (2006), and Banerjee et al. (2009) again suggest that for this entire class of models, inference on the regression function $\mu(x, z) \equiv E(Y|X = x, Z = z)$ can be made using asymptotically pivotal likelihood ratios that behave like $\mathbb{D}$, though once again rigorous results are lacking and a unified treatment would be a very welcome development. Secondly, and more ambitiously, one can think of the same models just considered, but now with convexity/concavity constraints on ψ. The MLE question for convex functions in the fully nonparametric scenario was settled in two landmark papers by Groeneboom et al. (2001a, 2001b) and since then there has been a spate

of publications dealing with nonparametric estimation in this area. There seems to be general agreement (and I mean to keep this as nebulous as it sounds) that the LRT for testing a convex function at a point should exhibit asymptotically pivotal behavior (as in the monotone function problems), in which case such advantages should be transferrable to the semiparametric domain, as in the monotone function scenario. However, all this remains, as yet, the proverbial 'pie in the sky', since little is known about the fully nonparametric LRT in the convex function setting and any progress in that direction will be a seminal development.

2 Graduate Education

MK's discussion on graduate education, to use a colloquialism, is very much 'on the money'. Indeed, the study of semiparametric theory and methods and empirical process theory ought to play a significantly enhanced role in graduate education in US schools, and also at an international level. In fact, modern empirical process theory is indispensable not only for answering complex theoretical questions in statistics, but also in several disciplines within engineering, like electrical engineering and theoretical computer science, as well as among quantitative social scientists, most notably econometricians who also use and have contributed to the development of semiparametrics actively. Unfortunately, the emphasis on such topics in graduate curriculums leaves much to be desired. It is heartening to know that University of Wisconsin and UNC, Chapel Hill have restructured their programs to emphasize these particular areas, but it would be more interesting to know in how many schools across the US, semiparametrics and empirical process theory are offered even semi-regularly as advanced level courses. I would predict that this number is lower for empirical processes than for semiparametrics in view of a general perception that empirical processes is somehow 'esoteric' and dispensable. At a fundamental level, it is not. Whether one is seeking to establish cutting-edge results in semiparametrics or nonparametrics or non-regular models (where the use of linearization arguments followed by an appropriate CLT application does not work) or machine learning or more generally high-dimensional learning, empirical process theory proves to be near-indispensable. Very few people would disagree with this statement at a superficial level, but my interactions with people leave me convinced that this message has not really sunk in.

It should be noted, though, that in several cases, the reasons behind such courses not being offered more regularly also have to do with logistical issues like resource allocation besides the issue of perception that I have raised. In statistics departments with relatively fewer FTEs and relatively heavy service teaching it often

becomes difficult to offer significant numbers of advanced level courses. Furthermore, there is also competition among different topics and often-times the 'hotter' topics in the popular imagination may get precedence, owing to the capacity to attract larger numbers of students. From personal experience, I would say that an applied special topics course that professes to say something about machine learning immediately attracts students of significantly larger orders of magnitude than a course-offering on empirical processes with applications. Part of this is confounded with an apathy, if not antipathy, towards 'theory' among a substantial proportion of graduate students that seems to be somewhat pronounced in the current day and age. A proper analysis of this syndrome is outside the scope of this discussion, though one tangible cause certainly has to be the fact that a Ph.D. in statistics or biostatistics has now largely become a conduit to a well-paid profession in industry and I think it is fairly safe to assume that apart from the few elite research-level tracks in industry, most other industry jobs pose intellectual challenges nowhere near what an aspiring academic may expect to face. Quite naturally, the motivation to grapple with a hard theoretical course takes a back seat under such circumstances, whereas a course in modern methodology with the right buzzwords is immediately seen as an opportunity to garner skills that may prove useful in the job-market. Invariably, to a certain extent, departments have also fallen prey to this attitude.

There are no easy solutions. However, one way to promote course-offerings in semiparametrics would be to offer them jointly between statistics, biostatistics and maybe economics departments, so that they may be more regularly offered with the instructor chosen on a rotating basis. In such a case, some degree of control needs to be exercised on the material, so that the somewhat differing needs of all groups of students are met. With empirical processes courses, one can also involve the more quantitatively inclined engineering groups. Such inter-departmental syneriges, that can cater to similar target groups across disciplines, might prove effective in raising the frequency of course-offerings in these areas.

3 Some Remaining Issues

I conclude the discussion with a brief look at a number of remaining interesting issues.

Semiparametric inference under non-regularity of the parametric component: Most of semiparametric literature is concerned, typically, with situations where the parametric finite dimensional component, say θ, can be estimated at rate $\sqrt{n}$ via a *regular* asymptotically efficient estimator, the (finite-dimensional) parameter itself being 'pathwise norm-differentiable' in the sense of van der Vaart (1991). The

efficient information is characterized in terms of the efficient influence function obtained by calculating the orthogonal projection of the ordinary score function for θ into the ortho-complement of the infinite-dimensional tangent space for the nuisance parameter, say η. Of course, the infinite dimensional parameter does not need to be (and is often not) pathwise norm-differentiable and may exhibit non-standard asymptotic behavior with slower than $\sqrt{n}$ rate of convergence. It is interesting to speculate what happens in semiparametric models where the finite-dimensional parameter itself fails to be pathwise norm-differentiable and has no regular asymptotically normal estimator. In a fully parametric setting, models exhibiting such non-standard asymptotics have been studied, for example, in a unified set-up by Kim and Pollard (1990) and more recently by Radchenko (2008) in an M-estimation framework, the first of these dealing exclusively with the case of cube-root asymptotics and the second being somewhat more flexible as it allows different parametric components to be estimable at different rates. What would happen in semiparametric versions of such models? Owing to space constraints, we confine ourselves to one such model: a semiparametric version of Rousseeuw's least median of squares (LMS) estimator. In the semiparametric formulation $Y_i = X_i^T \beta_0 + h_0(Z_i) + \epsilon_i$, for $i = 1, 2, \ldots, n$, with $X_i \in I\!R^d, Z_i \in I\!R$ and h_0 is a function assumed to belong to some shape-restricted or smoothness-restricted class of functions. Following Rousseeuw (1984), set: $(\hat{\beta}_n, \hat{h}_n) = \arg\min_{\beta,h} (\text{median}_{1\leq i\leq n} |Y_i - X_i^T \beta - h(Z_i)|^2)$, who prescribed the LMS estimator based on robustness considerations. The fully parametric version of this problem was addressed by Kim and Pollard (1990) and the LMS estimator was shown to be $n^{1/3}$ consistent with asymptotic distribution given by Chernoff's distribution. Clearly, the $n^{1/3}$ rate will not improve in presence of h but does it change? If not, is the asymptotic distribution still given by Chernoff's? If so, what is the loss of efficiency of the LMS of β owing to the need to estimate h, as compared to the situation in which h is known? What about the asymptotic behavior of h? To sum up, can we quantify how the nonparametric component affects estimation of the parametric component in such non-regular models in a unified way, as we can for semiparametric models with regularity in the parametric component? It is highly unlikely that efficiency losses will any longer be characterized through linear projections, as in standard semiparametric analysis. A study of the above proto-typical model may yield interesting insights, but in any case this genre of problems, whether at the level of the specific or the general, can be expected to be hard to resolve.

Boundary estimation problems in higher dimensions in a semiparametric framework: MK and other authors have studied the problem of likelihood based change-point estimation in semiparametric and nonparametric settings. As pointed out by MK in the paper under discussion, the Cox model for right-censored data

with a change-point in the regression, i.e. a model that posits that a regression parameter β has two different values depending on whether a continuous real variable W is above or below an unknown threshold, was studied in Pons (2003) and this work was extended to general transformation models by Kosorok and Song (2007) which posed significantly bigger challenges as the nuisance parameter could no longer be ignored, in contrast to what happens in the Cox model. A natural extension of such models is to incorporate situations where the regression regime changes dramatically with the change being driven by the interactions of multiple covariates. Consider, for example, a survival model where Y is the survival time of a patient afflicted with cancer, X a vector that captures information on whether the patient has received certain medical treatments, W gene-expression data (for multiple genes) from the biopsied carcinoma, and a Cox PH model for the conditional hazard $\lambda_Y(t \mid (X,W)) = \lambda_0(t)\, \exp(\alpha + \beta_1^T X 1(g_\gamma(W) \leq 0) + \beta_2^T X\, 1(g_\gamma(W) > 0))$, where $g_\gamma(W)$ is a smooth surface in W. For simplicity, we have written down a model with uncensored survival time, though realistic applications would, of course, need to incorporate some sort of censoring. The model then postulates a classifier (parametrized by γ) in terms of a smooth function of W, that divides W into two groups that are heterogeneous in terms of the relationship of the survival time to the treatment. The study of such models is extremely important from the scientific point of view, since complex gene interactions can be accounted for and can be expected to pose major challenges beyond the one-dimensional models hitherto considered. Obvious analogues can be formulated for regression models as well.

Availability of software: MK has addressed this issue very thoughtfully in his article and I concur whole-heartedly. I would argue that is one of the burning issues in semiparametrics and (more generally) an issue that afflicts statistics much more broadly. How does one induce non-statistical practitioners to use cutting-edge methodology that appears 'infernally complicated' to the layman? The intimidation factor often turns away scientists who may actually require the use of modern techniques for analyzing their data. The responsibility here probably falls more upon statisticians than a lack of willingness to learn on part of the layman. It is interesting to pose the question: to what extent have the advanced semi and nonparametric results and methodology that were introduced in the last decade been put to actual practice by the broader community? For example, even today, the temptation for non-statistical practitioners is to fall back on the so-called 'Wald-type' confidence intervals — estimate plus minus a factor times standard error — despite the repeatedly demonstrated advantages of using likelihood ratio based CI's over their peers (see, for example, Murphy and van der Vaart, 1997, 2000, Banerjee and Wellner, 2005, Banerjee, 2007). Inversion-based CI's (or confidence sets, more

generally) have never really captured the imagination of practitioners to the extent that the Wald-type intervals have. Lack of adequate readily accessible software and lack of proper dissemination of such ideas are both responsible for this. It is probably not unfair to say that statisticians often do not make a concerted effort to adequately convey their ideas to the broader community, beyond the ivory towers of the erudite. There is simply no systematic mechanism for converting cutting-edge methodology, methodology that provably 'delivers the goods', to readily accessible and comprehensible software. There needs to be serious thought on how this can be brought about as this is a major problem we have been living with for a long time now and a problem that has largely compromised the level of impact that state-of-the-art statistical results and methods could have had.

References

BANERJEE, M. and WELLNER, J.A. (2001). Likelihood ratio tests for monotone functions. *Ann. Statist.*, **29**, 1699–1731.

BANERJEE, M. and WELLNER, J.A. (2005). Confidence intervals for current status data. *Scand. J. Statist.*, **32**, 405–424.

BANERJEE, M., BISWAS, P. and GHOSH, D. (2006). A semiparametric binary regression model involving monotonicity constraints. *Scand. J. Statist.*, **33**, 673–697.

BANERJEE, M. (2007). Likelihood based inference for monotone response models. *Ann. Statist.*, **35**, 931–956.

BANERJEE, M., MUKHERJEE, D. and MISHRA, S. (2009). Semiparametric binary regression models under shape constraints with an application to Indian schooling data. *Journal of Econometrics*, **149**, 101–117.

GROENEBOOM, P. and WELLNER J.A. (1992). *Information Bounds and Nonparametric Likelihood Estimation.* Birkhäuser, Basel.

GROENEBOOM, P. (1996). Inverse problems in statistics. In *Proceedings of the St. Flour Summer School in Probability.* Lecture Notes in Math., **1648**, 67–164. Springer, Berlin.

GROENEBOOM, P. and WELLNER J.A. (2001) Computing Chernoff's distribution. *Journal of Computational and Graphical Statistics*, **10**, 388–400.

GROENEBOOM, P., JONGBLOED, G. and WELLNER, J.A. (2001a). A canonical process for the estimation of convex functions: the "invelope" of integrated Brownian motion + t^4. *Ann. Statist.*, **29**, 1620–1652.

GROENEBOOM, P., JONGBLOED, G. and WELLNER, J.A. (2001b). Estimation of a convex function: characterizations and asymptotic theory. *Ann. Statist.*, **29**, 1653–1698.

HUANG, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.*, **24**, 540–568.

HUANG, J. (2002). A note on estimating a partly linear model under monotonicity constraints. *Journal of Statistical Planning and Inference*, **107**, 343–351.

KIM, J. and POLLARD, D. (1990) Cube root asymptotics. *Ann. Statist.*, **18**, 191–219.

KOSOROK, M.R., LEE, B.L. and FINE, J.P. (2004). Robust inference for univariate proportional hazards frailty regression models. *Ann. Statist.*, **32**, 1448–1491.

KOSOROK, M.R. and SONG, R. (2007). Inference under right censoring for transformation models with a change-point based on a covariate threshold. *Ann. Statist.*, **35**, 957–989.

MURPHY, S.A. and VAN DER VAART, A.W. (1997). Semiparametric likelihood ratio inference. *Ann. Statist.*, **25**, 1471–1509.

MURPHY, S.A. and VAN DER VAART, A.W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.*, **95**, 449–485.

PONS, O. (2003). Estimation in a Cox regression model with a change-point according to a threshold in a covariate. *Ann. Statist.*, **31**, 442–463.

RADCHENKO, P. (2008). Mixed rate asymptotics. *Ann. Statist.*, **36**, 287–309.

ROUSSEEUW, P.J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.*, **79**, 871–880.

SCHICK, A. and YU, Q. (2000). Consistency of the GMLE with mixed case interval-censored data. *Scand. J. Statist.*, **27**, 45–55.

VAN DER VAART, A.W. (1991) On differentiable functionals. *Ann. Statist.*, **19**, 178–204.

MOULINATH BANERJEE
UNIVERSITY OF MICHIGAN
USA.
E-mail: moulib@umich.edu

Paper received May 2009; revised January 2010.