BULLETIN OF THE
INSTITUTE OF MATHEMATICS
ACADEMIA SINICA
Volume 7, Number 1, March 1979

ESTIMATION OF THE PRIOR DISTRIBUTION BY BEST APPROXIMATION IN UNIFORMLY CONVEX FUNCTION SPACES

BY

MICHAEL TORTORELLA AND THOMAS E. O'BRYAN

Abstract. We present a new technique based on some simple ideas from approximation theory for the solution of the problem of the estimation of the prior distribution. We first give the solution when the posterior density is known, and then when it is unknown but can be estimated. We show how the rate of convergence of the estimators of the prior is related to the rate of convergence of the estimators of the posterior and give a comprehensive example involving a location parameter family of $N(0, 1)$ distributions.

1. Introduction.

Let X be a random variable with density f. Suppose f is unknown but that the conditional densities of X given the values of an auxiliary random variable Y exist and are known. Put $g(x, u) = D_x P\{X \leq x | Y = u\}$ and suppose that the distribution of Y is G. Then f and G are related by the expression

$$f(x) = \int_{-\infty}^{\infty} g(x, u)\, dG(u)\,. \tag{1.1}$$

When we wish to emphasize the dependence on G we will write $f = f_G$. In the problem of "estimation of the prior distribution", g is known but G is not and f_G might not be. The aim in solving this problem is to obtain an estimate of G in the form of a sequence $\{G_n\}$ of *cdf*'s which converges weakly to G.

If independent observations $x_1, x_2, \cdots$ of X are available, the technique we propose for constructing such a sequential estimate of

Received by the editors October 1, 1977.

AMS (MOS) Subject Classifications (1970): Primary 60F05, secondary 41A50, 44A15, 62C99, 62G05.

Key words and phrases. Prior distribution, complete convergence, convexity, empirical Bayes.

the prior distribution G begins with a sequence of density estimators $\{f_n\}$ of f_G constructed from the observations. The nth member f_n of this sequence is a function of x which also depends on the first n observations $x_1,\cdots, x_n$. Having this sequence, what one would like to do is put f_n in on the left-hand side of (1.1), add a subscript n to the G, and treat the result as an integral equation for the unknown G_n in the hope that the sequence $\{G_n\}$ so obtained may be shown to be an estimator of G. This direct approach is in general doomed to failure because a sequence of density estimators is like an "approximate identity" and the identity operator is not representable as an integral operator in any of the "nice" function spaces. Thus in general, if we denote the integral operator by K,

$$KG(x) = \int_{-\infty}^{\infty} g(x, u)\, dG(u)\,, \tag{1.2}$$

f_n will fail to be in the image under K of the set $\mathscr{G}$ of cdf's and so the equation $KG_n = f_n$ will have no solution in $\mathscr{G}$. To overcome this difficulty we pose the problem in a uniformly convex function space (in our application we shall use $L^p(\mu)$ spaces, $1 < p < \infty$) and exploit the convexity properties of $\mathscr{G}$. What we then do is take G_n to be the best approximate solution (in the sense of approximation theory) of the equation $KG_n = f_n$. This means that $|KG_n - f_n| = \min\{|KG - f_n| : G \in \mathscr{G}\}$ (in the norm of the appropriate function space); in more descriptive terms we choose the $\bar{f}_n$ in the convex set $K(\mathscr{G})$ which is closest (in the norm) to f_n and then let G_n be the solution of the integral equation $KG_n = \bar{f}_n$. The method is simple and has great geometric appeal; moreover the construction of $\{G_n\}$ which we will describe is simple, straightforward, and effective.

The remainder of this paper is devoted to doing all of this carefully. In §2 is the main development of the procedure; in this section we suppose that f_G is known. In §3 we discuss the modifications which need to be made in case f_G is unknown but can be estimated by density estimators $\{f_n\}$. We discuss in §4 some basic results relating the rate of convergence of $\{G_n\}$ to G to the rate of convergence of $\{f_n\}$ to f_G.

We remark finally that this general framework can be applied to cover another situation, namely that in which $X = Y + Z$ with Y and Z independent. If Z has density k, then clearly (1.1) holds with $g(x, u) = k(x - u)$. The fifth and final section of the paper is devoted to a comprehensive example of this kind in which Z has the $N(0,1)$ distribution.

2. **Estimation of G when f is known.**

In this section we discuss the details of the construction of the estimates $\{G_n\}$ in the case where f is known. The method is an application of some elementary notions from approximation theory.

Let $\mathcal{D}$ be a class of *cdf*'s which is equicontinuous at infinity, that is, $1 - [F(a) - F(-a)]$ converges to zero as $a \to \infty$ uniformly over all F in $\mathcal{D}$. An example of such a class is the set $\mathcal{D}_R$ of *cdf*'s F which have the property that for some $R > 0$, $F(x) = 0$ for $x \leq -R$ and $F(x) = 1$ for $x > R$.

Let (S, Σ, μ) be a measure space and define the integral operator K on $\mathcal{D}$ by the expression (1.2), where

(A1) $g: S \times \boldsymbol{R} \to \boldsymbol{R}$ is Borel measurable and for each x in S the map $u \to g(x, u)$ is continuous and bounded.

K is well-defined because the boundedness of g entails the convergence of the integral for each x in S and G in $\mathcal{D}$. The function KG is also denoted by f_G.

A sequence $\{F_n\}$ of *cdf*'s is said to converge completely to a function F provided it converges weakly to F and F is itself a *cdf*. A necessary and sufficient condition for a sequence of *cdf*'s which converges weakly to a function to converge completely to the same limit is that the sequence be equicontinuous at infinity [11, p. 180]. In this section we shall show how to construct a sequence of *cdf*'s which converges completely to the prior distribution G. In order to do this we will require that g satisfy the following condition. Let m denote Lebesgue measure on the real line $\boldsymbol{R}$.

(A2) If G_1 and G_2 are nondecreasing functions on $\boldsymbol{R}$ and if KG_1 and KG_2 are defined and equal almost everywhere $[\mu]$, then G_1 and G_2 differ only by an additive constant.

We call this condition the *μ-identifiability condition.* For $\mu = m$ it is discussed by Teicher [16] and a characterization of this condition in spaces of continuous functions is given by Blum and Susarla [1]. Put briefly the condition says that if an equation of the form $KG = f$ has a nondecreasing solution, than it has exactly one nondecreasing solution, apart from an additive constant.

Let $\mathcal{H}$ denote the image $K(\mathcal{D})$ of $\mathcal{D}$ under the integral operator K. (A2) guarantees that K is injective on $\mathcal{D}$ so that K^{-1} is well-defined on $\mathcal{H}$. The following lemma shows that if $\{f_n\} \subset \mathcal{H}$ than the natural choice for an estimator of G works provided only that $\{f_n\}$ converge to f_G in measure $[\mu]$ on S.

LEMMA 2.1. *Let* (A1) *and* (A2) *obtain. Suppose* $\{f_n\} \subset \mathcal{H}$ *and* $\{f_n\}$ *converges to* f_G *in measure* $[\mu]$. *Put* $G_n = K^{-1}f_n$. *Then* $\{G_n\}$ *converges completely to* G.

Proof. G_n is well-defined by (A2) and $f_n \in \mathcal{H}$. Let $\{G_{n_k}\}$ be an arbitrary subsequence of $\{G_n\}$; by Helly's weak compactness theorem [11, p. 179] and the equicontinuity of $\mathcal{D}$ at infinity there is a subsequence $\{G_{n_{k_l}}\}$ and a *cdf* H in $\mathcal{D}$ with $\{G_{n_{k_l}}\}$ converging completely to H. Then by the Helly-Bray theorem [11, p. 182], $KG_{n_{k_l}}(x) \to KH(x)$ for each x in S. But $KH = KG$ a. e. $[\mu]$ because every sequence of functions which converges in measure has a subsequence with converges a. e. to the same limit. Then by (A2), $H = G + c$, but c must be zero, since both G and H are *cdf*'s. Thus every subsequence of $\{G_n\}$ has in turn a subsequence which converges completely to G, so $\{G_n\}$ converges completely to G.

In order to get a "good" choice of sequence $\{f_n\}$ in $\mathcal{H}$ we will exploit some convexity properties of $\mathcal{D}$. Note that the f_n are not yet related to any density estimators because f_G is assumed to be known. We require a third and final hypothesis about g. Put $b_p(x;\, G) = \int_{-\infty}^{\infty} g(x,\, u)^p \, dG(u)$. The condition imposed is

(A3) For some p in $]1, \infty[$, $|b_p|_{L^1(\mu)}$ is uniformly bounded over $\mathcal{D}$.

A sometimes useful different condition which is sufficient, but not necessary, for (A3) is

(A3′) $B_p(x) = \sup\{|b_p(x;\, G)| : G \text{ in } \mathcal{D}\}$ is integrable $[\mu]$ for some p in $]1, \infty[$.

A more readily verified condition which is sufficient for (A3) is

$$\text{(A3}''\text{)}\qquad \sup\left\{\int_{-\infty}^{\infty} |g(x, u)|^p \mu(dx) \colon u \text{ in } \mathscr{R}\right\} < \infty .$$

LEMMA 2.2. *Let* (A1) *and* (A3) *obtain. Then* $\mathscr{K}$ *is a compact convex subset of* $L^p(\mu)$.

Proof. The convexity of $\mathscr{K}$ is entailed by the convexity of $\mathscr{D}$ and the linearity of K. $\mathscr{K}$ is a subset of $L^p(\mu)$ by Jensen's inequality and (A3). To show $\mathscr{K}$ is compact, let $\{f_n\}$ be in $\mathscr{K}$, $f_n = KG_n$, G_n in $\mathscr{D}$. Helly's weak compactness theorem and the equicontinuity of $\mathscr{D}$ at infinity give the existence of a subsequence $\{G_{n_k}\}$ of $\{G_n\}$ and a *cdf* G with $\{G_{n_k}\}$ converging completely to G. By the Helly-Bray theorem, $f_{n_k}(x) \to KG(x)$ for each x in S. The inequality $|f_{n_k}(x) - KG(x)|^p \leq 2^{p-1}[|b_p(x; G_{n_k})| + |b_p(x; G)|]$ is used together with (A3) and the Lebesgue dominated convergence theorem to show that $f_{n_k} \to KG$ in $L^p(\mu)$, and this completes the proof.

Before proceeding, we wish to assuage the reader who has begun to wonder what all the fuss is about. After all if f is known and has the form (1.1), then all that need be done is solve (1.1) for G. If one is fortunate enough to be able to solve (1.1) in closed form then nothing more need be said. What is given here is a procedure for solving a type of equation (namely an integral equation of the first kind) which is notoriously difficult to solve. The procedure has the additional benefit that each of the approximate solutions is a *cdf*. An often-used method for solving such an equation is the eigenfunction expansion. This method is inappropriate here because the resulting sequence of functions which converges to G will not consist of *cdf*'s. We now make a final digression to discuss a few elementary properties of metric projections.

In a uniformly convex Banach space (such as $L^p(\mu)$, $1 < p < \infty$, [3]), a nonempty closed convex set contains a unique element of smallest norm [5, p. 74]. Hence for each f in $L^p(\mu)$ there is an $\bar{f}$ in $\mathscr{K}$ such that $|f - \bar{f}|_{L^p(\mu)} = \inf\{|f - h|_{L^p(\mu)} \colon h \text{ in } \mathscr{K}\}$. Define $P\colon L^p(\mu) \to \mathscr{K}$ by $P(f) = \bar{f}$. P is well-defined since $\bar{f}$ is uniquely determined by f, and is called the metric projection on $\mathscr{K}$. The following lemma gives some continuity properties of P.

LEMMA 2.3. *Let* $1 < p < \infty$. *Then* P *is a continuous map of* $L^p(\mu)$ *onto* $\mathcal{K}$. *If* $p = 2$, *then* P *is a nonexpansive map.*

Proof. The first assertion is Theorem 2 of [8]. For the second, see page 157 of [9].

We now return to the main thread of our development. The Krein-Milman theorem [5, p. 440] tells us that $\mathcal{K}$ is the closure of the convex hull of its extreme points. Since $L^p(\mu)$ is separable for $1 < p < \infty$, we may choose a countable dense subset $\{e_n\}$ of $\operatorname{ext}(\mathcal{K})$. In fact, the extreme points of $\mathcal{K}$ are the images of the extreme points of $\mathcal{D}$ under K because K is injective on $\mathcal{D}$ by (A2). Thus there are *cdf*'s E_n in $\mathcal{D}$ such that $e_n = KE_n$. If $\mathcal{D} = \mathcal{D}_R$, then the situation is particularly nice. Let V be the unit step (Heaviside) function; it can be shown directly or as a consequence of theorem V. 8. 6 of [5] that $\operatorname{ext}(\mathcal{D}_R) = \{V(x-u) \colon -R \le u \le R\}$. If $\mathcal{K}_R = K(\mathcal{D}_R)$ it follows that $\operatorname{ext}(\mathcal{K}_R) = \{g(x, u) \colon -R \le u \le R\}$. It is also easy to show that if $\{\eta_n\}$ is a countable dense subset of the interval $[-R, R]$ then $e_n = g(\cdot, \eta_n)$ will do for a countable dense subset of $\operatorname{ext}(\mathcal{K}_R)$ provided $D_2 g$ is in $L^p(\mu)$.

Let $S_n = \operatorname{co}\{e_1, \cdots, e_n\}$. It can be shown that as a consequence of (A2) the set $\{e_n\}$ is linearly independent, so that S_n is a simplex. The following lemma shows that the S_n's fill out $\mathcal{K}$ as $n \to \infty$.

LEMMA 2.4. $\mathcal{K} = \overline{\bigcup_{n=1}^{\infty} S_n}$.

Proof. Since $\mathcal{K}$ is closed and S_n is a subset of $\mathcal{K}$, we have the inclusion of the union in $\mathcal{K}$. For the reverse inclusion let f be an element of $\mathcal{K}$ and choose $\varepsilon > 0$. Then there exist $a_1, \cdots, a_n$ in $\operatorname{ext}(\mathcal{K})$ and nonnegative numbers $t_1, \cdots, t_n$ summing to one for which $|f - \sum_{i=1}^n t_i a_i| < \varepsilon/2$. For each i choose e_i so that $|e_i - a_i| < \varepsilon/2$. Then $|f - \sum_{i=1}^n t_i e_i| \le |f - \sum_{i=1}^n t_i a_i| + \sum_{i=1}^n t_i |e_i - a_i| < \varepsilon$. Since ε is arbitrary the result follows.

COROLLARY 2.5. *Let* f *be an element of* $L^p(\mu)$. *Then* $d(f, S_n) \to d(f, \mathcal{K})$ *as* $n \to \infty$.

Proof. Since S_n is included in S_{n+1} the sequence $\{d(f, S_n)\}$ is decreasing; since S_n is included in $\mathcal{K}$, $d(f, \mathcal{K})$ is a lower bound for it. Suppose $d(f, \mathcal{K}) + \delta$ were a lower bound for some $\delta > 0$. By the lemma there exist N and h with h in S_n for $n \geq N$ and $d(h, \mathcal{K}) < \delta/2$. Then $d(h, f) \leq d(h, \mathcal{K}) + d(f, \mathcal{K}) \leq d(f, \mathcal{K}) + \delta/2$. This contradiction shows that $d(f, \mathcal{K})$ is the greatest lower bound for $\{d(f, S_n)\}$, and from this the result follows.

COROLLARY 2.6. *Put* $P_n = P_{S_n}$. *If* f *is an element of* $L^p(\mu)$ *then* $P_n(f) \to P(f)$ *in* $L^p(\mu)$ *as* $n \to \infty$.

Proof. The proof of this is given on page 111 of [8].

COROLLARY 2.7. *If* f *is an element of* $\mathcal{K}$, *then* $P_n(f) \to f$ *in* $L^p(\mu)$ *as* $n \to \infty$.

Proof. This follows because $P(f) = f$.

The preparation for the final result is complete. Put $\bar{f}_n = P_n(f_G)$; note that $\bar{f}_n$ is an element of $\mathcal{K}$ and that as a consequence of Corollary 2.7 $\{\bar{f}_n\}$ converges in measure $[\mu]$ to f_G. Lemma 2.1 then applies to show that $\{G_n\}$ given by $G_n = K^{-1}\bar{f}_n$ is a desired sequential estimator of G. We collect these results as a theorem.

THEOREM 2.8. *Let* (A1), (A2), *and* (A3) *obtain. Then* $\{G_n\}$ *given by* $G_n = K^{-1} P_n(f_G)$ *converges completely to* G.

If it happens that $G \notin \mathcal{D}$ (i.e., we have not made a good choice of $\mathcal{D}$), Corollary 2.6 shows that the sequence $\{G_n\} \to G^* = K^{-1} P(f_G)$. G^* is a *cdf* in $\mathcal{D}$ and can be considered a "best approximation" to G out of $\mathcal{D}$ in the sense that for every $H \in \mathcal{D}$, $H \neq G^*$, $|f_G - KG^*|_p < |f_G - KH|_p$. In practical applications it often suffices to let $\mathcal{D} = \mathcal{D}_R$ for some sufficiently large R since a priori bounds on the support of the prior could be rationally chosen.

Since $P_n(f_G)$ is represented as a convex combination of $\{e_1, \cdots, e_n\}$, $K^{-1}P_n(f_G)$ is a convex combination (with the same coefficients) of $\{E_1, \cdots, E_n\}$. Once again if $\mathcal{D} = \mathcal{D}_R$ the situation is particularly nice because in that case the E_n's are unit step functions and G_n

is a step function with at most n points of discontinuity.

We have thus reduced the problem to the computation of the metric projection of f_G on S_n. This is the simplest possible kind of convex programming problem, and many good methods are available for its solution. When $p = 2$, it is possible to take advantage of the Hilbert space structure of $L^2(\mu)$. The method of [14] is applicable in this case; a similar but improved method is under development and will be published elsewhere [17]. This new method gives a procedure for the computation of the metric projection on a polytope in a Hilbert space by the solution of only a finite number of consistent systems of linear algebraic equations. Thus it enables one to compute $P_n(f_G)$ (in case $p = 2$) by a procedure which is linear and terminates after a finite number of steps, in contrast to many other iterative schemes for solving linear and nonlinear programming problems. Thus the procedure given here for estimation of G is certainly as simple as any other in the literature, and is often simpler. It has the further advantage that its development is completely obvious from elementary geometric considerations.

3. Estimation of G in case f is unknown.

In this section we shall discuss the modifications necessary to the procedure in §2 in case f_G is unknown. We shall suppose that there is a sequence $\{f_n\}$ of density estimators of f_G, constructed from the independent observations $x_1, x_2, \cdots$ from X. The modification is essentially to put $\bar{f}_n = P_n(f_n)$ instead of $P_n(f_G)$ and proceed naturally.

LEMMA 3.1. *Suppose* $\{f_n\}$ *converges to* f *in* $L^p(\mu)$. *Then* $P_n(f_n) \to P(f)$ *in* $L^p(\mu)$ *as* $n \to \infty$.

Proof. By Lemma A of [8] for each $\varepsilon > 0$ there is a $\delta > 0$ such that $|P_n(f) - P_n(f')| < \varepsilon$ whenever $|f - f'| < \delta$. The remainder of the proof uses Lemma 2.3.

LEMMA 3.2. *Suppose* (A3′) *holds,* $\{f_n\} \subset L^p(\mu)$, *and* $f_n \to f$ *in measure* $[\mu]$. *Put* $\hat{f}_n = \min\{f_n, B_p\}$. *Then* $\hat{f}_n \to f$ *in* $L^p(\mu)$.

Proof. This is a consequence of the Lebesgue dominated convergence theorem.

Observe that putting $\hat{f}_n$ equal to the minimum of f_n and the function of (A3″) would also work.

THEOREM 3.3. *Let* (A1), (A2), *and* (A3) *obtain. Suppose* $\{f_n\}$ *is a sequence of density estimates of* f_G *which converges to* f_G *in* $L^p(\mu)$ *almost surely. Then the sequence* $\{G_n\}$ *defined by* $G_n = K^{-1}P_n(f_n)$ *converges completely to* G *with probability* 1.

Proof. Let ω be a fixed element of the set of probability 1 on which $\{f_n\}$ converges to f_G. By Lemma 3.1 and the fact that $P(f_G) = f_G$ we obtain for the corresponding sequence $\{f_n(\cdot\,;\,\omega)\}$ the convergence of $\{P_n(f_n(\cdot\,;\,\omega))\}$ to f_G in $L^p(\mu)$, and hence in measure $[\mu]$. By Lemma 2.1 we obtain the complete convergence of $\{G_n\}$ defined by $G_n(x;\,\omega) = K^{-1}P_n(f_n(\cdot\,;\,\omega))(x)$ to G, and this completes the proof.

COROLLARY 3.4. *Let* (A1), (A2), *and* (A3′) *obtain. Suppose* $\{f_n\}$ *is a sequence of density estimates of* f_G *which converges to* f_G *in measure* $[\mu]$ *almost surely. Put* $\hat{f}_n = \min\{f_n, B_p\}$. *Then the sequence* $\{G_n\}$ *defined by* $G_n = K^{-1}P_n(\hat{f}_n)$ *converges completely to* G *with probability* 1.

Proof. This follows from Theorem 3.3 and Lemma 3.2.

Thus it is seen that the methods of §2 carry over completely to this case with only minor changes. The most significant difference is that the result obtained is probabilistic because the convergence of the density estimators to f_G is probabilistic.

Most often in practical applications this method will be employed with $S = \boldsymbol{R}$ and $\mu = m$, Lebesgue measure. The reason one might wish to use another measure space (S, Σ, μ) instead is that the treatment of various kinds of rate of convergence (of $\{G_n\}$ to G) problems may be facilitated by a suitable choice of (S, Σ, μ). We will discuss this problem in general in the next section and give an example in §5.

4. **Rate of convergence of the estimating sequence.** In this section we give a general framework for studying the rate of convergence of $\{G_n\}$ to G. We shall restrict ourselves to the case

$p=2$ in order to take advantage of the fact that in a Hilbert space the metric projection on a closed convex subset is Lipschitz with constant 1. Unfortunately this happy property does not carry over to even the nicest Banach spaces, even if the Lipschitz constant is allowed to vary from point to point [10]. We will first use the Lipschitz property of metric projections in $L^2(\mu)$ to get some information about the estimators $\{\bar{f}_n\}$ in terms of the estimators $\{f_n\}$.

THEOREM 4.1. *Suppose* $\{f_n\} \subset L^2(\mu)$ *converges to* f_G *in* $L^2(\mu)$. *Then* $|P(f_n)-f_G|_2 \leq |f_n - f_G|_2$.

Proof. $|P(f_n)-f_G|_2 = |P(f_n)-P(f_G)|_2 \leq |f_n - f_G|_2$.

COROLLARY 4.2. *Suppose that* $\sup\{|P_n(f)-P(f)|_2 \colon f \textit{ in } \mathcal{H}\} = \alpha_n$ *and that* (A3′) *holds with* $p=2$. *Let* $\{f_n\}$ *converge in measure* $[\mu]$ *to* f_G *and put* $\hat{f}_n = \min\{f_n, B_2\}$. *Then* $|P_n(\hat{f}_n) - f_G|_2 \leq |\hat{f}_n - f_G|_2 + \alpha_n$.

A result about the rate of convergence of $\{G_n\}$ to G is next.

THEOREM 4.3. *Let* (A1) *and* (A2) *obtain, and let* (A3) *hold with* $p=2$. *Let* X *be a normed linear space containing* $\mathcal{D}$. *Suppose* $K^{-1}\colon \mathcal{H} \to X$ *is bounded on* $\mathrm{span}(\mathcal{H})$, *i.e.* $\sup\{|K^{-1}f|_X |f^{-1}|_2 \colon f \textit{ in } \mathrm{span}(\mathcal{H}),\ f \neq 0\} = \kappa < \infty$. *Suppose* $\{f_n\}$ *converges to* f_G *in* $L^2(\mu)$ *with probability* 1. *Then* $\{G_n\}$ *defined by* $G_n = K^{-1}P(f_n)$ *has the property that for each* n, $|G_n - G|_X \leq \kappa |f_n - f_G|_2$ *with probability* 1.

Proof. Let ω be a fixed element of the set of probability one on which $\{f_n\}$ converges in measure to f_G. For the corresponding $f_n = f_n(\cdot;\ \omega)$ and $G_n = G_n(\cdot;\ \omega)$ we have $|G_n - G|_X = |K^{-1}(P(f_n) - f_G)|_X = |K^{-1}(P(f_n) - P(f_G))|_X \leq \kappa |P(f_n) - P(f_G)|_2 \leq \kappa |f_n - f_G|_2$, and this completes the proof.

COROLLARY 4.4. *Suppose that* $\sup\{|P_n(f)-P(f)|_2 \colon f \textit{ in } \mathcal{H}\} = \alpha_n$. *Then under the same conditions as Theorem* 4.3, $\{G_n\}$ *defined by* $G_n = K^{-1}P_n(f_n)$ *has the property that for each* n, $|G_n - G|_X \leq \kappa|f_n - f_G|_2 + \alpha_n$ *with probability* 1.

REMARKS 1. α_n is certainly always bounded by the diameter of

$\mathcal{K}$. Using the compactness of $\mathcal{K}$ it can be shown that in fact $\alpha_n \to 0$ as $n \to \infty$. If one can compute $P(f_n)$ directly then the inconvenience of the α_n is avoided.

2. Results analagous to Corollary 3.4 are obtained if we suppose that (A3′) holds instead of (A3).

3. The only difference between Corollary 4.2 and Theorem 3.3 is the hypothesis of the boundedness of K^{-1} on span($\mathcal{K}$). Certainly K^{-1} is well-defined on span($\mathcal{K}$) because (A2) implies that K is injective on span($\mathcal{D}$), and K^{-1} is linear. To get some boundedness of K^{-1} we may employ the freedom of choice we have of (S, Σ, μ) and X. Of course this freedom is not unlimited when g is given because (A1), (A2), and (A3) must continue to be satisfied. The example we will give in §5 shows how a balance can be struck among these conflicting requirements in case g arises from a location parameter family of $N(0, 1)$ densities. Broadly speaking what needs to be done is to make the topology of $L^2(\mu)$ strong enough (by choosing S, Σ, and μ) and the topology of X weak enough (by choosing the measure of the rate of convergence of $\{G_n\}$ to G) so that K becomes an open mapping on span ($\mathcal{D}$). Of course the choice of X depends also on what kind of rate of convergence result is desired for $\{G_n\}$ (e.g. $V(G_n - G) \to 0$, $|G_n - G|_\infty \to 0$, etc.), and the corresponding $L^2(\mu)$ tells us how strong the convergence of the estimators of f_G needs to be to obtain this rate of convergence.

5. **Example.** In this section we give an example which illustrates many of the points of the previous development. We consider (1.1) with a location parameter family of $N(0, 1)$ densities. Thus the equation we consider is

$$f(x) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} e^{-(x-u)^2/2}\, dG(u)\,. \tag{5.1}$$

To determine the set $\mathcal{D}$ we will use in this example we choose s linearly independent *cdf*'s $G^1, \cdots, G^s$ having finite second moments and let $\mathcal{D}_s = \mathrm{co}\{G^1, \cdots, G^s\}$. This is a kind of parametric assumption and is an expression of a belief that the prior distribution G is made up of the components $G^1, \cdots, G^s$ in varying amounts. We wish to stress that this kind of assumption is not required in general for the procedure in §§2 and 3, but was chosen in order to have a relatively simple example of the rate of convergence result in §4. As a rule the inverse of the integral operator K defined on the set of all *cdf*'s by

$$KG(x) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} e^{-(x-u)^2/2}\, dG(u)\,, \tag{5.2}$$

while it will be well-defined because the normal family is identifiable (see also Theorem VIII. 7. 5 of [7]), will not have good boundedness properties on any infinite-dimensional subset of its range. For an indication of why this should be so, consider K_1: $BC(\boldsymbol{R}) \to BC(\boldsymbol{R})$ defined by

$$K_1\phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} e^{-(x-u)^2/2}\, \phi(u)\, du\,.$$

Because of the identity

$$e^{-n^2/2}\, e^{inx} = (2\pi)^{-1/2} \int_{-\infty}^{\infty} e^{-(x-u)^2/2}\, e^{inu}\, du\,,$$

it is seen that for $n = 0, 1, \cdots$, $e^{-n^2/2}$ is an eigenvalue of K_1 of (real) multiplicity at least 2. These eigenvalues have a limit point at zero, and so zero must be in the spectrum of K_1. However, zero is not an eigenvalue of K_1 [7, Theorem VIII. 6. 3], so it must belong to either the continuous or the residual spectrum of K_1. In either case K_1^{-1} exists but is unbounded on the range of K_1.

In this example we shall measure the rate of convergence of $\{G_n\}$ to G by $V(G_n - G)$, where $V(G)$ stands for the total variation of G on the real line. Thus we shall consider K operating in the Banach space $NBV(\boldsymbol{R})$ of functions of bounded variation on the real line, normalized to zero at $-\infty$, with $|G| = V(G)$. We will also denote this norm by $|\ |_V$. Note that the set of *cdf*'s is a convex subset of the unit sphere of this space. We shall want the image $\mathcal{K}_s$ of $\mathcal{D}_s$ under K to lie in some $L^2(\mu)$, and this will be accomplished by a suitable choice of (S, Σ, μ). Before proceeding with this we wish to observe that while the above argument does not deal directly with the question of unboundedness of K^{-1}: $\mathrm{span}(\mathcal{K}_s) \to \mathrm{span}(\mathcal{D}_s)$, it does serve to indicate that K^{-1} will probably fail to be bounded on the range of K.

Choose $S = C$, the complex plane, Σ the σ-algebra of Borel-measurable subsets of C, and $\mu = w\, dm_2$ where m_2 is Lebesuge measure on C and w is a weight function given by $w(x, y) = (1 + x^2)(1 + y^2)^{-1}e^{-y^2}$. Then (A1) is clearly satisfied. Suppose KG_1 and KG_2 are equal a. e. $[\mu]$. Then they are equal everywhere on C because they are continuous, and in particular $KG_1(x) = KG_2(x)$ for all real x. It then follows from Theorem VIII. 7. 5 of [7] that (A2) is satisfied. For (A3), let $z = x + iy$ be a complex number. Then for every G in $\mathcal{D}_s$, there are nonnegative numbers $t_1, \cdots, t_s$ summing to 1 for which $G = \sum_{j=1}^{s} t_j G^j$, and

$$\int_C |b_2(z;\ G)|\, \mu(dz) \le (\sqrt{\pi}/2)\left[(5/4) + \max_{1\le j\le s} \int_{-\infty}^{\infty} u^2\, dG^j\,(u)\right],$$

which is finite and independent of G. Thus (A3) is also satisfied. Incidentally this is an example in which (A3′′) fails to be satisfied.

Let $\{f_n\} \subset L^2(\mu)$ be any sequence which converges to f_G in $L^2(\mu)$ (for example, $\{f_n\}$ might be some suitably chosen sequence of density estimators converging to f_G with probability 1). We may dispense with the approximating projections P_n in this example because $\mathcal{K}_s$ is itself a simplex and $P(f_n) \equiv P_{\mathcal{K}_s}(f_n)$ can be computed directly by the methods of [14] or [17]. Then Theorem 3.3 with $P_n = P$ for every n tells us that $\{G_n\}$ given by $G_n = K^{-1}P(f_n)$ converges completely to G (with probability 1 if appropriate). However, we are also interested in a rate of convergence result about $|G_n - G|_V$ (this is the reason for all the additional structure), and this, together with the inversion of K, is discussed below.

Let $\{\varepsilon_n\}$ be a sequence of positive numbers tending to zero as $n \to \infty$, and let $\{\bar{f}_n\} \subset \mathcal{K}_s$, $\bar{f}_n = KG_n$. Then for $|G_n - G|_V < \varepsilon_n$ it is necessary and sufficient that

$$|E_t(\bar{f}_n - f_G)|_{L^1(m)} < \varepsilon_n,\ 0 < t < 1, \tag{5.2}$$

[7, theorem VIII. 10. 3], where

$$E_t f(x) = (2\pi t)^{-1/2}\int_{-\infty}^{\infty} e^{-y^2/2t} f(x + iy)dy,\ 0 < t < 1. \tag{5.3}$$

$\mathcal{K}_s$ is a subset of the class A of Hirschman and Widder, so that if f is an element of $\mathcal{H}_s$ then $E_t f$ is given by (5.3) [7, p. 180]. The operator E_t is the e^{-tD^2} of [7, Chapter VIII], and [7, Theorem VIII. 7. 5] shows that $\lim_{t\to 1^-} E_t$, properly interpreted, inverts (5.1). With the additional structure we have imposed we will be able to show that (5.2) holds.

PROPOSITION 5.1. $\operatorname{span}(\mathcal{H}_s) \subset L^2(\mu)$.

Proof. It suffices to show that $KG^j \in L^2(\mu)$, $j = 1, \cdots, s$. We have

$$\begin{aligned}
&\int_C |KG^j(z)|^2 \mu(dz) \\
&\quad = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (2\pi)^{-1} \left| \int_{-\infty}^{\infty} e^{-(x+iy-u)^2/2}\, dG^j(u) \right|^2 \\
&\qquad\qquad\qquad \cdot (1+x^2)(1+y^2)^{-1} e^{-y^2}\, dx\, dy \\
&\quad \leq \frac{1}{2} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x-u)^2}\, dG^j(u)\, (1+x^2)\, dx \\
&\quad \leq \frac{\sqrt{\pi}}{2} \int_{-\infty}^{\infty} [\left(\frac{5}{4}\right) + u^2]\, dG^j(u) < \infty .
\end{aligned}$$

PROPOSITION 5.2. *For* $0 < t < 1$, E_t *is a bounded linear operator from* $L^2(\mu)$ *into* $L^1(m)$.

Proof. Let f be an element of $L^2(\mu)$. Then

$$\begin{aligned}
&|E_t f|_{L^1(m)} \\
&\quad = (2\pi t)^{-1/2} \int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} e^{-y^2/2t} f(x+iy)\, dy \right| dx \\
&\quad \leq (2\pi t)^{-1/2} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w(x,y)^{1/2}\, |f(x+iy)|\, e^{-y^2/2t}\, w(x,y)^{-1/2}\, dy\, dx \\
&\quad \leq |f|_{L^2(\mu)} (2\pi t)^{-1/2} \left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-y^2/t} e^{y^2} (1+x^2)^{-1} (1+y^2)\, dy\, dx \right)^{1/2} \\
&\quad = \text{constant} \times t^{-1/4} (1-t)^{-3/4}\, |f|_{L^2(\mu)} .
\end{aligned}$$

Thus $|E_t| \leq \text{constant} \times t^{-1/4}(1-t)^{-3/4}$ for $0 < t < 1$.

In order to satisfy (5.2) we need the bound on $|E_t|$ independent of t. This cannot be done over all of $L^2(\mu)$, but making use of the finite dimensionality of the subspace $\operatorname{span}(\mathcal{H}_s)$ this property can be obtained on $\operatorname{span}(\mathcal{H}_s)$. We shall write E_t^- for $E_t | \operatorname{span}(\mathcal{H}_s)$, i. e. the restriction of E_t to the subspace $\operatorname{span}(\mathcal{H}_s)$.

PROPOSITION 5.3. $\sup\{ |E_t^-| : 0 < t < 1\} = B < \infty$.

Proof.

$$\begin{aligned}
&|E_t\,KG^j|_{L^1(m)} \\
&\quad = (2\pi t^{1/2})^{-1}\int_{-\infty}^{\infty}\left|\int_{-\infty}^{\infty} e^{-y^2/2t}\int_{-\infty}^{\infty} e^{-(x+iy-u)^2/2}\,dG^j(u)\,dy\right| dx \\
&\quad = (2\pi t^{1/2})^{-1}\int_{-\infty}^{\infty}\left|\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-((1-t)/2t)y^2}\,e^{ivy}\,dy\,e^{-v^2/2}\,dG^j(x+v)\right| dx \\
&\quad = (2\pi(1-t))^{-1/2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x-u)^2/2(1-t)}\,dx\,dG^j(u) = 1\,.
\end{aligned}$$

Thus if $f = \sum_{j=1}^{s} \lambda_j\, KG^j$ we have $|E_t\, f\,|_{L^1(m)} \leq \sum_{j=1}^{s} |\lambda_j|$ for all t in $]0, 1[$. By Proposition 5.2 and because $\operatorname{span}(\mathscr{H}_s)$ is finite-dimensional, hence closed, the resonance theorem [5, Corollary II. 3. 21] applies to give the result.

The rate of convergence result we have been preparing for is the following.

THEOREM 5.4. *If* $\{f_n\}$ *converges to* f_G *in* $L^2(\mu)$ *(with probability* 1*) then* $G_n = K^{-1}P(f_n)$ *satisfies* $|G_n - G|_V \leq B|f_n - f_G|_{L^2(\mu)}$ *(with probability* 1*).*

Proof.
$$\begin{aligned}
|E_t(P(f_n) - f_G)|_{L^1(m)} &= |E_t(P(f_n) - P(f_G))|_{L^1(m)} \\
&\leq B\,|P(f_n) - P(f_G)|_{L^2(\mu)} \\
&\leq B\,|f_n - f_G|_{L^2(\mu)}.
\end{aligned}$$

The result now follows from (5.2).

REMARKS. 1. The restriction of finite second moments of $G^1, \cdots, G^s$ is not essential. With suitable modifications of the following development it can be replaced by conditions like finite rth moments for some $r > 1$ or compact support of all the $G^1, \cdots, G^s$.

2. Theorem 5.4 tells us that $V(G_n - G)$ is controlled by the rate of convergence of $\{f_n\}$ to f_G in $L^2(\mu)$. It would be of some interest to know exactly what B is, but this is not easy to discover in general because the proof of the resonance theorem is not constructive. Experiments with some special cases indicate that B behaves like $s^{1/2}$.

6. **Concluding remarks.** The principal advantages we wish to claim for our methods are two. The first is that only fairly "weak"

convergences of the estimators $\{f_n\}$ to f_G are required. In general we require only a convergence of averages in the x-domain (the a.s. $L^p(\mu)$ convergence), in contrast to, for example, the much stronger a.s. uniform convergence in the x-domain required for the method of [1]. In fact if (A3′) holds all that is required is convergence in measure (in the x-domain) almost surely. What we give up to gain this are simple proofs of the consistency and asymptotic unbiasedness of the estimators $P_n(\hat{f}_n)$. These problems are connected with some deeper unanswered questions in approximation theory, such as the uniform or a.e. convergence of $\{P_n(f_n)\}$ to f_G when $\{f_n\}$ converges uniformly or a.e. to f_G. The second advantage is ease of computation. The principal tool we use is the metric projection on a simplex. In a Euclidean or Hilbert space it is shown in [14] and [17] that this is a computationally reasonable procedure, with the method in [17] actually carrying out the solution (of this convex programming problem) by a completely linear procedure.

The applications of the estimation of the prior distribution are numerous in the literature. We mention only a few, including those found in [2]. [4], [6] and [15] and applications to the empirical Bayes decision theory problem of Robbins described in [1], [12], and [13]. Our method is applicable to all these cases, with appropriate adjustments.

References

1. J. Blum and V. Susarla, *Estimation of a mixing distribution function*, Ann. Probability **5** (1977), 200–209.

2. K. Choi and W. Bulgren, *An Estimation procedure for mixtures of distributions*, J. Roy. Statist. Soc. Ser. B. **30** (1968), 444–460.

3. J. Clarkson, *Uniformly convex spaces*, Trans. Amer. Math. Soc. **40** (1936), 396–414.

4. J.J. Deely and R.L. Kruse, *Construction of sequences estimatıng the mixing distributıon*, Ann. Math. Statist. **39** (1968), 286–288.

5. N. Dunford and J. Schwartz, *Linear operators*, part I, J. Wiley, New York 1957.

6. W. Gaffey, *A consistent estimator of a component of a convolution*, Ann. Math. Statist. **30** (1959), 198–205.

7. I. Hirschman and D. Widder, *The convolution transform*, Princeton University Press, Princeton, 1955.

8. R. Holmes, *Approximating best approximations*, Nieuw Arch. Wisk. **3** (1966), 106–113.

9. _______, *A course in optimization and best approximation*, Springer-Verlag Lecture Notes in Mathematics no. 257, Berlin, 1972.

10. R. Holmes and B. Kripke, *Smoothness of approximation*, Michigan Math. J. **15** (1968), 225–248.

11. M. Loeve, *Probability theory*, 3d edition, Van Nostrand, New York, 1963.

12. J.S. Maritz, *Smooth empirical Bayes estimation for continuous distributions*, Biometrika **54** (1967), 435–450.

13. _______, *Empirical Bayes methods*, Methuen and Co. Ltd. London, 1970.

14. B., Mitchell, V. Demyanov and V. Malozemov, *Finding the point of a polyhedron closest to the origin*, SIAM J. Control **12** (1974), 19–26.

15. P.E. Preston, *Estimating the mixing distribution by piecewise polynomial arcs*, Austral. J. Statist. **13** (1971), 64–76.

16. H. Teicher, *Identifiability of mixtures*, Ann. Math. Statist. **32** (1961), 244–248.

17. M. Tortorella, *Best approximation from a polytope*, (to appear).

BELL TELEPHONE LABORATORIES, HOLMDEL, NEW JERSEY 07733, U.S.A.
UNIVERSITY OF WISCONSIN, MILWAUKEE, WISCONSIN 53201, U.S.A.