Sankhyā : The Indian Journal of Statistics
2009, Volume 71-A, Part 2, pp. 364-368

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

Jon A. Wellner
University of Washington, USA

Abstract

I would like to thank Michael Kosorok for his stimulating and thought-provoking review of semiparametric methods, empirical processes, and some of the challenges for research and graduate education. His section 3 goes quite a way toward updating the (2006) review of semiparametric methods given in WKR, and his section gives some of the recent progress in empirical process techniques. Since I agree with most if not all of the points made in his review, I will confine my remarks here to pointing out some possible further issues and avenues for future research, mostly related to my own current interests.

AMS (2000) *subject classification.* Primary .
Keywords and phrases.

Section 2. I concur with Kosorok's point that much work remains concerning the behavior and interpretation of semi-parametric and other model-based methods under model miss-specification. The norm in publication has been to propose and study procedures under an assumed model, but it would be much more illuminating in many cases to propose estimators assuming models, but then study them in general, when the model fails, or at least when the model fails in directions that are important for scientific applications. In many cases semiparametric models are of relevance and interest as a way of defining stable and interpretable parameters to estimate, not necessarily in their own right.

Section 3. Another useful reference for students and researchers might be Aad van der Vaart's St. Flour Lectures on *Semiparametric Statistics*, van der Vaart (2002). In these notes van der Vaart also presents a lot of interesting material on the use of empirical process techniques in connection with properties of estimators for semiparametric models. Van der Vaart also

points out a number of challenging open problems: see pages 382, 419, 434, 435, 441.

One area of research intersecting both semiparametric models and empirical processes that Kosorok mentions only obliquely in connection with the work of Huang (1996) and Banerjee and Wellner (2001,2005) is that of shape-constrained inference. The theory of nonparametric shape constrained inference remains under active development, and offers a wide range of very appealing statistical problems and challenges together with very many interesting applications. This area has strong connections with semiparametric and nonparametric mixture models, latent variable models, and random effects models. Some of the recent developments include study of univariate and multivariate log-concave densities as potentially useful surrogates for parametric Gaussian models; see e.g. Duembgen and Rufibach (2009), Balabdaoui et al. (2009), Cule et al. (2007), Koenker and Mizera (2008), Walther (2001) and Seregin and Wellner (2009).

Sections 4–5. As noted by Kosorok, empirical process theory provides a valuable set of tools and techniques for dealing with asymptotic theory in many statistical problems, parametric, semiparametric, nonparametric, or the high-dimensional data and model selection problems discussed in Section 5. As useful as it has been, the "entry price" has remained high and somewhat forbidding for many students and researchers. One of the themes of my research with Aad van der Vaart, both in the writing of van der Vaart and Wellner (1996) and since, has been the development "preservation theorems" at the level of Glivenko-Cantelli and Donsker theorems — results which give ways of deducing further results "easily" without further entropy calculations. There seems to be considerable scope for further development in this direction.

In connection with models involving multiple rates, it seems worthwhile to note the interesting recent work of Radchenko (2008).

One of the many research directions in semiparametric models and empirical processes which I find particularly fascinating involves the interplay between classical sampling theory, empirical process theory, and semiparametric models with two-phase designs. The basic problems concerning empirical process theory were clearly delineated by Lin (2000). There are many central limit theorems for simple averages of data derived via finite sampling designs of various types: see, e.g. Rosén (1997), Hájek (1960) and Särndal et al. (1992). But the availability of uniform central limit theorems for

the empirical processes based on such sampling designs is still quite limited. Breslow and Wellner (2007) pointed out that the asymptotic theory for general empirical processes based on sampling without replacement from an i.i.d. superpopulation (and two-stage versions thereof) follows from the bootstrap central limit theory for exchangeably weighted bootstrapping as established by Praestgaard and Wellner (1993), but it is not at all clear how to extend this to more general sampling schemes.

Of the several directions not mentioned at all by Kosorok, the large current research area involving empirical processes of dependent data seem to me to one that should be mentioned at least in passing. The reader interested in getting a start in this direction should see the collection Dehling et al. (2002) and the valuable review paper therein, Dehling and Philipp (2002).

Section 6. I agree with Kosorok that the implementation of methods for semiparametric methods and development of new algorithms has lagged painfully and rather awkwardly behind the theory, and that these difficulties need serious attention if the promise of many of the new methods is to be realized in practice. Concerning computational problems connected with tuning parameters as mentioned in Kosorok's second point: one of the great attractions of shape constrained methods, at least in low-dimensional problems, is that no (or relatively few) tuning parameters are required. On the other hand, I suspect that this is an area which is ripe for further development of efficient algorithms via more interaction with optimization theory and convex analysis. From the theoretical perspective, empirical process theory can help with justifying and validating the complex procedures which do involve data-based choices of the tuning parameters; see e.g. Giné and Mason (2008) and Dony and Mason (2008).

Section 7. I am happy to learn of the successes elsewhere concerning modernization and updating of the graduate curriculum at the Universities of Wisconsin and North Carolina. Here at the University of Washington (the "other UW"), I have taken a more gradual approach by including a selection of empirical process topics and methods in the graduate statistical theory sequence, by offering special topics courses on empirical process theory approximately every second year, and arranging working groups in alternate years. I cannot say that this has accomplished a major revision of the graduate curriculum, so I look forward to learning more about how the graduate curriculum is changing and evolving at UNC.

Among the books that I find myself using frequently when I have taught courses on these topics, I would include Dudley (1999), de la Peña and Giné (1999) and Ledoux and Talagrand (1991).

In conclusion, it seems to me that the challenges faced by modern statistical theory in dealing with real problems will increasingly involve high-dimensional data, or high-dimensional models, or both. The tools of empirical process theory and the perspectives gained from research in semiparametric models seem likely to continue to play an important role in dealing with these challenges. Thanks again to Michael Kosorok for a stimulating review and reminder of the many challenges that remain, especially with regard to computation and education.

References

Balabdaoui, F., Rufibach, K. and Wellner, J.A. (2009). Limit distribution theory for maximum likelihood estimation of a log-concave density. *Ann. Statist.*, **37**, 1299–1331.

Breslow, N.E. and Wellner, J.A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scand. J. Statist.*, **34**, 86–102.

Cule, M., Samworth, R. and Stewart, M. (2007). Maximum likelihood estimation of a multidimensional log-concave density. Available at arXiv:0804.3989vl.

de la Peña, V.H. and Giné, E. (1999). *Decoupling: From Dependence to Independence*. Probability and its Applications. Springer-Verlag, New York.

Dehling, H., Mikosch, T. and Sørensen, M. (eds.) (2002). *Empirical process techniques for dependent data*. Birkhäuser, Boston.

Dehling, H. and Phillip, W. (2002). Empirical process techniques for dependent data. In *Empirical process techniques for dependent data*, 3–113. Birkhäuser, Boston.

Dony, J. and Mason, D.M. (2008). Uniform in bandwidth consistency of conditional U-statistics. *Bernoulli*, **14**, 1108–1133.

Dudley, R.M. (1999). *Uniform central limit theorems*. Cambridge Studies in Advanced Mathematics, **63**. Cambridge University Press.

Duembgen, L. and Rufibach, K. (2009). Maximum likelihood estimation of a log-concave density and its distribution function: basic properties and uniform consistency. *Bernoulli*, **15**, 40–68.

Giné, E. and Mason, D.M. (2008). Uniform in bandwidth estimation of integral functionals of the density function. *Scand. J. Statist.*, **35**, 739–761.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, **5**, 361–374.

Koenker, R. and Mizera, I. (2008). Quasi-concave density estimation. Technical Report. Department of Mathematical and Statistical Sciences, University of Alberta.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach spaces*. Ergebnisse der Mathematik und ihrer Grenzgebiete, **23**. Springer-Verlag, Berlin.

Lin, D.Y. (2000). On fitting Cox's proportional hazards models to survey data. *Biometrika*, **87**, 37–47.

Præstgaard, J. and Wellner, J.A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, **21**, 2053–2086.

Radchenko, P. (2008). Mixed-rates asymptotics. *Ann. Statist.*, **36**, 287–309.

Rosén, B. (1997). Asymptotic theory for order sampling. *J. Statist. Plann. Inference*, **62**, 135–158.

Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling*. Springer Series in Statistics. Springer-Verlag, New York.

Seregin, A. and Wellner, J.A. (2009). Nonparametric estimation of multivariate convex-transformed densities. Technical Report No. 562. Department of Statistics, University of Washington.

van der Vaart, A. (2002). Semiparametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1999)*, Lecture Notes in Math., **1781**, 331–457. Springer, Berlin.

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag.

Walther, G. (2001). Multiscale maximum likelihood analysis of a semiparametric model with applications. *Ann. Statist.*, **29**, 1297–1319.

Jon A. Wellner
Department of Statistics
University of Washington
USA.
E-mail: jaw@stat.washington.edu

Paper received May 2009; revised January 2010.