The Annals of Statistics
2003, Vol. 31, No. 1, 174–200

NONPARAMETRIC ESTIMATION OF CONVEX MODELS VIA MIXTURES[1]

BY PETER D. HOFF

University of Washington

We present a general approach to estimating probability measures constrained to lie in a convex set. We represent constrained measures as mixtures of simple, known extreme measures, and so the problem of estimating a constrained measure becomes one of estimating an unconstrained mixing measure. Convex constraints arise in many modeling situations, such as estimation of the mean and estimation under stochastic ordering constraints. We describe mixture representation techniques for these and other situations, and discuss applications to maximum likelihood and Bayesian estimation.

1. Introduction. The problem of estimation under convex constraints arises frequently in statistical inference: moment estimation, monotonicity restrictions, quantile estimation and stochastic orderings are all situations which involve constrained spaces of probability measures. These types of constrained inference problems have been studied before in a variety of contexts. For example, inference on the location of a distribution can be achieved by profiling over convex sets of measures having a fixed location parameter: Owen (1988, 1990) discusses constrained estimation as a method of constructing nonparametric likelihood-based confidence intervals for the mean of a sampling distribution; Doss (1985), Diaconis and Freedman (1986) and Brunner and Lo (1989) each discuss nonparametric Bayesian methods of estimating the location of a sampling distribution by taking a linear model approach, assuming each observation is equal to a common location parameter plus some error term. These authors put separate priors on the location parameter and the error distribution, the latter being constructed to be median zero (Doss), to be symmetric about zero (Diaconis and Freedman), or to be unimodal with mode zero (Brunner and Lo). Thus, each of these approaches involves putting a prior on a convex set of probability measures.

As a slightly different example, many authors have considered estimation of a collection of probability measures subject to a partial stochastic ordering constraint. A set of collections of measures defined by each element of the set satisfying the same partial ordering constraint is convex. Brunk, Franck, Hanson and Hogg (1966) give a closed form expression for the MLEs of two measures constrained to be stochastically ordered. For an arbitrary number of

Received March 2001; revised December 2001.

[1] Supported by NIH Grant T32 CA09565-09.

AMS 2000 subject classifications. Primary 62G07; secondary 62C10.

Key words and phrases. Choquet's theorem, convex constraints, stochastic ordering, nonparametric estimation, Bayesian inference.

measures there is no closed form expression for the MLEs, however, Dardanoni and Forcina (1998) give an iterative method of estimation when the ordering is linear (each measure is either stochastically larger or smaller than every other measure in the collection). For a partial (not necessarily linear) ordering, Dykstra and Feltz (1989) give an iterative method based on Fenchel's duality theorem. Bayesian inference for the two sample stochastic ordering problem has been discussed by Arjas and Gasbarra (1996), who construct a Markov chain which samples from the space of two ordered hazard functions having support on a grid.

In this article we discuss a method of inference for such problems, applicable whenever the constrained set of measures C is convex. The method is very general, and is useful in many contexts, including both Bayesian inference and maximum likelihood estimation. The method makes use of the special properties of convex sets and a generalization of Choquet's theorem: under mild conditions, each point in a convex set of measures can be written as a mixture over the extreme points, or vertices, of the convex set, that is,

$$P = T(Q) = \int_{\operatorname{ex} C} P^* \, dQ(P^*).$$

The function T is called the barycenter mapping, and is a mapping from the space of mixing measures to the constrained space C of interest. Inference over C can be made via unconstrained inference over the set of mixing measures.

The remainder of this article is as follows: Section 2 discusses some of the theory involved in mixture representations, such as continuity of the barycenter map, measurability issues and identification of the extreme points. The special cases of moment constraints, quantile constraints and partial stochastic orderings are developed in detail. The results of Section 2 are applied in Section 3 to provide a simple means of maximizing a constrained likelihood, and examples are given in the case of univariate moment constraints and partial stochastic orderings. In Section 4, mixture representations are used as a method of constructing priors on constrained spaces of probability measures. Specifically, priors for location, location-scale and stochastic ordering problems are constructed. Posterior distributions for the mean and median are calculated assuming a Dirichlet process prior on the mixing measure Q, and the results are compared to Doss' (1985) posterior for the median. Advantages and disadvantages of the mixture representation approach are discussed in Section 5. All proofs are given in the Appendix.

2. Mixture representations of probability measures.

The set of extreme points of a convex set C is the subset $\operatorname{ex} C$ of C whose elements cannot be written as a convex combination of any two other points in C. If C is a compact subset of a vector space, then Choquet's theorem says that each element of C can be written as a mixture over the extreme points. Mixture representations for convex sets of probability measures have been studied in a variety of contexts: Dynkin (1978) examines the relationship between sufficient σ-algebras and unique mixture representations of probability measures, Diaconis and Freedman (1980) give a representation for finitely exchangeable sequences, and de Finetti's theorem gives a representation for infinitely exchangeable sequences. Additionally, representations for monotone measures and unimodal measures with a fixed mode have been studied by several authors, including Dharmadhikari and Joag-Dev (1988) and Bertin, Cuculescu and Theodorescu (1997).

More generally, von Weizsäcker and Winkler (1979, 1980) give conditions for the existence of mixture representations for convex sets of probability measures, and two corollaries in von Weizsäcker and Winkler (1979) will suffice for the needs of this article. In what follows, we assume:

- $\mathcal{X}$ is a separable metric space and $\mathcal{B}$ the Borel sets;
- $\mathcal{P}$ is the set of probability measures on $\mathcal{B}$;
- for $A \subset \mathcal{P}$, σ_A is the smallest σ-algebra on A such that for every $B \in \mathcal{B}$, the function $P \to P(B)$ is measurable in $P \in A$.

COROLLARY 1 (Weakly closed sets). *Let C be a convex, weakly closed subset of $\mathcal{P}$. Then for every $P \in C$ there is a probability measure Q on $\sigma_{\mathrm{ex}\,C}$ such that*

$$P(B) = \int_{\mathrm{ex}\,C} P^*(B)\,dQ(P^*) \qquad \forall\, B \in \mathcal{B}.$$

COROLLARY 2 (Moment sets). *Let $\mathcal{F}$ be a countable set of measurable functions. For each $f \in \mathcal{F}$, let I_f be a closed, possibly degenerate real interval. Let*

$$C = \left\{P \in \mathcal{P} : \mathcal{F} \subset L^1(P),\ \int f\,dP \in I_f\ \forall\, f \in \mathcal{F}\right\}.$$

Then for every $P \in C$, there is a probability measure Q on $\sigma_{\mathrm{ex}\,C}$ such that

$$P(B) = \int_{\mathrm{ex}\,C} P^*(B)\,dQ(P^*) \qquad \forall\, B \in \mathcal{B}.$$

For most sample spaces of interest, the above σ-algebras on subsets of $\mathcal{P}$ can be related to the more familiar σ-algebras generated by the topology of weak convergence w: if $\mathcal{X}$ is separable, then the σ-algebra generated by w is equivalent to that generated by the functionals $P \to P(B)$, $B \in \mathcal{B}$ [Karr (1986)]. The restrictions of these σ-algebras to $A \subset \mathcal{P}$ are also equivalent, as $\sigma_A = \sigma \cap A$ [Ash (1972)].

Given a set of probability measures A, we define $\mathcal{Q} = \mathcal{Q}(A)$ as the set of probability measures on σ_A. For a given Q, the probability measure $T(Q) \equiv P_Q$ defined by $P_Q(B) = \int_A P^*(B)\,dQ(P^*)$ is called the barycenter of Q, and the function $T : \mathcal{Q} \to \mathcal{P}$ is called the barycenter map. For a convex set $C \subset \mathcal{P}$, one strategy for estimating a measure $P \in C$ is to assume P is the barycenter of some unknown $Q \in \mathcal{Q}(\operatorname{ex} C)$. An unconstrained estimate $\hat{Q}$ of Q is made, and our estimate of P is taken as $T(\hat{Q})$. With this in mind, we would like to ensure that $T(Q) \in C$ for each $Q \in \mathcal{Q}$. If this inclusion does not hold, we would at least hope $T(Q)$ is "close" to C in some regard:

PROPOSITION 1. *Let $A \subset \mathcal{P}$, and Q be a measure on σ_A. Then $T(Q) \in \overline{\mathcal{H}A}$, the weakly closed convex hull of A.*

Therefore, if C is a convex set for which an integral representation holds and $\mathcal{Q}$ is the set of measures on $\sigma_{\operatorname{ex} C}$, then $C \subset T(\mathcal{Q}) \subset \overline{C}$. Of course if C is closed, we have $T(\mathcal{Q}) = C$. This result provides a version of the Krein–Milman theorem for mixtures of probability measures.

2.1. *Closure.* Closure of a convex set C has important implications. A closed convex set of probability measures will have extreme points, which is not necessarily the case for nonclosed convex sets. For example, consider the set $\{P : P = p\delta_x + (1 - p)\delta_y, p \in (0, 1)\}$. This is a convex set, but is not closed, and does not have any extreme points. Closure of a set also implies an integral representation for each element of the set (Corollary 1). However, nonclosure does not preclude extreme points or an integral representation (Corollary 2).

Checking whether or not a convex set of probability measures is closed is typically quite easy: if the sample space $\mathcal{X}$ is a separable metric space, then the topological space $\mathcal{P}$ with the weak topology is metrizable as a separable metric space [Parthasarathy (1967), Theorem 2.6.2]. Therefore, if $P \in \overline{C}$, there exists a sequence $\{P_n\}_1^\infty \subset C$ converging to P. Thus, to check closure, we just need to check that every convergent sequence in C converges to a point in C.

Moment constraints. Let $C_\theta \subset \mathcal{P}$ be the set of mean-θ probability measures on a compact sample space $\mathcal{X}$, say, $\mathcal{X} \subset [-c, c]^K$ for some $c \in \mathbb{R}$ and integer K. To show C_θ is closed, let $\{P_n\}_1^\infty \subset C_\theta$ be a sequence weakly converging to $P \in \mathcal{P}$. Note that the K component functions $f_i(x) = x_i,\ i = 1, \dots, K$, are bounded continuous functions on $\mathcal{X}$, and by the properties of weak convergence we have $|\int x_i\, dP(x) - \int x_i\, dP_n(x)| \to 0$ as $n \to \infty$. Since this difference is equal to $|\int x_i\, dP(x) - \theta_i|$ for all n, we have $P \in C_\theta$ and so C_θ is closed. Therefore, by Proposition 1, $T(\mathcal{Q}) = C_\theta$, where $\mathcal{Q}$ is the space of mixing measures on the extreme points of C_θ.

The closure of C_θ does not necessarily hold for arbitrary sample spaces. If $\mathcal{X}$ is not compact, then in general we do not have closure. For example, let $\mathcal{X} = \mathbb{R}$, $\theta = 0$ and $P_n = \sum_1^n (\delta_{2^i} + \delta_{-2^i})2^{-i}/2(1 - 2^{-n})$. Then P_n converges weakly to $P = \frac{1}{2}\sum_1^\infty (\delta_{2^i} + \delta_{-2^i})2^{-i}$, but the first moment of P does not exist, so $P \notin C_\theta$. However, an integral representation for C_θ exists by Corollary 2, and so by Proposition 1, we have $C_\theta \subset T(\mathcal{Q}) \subset \overline{C_\theta}$.

Quantile constraints. When our sample space $\mathcal{X}$ is the real line, we say a probability measure P satisfies a set of K quantile constraints, given by $\theta \in \mathbb{R}^K$, $\alpha \in [0,1]^K$, if $P(-\infty, \theta_i) \leq \alpha_i \leq P(-\infty, \theta_i]$ for $i = 1, \dots, K$. We denote $C_{\theta,\alpha}$ as the set of all probability measures satisfying the constraints given by θ and α. As a simple example, the space of median-zero probability measures is $C_{\theta,\alpha}$ with $\theta = 0$ and $\alpha = 1/2$.

It is easy to check that $C_{\theta,\alpha}$ is convex. Closure is also easy to check: if $\{P_n\}_1^\infty$ is a sequence of measures in $C_{\theta,\alpha}$ converging to a measure P, then

$$\alpha_i \leq \limsup_{n\to\infty} P_n(-\infty, \theta_i] \leq P(-\infty, \theta_i],$$
$$\alpha_i \geq \liminf_{n\to\infty} P_n(-\infty, \theta_i) \geq P(-\infty, \theta_i),$$

which shows P also satisfies the constraints. Therefore $C_{\theta,\alpha}$ is weakly closed and so by Corollary 1, we have $T(\mathcal{Q}) = C_{\theta,\alpha}$, where $\mathcal{Q}$ is the space of mixing measures over the extreme points of $C_{\theta,\alpha}$.

2.2. *Finding the extreme points.* In order to make use of the integral representation theorems we need to identify the extreme points of a given convex set of probability measures. A strategy for finding the extreme points in some important cases is as follows:

DEFINITION 1. For a convex set of measures C, a set $s \in \mathcal{B}$ is an *extreme support* of C if there is one and only one probability measure $P_s \in C$ having a support set equal to s.

PROPOSITION 2. *Let $\mathcal{S}$ be the set of extreme supports of C. Then for each $s \in \mathcal{S}$:*

(i) *the measure P_s is extreme in C, and*
(ii) *there are no $P \in C$ with support equal to a proper subset of s.*

This suggests sets in $\mathcal{S}$ are "small" in the sense that no proper subset of a set in $\mathcal{S}$ can support a measure in C. As an example, consider $C = \mathcal{P}$, the set of all measures on a space $\mathcal{X}$. The "smallest" set that can be the support of a probability measure is a singleton $\{x\} \subset \mathcal{X}$, and every point-mass measure is extreme.

Moment constraints. First consider the simple case where the sample space $\mathcal{X}$ is some subset of the real line and our convex set of interest is C_θ, the space of mean-θ measures. We assume $\mathcal{X}$ includes at least one point greater than or equal to θ and at least one point less than or equal to θ. Using the above ideas, we look for a class of sets $\mathcal{S}$ so that each $s \in \mathcal{S}$ supports only one mean-θ measure. If a support set s consists of only one point, then clearly that point must be θ. Thus

δ_θ is an extreme point. If a support set contains only two points then clearly one of them is above θ and one below. In fact, for $s_1 < \theta < s_2$, there is one and only one mean-θ measure on $s = (s_1, s_2)$, given by

$$P_s = \frac{s_2 - \theta}{s_2 - s_1}\delta_{s_1} + \frac{\theta - s_1}{s_2 - s_1}\delta_{s_2}.$$

Every such two-point measure is therefore extreme in C_θ. Conversely, if P has support on more than two such points, it can be represented as a mixture of two other mean-θ distributions, and so is not extreme.

The situation is similar for $\mathcal{X} \subset \mathbb{R}^K$. In this case, given any set $s = \{s_1, \dots, s_k\}$ of $k \le K + 1$ points of $\mathcal{X}$ such that the vectors $s_i - \theta$ are affinely independent and θ is in their convex hull, there is one and only one mean-θ measure with support on s. As was proven by von Weizsäcker and Winkler, such measures constitute the extreme points of C_θ:

THEOREM 1 [von Weizsäcker and Winkler (1980)]. *A measure $P \in C_\theta$, the set of mean-θ measures, is extreme in C_θ if and only if its support lies in the set $\mathcal{S} = \{(s_1, \dots, s_k) : s_j \in \mathcal{X}$ for $j = 1, \dots, k \le K + 1$; $s_1 - \theta, \dots, s_k - \theta$ are affinely independent$\}$.*

For dimensions $K > 1$, an extreme point P can be constructed by first selecting $s_1, \dots, s_k, k \le K$, such that the vectors $s_1 - \theta, \dots, s_K - \theta$ are linearly independent. Then let $s_{k+1} = \theta + \sum_1^k \gamma_i(\theta - s_i)$ for some $\gamma_i > 0$, that is, let s_{k+1} lie in the strictly positive hull of the rays emanating from θ in directions away from $s_1, \dots, s_k$. The unique extreme point P with support on $s_1, \dots, s_{k+1}$ is then given by $P(s_i) = \gamma_i/(1 + \sum \gamma_j),\ i = 1, \dots, k$.

An interesting special case is the convex set $C_{(0,1)}$ of univariate mean-zero, variance-one distributions. To apply the above theorem, we can write our one-dimensional sample space as the curve $\{(x, x^2), x \in \mathcal{X}\}$, and our convex constraint becomes $(E(X), E(X^2)) = (0, 1)$. The extreme points of $C_{(0,1)}$ are thus:

- measures with support on two points, x and $-1/x$, and
- measures with support on three points x_1, x_2, x_3, satisfying

$$x_1 < -1/x_2 < 0 < -1/x_1 < x_2$$

and

$$-1/x_2 < x_3 < -1/x_1.$$

The measures in $C_{(0,1)}$ with such supports are given respectively by $P(x) = \frac{1}{1+x^2} = 1 - P(-1/x)$ and $P(x_i) = \frac{1+x_j x_k}{(x_i - x_j)(x_i - x_k)}$ for permutations i, j, k of $(1, 2, 3)$.

Quantile constraints. Consider first the space of median-θ distributions. It is clear that if a point mass measure is median-θ then it must be a point mass on θ, and so δ_θ is extreme. Now suppose a measure P has support on two points, s_1 and s_2, neither of which is θ. Such a measure is median-θ if and only if s_1 and s_2 are on opposite sides of θ and P puts equal mass $1/2$ on the two points. Therefore, $P_s = (\delta_{s_1} + \delta_{s_2})/2$ is extreme for each $s_1 < \theta < s_2$ by the above results. The situation is similar for more general quantile constraints, as shown by the following result:

PROPOSITION 3. *Let $\theta \in \mathbb{R}^K$, $\alpha \in [0, 1]^K$ such that $\theta_1 < \cdots < \theta_K$ and $\alpha_1 < \cdots < \alpha_K$. Then a measure P is extreme in $C_{\theta,\alpha}$ if and only if it can be written as*

$$P_s = \sum_{i=1}^{K+1} (\alpha_i - \alpha_{i-1})\delta_{s_i}$$

for $\alpha_0 = 0, \alpha_{K+1} = 1$ and $s_1 \le \theta_1 \le \cdots \le \theta_K \le s_{K+1}$.

With this notation it is possible that $s_i = \theta_i = s_{i+1}$, which may seem redundant. However, this notation allows us to identify each extreme point by a $(K + 1)$-dimensional vector $s = \{s_1, \dots, s_{K+1}\}$. The result above makes sense in light of our previous observations: extreme points of sets of probability measures tend to have small supports, and the above measures have the smallest supports possible while still putting a certain amount of mass in each of the $K + 1$ intervals, as required by the constraint.

2.3. *Reparametrizations.* As seen in the above examples, the set of extreme points $\operatorname{ex} C$ can often be indexed by a subset $\mathcal{S}$ of a finite-dimensional Euclidean space. In such cases, it may be desirable to reparametrize the set of mixing measures on $\operatorname{ex} C$ in terms of measures on $\mathcal{S}$. What is required is simply that the integral of P_s over $s \in \mathcal{S}$ is well defined, that is, the indexing function $s \to P_s \in \operatorname{ex} C$ satisfies a measurability condition. This and some of the preceding results are summarized in the following proposition:

PROPOSITION 4. *Let C be a convex set of probability measures on $(\mathcal{X}, \mathcal{B})$ for which there is an integral representation. Let $\mathcal{S}$ be a regular topological space and $P_{(\cdot)}$ be a measurable map from $\mathcal{S}$ onto $\operatorname{ex} C$. Let $T(Q) = \int P_s(\cdot)\, dQ(s)$ for each $Q \in \mathcal{Q}$, the set of probability measures on $\mathcal{B}(\mathcal{S})$. Then $C \subset T(\mathcal{Q}) \subset \overline{C}$.*

Note that $P_{(\cdot)}$ is a measurable function from $\mathcal{S}$ to $\operatorname{ex} C$ if and only if $P_{(\cdot)}(B)$ is measurable for each $B \in \mathcal{B}(\mathcal{X})$. From this, one can easily show that for both of our preceding examples, the indexing functions

$$P_{(\cdot)}\colon\ P_s = \frac{s_2 - \theta}{s_2 - s_1}\delta_{s_1} + \frac{\theta - s_1}{s_2 - s_1}\delta_{s_2}$$

and

$$P_{(.)}\colon\ P_s = \sum_1^{K+1} (\alpha_i - \alpha_{i-1})\delta_{s_i}$$

are Borel measurable for $\mathcal{S} = \{(s_1, s_2) \in R^2 : s_1 \le \theta < s_2\}$ and $\mathcal{S} = \{s \in \mathbb{R}^{K+1} : s_1 \le \theta_1 \le \cdots \le \theta_K \le s_{K+1}\}$, respectively, and thus satisfy the conditions of the proposition.

2.4. *Continuity.* Suppose a mixing measure Q is close to Q_0. Does this imply $T(Q)$ will be close to $T(Q_0)$? This is an important question in maximum likelihood inference, as the method of estimation is often based on a sequence of estimates $\{\hat{Q}_l\}_{l=1}^{\infty}$ converging to the MLE $\hat{Q}$. If T is continuous, then we can be assured that $\hat{Q}_n \to \hat{Q}$ implies $\hat{P}_n = T(\hat{Q}_n) \to T(\hat{Q}) = \hat{P}$. The question of continuity is also relevant in Bayesian inference: If a prior π for Q has support on the entire space of mixing measures, then continuity of T guarantees that the resulting induced prior for $P \in C$ will have support on all of C. This can be seen as follows: let A be a weakly open set in C. Then $\Pr(P \in A) = \Pr(T(Q) \in A) = \pi(T^{-1}A) > 0$, since $T^{-1}A$ is an open set by the continuity of T.

PROPOSITION 5. *Let $P_{(.)} : \mathcal{S} \to \mathrm{ex}\, C$ be a mapping from a space $\mathcal{S}$ to the extreme probability measures on C, a subset of probability measures on $\mathcal{B}(\mathcal{X})$. If $g(s) = \int_{\mathcal{X}} f(x)\, dP_s(x)$ is a bounded, continuous function of s for every $f \in C_b(\mathcal{X})$, then T is a weakly continuous mapping from $\mathcal{Q}$, the space of measures on $\mathcal{B}(\mathcal{S})$, to $\overline{C}$.*

Moment constraints. For any bounded, continuous function f on $\mathbb{R}$, we have

$$g_f(s) = \frac{s_2 - \theta}{s_2 - s_1} f(s_1) + \frac{\theta - s_1}{s_2 - s_1} f(s_2),$$

which is a bounded continuous function of $s \in \{(s_1, s_2) \in \mathbb{R}^2 : s_1 \le \theta < s_2\}$. Thus T mapping $\mathcal{S}$ to the mean-θ probability measures is weakly continuous.

Quantile constraints. Similarly, in the case of quantile constraints we have $g_f(s) = \sum_1^{K+1} (\alpha_i - \alpha_{i-1}) f(s_i)$ which is a bounded, continuous function of s. Thus the barycenter mapping is weakly continuous.

2.5. *Inversion of the barycenter map.* Maximum likelihood estimation of a measure P is often based on an iterative procedure which starts with a value $\hat{P}_0$ which we think is "close" to the MLE $\hat{P}$. Estimation of P via Q thus requires finding a $\hat{Q}_0 \in T^{-1}\hat{P}_0$. In a Bayesian analysis, prior information may suggest P is near some measure P_0. We would then want to make sure our prior for Q puts mass near a $Q_0 \in T^{-1}P_0$. Finding such a Q may be difficult in general, but closed form expressions exist for some simple special cases.

Univariate moment constraints. By the results of the previous sections, any mean-θ measure P can be written as $T(Q) = \int P_s \, dQ(s)$ for some measure Q on $\mathcal{B}(\mathcal{S})$. A closed form expression for one such Q is given as follows: let P be a mean-θ measure having density p with respect to a measure μ. Let:

- $2\gamma = \int_{\mathbb{R}} |x - \theta| p(x) \, d\mu(x)$,
- $q(s) = \frac{1}{\gamma}(s_2 - s_1) p(s_1) p(s_2)$,
- $\mu_{-+} = \mu|_{(-\infty,\theta]} \times \mu|_{(\theta,\infty)}$.

A simple integration shows that Q, given by $Q(A) = \int_A q(s) \, d\mu_{-+}$, is a probability measure such that $T(Q) = P$.

Quantile constraints. Let $P \in C_{\theta,\alpha}$ be a measure such that $P\{\theta_i\} = 0$, $i = 1, \dots, K$, that is, P has no atoms on the constraint points. In this case, $P(\theta_{i-1}, \theta_i] = \alpha_i - \alpha_{i-1}$, and any Q representing P must satisfy

$$P(B) = T(Q)(B) = \sum_{i=1}^{K+1} (\alpha_i - \alpha_{i-1}) Q(s_i \in B)$$

for each Borel set B. It suffices to solve this equation for sets $B \subset (\theta_{i-1}, \theta_i]$, $i = 1, \dots, K + 1$. For such a B, the condition becomes

$$P(B) = (\alpha_i - \alpha_{i-1}) Q(s_i \in B) \quad \Longrightarrow \quad Q_i(B) = P\bigl(B | (\theta_{i-1}, \theta_i]\bigr),$$

that is, the ith marginal of Q is the conditional probability of B given $(\theta_{i-1}, \theta_i]$. One possible representer Q of P is the product measure given by $Q(B_1 \times \cdots \times B_{K+1}) = \prod_1^{K+1} P(B_i | (\theta_{i-1}, \theta_i])$. A slight modification of this construction will be necessary if there are i's for which $P\{\theta_i\} > 0$, in which case the mass at θ_i may have to be shared by s_i and s_{i+1}.

2.6. *Convex collections of probability measures*: *Stochastic orderings.* Many of the above ideas are applicable when considering convex collections of measures. We illustrate this with the particular example of partial stochastic orderings. Given two measures P_1, P_2 on $\mathcal{X} \subset \mathbb{R}$, P_2 is said to be stochastically larger than P_1 if $P_1(x, \infty) \leq P_2(x, \infty)$ for all x, and we denote this relationship symbolically as $P_1 \preceq P_2$. A collection of measures $(P_1, \dots, P_K)$ is said to satisfy the partial ordering given by $E \subset (1, \dots, K)^2$ if $P_i \preceq P_j \ \forall (i, j) \in E$.

Let $\mathcal{P}$ be the set of Borel probability measures on $\mathcal{X} \subset \mathbb{R}$. For a set $E \subset (1, \dots, K)^2$, let $C_E = \{(P_1, \dots, P_K) \subset \mathcal{P} : P_i \preceq P_j \ \forall (i, j) \in E\}$.

PROPOSITION 6. *The set C_E is a weakly closed convex set.*

We now try to identify the extreme points of C_E, using the ideas presented in Section 2.2. First consider $E = \{(1, 2)\}$ so C_E is the set of all pairs of measures $P = (P_1, P_2)$ such that $P_1 \preceq P_2$. What is the nature of the smallest possible support of a pair P if it is to lie in C_E? If P is a pair of point-mass measures $(\delta_{s_1}, \delta_{s_2})$, then the stochastic ordering holds if and only if $s_1 \leq s_2$. It turns out that such pairs of point-mass measures constitute the set $\operatorname{ex} C_E$. For more general sets of ordered measures, we have the following similar result, proven in Hoff (2000):

PROPOSITION 7. *Let C_E be the set of measures on $\mathbb{R}^K$ which satisfy a partial stochastic ordering given by E. A collection of measures P is extreme in C_E iff $P = (\delta_{s_1}, \dots, \delta_{s_K})$ for a vector $s \in \mathcal{S} = \{(s_1, \dots, s_K) \in \mathcal{X}^K : s_i \leq s_j \ \forall (i, j) \in E\}$.*

As in the previous examples of convex sets, we see that the set of extreme points of C_E can be indexed by a finite-dimensional parameter. In this case, the indexing set is the set $\mathcal{S}$ of support vectors of the extreme points. Therefore, we can write a mixing measure over the extreme points as a measure over $\mathcal{S}$. For a probability measure Q on Borel sets of $\mathcal{S}$, the barycenter of Q is defined as

$$T(Q) = \left(\int_{\mathcal{S}} \delta_{s_1}\, dQ(s), \dots, \int_{\mathcal{S}} \delta_{s_K}\, dQ(s) \right) = (Q_1, \dots, Q_K) \in C_E,$$

and so the barycenter function T maps measures on $\mathcal{S}$ to their marginal distributions, which lie in C_E. Note that the barycenter mapping T is continuous in the product topology of weak convergence.

Our previous theorems on integral representations for probability measures do not directly apply to the class C_E, as elements of C_E are not probability measures, but collections of probability measures. However, if we assume our sample space $\mathcal{X}$ is compact (thus making C_E a compact subset of the product space of signed measures) we can use Choquet's theorem to prove the existence of integral representations. Without the compactness assumption, we can prove a representation theorem directly: if Q is a measure on $\operatorname{ex} C_E$, it is easy to show that the barycenter of Q is a collection of measures satisfying the ordering E. On the other hand, let $(P_1, \dots, P_K) \in C_E$ and for each $i = 1, \dots, K$, construct the functions $s_i(\omega) = F_i^{-1}(\omega)$, where $F_i(x) = P_i(-\infty, x]$ and $\omega \in [0, 1]$. Letting Q be the canonical measure on the vector $(s_1, \dots, s_K)$ induced by a uniform distribution on ω, it is seen that Q represents $(P_1, \dots, P_K)$. Thus we have the following:

PROPOSITION 8. *A collection of probability measures $\{P_1, \dots, P_K\}$ satisfies the partial stochastic ordering constraint given by E if and only if there exists a K-variate measure Q with ith marginal equal to P_i for $i = 1, \dots, K$ and $Q(s_i \leq s_j) = 1$ for all $(i, j) \in E$.*

This result can be seen as a generalization of the well-known result for a pair of stochastically ordered measures [Lehmann (1997)] and is similar to a result given by Kamae, Krengel and O'Brien (1977) for a linear stochastic ordering.

As described in previous sections, it may be useful to invert the barycenter map in order to specify a particular starting value in an iterative likelihood-maximizing procedure, or in order to induce a desired prior on C_E. One such barycenter inversion is given above, in which $s_i(\omega) = F_i^{-1}(\omega)$ for $i = 1, \dots, K$. Unfortunately this construction will not be of much use in applications, as the support of such a Q lies on the one-dimensional curve $s(\omega)$. In the context of maximum likelihood estimation, some common iterative methods have the property that the final estimate has support only on the support of the initial estimate. In a Bayesian analysis, the priors we construct for Q will typically have the same support as the prior predictive distribution of samples from Q. Therefore, what is often needed is a Q having mass on *all* of $\mathcal{S}$ with marginals $Q_k = P_k, k = 1, \dots, K$. A simple algorithm for obtaining such a Q is given in Kullback (1968). He presents an iterative method for finding a measure Q with fixed marginals so that Q minimizes $\int \ln \frac{q(s)}{q_0(s)}\, dQ(s)$, where $q_0(s)$ is the density of some measure Q_0 with support on $\mathcal{S}$. The algorithm proceeds by repeated sequential replacement of the marginals of Q_0 with the desired marginals, and so the measure being iterated is absolutely continuous with respect to Q_0 at each step.

3. Maximum likelihood inference. Constrained likelihood-based inference proceeds by examining the likelihood function

$$L(P|\mathbf{X}) = \prod_{i=1}^{n} P(X_i) \tag{3.1}$$

for various values of $P \in C$, where C is the set of measures satisfying the constraint in question. The value $\hat{P}$ which maximizes (3.1) subject to $\hat{P} \in C$ is the constrained MLE of P, and can be found via convex optimization. For example, El Barmi and Dykstra (1994, 1998) provide algorithms based upon solving a dual problem.

As discussed in Hoff (2000), if C is convex the results in Section 2 allow us to rewrite (3.1) as

$$\log L(Q|\mathbf{X}) = \sum_{\mathcal{X}} n(x) \log \int P_s(x)\, dQ(s), \tag{3.2}$$

where Q is a measure over an index set $\mathcal{S}$ of the extreme points of C, $\mathcal{X}$ is the empirical support and $n(x) = \sum_{i=1}^{n} \delta_x(X_i)$, the empirical count function. Every $P \in C$, including the MLE $\hat{P}$, can be represented by such a Q, and so if $\hat{Q}$ maximizes (3.2), then $\hat{P} = \int P_s\, d\hat{Q}$ maximizes (3.1) subject to $\hat{P} \in C$. Finding an unconstrained MLE of Q thus provides a method of finding the constrained MLE of P. In some cases, such as estimation under stochastic ordering constraints, finding the unconstrained MLE of a representing mixing measure may be easier than using a constrained optimization method. Note that although $\hat{P}$ is unique (at least on the observed data), $\hat{Q}$ may not be. See Lindsay and Roeder (1993) for a discussion on the uniqueness of $\hat{Q}$.

A review of methods for finding $\hat{Q}$ and $\hat{P}$ can be found in Böhning (1995) and Lindsay (1995). Asymptotic properties of the estimators are discussed by Pfanzagl (1988), Leroux (1992) and van de Geer (1995), among others. However, these three authors focus on models in which the mixture components are dominated by a σ-finite measure, a condition which is not met, for example, by the extreme points of moment constrained sets, or collections of stochastically ordered measures.

As shown by Peters and Coberly (1976) and Lindsay (1983, 1995), a mixing measure Q maximizes (3.2) if and only if

$$D_Q(s) \equiv \sum_{\mathcal{X}} n(x)\left(\frac{P_s(x)}{P_Q(x)} - 1\right) \le 0 \qquad \forall s \in \mathcal{S}, \tag{3.3}$$

where P_Q denotes the barycenter of Q. The left-hand side of (3.3) is the directional derivative of the log-likelihood at Q in the direction of the point mass measure on s, and is called the gradient function for Q. The EM algorithm for finding the MLE of Q is based on the gradient function, in that it proceeds by iteration of $Q_{l+1}(s) = Q_l(s)(1 + D_{Q_l}(s)/n)$. The gradient function can also be used to measure the discrepancy between a given estimate and the MLE: if $\check{Q}$ is any estimate of Q and $\hat{Q}$ is the MLE, then

$$\begin{aligned} n\ln\left(1 + \frac{\Delta(\check{Q})}{n}\right) &- (n - n_{\min})\ln\left(1 + \frac{\Delta(\check{Q})}{n - n_{\min}}\right) \\ &\le \ln\frac{L(\hat{Q}|\mathbf{X})}{L(\check{Q}|\mathbf{X})} \le n\ln\left(1 + \frac{\Delta(\check{Q})}{n}\right) \end{aligned}$$

where $\Delta(\check{Q}) = \max_{s\in\mathcal{S}} D_{\check{Q}}(s)$ and $n_{\min} = \min_{x\in\mathcal{X}} n(x)$ [Lindsay (1995), Theorem 23]. This result gives bounds on the difference between the log-likelihood of any mixing measure $\check{Q}$ and that of the MLE $\hat{Q}$. This result is quite useful, as it allows us to monitor the convergence of any iterative scheme for maximizing the likelihood.

The best approach for finding $\hat{Q}$ will largely depend on the nature of the set $\mathcal{S}$. If the number of possible support points of $\hat{Q}$ is small, then the EM algorithm provides a very simple method of likelihood maximization. When the support of $\mathcal{S}$ is large, the EM scheme is inefficient, as it needs to keep track of mass on all possible extreme points, even though the MLE will have support on n or fewer extreme points [Lindsay (1995)]. In these cases, methods such as the vertex direction method [Wu (1978)] or the intra-simplex direction method [Lesperance and Kalbfleisch (1992)] may be more appropriate.

Univariate moment constraints. Nonparametric likelihood-based confidence intervals for a mean can be derived from the empirical likelihood ratio function [Owen (1988, 2001)]

$$\lambda(\theta|x) = \frac{\sup_{P:E_P(X)=\theta} \prod_1^n P(X_i)}{\sup_P \prod_1^n P(X_i)}.$$

Calculation of the above profile likelihood requires maximizing the likelihood subject to the constraint $E_P(X) = \theta$ for a range of θ values.

For a univariate mean, Owen suggests using Lagrange multipliers and a zero-finding algorithm to do the constrained maximization. This approach is quite easy, but as an illustrative example we consider using the mixture representation approach. A measure P maximizing the likelihood will have mass only on the empirical support $\mathcal{X}$, and so we can restrict our search for a P to the set $C_\theta = \{P : P(\mathcal{X}) = 1,\ E_P(X) = \theta\}$, which is convex, compact and nonempty as long as θ is in the convex hull of $\mathcal{X}$. Without loss of generality, we assume $\theta = 0$. The set of extreme points can be enumerated as

$$\operatorname{ex} C = \left\{P : P = P_{(X_i,X_j)} = \frac{X_j}{X_j - X_i}\delta_{X_i} + \frac{X_i}{X_i - X_j}\delta_{X_j},\ X_i \le 0 < X_j\right\}$$

and therefore can be parametrized by the finite set $\{s \in \mathcal{X}^2 : s = (X_i, X_j),\ X_i \le 0 < X_j\}$. As described above, the constrained MLE of P can be calculated by finding an MLE of Q using the EM algorithm or some other gradient based method. The gradient can be written

$$D_Q(s) = \frac{n(s_1)}{P_Q(s_1)}\frac{s_2}{s_2 - s_1} + \frac{n(s_2)}{P_Q(s_2)}\frac{s_1}{s_1 - s_2} - n.$$

In an EM-sequence, the update $Q_{l+1}(s)$ is given by $Q_l(s)(1 + D_{Q_l}(s)/n)$, so an EM-sequence is found by iterating

$$Q_{l+1}(s) = Q_l(s)\left(\frac{\check{P}(s_1)}{P_{Q_l}(s_1)}\frac{s_2}{s_2 - s_1} + \frac{\check{P}(s_2)}{P_{Q_l}(s_2)}\frac{s_1}{s_1 - s_2}\right)$$

where $\check{P}$ is the empirical distribution of $X_1, \dots, X_n$.

Stochastic ordering. Consider K groups of data on the real line, with observations in the kth group $X_{(k,1)}, \dots, X_{(k,n_k)}$ modeled as i.i.d. according to some unknown measure P_k. Given a partial ordering E on $(1, \dots, K)$, our task is to maximize the log-likelihood $\sum_{k=1}^K \sum_{x\in\mathcal{X}_k} n_k(x) \log P_k(x)$ subject to the constraint $P_i \preceq P_j$ for $(i, j) \in E$. Each component P_k of the collection $(P_1, \dots, P_K)$ maximizing the constrained likelihood can be assumed to have mass on the empirical support of its sample, as well as potentially on $X_{\min}$ and $X_{\max}$, the largest and smallest observations from all groups, in order to ensure the ordering can be achieved. Writing each augmented empirical support as $\mathcal{X}_k =$

$\{X_{\min}, X_{\max}\} \cup \{X_{(k,1)}, \dots, X_{(k,n_k)}\}$, we can restrict our search for an MLE to the set $C_E = \{(P_1, \dots, P_K): P_j(\mathcal{X}_j) = 1, P_i \preceq P_j \ \forall (i,j) \in E\}$. Dykstra and Feltz (1989) discuss a method for finding the MLE by solving a convex dual problem. Alternatively, we represent the collection of ordered measures as a mixture, and maximize the unconstrained mixture likelihood.

As described in Section 2, the set C_E is closed and convex, with extreme points given by collections of point-mass measures whose supports can be indexed by elements of the set $\mathcal{S} = \{s \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_K : s_i \le s_j \ \forall (i,j) \in E\}$. A mixing measure Q over the extreme points can be written as a measure over the set $\mathcal{S}$.

Although this is a multi-sample problem, the aforementioned results are applicable with respect to the appropriate gradient function. The directional derivative of the log-likelihood at Q in the direction of δ_s is given by

$$D_Q(s) = \sum_{k=1}^{K} \sum_{x \in \mathcal{X}_k} n_k(x) \left(\frac{P_{s,k}(x)}{P_{Q,k}(x)} - 1 \right) = \left(\sum_{k=1}^{K} \frac{n_k(s_k)}{P_{Q,k}(x)} \right) - n,$$

where $P_{Q,k}$ is the kth marginal of $P_Q = T(Q)$. As in the previous example, an EM-sequence is constructed by computing iterates of the function

$$Q_{l+1}(s) = Q_l(s)\big(1 + D_{Q_l}(s)/n\big) = Q_l(s) \sum_{k=1}^{K} \frac{n_k}{n} \frac{\check{P}_k(s_k)}{P_{Q_l,k}(s_k)},$$

where $\check{P}_k$ and n_k are the empirical distribution and the number of observations in the kth group, respectively. A slightly modified version of this EM algorithm is used in Hoff (2000) to estimate four partially ordered distributions with the additional complication of missing data.

4. Bayesian inference.

A nonparametric prior is one which puts mass in all open neighborhoods of the space of probability measures, relative to some sample space and topology. One of the simplest of such priors is the Dirichlet process prior [Ferguson (1973, 1974) and Blackwell and MacQueen (1973)] having weak support on an entire space of probability measures. Putting priors on large but proper subsets of a space is less straightforward. Doss (1985) discusses a technique of putting a prior on the set of all median-zero measures, Diaconis and Freedman (1986) construct a prior on the set of symmetric measures, and Brunner and Lo (1989) present a method of putting a prior on the space of symmetric, unimodal probability measures. We show these techniques are special cases of the mixture representation method described above, and extend the method to Bayesian estimation of the mean, given a nonparametric prior on the error distribution. We also discuss the mixture representation approach for collections of measures satisfying a partial stochastic ordering, and comment on methods of posterior approximation using MCMC.

4.1. *The mixture model.* A random variable X is modeled as having a sampling distribution P, which is assumed to lie in a convex set C. Under conditions given in Section 2, such a measure can be expressed as

$$P(A) = \int_{\mathcal{S}} P_s(A)\, dQ(s)$$

where Q is a mixing measure over a set $\mathcal{S}$ which indexes the extreme points of C. Prior uncertainty about an appropriate P can be quantified by a prior probability measure π on $\mathcal{Q}$, the space of measures on $\mathcal{S}$. This gives rise to the following model for observations $X_1, \dots, X_n$ distributed according to an unknown $P \in C$:

$$\begin{aligned} Q &\sim \pi, \\ S_1, \dots, S_n \mid Q &\sim \text{ i.i.d. } Q, \\ X_i \mid S_i &\sim P_{S_i} \qquad \text{independently over } i. \end{aligned}$$

Note that the latent data $\mathbf{S} = \{S_1, \dots, S_n\}$ are unobserved.

A convenient choice for π is the Dirichlet process prior, for which there exist closed form expressions of various quantities of interest; see Lo (1984) for some results on general kernel density estimation via Dirichlet mixtures. Many posterior quantities of interest can be derived via calculation of the marginal density of the latent variables $S_1, \dots, S_n$, unconditional on Q. This marginal density can be derived by using Blackwell and MacQueen's (1973) result concerning samples from a Dirichlet-distributed probability measure: an i.i.d. sample $S_1, \dots, S_n$ from $Q \sim \mathcal{D}(\alpha Q_0)$ can be generated by first sampling $S_1 \sim Q_0$. Then, with probability $\alpha/(\alpha+1)$, S_2 is sampled from Q_0, and with probability $1/(\alpha+1)$, S_2 is set equal to S_1. Proceeding sequentially, S_{i+1} is sampled from Q_0 with probability $\alpha/(\alpha+i)$ and is sampled from the empirical distribution of $S_1, \dots, S_i$ with probability $i/(\alpha+i)$.

Antoniak (1974) uses Blackwell and MacQueen's result to show that the density of a sample from a Dirichlet process with parameter αQ_0, Q_0 having a continuous density q_0, is given by

$$p(s_1, \dots, s_n) = \alpha^m \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \prod_1^m q_0(s_{(j)})\Gamma(n_j) \tag{4.1}$$

where m is the number of unique values of $s_1, \dots, s_n$ and n_j is the number of observations equal to the jth unique value $s_{(j)}$. From this it is easy to see that, conditional on m, the unique values of the latent variables are i.i.d. according to Q_0, as was shown by Korwar and Hollander (1973). The situation for a discrete Q_0 is similar. In this case, the density of the S_i's is given by

$$p(s_1, \dots, s_n) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \prod_{j=1}^m \frac{\Gamma(\alpha q_0(s_{(j)}) + n_j)}{\Gamma(\alpha q_0(s_{(j)}))}.$$

As shown by Petrone and Raftery (1997), the density (4.1) is absolutely continuous with respect to a sum of measures, having supports on various hyperplanes of $\mathcal{S}^n$. To see this, consider again the construction of the sample $S_1, \dots, S_n$: define a function g on $(1, \dots, n)$ by setting $g(1) = 1$ and set $g(2) = 2$ if S_2 is sampled from Q_0 and $g(2) = 1$ if $S_2 = S_1$. For $j \le n$, let $g(j) = 1 + \max_{i<j} g(i)$ if the value of S_j is sampled from Q_0 and set $g(j)$ equal to the unique value of $\{g(k) : S_j = S_{g(k)}\}$ otherwise. The function g is thus a mapping from the indices of the observations to the indices of the unique values, where the number of unique values is given by $m = \max g(i)$. We denote by $\mathcal{G}_m$ the set of all possible mappings $g : (1, \dots, n) \to (1, \dots, m)$ such that $g(1) = 1$, $\max\{g(1), \dots, g(n)\} = m$ and $g(j) \le 1 + \max_{i<j} g(i)$. Thus $\mathcal{G}_m$ corresponds to the set of partitions of $(1, \dots, n)$ into m groups.

For $B \in \mathcal{B}(\mathcal{S})^n$, let

$$\lambda(B) = \sum_{m=1}^{n} \sum_{g \in \mathcal{G}_m} \mu_{m,g}(B), \tag{4.2}$$

where $\mu_{m,g}$ is m-dimensional Lebesque measure on the hyperplane $\{(s_1, \dots, s_n) : g(i) = g(j) \Rightarrow s_i = s_j\}$. With this notation, the density $p(s_1, \dots, s_n)$ given in (4.1) is absolutely continuous with respect to λ, and so

$$\begin{aligned} &\Pr(S_1 \in B_1, \dots, S_n \in B_n) \\ &\quad = \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \sum_{m=1}^{n} \alpha^m \sum_{\mathcal{G}_m} \prod_{j=1}^{m} \Gamma(n_j) Q_0\biggl(\bigcap_{i:g(i)=j} B_i \biggr). \end{aligned} \tag{4.3}$$

Returning to the mixture model, the predictive distribution of the observed data $X_1, \dots, X_n$ is then

$$\begin{aligned} &\Pr(X_1 \in A_1, \dots, X_n \in A_n) \\ &\quad = \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \sum_{m=1}^{n} \alpha^m \sum_{\mathcal{G}_m} \prod_{j=1}^{m} \biggl\{ \Gamma(n_j) \int_{\mathcal{S}} \biggl(\prod_{i\,:\,g(i)=j} P_s(A_i) \biggr) q_0(s)\, ds \biggr\}. \end{aligned} \tag{4.4}$$

Equation (4.4) is used in what follows to calculate posterior distributions. For large n, however, (4.4) and expressions derived from it may be difficult to work with, as the size of the set $\mathcal{G}_m$ grows quite quickly with n. In such cases, posterior quantities may be calculated using MCMC methods, as discussed at the end of this section.

4.2. *Location and location-scale families.* In a location problem we have observations of the form

$$X_i = \theta + \varepsilon_i, \qquad i = 1, \dots, n,$$

where θ is the unknown location parameter and $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. according to some unknown error distribution P. A Bayesian procedure puts a prior on both

θ and P, the prior on P having support on some set of measures with a fixed location, that is, mean, median or mode equal to zero. A nonparametric prior for P is a prior with support on an entire space of fixed-location measures.

Median. Let C be the space of median-zero measures. Doss (1985) suggests if the error distribution $P \in C$ is known to be close to a measure $P_0 \in C$, then we can model P as follows:

- sample $P_- \sim \mathcal{D}[\alpha\{\frac{1}{2}P_0(\cdot \cap \{0\}) + P_0(\cdot \cap (-\infty, 0))\}]$;
- sample $P_+ \sim \mathcal{D}[\alpha\{\frac{1}{2}P_0(\cdot \cap \{0\}) + P_0(\cdot \cap (0, \infty))\}]$;
- let $P = \frac{1}{2}(P_- + P_+)$.

Note the resulting measure P is median zero.

In general, a method of sampling a median-zero measure is obtained by noting that C is weakly closed and that the set of extreme points of C is

$$\operatorname{ex} C = \{P_s \in C : P_s = (\delta_{s_1} + \delta_{s_2})/2,\ s_1 \le 0 \le s_2\}.$$

By the results of Section 2, any median-zero measure P can be written as $P = \int_{\mathcal{S}} P_s\, dQ(s)$ for some measure Q on $\mathcal{S} = \{s \in \mathbb{R}^2 : s_1 \le 0 \le s_2\}$. A measure $P \in C$ can be generated by:

- sampling $Q \sim \pi$, where π is a prior on $\mathcal{Q}$, the set of distributions on $\mathcal{S}$;
- setting $P = T(Q) = \int (\delta_{s_1} + \delta_{s_2})/2\, dQ(s)$.

The barycenter map T, mapping $\mathcal{Q}$ to C, is weakly continuous. Thus if the prior π has weak support on $\mathcal{Q}$, then the induced prior for P has weak support in C.

Doss' procedure is an example of this general method: Doss' prior for $P \in C$ is equivalent to sampling P_- and P_+ from the Dirichlet distributions above, and setting $Q(S \in A \times B) = P_-(S_1 \in A)P_+(S_2 \in B)$. Diaconis and Freedman's (1986) model is simpler still: their procedure is equivalent to sampling $P \sim \mathcal{D}(\alpha P_0)$ where P_0 has mass on $\mathbb{R}$, and letting $Q(S \in A \times B) = \frac{1}{2}(P(-A \cap B) + P(A \cap -B))$, that is, $S_1 = -S_2$ a.s. Q. This further restricts their error distribution P to be symmetric.

Mean. The set of extreme points of the mean-zero distributions C is given by

$$\operatorname{ex} C = \left\{P_s = \frac{s_2}{s_2 - s_1}\delta_{s_1} + \frac{s_1}{s_1 - s_2}\delta_{s_2},\ s_1 \le 0 < s_2\right\}.$$

As above, a prior on C can be constructed by putting a prior π on $\mathcal{Q}$, the space of measures on $\mathcal{S}$. A P is sampled by first sampling $Q \sim \pi$ and letting $P = T(Q) = \int P_s\, dQ(s)$. Recall that C is not a closed convex set unless the sample space is compact and so $T(Q)$ may not have a mean for some Q. However, such a $T(Q)$ will be the weak limit of a sequence of mean-zero measures.

Posterior quantities for the median and mean. The extreme points in the above examples are indexed by the same set $\mathcal{S}$, and as a result, calculations of posterior quantities are quite similar. Suppose we wish to estimate the location θ, assuming a Dirichlet process prior for Q and an absolutely continuous prior π for θ. Letting $\mathbf{x} = x_1, \dots, x_n$ be the observed values of $X_1, \dots, X_n$, the posterior of θ is proportional to $\pi(\theta)p(\mathbf{x}|\theta)$, where the latter term can be derived from (4.4). This density in general is quite complicated due to the possibility of ties among the latent data, but can be greatly simplified if there are no ties in the observed data. We compute this simplified density in a manner similar to that in Doss (1985): for $\eta > 0$ we consider the conditional probability

$$\begin{aligned}
&\Pr\bigl(\theta \in A | X_i \in N_\eta(x_i), i = 1, \dots, n\bigr)\\
&\qquad = \frac{\int_A \pi(\theta) \Pr(\varepsilon_i \in N_\eta(x_i - \theta), i = 1, \dots, n)\, d\theta}{\int_{\mathbb{R}} \pi(\theta) \Pr(\varepsilon_i \in N_\eta(x_i - \theta), i = 1, \dots, n)\, d\theta},
\end{aligned}$$

where $N_\eta(x) = (x - \eta/2, x + \eta/2)$. Since the limit

$$p(\mathbf{x}|\theta) = \lim_{\eta \to 0} \frac{1}{\eta^n} \Pr\bigl(\varepsilon_i \in N_\eta(x_i - \theta), i = 1, \dots, n\bigr)$$

exists for θ a.e. and can be passed through the integral above, then by a theorem of Pfanzagl (1979),

$$\Pr(\theta \in A | X_1 = x_1, \dots, X_n = x_n) = \frac{\int_A \pi(\theta) p(\mathbf{x}|\theta)\, d\theta}{\int_{\mathbb{R}} \pi(\theta) p(\mathbf{x}|\theta)\, d\theta}$$

is a conditional probability distribution of θ given $X_1 = x_1, \dots, X_n = x_n$, absolutely continuous with respect to Lebesque measure with density given by $\pi(\theta|\mathbf{x}) \propto \pi(\theta)p(\mathbf{x}|\theta)$.

PROPOSITION 9. *Using the above notation, if Q_0 is absolutely continuous with respect to Lebesque measure, with bounded density q_0 and marginal densities q_{01}, q_{02}, then*

$$\begin{aligned}
p(\mathbf{x}|\theta) \propto \sum_{m = n_- \vee n_+}^{n} \alpha^m \sum_{A \in \mathcal{A}(n-m)} &\Biggl(\prod_{ij \in A} q_0(\varepsilon_i, \varepsilon_j) P_{(\varepsilon_i, \varepsilon_j)}(\varepsilon_i) P_{(\varepsilon_i, \varepsilon_j)}(\varepsilon_j)\Biggr)\\
&\times \Biggl(\prod_{i \notin A, \varepsilon_i \le 0} q_{01}(\varepsilon_i) E_{Q_0}\bigl(P_s(\varepsilon_i) | s_1 = \varepsilon_i\bigr)\Biggr)\\
&\times \Biggl(\prod_{i \notin A, \varepsilon_i > 0} q_{02}(\varepsilon_i) E_{Q_0}\bigl(P_s(\varepsilon_i) | s_2 = \varepsilon_i\bigr)\Biggr),
\end{aligned}$$

where $\varepsilon_i = x_i - \theta$ are the residuals, n_-, n_+ are the numbers of negative and positive residuals respectively, and $\mathcal{A}(k)$ is the set of all possible pairings of k indices of negative residuals with k indices of positive residuals.

The sum over the set $\mathcal{G}_m$ in (4.4) reduces to a sum over the set $\mathcal{A}$ as follows: We assume there are no ties among the observed data, but this does not preclude ties in the latent data. However, a tie between two latent variables S_i and S_j is only possible if ε_i and ε_j are of opposite sign, otherwise we would have a tie in the observed data. By this reasoning, there must be at least $n_- \vee n_+$ unique values of the latent variables. For a given number of unique values m, the set $\mathcal{A}(n-m)$ represents all possible pairings of the latent variables, assuming $n-m$ pairs. The above formula simplifies greatly if Q_0 has a certain structure:

COROLLARY 3. *Suppose $\{P_s : s \in \mathcal{S}\}$ are the extreme points of the set of mean-zero distributions. If the base measure Q_0 has a density $q_0(s_1, s_2) = \gamma^{-1}(s_2 - s_1)p_0(s_1)p_0(s_2)$, where p_0 is the density of a mean-zero measure P_0 and $2\gamma = \int_{\mathbb{R}} |\varepsilon|\, dP_0(\varepsilon)$, then $P_0 = T(Q_0)$ and*

$$\begin{aligned} p(\theta|\mathbf{x}) &\propto \pi(\theta)\left(\prod_1^n p_0(x_i - \theta)\right) \\ &\quad \times \sum_{n_- \vee n_+}^{n} (\alpha\gamma)^m \sum_{\mathcal{A}(n-m)} \prod_{ij \in A} \frac{(\theta - x_i)(x_j - \theta)}{x_j - x_i}. \end{aligned} \tag{4.5}$$

COROLLARY 4. *Suppose $\{P_s : s \in \mathcal{S}\}$ are the extreme points of the set of median-zero distributions. If the base measure Q_0 has a density $q_0(s_1, s_2) = p_0(s_1)p_0(s_2)$, where p_0 is the density of a median-zero measure P_0, then $P_0 = T(Q_0)$ and*

$$p(\theta|\mathbf{x}) \propto \pi(\theta)\left(\prod_1^n p_0(x_i - \theta)\right) \sum_{n_- \vee n_+}^{n} \alpha^m \frac{n_-!n_+!}{(m - n_-)!(m - n_+)!(n - m)!}. \tag{4.6}$$

Note in both cases, the first two terms form the joint density of θ and $X_1, \ldots, X_n$ under a parametric model using P_0 as the error distribution. The remaining term reflects the effect of the Dirichlet prior. For each model, as α gets large the $m = n$ term in the sum dominates the posterior, and so the nonparametric posterior concentrates around the parametric posterior based on P_0. This is more or less what we would expect: for large α, Q should be close to Q_0, and so P should be close to P_0. On the other hand, when α is much less than 1, then the $m = n_- \vee n_+$ term tends to dominate the posterior. In fact, for fixed n and small enough α the posterior will be largest at θ-values for which $n_- \vee n_+$ is a minimum, that is, θ is a sample median. This is due to the large probability of ties in the latent data for small α.

The posterior (4.6) is very similar to Doss' (1985) posterior for the median: both Doss' posterior and the above posterior can be written as $\pi(\theta)\{\prod_{i=1}^n p_0 \times (x_i - \theta)\} f(\mathbf{x}, \theta)$, where in Doss' case $f(\mathbf{x}, \theta) = \{\Gamma(\alpha/2 + n_-)\Gamma(\alpha/2 + n_+)\}^{-1}$.

The f corresponding to (4.6) concentrates a bit more about the empirical median than Doss' f, especially for small α, although this seems to be partially an artifact of the probability of ties among the latent data.

Mode. Mixture representations for unimodal measures have been studied by Dharmadhikari and Joag-Dev (1988) and Bertin, Cuculescu and Theodorescu (1997). The set C of unimodal, mode-zero distributions is weakly closed and convex, and the set of extreme points of C is given by $\operatorname{ex} C = \{P_s : P_s = (s_2 - s_1)^{-1}\delta_{(s_1,s_2)}, s_1 \le 0 \le s_2\}$, the set of uniform densities which include the origin. As above, a mode-zero measure can be sampled by first sampling $Q \sim \pi$ and setting $P = T(Q) = \int P_s \, dQ(s)$. Brunner and Lo (1989) restrict themselves to the class of symmetric unimodal densities by generating $P_+ \sim \mathcal{D}(\alpha P_0)$, where P_0 is a measure on $\mathbb{R}^+$, and then letting $Q(S \in A \times B) = P_+(-A \cap B)$. Modeling unimodal measures which are not necessarily symmetric can be achieved by letting π have support on mixing measures having mass on all extreme distributions of C. For example, if $Q \sim \mathcal{D}(\alpha Q_0)$ where Q_0 is a measure with support on all of $\{s \in R^2 : s_1 \le 0 \le s_2\}$, then since the mapping $T : \mathcal{Q} \to C$ is weakly continuous, the induced prior on P will have weak support in C.

Location-scale families via quantile constraints. As in the location problem, suppose we are modeling a set of random variables where our prior information is in terms of the center θ and scale σ of the set of observations. Our model is:

- $\theta \sim \pi(\theta)$, $\sigma \sim \pi(\sigma)$, $P \sim \pi(P)$,
- $\varepsilon_1, \dots, \varepsilon_n \mid P \sim$ i.i.d. P,
- $X_i = \theta + \sigma\varepsilon_i$,

where P is constrained to be an error distribution with a fixed center and scale. One possibility is to constrain P to lie in the convex set C of measures having a median of zero and an interquartile range of two. As discussed in Section 2, the extreme points of this set are measures with equal mass on four support points, one in each quartile:

$$\operatorname{ex} C = \{P_s = (\delta_{s_1} + \delta_{s_2} + \delta_{s_3} + \delta_{s_4})/4 : s_1 \le -1 \le s_2 \le 0 \le s_3 \le 1 \le s_4\}.$$

Our full model is:

- $\theta \sim \pi(\theta)$, $\sigma \sim \pi(\sigma)$, $Q \sim \pi(Q)$,
- $S_1, \dots, S_n \mid Q \sim$ i.i.d. Q,
- $\varepsilon_i \mid S_i \sim \frac{1}{4}\sum_{j=1}^{4} \delta_{S_{i,j}}$,
- $X_i = \theta + \sigma\varepsilon_i$.

Newton, Czado and Chappell (1996) propose a similar model: from a given prior, they select three points at random to represent the quantiles of the error distribution. These three points induce a partition of the real line into four regions.

The error distribution P is then modeled as the average of four independent Dirichlet processes, each one with a base measure proportional to a given measure P_0 restricted to one of the four partitions. In our formulation this is equivalent to the following prior on $\mathcal{Q}$:

- $Q_i \sim \mathcal{D}(\alpha P_0(\cdot|A_i)),\ i = 1, 2, 3, 4,$
- $Q = Q_1 \times Q_2 \times Q_3 \times Q_4,$

where A_i is one of the four sets of the partition. Newton, Czado and Chappell then use the resulting measure P as an inverse-link function for binary regression.

The posterior distribution for θ and σ can be obtained in a manner similar to that in estimating the median alone, although the possibility of ties among four groups of latent variables makes the posterior more complex.

4.3. *Stochastic orderings.* For K groups of data, we assume observations in the kth group are i.i.d. according to some unknown probability measure P_k. We wish to impose a stochastic ordering constraint given by the set $E \subset (1, \dots, K)^2$ so that $P_i \preceq P_j\ \forall (i, j) \in E$. As discussed in Section 2 the extreme points of this set C_E are collections of point-mass measures on vectors satisfying the ordering E:

$$\operatorname{ex} C_E = \{(\delta_{s_1}, \dots, \delta_{s_K}) : s_i \le s_j\ \forall (i, j) \in E\}.$$

The extreme points can be indexed by the set $\mathcal{S} = \{s \in \mathbb{R}^K : s_i \le s_j\ \forall (i, j) \in E\}$, and any collection in C_E can be expressed as a mixture over these extreme points. As in previous examples, a prior on C_E can be induced by a prior π on the space $\mathcal{Q}$ of mixing measures over the set $\mathcal{S}$.

Given a prior π on $\mathcal{Q}$, a set of observations $X_1, \dots, X_n$ can be sampled as follows:

- $Q \sim \pi(Q)$,
- $S_1, \dots, S_n \mid Q \sim$ i.i.d. Q,
- $X_i \sim P_{S_{i,y_i}} \Rightarrow X_i = S_{i,y_i}$,

where y_i is the group membership of the ith observation and $P_{s,j}$ denotes the jth component measure of the collection P_s (thus $P_{S_{i,y_i}}$ is the point-mass measure on S_{i,y_i}).

Observing $X_i = x_i$ is equivalent to observing $S_i \in B_i$, where $B_i = \mathcal{S} \cap \{s_{y_i} = x_i\}$. From (4.3), the posterior probability that a new value of $S \sim Q$ lies in a set B is seen to be

$$\frac{\alpha}{\alpha + n} Q_0(B) + \frac{1}{\alpha + n} \sum_{i=1}^{n} \Pr(S_i \in B | S_j \in B_j, j = 1, \dots, n).$$

The sum in the above is equivalent to the expected number of latent observations in the set B, conditional on the observed data. Although this simplifies somewhat in some cases (e.g., if there are no within-group ties in the observed data), the expectation is still quite complicated due to the possibility of ties in the latent data (as in the examples involving the mean and median) and the additional complication that the possibility of ties among the latent data is dependent on the assumed stochastic ordering.

4.4. *Posterior approximation methods.* Due to the complicated nature of many of the posteriors presented above, practical use of the aforementioned mixture models may require MCMC methods for posterior approximation. When the number of extreme points is finite and the dimension of $\mathcal{S}$ is small, posterior approximation is made relatively easy by use of the Gibbs sampler for the unknown quantities $\mathbf{S}$ and Q. Given current values $(\mathbf{S}^b, Q^b)$, one scan of the Markov chain consists of:

- sampling $S_i^{b+1} \sim p(s|Q^b, X_i)$, independently for $i = 1, \dots, n$,
- sampling $Q^{b+1} \sim \mathcal{D}(\alpha Q_0 + n\hat{Q}_{\mathbf{S}^{b+1}})$,

where $\hat{Q}_{\mathbf{S}^{b+1}}$ is the empirical distribution of $(S_1, \dots, S_n)^{b+1}$. This sampling scheme is straightforward to implement, as the conditional distribution of S_i given Q and the observed data is simply $\Pr(S_i = s|Q, X_i = x_i) = Q(s)P_s(x_i)/\{\sum_{s'} Q(s')P_{s'}(x_i)\}$. Sampling Q given $S_1, \dots, S_n$ is also straightforward. It is well known [Ferguson (1973)] that a sample Q from a Dirichlet distribution $\mathcal{D}(\alpha Q_0)$ can be obtained by independently sampling $z_i \sim \text{Gamma}(\alpha Q_0(s_i), 1)$, $i = 1, \dots, K$, and setting $Q(s_i) = z_i / \sum z_j$, where $s_1, \dots, s_K$ are the points of the finite space indexing the extreme points.

Covariate information or hyperparameters can be incorporated into the estimation procedure by adding appropriate Metropolis–Hastings steps to the above scheme. Such a sampling scheme was implemented in Hoff et al. (2001) in the context of estimating stochastically ordered measures in the presence of missing data and other model parameters.

MCMC methods are more difficult when the number of extreme points is not finite or the dimension of $\mathcal{S}$ is large. In such cases, incorporating a sequence of Q's into the Markov chain is impractical, and posterior inference must be made via the posterior distribution of $S_1, \dots, S_n$. West, Müller and Escobar (1994) and MacEachern and Müller (1998) discuss Gibbs sampling of $S_1, \dots, S_n$, although such Gibbs methods are typically either inefficient or unworkable for the models presented here: the main difficulty is that in many cases the extreme points P_s do not have common support, resulting in poor mixing. Alternatively, a feasible approach is to construct a Markov chain using Metropolis–Hastings updates which alternately resamples the tie-structure information g and the unique values of the latent variables $S_1, \dots, S_n$. This is tractable, as g has a fairly simple structure and the unique values of the S_i's are i.i.d. Q_0, given g. Such an approach is discussed in Neal (2000), and efficient application of this approach to stochastic ordering problems is a topic of current research of the author.

5. Discussion. Mixture representations allow one to turn a possibly difficult constrained problem into an unconstrained mixture problem, at a cost of introducing a high-dimensional and difficult to interpret mixing distribution. In terms of maximum likelihood estimation, one can expect that mixture model estimation methods (such as EM) will be less efficient than other methods for direct optimization of the constrained likelihood, at least in terms of raw computation time. On the other hand, estimation of a representing mixture may be much easier to implement.

Additionally, one should keep in mind that a representing Q is typically not unique, and one should be cautious about inference on Q beyond the measure P it represents. This point is underscored by the fact that the maximum likelihood estimate $\hat{P}$ is unique, whereas $\hat{Q}$ is generally not. Similarly, in a Bayesian analysis one should take care that one's prior distribution for Q induces a desired prior distribution on P. What may seem like a "noninformative" prior for Q may not be noninformative for P. For example, in the two sample stochastic ordering problem on a finite sample space, a prior for Q which is uniform on the simplex induces a prior on P with highly separated component measures.

APPENDIX: PROOFS

PROOF OF PROPOSITION 1. Since $\mathcal{X}$ is a separable metric space, so is $(\mathcal{P}, w)$ [Parthasarathy (1967), Theorem 2.6.2] and thus also $(A, w \cap A)$ for $A \subset \mathcal{P}$. The set $\mathcal{Q}$ of probability measures on (A, σ_A) is a separable metric space, and by Parthasarathy (1967), Theorem 2.6.3, the set of measures with finite support is weakly dense in $\mathcal{Q}$. Therefore, for each $Q \in \mathcal{Q}$ there is a sequence $Q_n \to Q$ such that Q_n has support on no more than n points. Now the weak topology w on $\mathcal{P}$ is the weakest topology such that $\int f\,dP$ is continuous for each $f \in C_b(\mathcal{X})$. Therefore $\int f\,dP$ is bounded and continuous in $P \in A$ in the inherited topology $w \cap A$. Since $Q_n \to Q$ and f is bounded, we have

$$\int f\,dP_{Q_n} = \int \biggl(\int f\,dP\biggr)\,dQ_n \quad \Longrightarrow \quad \int \biggl(\int f\,dP\biggr)\,dQ = \int f\,dP_Q,$$

and so $P_{Q_n} \to P_Q$ in the weak topology on $\mathcal{P}$. Now $P_{Q_n} = \sum_1^n P_i Q_n(P_i) \in \mathcal{H}A$, and so P_Q is the weak limit of a sequence of elements of $\mathcal{H}A$. □

PROOF OF PROPOSITION 2. Let $s \in \mathcal{S}$ and suppose $P_s = \alpha_0 P_1 + (1-\alpha_0)P_2$ where P_1 and P_2 are in C and $\alpha_0 \in (0,1)$. This implies $\alpha P_1 + (1-\alpha)P_2$ is in C and has support on s for all $\alpha \in (0,1)$. By the uniqueness of P_s, we must have $P_1 = P_2$, and so P_s is extreme. Now suppose $s \in \mathcal{S}$ and let s_0 be a proper subset of s. Then there can be no probability measure $P_{s_0} \in C$ with support on s_0, since if there was then $\alpha P_s + (1-\alpha)P_{s_0}$ would have support on s and be in C for all $\alpha \in (0,1]$. □

PROOF OF PROPOSITION 3. Suppose P has the given form, and so the corresponding CDF has the form $F(x) \in \{\alpha_i, \alpha_{i+1}\}\ \forall x \in [\theta_i, \theta_{i+1})$. Now suppose $F(x) = \alpha F_1(x) + (1-\alpha)F_2(x)$ for some $\alpha \in (0,1)$ and $F_1, F_2 \in C$. The quantile constraint implies

$$\alpha_i \le F_1(x), F(x), F_2(x) \le \alpha_{i+1} \qquad \forall x \in [\theta_i, \theta_{i+1}).$$

Letting x be any point in $[\theta_i, \theta_{i+1})$, if $F(x) = \alpha_i$ then so must $F_1(x)$ and $F_2(x)$. The same holds for the possibility $F(x) = \alpha_{i+1}$, and thus $F_1 = F_2 = F$, proving extremity of the measure P. On the other hand, suppose P is in C and let F be the corresponding CDF. Suppose there is an $x_0 \in [\theta_i, \theta_{i+1})$ such that $\alpha_i < F(x_0) < \alpha_{i+1}$. We can construct two CDFs $F^+, F^- \in C$ such that $F^+ \neq F^-$ as follows:

$$F^+(x) = F(x) + \min\{F(x) - \alpha_i, \alpha_{i+1} - F(x)\}/2,$$
$$F^-(x) = F(x) - \min\{F(x) - \alpha_i, \alpha_{i+1} - F(x)\}/2$$

for $x \in [\theta_i, \theta_{i+1})$, and $F^+(x) = F^-(x) = F(x)$ otherwise. Since $F = (F^- + F^+)/2$, such an F cannot be extreme. Thus for F to be extreme we must have $F(x) \in \{\alpha_i, \alpha_{i+1}\}\ \forall x \in [\theta_i, \theta_{i+1})$. This condition, together with the quantile constraint, implies the parametrization. □

PROOF OF PROPOSITION 4. Let $P \in C$ and Q_0 represent P. Since $P_{(\cdot)}$ is measurable, by a theorem of Yershov (1973), any measure Q_0 on $\operatorname{ex} C$ can be extended to a measure Q on $\mathcal{B}(\mathcal{S})$, and so $C \subset T(\mathcal{Q})$. On the other hand, if $P = T(Q)$ then we can define Q_0 so that $Q_0(A) = Q(s : P_s \in A) = Q(P^{-1}A)$. Therefore $T(Q)(B) = T(Q_0)(B)$ for all $B \in \mathcal{B}$, and since $T(Q_0) \in \overline{C}$ (Proposition 1), so is $T(Q)$. □

PROOF OF PROPOSITION 5. Let $f \in C_b(\mathcal{X})$, $Q_n \overset{w}{\to} Q$, $P_n = T(Q_n)$ and $P = T(Q)$. Then

$$E(f|P_n) = \int_{\mathcal{X}} f(x)\,dP_n(x) = \int_{\mathcal{S}} \left(\int_{\mathcal{X}} f(x)\,dP_s(x) \right) dQ_n(s).$$

If the term in the parentheses is a bounded continuous function of s for all bounded continuous f, then by the weak convergence assumption we have $E(f|P_n) \to E(f|P)$, that is, $P_n \overset{w}{\to} P$. Therefore, $Q_n \overset{w}{\to} Q \Rightarrow T(Q_n) \overset{w}{\to} T(Q)$. Since the weak topology is metrizable, this convergence property implies continuity of T. □

PROOF OF PROPOSITION 6. Let $P \in \overline{C}_E$, and let $\{P^{(n)}\}_{n=1}^{\infty} \subset C_E$ be a sequence weakly converging to P (i.e., converging in the product topology of weak convergence). Pick any $(i, j) \in E$ and let F_i, F_j be the CDFs of the corresponding component measures of P. Let $\mathcal{A} = \{x \in \mathbb{R} : F_i^{(n)}(x) \to F_i(x),\ F_j^{(n)}(x) \to F_j(x)\}$. By the definition of weak convergence, $\mathcal{A}$ is the intersection of $\mathcal{A}_i$, the set of continuity points of F_i, and $\mathcal{A}_j$, the set of continuity points of F_j. Note that $\mathcal{A}_i^c$ and $\mathcal{A}_j^c$ are both countable, and so $\mathcal{A}^c$ is countable and $\mathcal{A}$ is dense in $\mathbb{R}$. For $x \in \mathcal{A}$, we clearly have $F_j(x) \leq F_i(x)$ by the definition of $\mathcal{A}$ and the stochastic ordering constraint. Since F_i, F_j are right continuous, for $x \notin \mathcal{A}$ and any $\varepsilon > 0$, a $y > x$ can be chosen from $\mathcal{A}$ such that $F_i(y) - F_i(x) < \varepsilon$ and $F_j(y) - F_j(x) < \varepsilon$. Thus

$$F_i(x) - F_j(x) = [F_i(x) - F_i(y)] + [F_j(y) - F_j(x)] + [F_i(y) - F_j(y)] \geq -\varepsilon.$$

Since ε is arbitrary, we have $F_i(x) \geq F_j(x)$ and the stochastic ordering is preserved. Therefore $P \in C_E$, and so C_E is a weakly closed, convex set. □

PROOF OF PROPOSITION 8. If Q has the specified support then $\{s : s_j \leq x\} \subset \{s : s_i \leq x\}$ and so $F_j(x) \leq F_i(x)$. Conversely, construct the K functions $s_i(\omega) = F_i^{-1}(\omega)$ for $i = 1, \dots, K$ and $\omega \in [0, 1]$, where $F^{-1}(\omega) = \inf\{x : F(x) \geq \omega\}$. Then if $P_i \preceq P_j$ we have $\{x : F_j(x) \geq \omega\} \subset \{x : F_i(x) \geq \omega\}$ and so $s_i(\omega) \leq s_j(\omega)\ \forall\,\omega$.

Let ω be uniformly distributed and let $\hat{F}_i$ be the canonical distribution of s_i. We need to show $\hat{F}_i(s) = F_i(s)$, or equivalently, that the two sets $\{\omega : \inf\{x : F_i(x) \geq \omega\} \leq c\}$ and $\{\omega : \omega \leq F_i(c)\}$ are equal. Suppose $\omega \leq F_i(c)$, in which case $c \in \{x : F_i(x) \geq \omega\}$ and so $\inf\{x : F_i(x) \geq \omega\} \leq c$ trivially. On the other hand, if $\inf\{x : F_i(x) \geq \omega\} \leq c$, then $w \leq F(c)$ by the monotonicity of F and the equivalence of the two sets is shown. □

Acknowledgments. The author thanks Michael Newton and Tom Kurtz of the University of Wisconsin–Madison for their helpful guidance and instruction over the years. The author also thanks the two referees for their helpful comments.

REFERENCES

ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174.

ARJAS, E. and GASBARRA, D. (1996). Bayesian inference of survival probabilities, under stochastic ordering constraints. *J. Amer. Statist. Assoc.* **91** 1101–1109.

ASH, R. B. (1972). *Real Analysis and Probability*. Academic Press, New York.

BERTIN, E. M. J., CUCULESCU, I. and THEODORESCU, R. (1997). *Unimodality of Probability Measures*. Kluwer, Dordrecht.

BLACKWELL, D. and MACQUEEN, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1** 353–355.

BÖHNING, D. (1995). A review of reliable maximum likelihood algorithms for semiparametric mixture models. *J. Statist. Plann. Inference* **47** 5–28.

BRUNK, H. D., FRANCK, W. E., HANSON, D. L. and HOGG, R. V. (1966). Maximum likelihood estimation of the distributions of two stochastically ordered random variables. *J. Amer. Statist. Assoc.* **61** 1067–1080.

BRUNNER, L. J. and LO, A. Y. (1989). Bayes methods for a symmetric unimodal density and its mode. *Ann. Statist.* **17** 1550–1566.

DARDANONI, V. and FORCINA, A. (1998). A unified approach to likelihood inference on stochastic orderings in a nonparametric context. *J. Amer. Statist. Assoc.* **93** 1112–1123.

DHARMADHIKARI, S. and JOAG-DEV, K. (1988). *Unimodality, Convexity, and Applications*. Academic Press, Boston.

DIACONIS, P. and FREEDMAN, D. (1980). Finite exchangeable sequences. *Ann. Probab.* **8** 745–764.

DIACONIS, P. and FREEDMAN, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14** 1–26.

DOSS, H. (1985). Bayesian nonparametric estimation of the median. I. Computation of the estimates. *Ann. Statist.* **13** 1432–1444.

DYKSTRA, R. L. and FELTZ, C. J. (1989). Nonparametric maximum likelihood estimation of survival functions with a general stochastic ordering and its dual. *Biometrika* **76** 331–341.

DYNKIN, E. B. (1978). Sufficient statistics and extreme points. *Ann. Probab.* **6** 705–730.

EL BARMI, H. and DYKSTRA, R. L. (1994). Restricted multinomial maximum likelihood estimation based upon Fenchel duality. *Statist. Probab. Lett.* **21** 121–130.

EL BARMI, H. and DYKSTRA, R. (1998). Maximum likelihood estimates via duality for log-convex models when cell probabilities are subject to convex constraints. *Ann. Statist.* **26** 1878–1893.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230.

FERGUSON, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2** 615–629.

HOFF, P. D. (2000). Constrained nonparametric maximum likelihood via mixtures. *J. Comput. Graph. Statist.* **9** 633–641.

HOFF, P. D., HALBERG, R. B., SHEDLOVSKY, A., DOVE, W. F. and NEWTON, M. A. (2001). Identifying carriers of a genetic modifier using nonparametric Bayes methods. *Case Studies in Bayesian Statistics* **5**. *Lecture Notes on Statist.* **162** 327–342.

KAMAE, T., KRENGEL, U. and O'BRIEN, G. L. (1977). Stochastic inequalities on partially ordered spaces. *Ann. Probab.* **5** 899–912.

KARR, A. F. (1991). *Point Processes and Their Statistical Inference*, 2nd ed. Dekker, New York.

KORWAR, R. M. and HOLLANDER, M. (1973). Contributions to the theory of Dirichlet processes. *Ann. Probab.* **1** 705–711.

KULLBACK, S. (1968). Probability densities with given marginals. *Ann. Math. Statist.* **39** 1236–1243.

LEHMANN, E. L. (1997). *Testing Statistical Hypotheses*, 2nd ed. Springer, New York.

LEROUX, B. G. (1992). Consistent estimation of a mixing distribution. *Ann. Statist.* **20** 1350–1360.

LESPERANCE, M. L. and KALBFLEISCH, J. D. (1992). An algorithm for computing the nonparametric MLE of a mixing distribution. *J. Amer. Statist. Assoc.* **87** 120–126.

LINDSAY, B. G. (1983). The geometry of mixture likelihoods: A general theory. *Ann. Statist.* **11** 86–94.

LINDSAY, B. G. (1995). *Mixture Models*: *Theory, Geometry and Applications*. IMS, Hayward, CA.

LINDSAY, B. G. and ROEDER, K. (1993). Uniqueness of estimation and identifiability in mixture models. *Canad. J. Statist.* **21** 139–147.

LO, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12** 351–357.

MACEACHERN, S. N. and MÜLLER, P. (1998). Estimating mixture of Dirichlet process models. *J. Comput. Graph. Statist.* **7** 223–238.

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265.

NEWTON, M. A., CZADO, C. and CHAPPELL, R. (1996). Bayesian inference for semiparametric binary regression. *J. Amer. Statist. Assoc.* **91** 142–153.

OWEN, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249.

OWEN, A. B. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18** 90–120.

OWEN, A. B. (2001). *Empirical Likelihood.* Chapman and Hall, London.

PARTHASARATHY, K. R. (1967). *Probability Measures on Metric Spaces.* Academic Press, New York.

PETERS, C. and COBERLY, W. A. (1976). The numerical evaluation of the maximum-likelihood estimate of mixture proportions. *Comm. Statist. Theory Methods* **5** 1127–1135.

PETRONE, S. and RAFTERY, A. E. (1997). A note on the Dirichlet process prior in Bayesian nonparametric inference with partial exchangeability. *Statist. Probab. Lett.* **36** 69–83.

PFANZAGL, J. (1979). Conditional distributions as derivatives. *Ann. Probab.* **7** 1046–1050.

PFANZAGL, J. (1988). Consistency of maximum likelihood estimators for certain nonparametric families, in particular: Mixtures. *J. Statist. Plann. Inference* **19** 137–158.

VAN DE GEER, S. (1995). Asymptotic normality in mixture models. *ESAIM Probab. Statist.* **1** 17–33.

VON WEIZSÄCKER, H. and WINKLER, G. (1979). Integral representation in the set of solutions of a generalized moment problem. *Math. Ann.* **246** 23–32.

VON WEIZSÄCKER, H. and WINKLER, G. (1980). Noncompact extremal integral representations: Some probabilistic aspects. In *Functional Analysis*: *Surveys and Recent Results II* (K.-D. Bierstedt and B. Fuchssteiner, eds.) 115–148. North-Holland, Amsterdam.

WEST, M., MÜLLER, P. and ESCOBAR, M. D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In *Aspects of Uncertainty* (P. R. Freeman and A. F. M. Smith, eds.) 363–386. Wiley, Chichester.

WU, C. F. (1978). Some iterative procedures for generating nonsingular optimal designs. *Comm. Statist. Theory Methods* **7** 1399–1412.

YERSHOV, M. P. (1973). Extensions of measures. Stochastic equations. *Proc. Second Japan–USSR Symposium on Probability Theory. Lecture Notes in Math.* **330** 516–526. Springer, Berlin.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
BOX 354322
SEATTLE, WASHINGTON 98195-4322
E-MAIL: hoff@stat.washington.edu