On Profile Likelihood: Rejoinder

Author(s): S. A. Murphy and A. W. van der Vaart

Source: *Journal of the American Statistical Association*, Vol. 95, No. 450 (Jun., 2000), pp. 482-485

Published by: American Statistical Association

Stable URL: http://www.jstor.org/stable/2669392

Accessed: 14/06/2014 03:14

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

American Statistical Association is collaborating with JSTOR to digitize, preserve and extend access to *Journal of the American Statistical Association.*

http://www.jstor.org

This content downloaded from 62.122.73.250 on Sat, 14 Jun 2014 03:14:50 AM
All use subject to JSTOR Terms and Conditions

MLE of θ is also locally efficient in model $M(\mathcal{K})$. They used as their sieve the fixed model $M(\mathcal{K}_{\text{sub}})$ for the observed data X induced by the model $M_{\text{ful}}(\mathcal{K}_{\text{sub}})$ characterized by the linear logistic model $\text{pr}(Y=1|V;\kappa_1,\gamma^*)=\{1+\exp[-d_1(V;\kappa_{1,1})-\kappa_{1,2}/\pi(V;\gamma^*)]\}^{-1}$, where $d_1(V;\kappa_{1,1})$ is a known function of an unknown p_1-1–dimensional parameter $\kappa_{1,1}$ and $\kappa_{1,2}$ is an unknown scalar. The sieve MLE $\hat{\theta}_{\text{sieve}}(\gamma^*)\equiv n^{-1}\sum_{i=1}^{n}\text{pr}(Y=1|V_i;\hat{\kappa}_1(\gamma^*),\gamma^*)$ is locally efficient in model $M(\mathcal{K})$ at the submodel $M(\mathcal{K}_{\text{sub}})$, because $\hat{\theta}_{\text{sieve}}(\gamma^*)$ is algebraically identical to $\tilde{\theta}_{\text{loceff}}(\gamma^*)$ based on model $M(\mathcal{K}_{\text{sub}})$. Here $\hat{\kappa}_1(\gamma^*)$ is the κ_1 maximizing the linear logistic likelihood among units with $R=1$.

However, we conjecture without proof that for many models, likelihood inference cannot be "saved" even in this narrow sense. This conjecture comes from our pondering the following examples. Consider the model $M(\mathcal{K}\times\Gamma_{\text{sub}})$ in Example 2a of Section 6. Scharfstein et al. (1999) showed that $\hat{\theta}_{\text{sieve}}(\hat{\gamma})$, with $\hat{\gamma}$ maximizing $\mathcal{L}^{\dagger}_{n2}(\gamma)\equiv\mathcal{L}_{n2}(\gamma)$ over Γ_{sub}, is identically equal to $\tilde{\theta}_{\text{loceff}}(\hat{\gamma})$ and thus is locally efficient in model $M(\mathcal{K}\times\Gamma_{\text{sub}})$ at the submodel $M(\mathcal{K}_{\text{sub}}\times\Gamma_{\text{sub}})$. However, we do not regard $\hat{\theta}_{\text{sieve}}(\hat{\gamma})$ as likelihood-based, because it is not obtained by maximizing a likelihood function. Rather, in our fixed sieve model, the sieve MLE is $\theta\{\hat{\kappa}_{\text{sub}}\}$, where $\hat{\kappa}_{\text{sub}}$ is the value of κ obtained by maximizing $\mathcal{L}^{\dagger}_{n1}(\kappa,\gamma)\mathcal{L}^{\dagger}_{n2}(\gamma)$ over (κ,γ) in $\mathcal{K}_{\text{sub}}\times\Gamma_{\text{sub}}$. Unfortunately, $\theta\{\hat{\gamma}_{\text{sub}}\}$ fails in the sense that it is inconsistent if either $\kappa^*\notin\mathcal{K}_{\text{sub}}$ or $\gamma^*\notin\Gamma_{\text{sub}}$. Next, consider the extension of model $M(\mathcal{K})$ of Example 2a to a two-stage stratified random sampling design. Specifically, redefine $\mathbf{V}=(V_0,V_1),\mathbf{R}=(R_1,R_2)$, with $V_j, j=1,2$, highly multivariate and continuous; R_j Bernoulli, $R_2=R_1R_2$; and $c_{\mathbf{R}}(\mathbf{L})=(V_0,R_1V_1,R_2Y)$. Here CAR implies that $\text{pr}(R_2=1|\mathbf{L},R_1=1;\gamma^*)=\pi_2(\mathbf{V};\gamma^*)$ and $\text{pr}(R_1=1|\mathbf{L};\gamma^*)=\pi_1(V_0;\gamma^*)$. In contrast with the one-stage design discussed earlier, we have failed to find a sieve likelihood for which the MLE $\hat{\theta}_{\text{sieve}}(\gamma^*)$ is a locally efficient estimator of θ^* (see Robins 2000 for further discussion).

ADDITIONAL REFERENCES

Dabrowska, D. M. (1988), "Kaplan–Meier Estimator on the Plane," *The Annals of Statistics*, 16, 1475–1489.

Gill, R. D., van der Laan, M. J., and Robins, J. M. (1997), "Coarsening at Random: Characterizations, Conjectures and Counterexamples," in *Proceedings of the First Seattle Symposium on Survival Analysis*, eds. D. Y. Lin and T. R. Fleming, New York: Springer, pp. 255–294.

Hastie, T. J., and Tibshirani, R. J. (1995), *Generalized Additive Models*, New York: Chapman and Hall.

Heitjan, D. F., and Rubin, D. B. (1991), "Ignorability and Coarse Data," *The Annals of Statistics*, 19, 2244–2253.

Jacobsen, N., and Keiding, N. (1995), "Coarsening at Random in General Sample Spaces and Random Censoring in Continuous Time," *The Annals of Statistics*, 23, 774–786.

Kress, R. (1989), *Linear Integral Equations*, Berlin: Springer-Verlag.

Lin, D. Y., Sun, W., and Ying, Z. (1999), "Estimation of the Gap Distribution," *Biometrika*, 86, 350–359.

Prentice, R. L., and Cai, J. (1992), "Covariance and Survivor Function Estimation Using Censored Multivariate Failure Time Data," *Biometrika*, 79, 495–512.

Robins, J. M. (1999), "Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference," in *Statistical Models in Epidemiology: The Environment and Clinical Trials,* eds. M. E. Halloran and D. Berry, NY: Springer-Verlag, pp. 95-134.

———, (2000), "Robust Estimation in Sequentially Ignorable Missing Data and Causal Inference Models," *Proceedings of the 1999 Joint Statistical Meetings* (to appear).

Robins, J. M., and Rotnitzky, A. (1992), "Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers," in *AIDS Epidemiology—Methodological Issues*, eds. N. Jewell, K. Dietz, and V. Farewell, Boston: Birkhäuser, pp. 297–331.

Robins, J. M., and Wang, N. (1998), Discussion of the articles by Forster and Smith and Clayton et al., *Journal of the Royal Statistical Society*, Ser. B, 60, 91–93.

Robins, J. M., and Wasserman, L. (2000), "Conditioning, Likelihood, and Coherence: A Review of Some Foundational Concepts," *Journal of the American Statistical Association*, (to appear).

Satten, G., and Datta, S. (2000), "Marginal Estimation for Multistage Models: Waiting Time Distributions and Competing Risks Analysis," University of Georgia, Department of Statistics Tech Report #STA 00-04.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999), Rejoinder to Comments on "Adjusting for Nonignorable Dropout Using Semiparametric Nonresponse Models," *Journal of the American Statistical Association* 94, 1135–1146.

van der Laan, M. (1996a), "Efficient Estimator of the Bivariate Survival Function and Repairing NPMLE," *The Annals of Statistics*, 24, 596–627.

——— (1996b), "Efficient and Inefficient Estimation in Semiparametric Models," CWL tract 114, Centre of Computer Science and Mathematics, Amsterdam.

van der Laan, M. J., Hubbard, A. E., and Robins, J. M. (1999), "Locally Efficient Estimation of a Multivariate Survival Function in Longitudinal Studies," *Journal of the American Statistician*, (submitted).

van der Laan, M., Robins, J. M., and Gill, R. D. (2000), "Locally Efficient Estimation in Censored Data Models: Theory and Examples," technical report, University of California, Berkeley, Dept. of Biostatistics.

Rejoinder

S. A. MURPHY and A. W. VAN DER VAART

We thank all discussants for helping us view this work from a variety of perspectives. Their contributions are helpful for understanding how our article fits into the general development of methodology for models with high-dimensional parameter spaces.

Likelihood-based inference for parametric models is well established, and hence the purpose of our article to give a justification for a similar use of profile likelihoods in the semiparametric context hardly needs defense. We think that this is true irrespective of whether or not one feels that likelihood should be the principle of choice. Probably the discussants would agree with this general statement, even though they argue that in certain situations other methods are more natural or even better.

© 2000 American Statistical Association
Journal of the American Statistical Association
June 2000, Vol. 95, No. 450, Theory and Methods

This content downloaded from 62.122.73.250 on Sat, 14 Jun 2014 03:14:50 AM
All use subject to JSTOR Terms and Conditions

Our work was particularly motivated by situations in which $\sqrt{n}$-efficient estimation is feasible, but practical use of the efficient score as an estimating function for θ is difficult. This difficulty may be due to only an implicit definition of the efficient score or due to the need to estimate functionals such as densities or conditional densities to use the efficient score. It should be stressed that our article uses the efficient score only as part of our proof technique. The statistician need not calculate tangent spaces or efficient scores to use the likelihood methodology. The take-home message is that if the graph of the profile likelihood looks approximate quadratic, then we can expect that methods such as those given in our article can be used to justify the the classical properties of maximum likelihood estimation. Furthermore, as in the parametric case, we can use the curvature to estimate standard errors and the likelihood ratio statistic for testing purposes. As pointed out by all discussants, these methods need to be altered and generalized to deal with a greater variety of semiparametric models. In particular, we wholeheartedly agree with Fan and Wong's point in their Sections 2 and 3 that one must also include sieve profile likelihoods in this effort. Similarly, profile likelihood resulting from a penalized likelihood should be included. As an example, in earlier work (Murphy and van der Vaart 1999) we considered a semiparametric logistic regression model and derived a quadratic approximation to the profile of a penalized likelihood. This leads us to believe that our results extend to these cases, at the expense of a more complex notation.

SEMIPARAMETRIC MODELS IN WHICH EITHER $\sqrt{n}$ OR EFFICIENT ESTIMATION OF θ IS INFEASIBLE

Both Bickel and Ritov and Robins, Rotnitzky, and van der Laan point out that there are many models in which one would not want to use maximum likelihood estimation for estimation of the parametric component of a semiparametric model. Sometimes the MLE is not even well defined (such as in the censored regression setting). Estimating functions for θ are a natural alternative in this case.

Bickel and Ritov give a valuable overview of technical conditions from their book that allow the use of estimating functions. However, they do not address the issue of our article, the profile likelihood function. Our understanding of their contribution was somewhat hampered by a large number of typos and small inaccuracies. As far as we can see, however, their method of proof (given in Bickel, Klassen, Ritov, and Wellner 1994; BKRW) has the same structure as ours, if specialized to a proof of asymptotic normality only. A difference is Bickel and Ritov's use of an intermediate quantity η_θ, which does not appear in our presentation. Our paper did not have the intention of discussing regularity conditions at length and this discussion is also not the place to spell out such details. However, we would like to respond to the claim of Bickel and Ritov that they "give a weaker version of the conditions of MvV which yet permits their applicability" by saying that this is not clear to us. In particular, Bickel and Ritov's conditions D1–D2 look restrictive and unnatural, and their condition D5 has no counterpart in our article, simply because we do not have the quantities η_θ. [In the version of the discussion that we have received, condition D5 contains the quantity $P_0\psi(\theta, \eta_0)$, but this should read $P_\psi(\theta, \eta_\theta)$, as is clear from section 7.7 of BKRW.] Here the use of empirical process techniques as mentioned briefly in our article would help clarify the argument, and probably lead to conditions that are both weaker and easier to verify. Specifically, one possibility would be to assume that the functions $\psi(\theta, \eta)$ belong to a Donsker class. [More generally, we could require bounds on the entropy numbers of the classes of functions $\psi(\theta, \hat{\eta}_\theta)$ and $\psi(\theta, \eta_\theta)$.] The present conditions D1–D2 are close to requiring that the processes $x \mapsto \sqrt{n}(\psi(x, \theta, \hat{\eta}_\theta) - \psi(x, \theta, \eta_\theta))$ converge in distribution relative to the uniform norm (and thus are asymptotically continuous in x, which is implicitly assumed to range over a Euclidean compact), and this does not seem to be of prime relevance for the argument. Of course, this does not mean that this approach cannot be useful or cannot work in cases where other approaches may fail.

The possible misbehavior of likelihood-based inference under misspecification is a serious objection to using likelihood in situations where a given model cannot be trusted. Of course, the purpose of semiparametric modeling is to alleviate this problem. However, the discussants point out several examples where this problem can be serious; for instance, when there are high-dimensional covariates. Repairs can be made to likelihood, as has been shown elsewhere by one of us. In particular, we shall respond to Robins and Rotnitzky's Section 8 elsewhere. However, we are not blind to the flexibility of estimating equations. Particularly, if one is able to capture the question of interest as a simple operation on the data, such as taking a mean or another moment, then appropriate estimating equations may be very attractive. The discussants have made very interesting contributions to extend this to missing- and censored-data situations. In general, it appears that the behavior of semiparametric likelihood–based procedures under the wrong model has not been studied enough, as is their robustness theory. This is likely to be a richer theory than for parametric models.

Fan and Wong also discuss (in sec. 4) an example in which θ is not estimable at a parametric rate and thus one can not expect a quadratic approximation to the profile likelihood to hold. It is interesting that despite the nonstandard nature of this example, Fan and Wong were able to use the likelihood method (using an appropriate sieve) to construct an optimal test for $\theta = 0$. We note that even when the nuisance parameter is finite-dimensional, the statistician must always be concerned that the distribution of $\hat{\theta}$ may not be well (or at all) approximated by a normal distribution. A first check for this is to graph the profile likelihood against θ to see whether the graph is approximately quadratic. The methods presented in the article lend support to this ad hoc check.

Robins, Rotnitzky, and van der Laan discuss another situation in which maximum likelihood estimation appears infeasible. Their examples are characterized by high-dimensional covariates. In their examples, the statistician is confronted with a choice. One either makes or is given

This content downloaded from 62.122.73.250 on Sat, 14 Jun 2014 03:14:50 AM
All use subject to JSTOR Terms and Conditions

smoothness assumptions to reduce the size of the parameter space $\mathcal{K}$, or one makes or is given smoothness assumptions to reduce the size of the parameter space Γ. In Examples 1 and 2a discussed by Robins, Rotnitzky, and van der Laan, $\mathcal{K}$ includes the parameters in the conditional distribution of the response Y given the high-dimensional covariate V, and Γ includes the parameters in the conditional distribution of the indicator R given the high-dimensional covariate V. In both cases the high dimensionality of the covariate V is problematic.

Statisticians following the likelihood principle are naturally inclined to make assumptions on the parameter space $\mathcal{K}$, because the parameter of interest is $\theta = \theta(\kappa), \kappa \in \mathcal{K}$, and the likelihood factorizes into a part depending only on κ and a part depending only on γ. On the other hand, statisticians who are accustomed to randomization trials are often in the enviable position of knowing the true value of $\gamma \in \Gamma$. (R is the randomization indicator, and γ indexes $P[R = 1|V]$.) Thus it is natural, even when γ is unknown, for a statistician with this background to make smoothness assumptions to reduce the size of the parameter space Γ. In both cases statisticians must resort to making gross smoothness (or even parametric) assumptions on a conditional distribution given high-dimensional V. Nothing will come cheap here.

Robins, Rotnitzky, and van der Laan try to prove that likelihood-based inference in high-dimensional situations must fail. Their argument appears to be as follows. Given a partitioned likelihood $l_n(\kappa, \gamma) = l_{n1}(\kappa)l_{n2}(\gamma)$, where the parameter of interest $\theta = \theta(\kappa)$ depends on κ only, (a) a "true likelihoodist" uses a procedure for θ based on l_n, and hence the procedure will be the same function of the data, irrespective of the value of γ, by the factorization of l_n; and (b) such a procedure cannot behave well simultaneously for all values of γ (at least if both parameters are appropriately high-dimensional); thus (c) if γ is unknown, then a "true likelihoodist" must go astray, (d) whereas estimating equations can give good procedures provided that the true value of γ is used in their construction.

This is an interesting argument, but full proof it is not. A general lesson to be learned from this interesting example is that we should not take everything that we have learned from the application of parametric models for granted when applying similar techniques to semiparametric models. A "pragmatic likelihoodist" should use the knowledge of the true γ_0 if available, thus violating the premise (a). This is possible by making the parameter set for κ dependent on γ_0, or by including a penalty that is dependent on γ_0 into the likelihood. This may seem odd, but it is not, as follows if we consider again the factorized likelihood $l_n(\kappa, \gamma) = l_{n1}(\kappa)l_{n2}(\gamma)$, where both parameters are unknown. In the situations considered by Robins, Rotnitzky, and van der Laan we must make a priori assumptions on one of the two parameters κ or γ, or both, to ensure the existence of good estimators of $\theta = \theta(\gamma)$. In their example Robins, Rotnitzky, and van der Laan put all assumptions on γ by assuming this to be known. A "pragmatic likelihoodist" could make a priori assumptions on the pair κ and γ jointly, thus creating a trade-off between smoothness assumptions on κ or on γ, by choosing the parameter set for (κ, γ) not equal to a product set. The case that γ is known is then an extreme case of this trade-off, and will lead to a parameter set for κ that depends on γ_0. However we have no general results that imply that the foregoing strategy always works. See also Robins, Rotnitzky, and van der Laan's sections 7 and 8 for further discussion.

It is interesting that the Robins, Rotnitzky, and van der Laan argument attacks not only the frequentist use of the likelihood (through maximum likelihood and the likelihood ratio statistic), but also the Bayesian use. As applied to Bayesian inference, the defense given in the preceding paragraph also is valid and is perhaps more readily acceptable. It shows that we should not necessarily make the parameters κ and γ independent under the prior, and in the extreme case that γ_0 is known and hence all prior uncertainty is put on κ, the prior for κ should depend on γ_0. That sounds very reasonable to us, and hence we are not necessarily led into using Robins, Rotnitzky, van der Laan's estimating equations (although we have nothing against them). By the usual argument, a prior on κ depending on γ_0 is somewhat similar to a penalty depending on γ_0, and this in turn is somewhat similar to a sieved parameter set for κ depending on γ_0.

Robins, Rotnitzky, and van der Laan make the point that the existing asymptotic information bounds are not always relevant here. We agree. Because the existing bounds take only the local properties of a model into account, they may be pure "asymptopia." Our favorite example of this situation is the higher-dimensional version of the oldest semiparametric model: estimating a center of symmetry. This problem is adaptive in the sense of Bickel (1981), meaning that the center of symmetry can be estimated just as well knowing the (symmetric) shape of the underlying density as not knowing it. The construction proving this first estimates the shape of the underlying density nonparametrically, next symmetrizes the density estimator around the unknown center, and finally roughly does maximum likelihood for the center of symmetry given the estimated shape. This procedures "works" in the sense of achieving the information bounds (as in Bickel et al. 1993) for any dimension of the observations (see, e.g., van der Vaart 1988, sec. 5.7.2). However, even for dimensions as low as four or five, one should seriously question the relevance of this theoretical result. Presumably a more relevant definition of information for such examples should involve some type of (global) minimax bounds, which should also show the extent to which adaptation to given smoothness of the nuisance parameter is possible. Thus we would obtain, for instance, relevant lower bounds for the maximal mean squared error of an estimator over a prespecified smoothness class and would require an estimator to attain these bounds for many smoothness classes simultaneously. The bounds would be bigger than the maximal existing lower bounds. Robins, Rotnitzky, and van der Laan and Robins and Ritov (1997) have taken a first step in this direction by stressing uniformity in convergence of distribution, but much remains to be done. The recent work on model selection, for instance by Birgé and

This content downloaded from 62.122.73.250 on Sat, 14 Jun 2014 03:14:50 AM
All use subject to JSTOR Terms and Conditions

Massart, should be relevant here, even though it is not directly applicable.

THE NO-BIAS CONDITION

We are interested in restatements of the no-bias condition (11), and thank the discussants for their remarks on this. At this time only the form (11) appears to cover all examples. The condition can certainly not be removed, but it would be interesting to have general, easy to verify conditions that imply it. That some form of the condition is necessary under regularity conditions as in our article can be proved as a theorem (see, e.g., van der Vaart 1988, thms. 25.59 and 5.31).

The methodology discussed by Shen offers hope that the no-bias condition can be replaced by other conditions. His argument is based on a quadratic approximation to the difference in Kullback–Leibler informations [(4) or (7)]. Just as we had to generalize the notion of a likelihood to include empirical likelihoods in our examples, we will need to to generalize the definition of the Kullback–Leibler information. This generalization is needed because the classical Kullback–Leibler information, $K(\Phi_0, \Phi)$, is often undefined for Φ the MLE (see the case-control studies with missing covariate and the shared gamma frailty model examples). This occurs when we use an empirical version of the likelihood, a situation of great practical relevance.

A generalization of the Kullback–Leibler information is also needed in the proportional hazards model for current status data example. The MLE of the baseline hazard, $\hat{\eta}(y)$, can take the value 0 with positive probability for values of y with $\eta_0(y) > 0$ (see Huang 1996 for a discussion). This means that the Kullback–Leibler information may be minus infinity, and the left side of Shen's equation (2) with $\Phi_1 = T_n + \gamma_n u^*$ and $\Phi_2 = T_n$ will be undefined with positive probability. A simple generalization would be to replace $K(\Phi_0, \Phi_1)$ by $-P_0 \log\{[p_{\Phi_0}(x) + p_{\Phi_1}(x)]/[2p_{\Phi_0}(x)]\}$. This quantity has many of the same properties as the Kullback–Leibler information and $-\mathbb{P}_n \log\{[p_{\Phi_0}(x) + p_{\hat{\Phi}}(x)]/[2p_{\Phi_0}(x)]\} \leq 0$ if $\hat{\Phi}$ is the MLE.

Once we have an appropriate generalization of the Kullback–Leibler information, we must verify (4). We view equation (4) as a no-bias condition, as it seeks to approximate the Kullback–Leibler information evaluated at estimators by a quadratic, with coefficients depending only on the true parameter values; the definition of the norm in (2) depends on the true parameter values. It is in the verification of (4) that one generally must make use of convergence rates of the estimators just as it is only in the verification of our no-bias condition (11) that we generally need to use convergence rates of the estimators. In our article we point out the potential drawbacks of quadratic approximations, which may insufficiently reflect the structure of the model.

Bickel and Ritov also propose an alternative to the no-bias condition. They replace our (11) with their assumptions (6) and (9). [Their condition (6) is condition D4 as specialized from general estimating functions to efficient estimating functions.] Indeed, if we only want to derive the asymptotic distribution of $\hat{\theta}$, then we can replace (11) with the simpler assumption (6). We used the stronger condition (11) to justify both the quadratic approximation to the profile likelihood and the standard error estimator in our Corollary 3. We do not see a major difference between their (6) and our (11). In all examples, both conditions would follow by exactly the same techniques.

GENERALIZATIONS TO CRITERION FUNCTIONS

We welcome Li's discussion about how one can extend this method to a general criterion function, such as quasi-likelihood. In particular we appreciate his substitute for the least favorable path. Because the information equality does not generally hold, the asymptotic variance of $\hat{\theta}$ will be a sandwich variance, for example, of the form $\dot{\mathcal{W}}^{-1}\Sigma\dot{\mathcal{W}}^{-1}$. The $\dot{\mathcal{W}}$ corresponds to minus the expectation of the second derivative of the criterion function along Li's substitute for the least favorable path. We believe that this quantity can be estimated by the curvature of the profile criterion function much as in our Corollary 3. The Σ corresponds to the variance of Li's equation (4). Equation (4) is the first derivative of the criterion function along Li's substitute for the least favorable path. This variance appears hard to estimate unless the formula for the score in (4) is simple. One possible indirect estimator would be

$$\mathbb{P}_n \left(\frac{\log \mathrm{l}(\hat{\theta} + h_n v_n, \hat{\eta}_{\hat{\theta}+h_n v_n}) - \log \mathrm{l}(\hat{\theta}, \hat{\eta}_{\hat{\theta}})}{h_n} \right)^2 .$$

However, at this time we have not seen a proof that this estimator is consistent.

ADDITIONAL REFERENCES

van der Vaart, A. W. (1988), *Statistical Estimation in Large Parameter Spaces*, Amsterdam: CWI.

——— (1998), *Asymptotic Statistics*, Cambridge, U.K.: Cambridge University Press.

This content downloaded from 62.122.73.250 on Sat, 14 Jun 2014 03:14:50 AM
All use subject to JSTOR Terms and Conditions