Statistica Sinica **6**(1996), 917-923

# ON CONSISTENCY IN PARAMETER SPACES OF EXPANDING DIMENSION: AN APPLICATION OF THE INVERSE FUNCTION THEOREM

Robert L. Strawderman and Anastasios A. Tsiatis

*University of Michigan and Harvard University*

*Abstract:* Foutz (1977) uses the Inverse Function Theorem to prove the existence of a unique and consistent solution to the likelihood equations. This note extends his results in three useful directions. The first is to remark that with minor modification the same proof may be used to show that the solution to the likelihood equations converges asymptotically to the least-false parameter (Hjort (1986, 1992)) when the true probability distribution of the data differs from the parametric family of models under consideration. The second is to extend his results to certain situations in which the dimension of the parameter space is not fixed but expanding at some rate less than the sample size. Lastly, we indicate how this result may be applied in more general $M$-estimation problems. An application of these results to proving consistency in problems involving splines is discussed.

*Key words and phrases:* Least-false parameter, maximum likelihood, regression spline, uniform convergence.

## 1. Introduction

The Inverse Function Theorem (*IFT*; see Rudin (1964)) specifies conditions under which (i) a function $f(u)$ is one-to-one for $u$ in an open $\varepsilon$-neighborhood about a point $u_0$, say $U_\varepsilon$; and, (ii) $f^{-1}(\cdot)$ is well-defined on the image of $U_\varepsilon$ under $f(\cdot)$. This result provides a simple and useful way to prove the existence and uniqueness of a consistent root to the likelihood equations. Let $l_n(\beta)$ denote the likelihood for a sample of $n$ observations under a parametric family of models $\{\mathcal{P}_\beta, \beta \in \mathcal{B}\}$, where $\beta$ is a finite-dimensional parameter. Let $S_n(\beta)$ be the associated score function, and define $\beta_0$ as the unique solution to $E\left(S_n(\beta)\right) = \mathbf{0}$; that is, $\beta_0$ denotes the underlying parameter being estimated. Here, $\dim(\mathbf{0})$=$\dim(\beta)$ and is a convention which is used throughout this paper. Under sufficient regularity, Foutz (1977) uses the *IFT* to prove that with probability going to one a unique sequence of solutions $\{\hat{\beta} : S_n^{-1}(\mathbf{0}) = \hat{\beta}\}$ exists in an $\varepsilon$-neighborhood of $\beta_0$ such that $\hat{\beta} \xrightarrow{p} \beta_0$. This method for proving the consistency of $\hat{\beta}$ is particularly useful when $\hat{\beta}$ only has an implicit representation as the solution to the likelihood equations i.e. $S_n(\beta) = \mathbf{0}$.

Foutz (1977) assumes that $\hat{\beta}$ is the maximum likelihood estimate, that the underlying probability distribution of the data is a member of the parametric family $\{\mathcal{P}_\beta, \beta \in \mathcal{B}\}$, and that the dimension of $\beta$ is fixed. The intent of this paper is to generalize his results with respect to these assumptions. Suppose that the true probability distribution of the data is given by $\mathcal{P}$, where $\mathcal{P}$ is not necessarily a member of $\{\mathcal{P}_\beta, \beta \in \mathcal{B}\}$, and define the "least-false parameter" $\beta^\star$ as the solution to $E_{\mathcal{P}}(S_n(\beta)) = \mathbf{0}$. The value $\beta^\star$ minimizes the Kullback-Leibler distance between $\{\mathcal{P}_\beta, \beta \in \mathcal{B}\}$ and $\mathcal{P}$ (see Hjort (1986, 1992) for additional results). Assuming $\beta^\star$ exists, a proof that the maximum likelihood estimator $\hat{\beta}_n$ exists, is locally unique, and consistent for $\beta^\star$ can be obtained via the *IFT*. The proof is essentially identical to that of Foutz (1977), except that all expectations are taken over $\mathcal{P}$ instead of $\mathcal{P}_{\beta^\star}$.

In understanding the importance of generalizing the fixed dimension assumption, it is useful to consider an example in which the dimension of the parameter space grows with the sample size. Many problems in statistics involve modeling an unknown continuous real-valued function $h(t)$ (e.g. a density or hazard function). A common approach is to use polynomial splines, which induce parametric models that generally take the form $g(t; \beta_n, \tau_n)$, where $\beta_n$ are parameters to be estimated and $\tau_n$ is a vector of knots, or breakpoints. Spline functions can be made more flexible by increasing the number of knots. One trade-off, however, is that the dimension of $\beta_n$ also gets larger. Determining the theoretical behavior of the sequence of parametric families generated by $g(t; \beta_n, \tau_n)$ as $n \to \infty$ is often of interest when the number of knots (and hence the dimension of $\beta_n$) is allowed to grow larger with $n$, but at some lesser rate. Murphy and Sen (1991) investigate the consistency of time-dependent coefficients (modeled by step functions) in the Cox model. Stone (1994), unifying much of his work over the last 15 years, provides a general framework for proving consistency in flexible exponential family-type models for density estimation and generalized regression problems. Strawderman and Tsiatis (1996) model the dependence of a hazard function on a stochastic process using $B$-splines, and investigate the consistency and asymptotic normality of the resulting hazard estimator. Examples of other problems where the parameter space expands with the sample size can be found in the literature on contingency tables. Portnoy (1988) discusses the general problem of expanding parameter spaces in the context of exponential families. Several other examples can be found in Bickel et al. (1993).

The main result of this paper is given in Section 2 as Theorem 1, and constitutes a stochastic reformulation of the *IFT* in terms of the supremum norm $\|\cdot\|_\infty$. Theorem 1 provides a systematic method for verifying consistency in a wide variety of problems, and reduces the general problem to demonstrating the validity of three conditions. In Section 3, we describe how these results may be

used to prove consistency in problems involving polynomial spline approximations, and close the paper with a few remarks on the application of Theorem 1 to more general estimating functions.

## 2. Main Result

Let $X$ be a random vector defined on a probability space $\{\Omega, \mathcal{F}, \mathcal{P}\}$, and let $X_1, X_2, \ldots$ denote independent and identically distributed copies of $X$. For each $n$, let $\mathcal{B}_n \subset \mathbb{R}^{k_n}$ be open, and let $\{\mathcal{P}_{n,\beta_n},\ \beta_n \in \mathcal{B}_n\}$ denote a family of parametric probability distributions indexed by a $k_n \times 1$ parameter vector $\beta_n$. It is assumed that $\dim(\beta_n) = k_n$ grows with $n$ at some slower rate $n^\theta$, where $0 \leq \theta < 1$ and $\theta = 0$ corresponds to a fixed dimensional parameter. Unless otherwise specified, all probabilities and expectations to follow are with respect to $\mathcal{P}$.

Let $l_n(\beta_n)$ denote the log-likelihood for the observations $X_1, \ldots, X_n$ under $\mathcal{P}_{n,\beta_n}$. Define

$$S_n(\beta_n) \ = \ \frac{k_n}{n}\frac{\partial l_n(\beta_n)}{\partial \beta_n}$$

as the corresponding $k_n \times 1$ normalized score vector. If $A \subseteq \mathbb{R}^{k_n}$, then the notation $S_n(A)$ is understood to mean the image of $A$ under the transformation $S_n(\cdot)$. Let $-I_n(\beta_n)$ be the first derivative of $S_n(\beta_n)$. Define $\beta_n^\star$ as the solution to $E\left(S_n(\beta_n)\right) = \mathbf{0}$ and let $\mathcal{I}_n(\beta_n) = E\left(I_n(\beta_n)\right)$. It is assumed that $\beta_n^\star$ exists and is unique for $n \geq N_b$, where $N_b < \infty$.

**Theorem 1.** *For $n \geq N_b$, suppose that $S_n(\beta_n)$ is a continuously differentiable mapping from $\mathbb{R}^{k_n}$ to $\mathbb{R}^{k_n}$ in a neighborhood of $\beta_n^\star$. In addition, suppose that*

(a) *there exists a constant $0 < c < \infty$ such that $\|\mathcal{I}_n^{-1}(\beta_n^\star)\|_\infty \leq c$ for $n \geq N_b$;*

(b) *there exists an $\varepsilon > 0$ that may depend only on $c$ such that for all $\delta > 0$ there exists an $N_\delta \geq N_b$ such that for all $n > N_\delta$,*

$$Pr\Big\{\sup_{\|\beta_n - \beta_n^\star\|_\infty < \varepsilon} \|I_n(\beta_n) - \mathcal{I}_n(\beta_n^\star)\|_\infty > \frac{1}{2c}\Big\} < \delta;$$

(c) $\|S_n(\beta_n^\star)\|_\infty \xrightarrow{p} 0$.

*Then, as $n \to \infty$, a unique solution $\{\hat{\beta}_n : S_n(\hat{\beta}_n) = \mathbf{0}\}$ exists in a neighborhood about $\beta_n^\star$ with probability going to one, and $\|\hat{\beta}_n - \beta_n^\star\|_\infty = O_p(\|S_n(\beta_n^\star)\|_\infty)$.*

The reader may recognize that conditions (a) and (b) given in the statement of the theorem are essentially those needed to invoke the *IFT*. Taken together, these two conditions simply require that $S_n(\beta_n)$ satisfies the conditions of the *IFT* with probability going to one for $\beta_n$ in an $\varepsilon$-neighborhood about $\beta_n^\star$. Condition (c) guarantees that $\mathbf{0} \in S_n(\{\beta_n : \|\beta_n - \beta_n^\star\|_\infty < \varepsilon\})$ with probability going to

one, and is necessary for proving existence and consistency of $\hat{\beta}_n$ using the *IFT*. The proof of Theorem 1 is straightforward and is given below.

**Proof of Theorem 1.** Let $Y_n = (X_1, \ldots, X_n)' \in \Omega^n$, and consider the sets $A^{(n)} \subset \Omega^n$ and $B^{(n)} \subset \Omega^n$, where

$$A^{(n)} = \Big\{ Y_n : \sup_{\|\beta_n - \beta_n^\star\|_\infty < \varepsilon} \|I_n(\beta_n) - \mathcal{I}_n(\beta_n^\star)\|_\infty \leq \frac{1}{2c} \Big\}$$

and

$$B^{(n)} = \Big\{ Y_n : \|S_n(\beta_n^\star)\|_\infty < \frac{\varepsilon}{4c} \Big\}.$$

Let $N_b$, $c$, and $\varepsilon$ be chosen to satisfy suppositions (a) and (b) of the theorem. In conjunction with supposition (c), this implies that we may fix any $\delta > 0$ and $\delta^\star > 0$ and find $N_\delta$ and $N_{\delta^\star}$ such that for $n > \max\{N_b, N_\delta, N_{\delta^\star}\}$, $P\{A^{(n)} \cap B^{(n)}\} > 1 - 2\max\{\delta, \delta^\star\}$. Since $\max\{\delta, \delta^\star\}$ may be made arbitrarily small, it follows that $\lim_{n\to\infty} P\{Y_n \in A^{(n)} \cap B^{(n)}\} = 1$.

Define the sets $C_{\varepsilon,n} = \{\beta_n : \|\beta_n - \beta_n^\star\|_\infty \leq \varepsilon\}$, and $D_{\varepsilon,n} = \{y : \|y - S_n(\beta_n^\star)\|_\infty < \varepsilon/(4c)\}$. Then, for $n > N_b$, it can be shown that $S_n(\beta_n)$ is one-to-one from $C_{\varepsilon,n}$ onto $D_{\varepsilon,n} \subseteq S_n(C_{\varepsilon,n})$ and $\mathbf{0} \in D_{\varepsilon,n}$ whenever $Y_n \in A^{(n)} \cap B^{(n)}$. The proof that this is so is straightforward, and entails a simple modification of the proof of the *IFT* found in Rudin (1964).

Together, these results imply that with probability going to one a (locally) unique solution $\hat{\beta}_n \in C_{\varepsilon,n}^o$ (the interior of $C_{\varepsilon,n}$) exists such that $S_n(\hat{\beta}_n) = \mathbf{0}$ and $S_n^{-1}(\mathbf{0}) = \hat{\beta}_n$. The fact that $\varepsilon$ may be chosen independently of $n > N_b$ further implies that $\|\hat{\beta}_n - \beta_n^\star\|_\infty = O_p(\|S_n(\beta_n^\star)\|_\infty)$, completing the proof.

## 3. An Application to Problems Involving Splines

Consider the problem in which an unknown continuous real-valued function $h(t)$ is modeled using $B$-splines. An introduction to $B$-splines and their properties can be found in de Boor (1978). Stone (1994) develops a rather general framework in which the consistency of $B$-spline-based estimators in certain density estimation, regression, and generalized regression problems (e.g. generalized linear models with flexible link functions) can be investigated for uncensored data. Strawderman and Tsiatis (1996) consider the questions of consistency and asymptotic normality when the dependence of the log-hazard function on a stochastic time-dependent covariate is modeled via $B$-splines. Many other examples can be found in the statistical literature. Often, the general parametric model takes the form $g(t; \beta_n, \tau_n)$, where for each $n$ the deterministic set of knots $\tau_n$ is such that $\dim(\tau_n) = O(n^\theta)$ for $\theta \in [0, 1)$ and $\beta_n$ are parameters to be estimated, with $\dim(\beta_n) = O(\dim(\tau_n)) = k_n$.

For a suitably chosen sequence of knots $\tau_n$, we assume that one can find a deterministic sequence of solutions, say $\beta_n^{\star\star}$, such that $\|g(\cdot;\beta_n^{\star\star},\tau_n)-h\|_\infty \to 0$ as $n\to\infty$. Such a sequence is assumed to exist, for establishing the consistency of the estimator $g(\cdot;\hat{\beta}_n,\tau_n)$ is futile otherwise. In the case of polynomial splines, the fact that $\|g(\cdot;\beta_n^{\star\star},\tau_n)-h\|_\infty \to 0$ can be established using the analytical properties of spline functions. Such results may be found, for example, in de Boor (1978) or Schumaker (1981).

Let $\hat{\beta}_n$ denote the maximum likelihood estimator of $\beta_n$ based on a sample of size $n$. To verify that $\|g(\cdot;\hat{\beta}_n,\tau_n)-h\|_\infty \xrightarrow{p} 0$ as $n\to\infty$, we must show that $\|g(\cdot;\hat{\beta}_n,\tau_n)-g(\cdot;\beta_n^{\star\star},\tau_n)\|_\infty \xrightarrow{p} 0$. However, it may be quite difficult to establish a direct link between $\hat{\beta}_n$, the stochastic solution to the likelihood equation, and $\beta_n^{\star\star}$, a deterministic vector with properties that are primarily governed by the relevant function approximation theory and that ostensibly has little to do with the estimating function which defines $\hat{\beta}_n$. Therefore, it may be necessary to introduce an intermediate quantity, say $\beta_n^{\star}$, in order to establish a link between $\hat{\beta}_n$ and $\beta_n^{\star\star}$. In view of Theorem 1, a natural choice is the sequence of least-false parameters $\{\beta_n^{\star} : E_{\mathcal{P}}\left(S_n(\beta_n^{\star})\right)=\mathbf{0}\}$.

To see this, we first apply the triangle inequality to further reduce the problem to demonstrating that $\|g(\cdot;\hat{\beta}_n,\tau_n)-g(\cdot;\beta_n^{\star},\tau_n)\|_\infty \xrightarrow{p} 0$ and $\|g(\cdot;\beta_n^{\star},\tau_n)-g(\cdot;\beta_n^{\star\star},\tau_n)\|_\infty \to 0$ for the two deterministic sequences $\beta_n^{\star}$ and $\beta_n^{\star\star}$. Under mild conditions on $g$ (e.g. bounded and continuous), it is then sufficient to prove that $\|\hat{\beta}_n-\beta_n^{\star}\|_\infty \xrightarrow{p} 0$ and $\|\beta_n^{\star}-\beta_n^{\star\star}\|_\infty \to 0$. These results will follow if the conditions of Theorem 1 (and an appropriately modified deterministic version thereof) are respectively satisfied by $\beta_n^{\star}$ and $\beta_n^{\star\star}$. More specifically, the existence of $\beta_n^{\star}$ and the fact that $\|\beta_n^{\star}-\beta_n^{\star\star}\|_\infty \to 0$ will be established if it can be demonstrated that there exists an $N$ such that for $n>N$,

- $\|\mathcal{I}_n^{-1}(\beta_n^{\star\star})\|_\infty \le M_1$, where $M_1<\infty$;
- there exists an $\varepsilon>0$ that may depend on $M_1$ such that

$$\sup_{\|\beta_n-\beta_n^{\star\star}\|_\infty<\varepsilon} \|\mathcal{I}_n(\beta_n)-\mathcal{I}_n(\beta_n^{\star\star})\|_\infty \le \frac{1}{2M_1};$$

and

- $\lim_{n\to\infty}\|E\left(S_n(\beta_n^{\star\star})\right)\|_\infty = 0$.

If these conditions hold, Theorem 1 may then be applied directly in order to prove that $\|\hat{\beta}_n-\beta_n^{\star}\|_\infty \xrightarrow{p} 0$. Strawderman and Tsiatis (1996) and Rossini and Tsiatis (1996) have obtained convergence results for specific problems using this general approach.

There is some overlap here with the elegant results of Stone (1994). In particular, under suitable regularity, Theorem 1 may be sufficient to guarantee

many of the results Stone must prove. For this reason, we suspect that the conditions which must be verified in order to use the $IFT$ may be moderately more restrictive. It would be interesting and useful to further identify the extent to which the two approaches are similar. The proof of consistency found in Murphy and Sen (1991) shares some similarities with both the method described here and that found in Stone (1994), and may help to further connect the two.

We remark that Theorem 1 has broad applicability, and is not specific to a particular application (e.g. splines). The most common use of this result is likely to be the case where $\hat{\beta}_n$ is the maximum likelihood estimator. However, it is important to note that all that we really require, subject to regularity, is an estimating function for $\hat{\beta}_n$. Thus, Theorem 1 may be used to prove consistency in the general $M$-estimation problem as long as the estimating function behaves nicely (i.e. as defined through the conditions of Theorem 1) as a function of $\beta_n$.

## Acknowledgements

Portions of this work were completed as part of RL Strawderman's dissertation (Strawderman (1992)). AA Tsiatis' research was partially supported by NIH grants AI-31789 and CA-51962. The authors would like to thank the Editor, Associate Editor and two reviewers for their helpful comments and a prompt review.

## References

Bickel, P., Klaassen, C., Ritov, Y. and Wellner, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore, MD.

de Boor, C. (1978). *A Practical Guide to Splines.* Springer-Verlag, New York.

Foutz, R. V. (1977). On the unique consistent solution to the likelihood equations. *J. Amer. Statist. Assoc.* **72**, 147-148.

Hjort, N. L. (1986). Bayes estimators and asymptotic efficiency in parametric counting process models. *Scand. J. Statist.* **13**, 63-85.

Hjort, N. L. (1992). On inference in parametric survival data models. *Internat. Statist. Rev.* **60**, 355-387.

Murphy, S. A. and Sen, P. K. (1991). Time-dependent coefficients in a Cox-type regression model. *Stochastic Process. Appl.* **39**, 153-180.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16**, 356-366.

Rossini, A. and Tsiatis, A. A. (1996). A semiparametric proportional odds model regression model for the analysis of current status data. *J. Amer. Statist. Assoc.* **91**, 713-721.

Rudin, W. (1964). *Principles of Mathematical Analysis.* McGraw-Hill, New York.

Schumaker, L. (1981). *Spline Functions: Basic Theory.* John Wiley and Sons, New York.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.* **22**, 118-171.

Strawderman, R. L. (1992). Statistical methods in the surrogate marker problem. Doctoral Dissertation, Dept. of Biostatistics, Harvard School of Public Health.

Strawderman, R. L. and Tsiatis, A. A. (1996). On the asymptotic properties of a flexible hazard estimator. *Ann. Statist.* **24**, 41-63.

Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, U.S.A.

Department of Biostatistics, Harvard University, School of Public Health, 677 Huntington Avenue, Boston MA 02115, U.S.A.

(Received November 1994; accepted January 1996)