The Annals of Statistics
1974, Vol. 2, No. 4, 703–712

POSTERIOR CONSISTENCY FOR COEFFICIENT ESTIMATION AND MODEL SELECTION IN THE GENERAL LINEAR HYPOTHESIS

BY ELKAN F. HALPERN

Rutgers—The State University

Berk (1970), LeCam (1953) and others have given conditions for the consistency of posterior distributions from a sequence of random variables. They have required that the sequence be i.i.d. We show that their results, Berk's in particular, may be extended to the general linear hypothesis with normal errors model (where the sequence of observations of the dependent variable need not be i.i.d.). We do not assume that the distribution governing the sequence of dependent variables has a regression function which satisfies the assumed model nor do we assume its errors are normal. Consistency is shown for both fixed and random sampling designs. We show that the convergence is to a projection of only the true regression function upon the space of regression functions given by the model. Finally, we assume that several such models are under consideration, each with a prior probability. We determine conditions for the a.s. convergence of their posterior probabilities to a degenerate distribution. Not all these conditions may be derived by any simple extension of Berk's results.

1. Introduction. Let $(X_1, Y_1), (X_2, Y_2), \cdots$ be a sequence of sample observations. Consider $Y_1, Y_2, \cdots$ to be the sequence of the values of a random dependent variable and $X_1, X_2, \cdots$ to be the sequence of corresponding independent variables. X_i may be scalar or vector-valued. Let the actual distribution of Y_i given X_i be determined by

$$Y_i = f(X_i) + \delta_i \,, \tag{1.1}$$

where $f(\cdot)$ is some arbitrary function and $\delta_1, \delta_2, \cdots$ is a sequence of i.i.d. random variables with mean zero and finite variance v^2. We will use Z to denote the common probability distribution of the δ's.

The sequence $X_1, X_2, \cdots$ will be considered to be either a fixed sequence of numbers (vectors) as in Section 4 or a sequence of i.i.d. random variables with common probability distribution W as in Section 3.

Assume that the family of models for the distribution of Y_i given X_i is determined by

$$Y_i = \sum_1^r \gamma_j g_j(X_i) + \varepsilon_i \,, \tag{1.2}$$

where $g_1(\cdot), g_2(\cdot), \cdots, g_r(\cdot)$ are known functions, $\varepsilon_1, \varepsilon_2, \cdots$ is a sequence of i.i.d. $N(0, \sigma^2)$ random variables and $\gamma_1, \gamma_2, \cdots, \gamma_r$ and σ^2 are the parameters of the

Received March 1972; revised May 1973.

AMS 1970 *subject classifications.* Primary 62J05, 62F10, 62F20; Secondary 62F15, 62F10, 62C10.

Key words and phrases. Consistency, estimation, model selection, general linear hypothesis, Bayesian.

family. In Section 5, we shall assume that there are T such families differing only in r and the functions $g_1(\cdot), \cdots, g_r(\cdot)$.

Let $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_r)$ and let $\mathbf{g}(\cdot)$ be the column vector defined by $\mathbf{g}(\cdot)' = (g_1(\cdot), g_2(\cdot), \cdots, g_r(\cdot))$. Our model may be reexpressed as

$$Y_i = \boldsymbol{\gamma}\mathbf{g}(X_i) + \varepsilon_i \,. \tag{1.3}$$

Let P_0 denote the (possibly improper) prior measure on the parameter space. Let $C_{\gamma,\sigma}$ denote the carrier of P_0 on the parameter space $(\boldsymbol{\gamma}, \sigma^2)$. Assume that there exists a $C_\gamma \subset R^r$ and a $C_\sigma \subset (0, \infty)$ such that $C_{\gamma,\sigma} = C_\gamma \otimes C_\sigma$. This assumption is included merely to assure that, given σ^2, the a.s. convergence of the conditional posteriors of $\boldsymbol{\gamma}$ is to a distribution of $\boldsymbol{\gamma}$ which is degenerate at a value of $\boldsymbol{\gamma}$, say $\boldsymbol{\gamma}^*$, which is independent of σ^2. The consequence is that the a.s. convergence of the marginal posteriors of $\boldsymbol{\gamma}$ is to this same $\boldsymbol{\gamma}^*$. Without this asumption, the vector $\boldsymbol{\gamma}^*$ given σ^2 may depend on σ^2 simply because some $\boldsymbol{\gamma}$'s are excluded from consideration for given σ^2. If so, we retain the a.s. convergence but do not have the convenient expression of $\boldsymbol{\gamma}^*$ as simply a projection of $f(x)$ onto a space spanned by $\mathbf{g}(x)$.

Let A be any set of the σ-field of C_γ. Let $P_n A$ denote the marginal posterior probability of the event $\boldsymbol{\gamma} \in A$ given $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$.

$$P_n A = \frac{\int_{A\otimes C_\sigma} \prod_1^n \phi[(Y_i - \boldsymbol{\gamma}\mathbf{g}(X_i))/\sigma]\, dP_0}{\int_{C_\gamma\otimes C_\sigma} \prod_1^n \phi[(Y_i - \boldsymbol{\gamma}\mathbf{g}(X_i))/\sigma]\, dP_0} \tag{1.4}$$

where

$$\phi(u) = \exp\{-\tfrac{1}{2}u^2\}/(2\pi)^{\frac{1}{2}} \,. \tag{1.5}$$

In Section 3, we assume that the sequence $X_1, X_2, \cdots$ is, itself, a sequence of sample observations of i.i.d. random variables. (That is, the sequence (X_1, Y_1), $(X_2, Y_2), \cdots$ is a sequence of i.i.d. random variables.) We use W to denote the common distribution function. Let

$$\lambda(\boldsymbol{\gamma}) = \int (f(x) - \boldsymbol{\gamma}\mathbf{g}(x))^2\, dW(x) \,. \tag{1.6}$$

By this definition, $\lambda(\boldsymbol{\gamma})$ is the distance of $f(x)$ from $\boldsymbol{\gamma}\mathbf{g}(x)$ in the $L_2(W)$ norm. In Theorem 3.1, we give conditions for the a.s. $[Z, W]$ convergence of P_n to a degenerate distribution. When these conditions are satisfied, the convergence is to the vector $\boldsymbol{\gamma}^*$ such that

$$\lambda(\boldsymbol{\gamma}^*) = \min_{\gamma\in C_\gamma} \{\lambda(\boldsymbol{\gamma})\} \tag{1.7}$$

provided the solution is unique. Thus, $\boldsymbol{\gamma}^*$ is the vector of coefficients of the projection of $f(x)$ onto the space spanned by $\mathbf{g}(x)$ with coefficients in C_γ. All that is needed to show the consistency is no more than proving that the assumptions made in Berk's paper are satisfied.

In Section 4, we assume that the sequence $X_1, X_2, \cdots$ is the non-random sequence of values of the independent variable(s) for a fixed sampling design. Treating $X_1, X_2, \cdots, X_n$ as a sample of size n, let W_n be the empirical cumulative sampling distribution. Assume $W_n(x) \to W(x)$ pointwise, where $W(x)$ is a probability measure. In Theorem 4.1, we give conditions for the a.s. $[Z]$ convergence of P_n to a degenerate distribution. We find that the convergence is to γ^* as defined by (1.6) and (1.7) again, where, of course, W has this new meaning. In this situation where the values of the dependent variable are the only random part of the sample, the assumptions in Berk's paper are not satisfied because the sequence $Y_1, Y_2, \cdots$ is not i.i.d. However, Berk used the assumption of identical and independent observations in proving only one lemma (his Lemma 2.1). We show that a result equivalent to this lemma holds for our problem too and that, as a consequence, all his other lemmas and theorems may be applied to prove consistency.

In Section 5, we assume that instead of one model, we have T alternative models, each with a prior probability. We find some additional constraints that guarantee by Berk's method of proof that when there is a.s. convergence for the coefficients for each of the models, there is a.s. convergence of the posterior probabilities of the models to a degenerate distribution. We conclude this paper by showing that, for at least a restricted set of priors, there will be a.s. convergence of these model probabilities even when all the additional constraints are not satisfied. These last more general constraints may not be derived by any direct extension of Berk's methodology.

To conclude this introduction, let us stress that neither of the assumptions concerning the form of the model are crucial for all of our results. The linear combination $\gamma\mathbf{g}(x)$ may be replaced by $g(\gamma, x)$ in all results except for Theorem 5.2. One need only assume that γ is estimable. In addition, the assumption of normal errors for the model is not necessary for the proofs of consistency in all theorems except 5.2. It may not even be necessary there. However, even though there may be consistency without normality, the convergence need not be to γ^* as we defined it.

2. A summary of Berk's results.

Berk assumed that $X_1, X_2, \cdots$ was an i.i.d. sample sequence with common distribution F. F also denoted the joint distribution of the sequence. He let $p(x \mid \theta)$ be the family of probability densities for $\theta \in (\Theta, \mathscr{A})$, the parameter space, which served as the model for the common distribution of the sequence. He assumed that there was a prior measure, P_0, on $(\Theta, \mathscr{A})$.

For all $A \in \mathscr{A}$, Berk let $P_n A$ denote the posterior probability of $\theta \in A$ given $X_1, X_2, \cdots, X_n$. He defined the function $l_n(\theta)$ of θ and $X_1, X_2, \cdots, X_n$ by

$$(2.1) \qquad l_n(\theta) = n^{-1} \textstyle\sum_1^n \ln [p(X_i \mid \theta) p^*(X_i)]$$

where p^* was some positive function. By this definition

$$(2.2) \qquad P_n A = \textstyle\int_A \exp\{n l_n\}\, dP_0 / \int_\Theta \exp\{n l_n\}\, dP_0 \,.$$

Berk then defined $\lambda(\theta)$ by $\lambda(\theta) = E \ln [p(X \mid \theta) p^*(X)]$ where the expectation was with respect to F. He showed in his Lemma 2.1 by means of the Strong

Law of Large Numbers for i.i.d. random variables that, if p^* was such that $P_0\{\lambda(\theta)$ exists$\} = 1$, then $F(l_n(\theta) \to \lambda(\theta)[P_0]) = 1$.

Berk then demonstrated that if $\lambda(\theta)$ is substituted for $l_n(\theta)$ in (2.2) for all n, $P_n A \to 0$ iff ess sup $\{\lambda(\theta)\colon \theta \in A\} <$ ess sup $\{\lambda(\theta)\colon \theta \in \Theta\} = \lambda^*$. Consequently, if $l_n(\theta) = \lambda(\theta)$ for all n and "if λ achieves its essential supremum λ^* at some point θ^* and is essentially bounded below λ^* off (open) neighborhoods of θ^*, then $\cdots P_n$ converges weakly to the distribution degenerate at θ^*."

Since, in general, one does not have $l_n(\theta) = \lambda(\theta)$ but rather a.s. convergence of $l_n(\theta)$ to $\lambda(\theta)$, more was needed to prove the a.s. convergence of P_n to the "distribution degenerate at θ^*." In this Theorem 3.3, Berk gave general conditions for this convergence. The material from his Lemma 2.1 through his point within these Theorem 3.3 contained the proofs of the sufficiency of these conditions. At no point within these proofs did he need or use any assumption of i.i.d. sample observations $X_1, X_2, \cdots$.

3. Random sampling designs. Let $X_1, X_2, \cdots$ be a sequence of sample observations of i.i.d. random variables with common distribution W. Then, by (1.1), the sequence $(X_1, Y_1), (X_2, Y_2), \cdots$ is also a sequence of observations of i.i.d. random variables. Thus, we may apply all of Berk's results directly.

Let us fix σ^2. We intend to show that the conditional posterior distributions of γ given σ^2 tend a.s. $[Z, W]$ to the distribution degenerate at γ^* (as defined in (1.6) and (1.7)) for all σ^2. Since γ^* is independent of σ^2, and v^2 is finite, it is immediate that the marginal posteriors $P_1, P_2, \cdots$, for the normal regression model also converge a.s. $[Z, W]$ to the same distribution.

Assume that under the model, $w_a(x)$ is the density of the distribution of X_i for all i. Let $p^*(x, y) = \{\phi[(y - f(x))/\sigma]w_a(x)\}^{-1}$. Then our analogue, $l(X, Y \mid \gamma, \sigma^2)$, of Berk's $\ln [p(X \mid \theta)p^*(X)]$ is given by

$$(3.1) \qquad \begin{aligned} l(X, Y \mid \gamma, \sigma^2) &= \ln \{\phi[(Y - \gamma g(X))/\sigma]w_a(X)p^*(X, Y)\} \\ &= \ln \{\phi[(Y - \gamma g(X))/\sigma]/\phi[(Y - f(X)/\sigma]\}\,. \end{aligned}$$

If, according to the model, the common distribution of $X_1, X_2, \cdots$ is not assumed to have a density, there still exists a $p^*(x, y)$ so that (3.1) holds.

Since by (1.5), $\phi(\cdot)$ is the density of a $N(0, 1)$ random variable, we may write

$$(3.2) \qquad \begin{aligned} -l(x, y \mid \gamma, \sigma^2) &= \{2y(f(x) - \gamma g(x)) + (\gamma g(x))^2 - (f(x))^2\}/2\sigma^2 \\ &= \{2(y - f(x))(f(x) - \gamma g(x)) + (f(x) - \gamma g(x))^2\}/2\sigma^2\,. \end{aligned}$$

If we take the expectation of $l(x, y \mid \gamma, \sigma^2)$ with respect to the actual distribution governing the data, we find that since Z is a probability measure with mean zero,

$$(3.3) \qquad \begin{aligned} E[l(X, Y \mid \gamma, \sigma^2)] &= \int\int l(x, y \mid \gamma, \sigma^2)\, dZ(y - f(x))\, dW(x) \\ &= -\int (f(x) - \gamma g(x))^2/2\sigma^2\, dW(x) \\ &= -\lambda(\gamma)/2\sigma^2\,. \end{aligned}$$

Thus, for our problem $-\lambda(\gamma)/2\sigma^2$ serves the role of $\lambda(\theta)$ in Berk's paper.

We have shown that if the conditions of Berk's Theorem 3.3 are satified, then the conditional posterior distributions of γ given σ^2 converge a.s. to the single degenerate distribution, which is degenerate at γ^* as defined by (1.7). The immediate consequence is that the marginal posteriors of γ, namely $P_1, P_2, \cdots$ also converge a.s. to this distribution.

Because we have assumed normal errors for our model, many of the conditions needed to prove Berk's Theorem 3.3 are automatically satisfied. What remains is given in the following theorem.

THEOREM 3.1. *When $X_1, X_2, \cdots$ are a random sample from a population with distribution W, then P_n converges* a.s. $[Z, W]$ *to the distribution degenerate at γ^* whenever*

(3.4a) $$C_{\gamma,\sigma} = C_\gamma \otimes C_\sigma$$

(3.4b) $$P_0\{\lambda(\gamma)/\sigma^2 < \infty\} > 0$$

(3.4c) $$\lambda(\gamma^*) < \lambda(\gamma) \qquad \forall \gamma \neq \gamma^*, \gamma \in C_\gamma$$

and (3.4d) P_n *becomes proper* a.s. $[Z, W]$..

These conditions are not very restrictive. Condition (3.4b) will be satisfied if all of $f(x), g_1(x), g_2(x), \cdots, g_r(x)$ are in $L_2(W)$. Condition (3.4c) requires that there be a unique γ^* so that $\gamma^*\mathbf{g}(x)$ is "closest" to $f(x)$. If C_γ is the entire r dimensional space, condition (3.4c) will be trivially satisfied.

Condition (3.4d) is trivially satisfied if P_0, itself, is proper. It is easy to show that if P_0 is the improper measure with density proportional to $\sigma^{-a}\, d\sigma\, d\gamma$, then the joint posteriors of γ and σ^2 (and, hence P_n also) become proper whenever $n \geqq r + a$ and the $r \times r$ matrix $\sum_1^n \mathbf{g}(X_i)'\mathbf{g}(X_i)$ is nonsingular. For the special case of polynomial regression ($g_k(X_i) = X_i^{(k-1)}$), we have shown (Halpern (1973a)) that the matrix is nonsingular if there are at least r distinct values among X_1, $X_2, \cdots, X_n$. For the special case of linear spline regression with $r - 2$ knots, we have shown (Halpern (1973b)) that nonsingularity follows from having at least one X_1, of $X_2, \cdots, X_n$ on each interval between successive knots.

Condition (3.4a) certainly may be relaxed. We have given it in this form rather than a more general one because we feel that any generalization is of only mathematical interest. Let us remark again that if (3.4a) is not satisfied one may use another $p^*(x)$ to get a $\lambda(\theta)$, restate the other conditions, and still have consistency. However, one no longer can prove that the convergence for the coefficients is to a function of $f(x)$ alone.

4. Fixed sampling designs. Let $X_1, X_2, \cdots$ be a sequence of fixed numbers (vectors). Then, neither sequence, $Y_1, Y_2, \cdots$ nor $(X_1, Y_1), (X_2, Y_2), \cdots$, is a sequence of i.i.d. random variables (unless $f(X_i) = f(X_j)$ for all i and j).

However, as we stated in Section 2, Berk used the fact that the sequence of observations was i.i.d. in only one proof, that of his Lemma 2.1. If we can prove that the consequences of Berk's Lemma 2.1 hold in this situation we may apply the argument of Section 3 to reach the desired conclusions.

In our Lemma 4.1, we give sufficient conditions for the equivalent results for our situation. Let $l(x, y\,|\,\gamma, \sigma^2)$ be as defined in (3.1). Let

$$l_n(\gamma, \sigma^2) = n^{-1} \textstyle\sum_1^n l(X_i, Y_i\,|\,\gamma, \sigma^2)\,. \tag{4.1}$$

Thus, our $l_n(\gamma, \sigma^2)$ is the analogue of Berk's $l_n(\theta)$ defined in (2.1). Let W_n be the empirical cumulative sampling distribution of $X_1, X_2, \cdots, X_n$ when they are treated as a sample of size n. Then, we may write

$$n^{-1} \textstyle\sum_1^n (f(X_i) - \gamma \mathbf{g}(X_i))^2 = \int (f(x) - \gamma \mathbf{g}(x))^2\, dW_n(x)\,.$$

LEMMA 4.1. *If there exists a probability measure W such that*

$$\int \mathbf{g}(x)'\mathbf{g}(x)\, dW_n(x) \to \int \mathbf{g}(x)'\mathbf{g}(x)\, dW(x)\,, \tag{4.2a}$$

$$\int f(x)\mathbf{g}(x)\, dW_n(x) \to \int f(x)\mathbf{g}(x)\, dW(x)\,, \tag{4.2b}$$

$$\int (f(x))^2\, dW_n(x) \to \int (f(x))^2\, dW(x)\,, \tag{4.2c}$$

and, if

$$\lambda(\gamma) < \infty\,, \qquad \textit{for all } \gamma \tag{4.2d}$$

then, if v^2 is finite,

$$Z(l_n(\gamma, \sigma^2) \to -\lambda(\gamma)/\sigma^2) = 1 \tag{4.3}$$

for all $(\gamma, \sigma^2) \in C_{\gamma,\sigma^2}$.

REMARK. Before we prove the lemma, note that each integral on the left-hand side of the conditions given in (4.2) may be written as an average for the first nX_i's. Note, also these conditions are equivalent to: for all $\gamma \in C_\gamma$,

$$\int (f(x) - \gamma\mathbf{g}(x))^2\, dW_n(x) \to \lambda(\gamma) = \int (f(x) - \gamma\mathbf{g}(x))^2\, dW(x) \tag{4.4}$$

and

$$\lambda(\gamma) < \infty\,.$$

In addition, they are implied by $W_n \to W$ weakly and $\limsup_{n\to\infty} \int [f^2(x) + \sum g_i^2(x)]\, dW_n(x) < \infty$ whenever $g_1, g_2, \cdots, g_r$ and f are all continuous functions.

PROOF OF LEMMA 4.1. We would like to apply Kolmogorov's Strong Law of Large Numbers for independent variables.

Let $a_i = f(X_i) - \gamma\mathbf{g}(X_i)$ and let $b_n = n^{-1}$. Define $U_i = a_i(Y_i - f(X_i))$. Then, by (1.1) and the assumption that Var $(Y_i - f(X_i)) = v^2 < \infty$ for all i, $U_1, U_2, \cdots$ is a sequence of independently distributed random variables such that $E(U_i) = 0$ and $E(U_i^2) = v^2a_i^2$ ($< \infty$, for all i).

Use (3.2) to write

$$-l(X_i, Y_i\,|\,\gamma, \sigma^2) = (2U_i + a_i^2)/2\sigma^2\,.$$

By (4.1), write

$$-l_n(\gamma, \sigma^2) = (2 \textstyle\sum_1^n U_i/n + \sum_1^n a_i^2/n)/2\sigma^2\,.$$

By (4.4), $n^{-1} \sum_1^n a_i^2 \to \lambda(\gamma)$. Hence by the Law of Large Numbers a sufficient condition for $n^{-1} \sum_1^n U_i \to 0$ a.s. $[Z]$ is $\sum_1^\infty a_j^2/j^2 < \infty$.

It is merely a matter of series manipulation to show that for all sequences of nonnegative numbers $a_1^2, a_2^2, \cdots, \lim_{n\to\infty} n^{-1} \sum_1^n a_j^2 < \infty$ implies that

$$\lim_{n\to\infty} \textstyle\sum_1^n a_j^2/j^2 < \infty .$$

Since (4.4) explicitly states that $\lim_{n\to\infty} n^{-1} \sum_1^n a_i^2 = \lambda(\gamma) < \infty$, we have shown that for all $\gamma \in C_\gamma$,

$$l_n(\gamma, \sigma^2) \to -\lambda(\gamma)/\sigma^2 \qquad \text{a.s.} \quad [Z] .$$

With this lemma proven, we may follow Berk's method through his Theorem 3.3, eliminating conditions our model automatically satisfies, and prove the following theorem.

THEOREM 4.1. *When $X_1, X_2, \cdots$ is the sequence of values of the independent variables for a fixed sampling design, when W_n is the empirical cumulative sampling distribution of $X_1, X_2, \cdots, X_n$ treated as a sample of size n, and when the probability measure W exists then P_n converge* a.s. $[Z]$ *to a distribution degenerate at γ^* whenever*

$$C_{\gamma,\sigma} = C_\gamma \otimes C_\sigma , \tag{4.5a}$$

(4.5b) *equations* (4.2a), (4.2b) *and* (4.2c) *are satisfied,*

$$\lambda(\gamma^*) < \lambda(\gamma) \qquad \forall \gamma \neq \gamma^*, \gamma \in C_\gamma , \tag{4.5c}$$

and

(4.5d) P_n *becomes proper* a.s. $[Z]$.

5. Model selection. Assume that there are T alternative models for the distribution of the sequence $(X_1, Y_1), (X_2, Y_2), \cdots$ instead of only one as assumed in the previous two sections. Assume, further, that each of the models fits the structure assumed for one model in Section 1. The only change in notation we need is the addition of the subscript t to any symbol that refers to a quantity or vector associated with a specific model, the tth. Finally, let $q_{0,t}$ denote the prior probability of the tth model and let $q_{n,t}$ denote the posterior probability of that model given $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$, where $\sum_t q_{n,t} = 1$.

We wish to determine when the a.s. convergence of $P_{n,t}$ implies that there exists an index, m, such that $q_{n,m} \to 1$ a.s. A very general theorem of this form may be stated as follows.

THEOREM 5.1. *If $X_1, X_2, \cdots$ is an* i.i.d. *random sequence and conditions* (3.4) *are satisfied for all t or if it is a fixed sequence and conditions* (4.5) *are satisfied, then, whenever*

$$C_{\sigma,t} = C_\sigma \qquad \textit{for all} \quad t \tag{5.1a}$$

and

$$\lambda_m(\gamma_m^*) < \lambda_t(\gamma_t^*) \quad \forall t \neq m \tag{5.1b}$$

are satisfied,

$$q_{n,m} \to 1 \qquad \text{a.s.} \quad [Z, W] \quad \textit{or} \quad [Z] . \tag{5.2}$$

This theorem is really a corollary to the two previous theorems. To prove it, note only that if we create the new model

$$Y_i = \sum_1^T \gamma_t \mathbf{g}_t(X_i) + \varepsilon_i$$

we have returned to the one model problem with the parameter space being the space of vectors of the form $(\gamma_1, \gamma_2, \cdots, \gamma_T, \sigma^2) = (\gamma, \sigma^2)$. The carrier of the implied prior of this parameter space is contained within the union of appropriate subspaces. Hence, with assumption (5.1a) to guarantee (3.4a) or (4.5a) for the new parameter space, all we need is that there exists a unique minimum to $\lambda(\theta)$ over the carrier to assure a.s. convergence to a unique θ contained within one of the subspaces. Since each $q_{n,t}$ is the total mass of a subspace for the nth posterior distribution, we have our result.

This theorem is almost as satisfactory as the two preceding ones. The only discomfort we feel with it is that we are more willing to assume (3.4c) or (4.5c) than we are to assume (5.1b). Any P_0 for which C_γ equals the entire r dimensional space will automatically guarantee a unique minimum to $\lambda(\gamma)$ at the projection. No equivalent type of statement may be made for more than one model. It is too easy to think of examples such as T nested models with an $f(x)$ orthogonal to them all. Theorem 5.1 would not imply posterior consistency here since $\lambda_t(\mathbf{0}) = 0$ for all t and, therefore, (5.1b) is not satisfied.

Yet there may well be posterior consistency, at least for some forms of $P_{0,t}$. The remainder of this paper is an attempt to show posterior consistency when (5.1b) is not satisfied. To do so, we restrict the type or priors we consider.

We shall assume the following: given σ^2, the prior distribution of γ_t is MVN $(\mu_{0,t}, \sigma^2 \Sigma_{0,t})$ for all t or it is the improper uniform on r_t dimensional space for all t. The marginal prior distribution of σ^2 is prior inverted gamma for all t, improper with density proportional to $\sigma^{-a}\, d\sigma$ for all t or degenerate at some fixed value σ_0^2 for all t. Thus, the priors we allow are the natural conjugate distributions given the model or the improper distributions representing the limits as the prior certainty about either γ_t or σ^2 or both approaches zero for all t.

With these priors $P_{n,t}$ may be explicitly expressed. Assume all assumptions except (5.1b) are satisfied. Then since, for all t, $P_{n,t}$ converges to the degenerate distribution with a spike at γ_t^*,

$$(5.3) \qquad q_{n,t} \sim b_1 |M_{n,t}|^{-\frac{1}{2}} \exp\{-b_2 \sum_1^n [Y_i - \gamma_t^* \mathbf{g}_t(X_i)]^2\}$$

for some positive constants b_1 and b_2 where $M_{n,t}$ is the $r_t \times r_t$ matrix defined earlier by $M_{n,t} = \sum_1^n \mathbf{g}_t(X_i)'\mathbf{g}_t(X_i)$. (A result similar to (5.3) is contained in Section 11.12 of DeGroot (1970)).

With the convergence of $P_{n,t}$ for all t implied by our assumptions, we also have that a.s. $\sum_1^n [Y_i - \gamma_t^* \mathbf{g}_t(X_i)]^2$ is asymptotically of the order $n[v^2 + \lambda_t(\gamma_t^*)]$, where $v^2 = E(Y_i - f(X_i))^2 < \infty$. This follows from $n^{-1} \sum_1^n [Y_i - \gamma_t^* \mathbf{g}_t(X_i)]^2 = \sum_1^n (Y_i - f(X_i))^2/n - 2\sigma^2 l_n(\gamma_t^*, \sigma^2)$.

The asymptotic order of $|M_{n,t}|^{-\frac{1}{2}}$ equals $n^{-\frac{1}{2}r_t}$. This follows from the fact that a typical element of $M_{n,t}$ equals $\sum_1^n g_{j,t}(X_i)g_{k,t}(X_i)$. When $X_1, X_2, \cdots$ is fixed, by definition and assumption (4.2a), $n^{-1}\sum_1^n g_{j,t}(X_i)g_{k,t}(X_i) \rightarrow \int g_{j,t}(x)g_{k,t}(x)\,dW(x)$. If the sequence is random, the convergence is a.s. $[W]$. The matrix $\int \mathbf{g}_t(x)'\mathbf{g}_t(x)\,dW(x)$ is nonsingular by the assumption that eventually $P_{n,t}$ becomes proper.

We have proven the following theorem.

THEOREM 5.2. *Assume all conditions of Theorem* 5.1 *are met except for* (5.1b). *Assume also that, for all t, $P_{0,t}$ is the natural conjugate or related improper form specified earlier. Let S be the largest subset of* $\{1, 2, \cdots, T\}$ *such that*

$$\lambda_s(\gamma_s^*) = \min_{1\leq t\leq T} \lambda_t(\gamma_t^*) \qquad \forall s \in S. \tag{5.4}$$

If there exists an $m \in S$ such that

$$r_m < r_s \qquad \forall s \in S,\ s \neq m, \tag{5.5}$$

then

$$q_{n,m} \rightarrow 1 \quad \text{a.s.} \quad [Z, W] \quad \text{or} \quad [Z]. \tag{5.6}$$

The implications of this theorem, especially for nested models, are pleasing. They seem to be a mathematical form of Occam's Razor. If, for instance, the tth model for the regression curve is of a $t-1$st degree polynomial in the univariate independent variable x and $f(x)$ is an $m-1$st degree polynomial in x, then under the weak assumptions given above, $q_{n,m} \rightarrow 1$ a.s. This convergence is independent of the actual form of the errors as long as $v^2 < \infty$. It is independent of the priors; as long as they are of the correct form they can be as misleading as possible and still the convergence holds. Finally, it is independent of the sampling design as long as the posterior distributions become proper.

When the models are not nested, this form of Occam's Razor still follows. If two models are equidistant from $f(x)$ in terms of $\lambda_t(\gamma_t^*)$, eventually one will believe in the model with fewer terms. Intriguingly, here though, the word fewer is used in reference to r_m, the number of terms in the mth model and not in reference to numbers of nonzero elements of γ_m^*.

Acknowledgment. The author is deeply indebted to Professor Robert Berk for his critical reading of early drafts and his many helpful suggestions which have been incorporated into this study.

REFERENCES

[1] BERK, R. H. (1970). Consistency a posteriori. *Ann. Math. Statist.* **41** 894–906.

[2] DEGROOT, M. (1970). *Optimal Statistical Decisions.* McGraw-Hill, New York.

[3] HALPERN, E. F. (1973a). Polynomial regression from a Bayesian approach. *J. Amer. Statist. Assoc.* **68** 137–143.

[4] HALPERN, E. F. (1973b). Bayesian spline regression when the number of knots is unknown. *J. Roy. Statist. Soc. Ser. B* **35** 347–360.

[5] LECAM, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *Univ. Calif. Publ. Statist.* **1** 277-300.

STATISTICS CENTER
RUTGERS UNIVERSITY
NEW BRUNSWICK, N. J. 08903