Nonparametric hierarchical Bayesian quantiles

Luke Bornn
Department of Statistics and Actuarial Science, Simon Fraser University
bornn@fas.harvard.edu

Neil Shephard
Department of Economics and Department of Statistics, Harvard University
shephard@fas.harvard.edu

Reza Solgi
Department of Statistics, Harvard University
rezasolgi@fas.harvard.edu

June 27, 2017

Abstract

Here we develop a method for performing nonparametric Bayesian inference on quantiles. Relying on geometric measure theory and employing a Hausdorff base measure, we are able to specify meaningful priors for the quantile while treating the distribution of the data otherwise nonparametrically. We further extend the method to a hierarchical model for quantiles of subpopulations, linking subgroups together solely through their quantiles. Our approach is computationally straightforward, allowing for censored and noisy data. We demonstrate the proposed methodology on simulated data and an applied problem from sports statistics, where it is observed to stabilize and improve inference and prediction.

Keywords: Censoring; Hausdorff measure, Hierarchical models; Nonparametrics; Quantile.

1 Introduction

Consider learning about β, the $\tau \in (0,1)$ quantile of the random variable Z. This will be based on data $\mathcal{D} = \{z_1, ..., z_n\}$, where we assume z_i, $i = 1, 2, ..., n$, are scalars and initially that they are independent and identically distributed. We will perform nonparametric Bayesian inference on β given $\mathcal{D}$, a problem emphasized by, for example, Parzen (1979, 2004), Koenker and Bassett (1978) and Koenker (2005). By solving this problem we will also deliver a nonparametric Bayesian hierarchical quantile model, which allows us to analyze data with subpopulations only linked through quantiles. The proposed methods extend to censored and partially observed data.

1.1 Background

In early work on Bayesian inference on quantiles, Section 4.4 of Jeffreys (1961) used a "substitute likelihood" $s(\beta) = \binom{n}{n_\beta} \tau^{n_\beta} (1-\tau)^{n-n_\beta}$, where $n_\beta = \sum_{i=1}^{N} 1(z_i \leq \beta)$. See also Boos and Monahan

(1986), Lavine (1995) and Dunson and Taylor (2005). This relates to other approximations to the likelihood suggested by Lazar (2003), Lancaster and Jun (2010) and Yang and He (2012), who use empirical likelihoods, and Chernozhukov and Hong (2003) who connect with M-estimators. Chamberlain and Imbens (2003) use a Bayesian bootstrap (Rubin (1981)) to carry out Bayesian inference on a quantile but have no control over the prior for β.

Yu and Moyeed (2001) carried out Bayesian analysis of quantiles using a likelihood based on an asymmetric Laplace distribution for the regression residuals $e_i = y_i - x_i'\beta$ (see also Koenker and Machado (1999) and Tsionas (2003)), $L(\mathcal{D}|\beta) = \exp\{-\sum_{i=1}^n \rho_\tau(e_i)\}$ where $\rho_\tau(\cdot)$ is the "check function" (Koenker and Bassett (1978)),

$$\rho_\tau(e) = |e| \left\{(1-\tau)1_{e<0} + \tau 1_{e\geq 0}\right\}, \quad e \in R. \tag{1}$$

Here $\rho_\tau(e)$ is continuous everywhere, convex and differentiable at all points except when $e = 0$. This Bayesian posterior is relatively easy to compute using mixture representations of Laplace distributions. Papers which extend this tradition include Kozumi and Kobayashi (2011), Li et al. (2010), Tsionas (2003), Kottas and Krnjajic (2009) and Yang et al. (2016). Unfortunately the Laplace distribution is a misspecified distribution and so typically yields inference which is overly optimistic. Yang et al. (2016) and Feng et al. (2015) discuss how to overcome some of these challenges; see the related works by Chernozhukov and Hong (2003) and Muller (2013).

Closer to our paper is Hjort and Petrone (2007) who assume the distribution function of Z is a Dirichlet process with parameter aF_0, focusing on when $a \downarrow 0$. Hjort and Walker (2009) write nonparametric Bayesian priors on the quantile function. Our focus is on using informative priors for β, but our use of a non-informative prior for the distribution of Z aligns with that of Hjort and Petrone (2007).

Our paper is related to Bornn et al. (2016), who develop a Bayesian nonparametric approach to moment based estimation. Though their methods do not cover our case, the intellectual root is similar: the quantile model only specifies a part of the distribution, so we complete the model by using Bayesian nonparametrics.

Hierarchical models date back to Stein (1966) and Lindley and Smith (1972). Discussions of the literature include Morris and Lysy (2012) and Efron (2010). Our focus is on developing models where the quantiles of individual subpopulations are thought of as drawn from a common population-wide mixing distribution, but where all other features of the subpopulations are nonparametric and uncommon across the populations. The mixing distribution is also nonparametrically specified. There is some linkages with deconvolution problems (e.g. Butucea and Comte

(2009) and Cavalier and Hengartner (2009)), but our work is discrete and not linear. It is more related to, for example, Robbins (1956), Carlin and Louis (2008), McAuliffe et al. (2006) and Efron (2013) on empirical Bayes methods.

Here we report a simple to use method for handling this problem, which scales effectively with the sample size and the number of subpopulations, and allows for censored data. Our hierarchical method is illustrated on an example drawn from sports statistics.

1.2 Outline of the paper

In Section 2 we discuss our modelling framework and how we define Bayesian inference on quantiles, with a focus on uniqueness and priors. A flexible way of building tractable models is developed which gives an analytic expression for the posterior on a quantile. A Monte Carlo analysis is carried out to study the bias, precision and coverage of our proposed method, which also compares the results to that seen for sample quantiles. In Section 3 we extend the analysis by introducing a nonparametric hierarchical quantile model and show how to handle it using very simple simulation methods. A detailed study is made of individual sporting careers using the hierarchical model, borrowing strengths across players when careers are short and data is limited. In Section 4 we extend the analysis to data which is censored and extend our sporting career example in this context. Section 5 concludes, while an online Appendix contains various proofs of results stated in the paper.

2 A Bayesian nonparametric quantile

2.1 Definition of the problem

We use the conventional modern definition of the τ quantile β, that is

$$\beta = \underset{b}{\operatorname{argmin}} \ \mathrm{E}\left\{\rho_\tau(Z-b)\right\}.$$

To start, suppose Z has known finite support $\mathcal{S} = \{s_1, ..., s_J\}$, and write

$$\Pr(Z = s_j|\theta) = \theta_j, \quad \text{for } 1 \leq j \leq J,$$

with $\theta = (\theta_1, \theta_2, ..., \theta_{J-1})' \in \Theta_\theta$, and $\Theta_\theta \subseteq \Delta$, where Δ is the simplex, $\Delta = \{\theta;\ \iota'\theta < 1 \text{ and } \theta_j > 0\}$, and define $\theta_J = 1 - \iota'\theta$, in which ι is a vector of ones. The function

$$\Psi(b, \theta) = \mathrm{E}_\theta\left\{\rho_\tau(Z-b)\right\} = \sum_{j=1}^{J} \theta_j \rho_\tau(s_j - b),$$

is continuous everywhere, convex and differentiable at all points except when $b \in \mathcal{S}$.

We define the "Bayesian nonparametric quantile" problem as learning from data the unknowns

$$(\beta, \theta')' \in \Theta_{\beta,\theta}, \quad \text{where} \quad \Theta_{\beta,\theta} \subseteq \mathbb{R} \times \Delta \subset \mathbb{R}^J.$$

Each point within $\Theta_{\beta,\theta}$ is a pair (β, θ) which satisfies both the probability axioms and

$$\beta = \underset{b}{\operatorname{argmin}} \sum_{j=1}^{J} \theta_j \rho_\tau (s_j - b).$$

Every θ, up to a set of 0 Lebesgue measure, uniquely determines β. This will be formalized in Proposition 1.

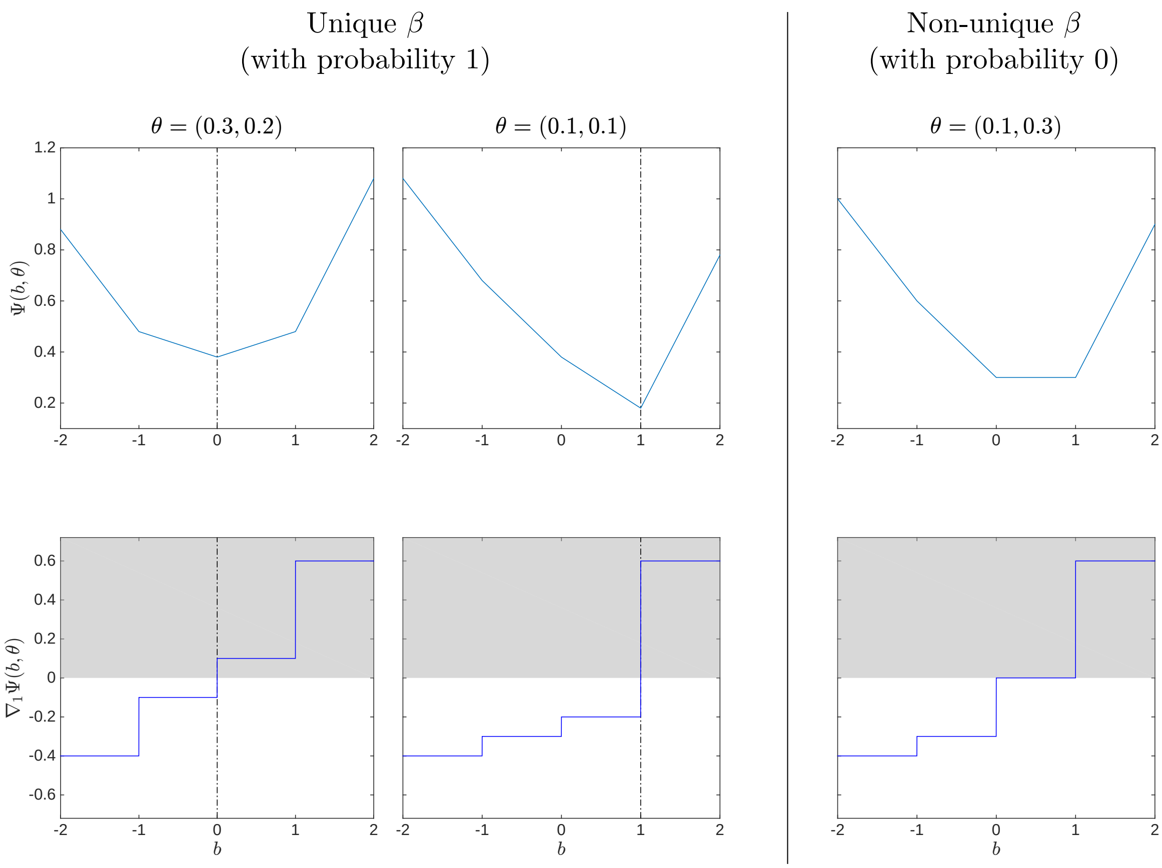

Figure 1: This plot shows the 0.4-quantile with support $S = \{-1, 0, 1\}$. Plotted is $\Psi(b, \theta)$ and its directional derivatives with respect to b, $\nabla_1 \Psi(b, \theta)$. Left hand has $\theta = (0.3, 0.2)'$, the center is $\theta = (0.1, 0.1)'$, and the right hand is $\theta = (0.1, 0.3)'$. In the left and center, the quantiles are 0 and 1, respectively, while, in the right the optimization does not have a unique solution.

Example 1 *Figure 1*[1] *sets* $\tau = 0.4$, *and* $\mathcal{S} = \{-1, 0, 1\}$. *In the left panel, for* $\theta = (0.3, 0.2)'$, *we*

[1] Figure 1 demonstrates that Bayesian nonparametric quantile estimation is not a special case of Bayesian nonparametric ψ type M-estimators, and so not a special case of moment estimation. This means we are outside the framework developed by Bornn et al. (2016).

plot $\Psi(b,\theta)$ and its directional derivatives[2] *with respect to b, $\nabla_1\Psi(b,\theta)$. The resulting quantile is $\beta=0$. In the center panels, $\theta=(0.1,0.1)'$, implying $\beta=1$. β is not unique iff $\theta_1=0.4$ or $\theta_1+\theta_2=0.4$, a 0 probability event under a continuous distribution for θ. An example of the latter case is $\theta=(0.1,0.3)'$, which is shown in the right panel . Here $\Psi(b,\theta)$ is minimized on $[0,1]$*[3].

Proposition 1 *Without loss of generality, assume $s_1<\cdots<s_J$. Then β is unique iff $\tau\notin\{\theta_1,\theta_1+\theta_2,....,\theta_1+\cdots+\theta_{J-1}\}$. If θ has a continuous distribution with respect to the Lebesgue measure, then with probability 1, for each θ there is a unique quantile $\beta\in\mathcal{S}$ and with probability 1*

$$\frac{\partial\beta}{\partial\theta'}=0. \tag{2}$$

Proposition 1 means we can partition the simplex in $J+1$ sets, $\Delta=\left(\bigcup_{k=1}^{J}\mathcal{A}_k\right)\cup\mathcal{N}$, where $\mathcal{N}$ is a zero Lebesgue measure set and the sets $\mathcal{A}_k=\{\theta\in\Delta;\ s_k=\underset{b}{\operatorname{argmin}}\ \Psi(b,\theta)\}$, $1\le k\le J$, contain all the values of θ which deliver a quantile $\beta=s_k=\underset{b}{\operatorname{argmin}}\ \Psi(b,\theta)$. We write this compactly as $\beta=t(\theta)$, $\beta\in\mathcal{S}$, $\theta\in\Delta$, and the corresponding set index $k=k(\theta)$, $1\le k\le J$, $\theta\in\Delta$, so $\beta=s_{k(\theta)}$.

Example 1 (Continued) *Figure 2 is a ternary plot showing all possible values of $\theta=(\theta_1,\theta_2)'$ and $\theta_3=1-\theta_1-\theta_2$ and the implied value of β overlaid for $\tau=0.4$. The values of θ which contain distinct values of β are collected into the sets $\mathcal{A}_1$ (where $\beta=s_1$), $\mathcal{A}_2$ (where $\beta=s_2$), $\mathcal{A}_3$ (where $\beta=s_3$). The interior lines marking the boundaries between these sets are the zero measure events collected into $\mathcal{N}$. The union of the disjoint sets $\mathcal{A}_1,\mathcal{A}_2,\mathcal{A}_3$, and $\mathcal{N}$, make up the simplex Δ.*

2.2 The prior and posterior

The set of admissible pairs (β,θ) is denoted by $\Theta_{\beta,\theta}\subseteq\mathcal{S}\times\Delta$. Now $\Theta_{\beta,\theta}$ is a lower dimensional space as $\beta=t(\theta)$. Using the Hausdorff measure, we are able to assign measures to the lower dimensional subsets of $\mathcal{S}\times\Delta$, and therefore we can define probability density functions with respect to Hausdorff measure on manifolds within $\mathcal{S}\times\Delta$.

One approach to building a joint prior $p(\beta,\theta)$ is to place a prior on $\beta\in\mathcal{S}$, which we write as $p(\beta)$, and then build a conditional prior density, $p(\theta|\beta=s_k)$, $\theta\in\mathcal{A}_k$, recalling $\mathcal{A}_k\subseteq\Delta$. Then the joint density with respect to Hausdorff measure on $\Theta_{\beta,\theta}$ is

$$p(\beta,\theta)=p(\beta)p(\theta|\beta).$$

[2] Recall, for the generic function $f(b)$, the corresponding directional derivative is $\nabla_v f(b)=\lim_{h\downarrow 0}\frac{f(b+hv)-f(b)}{h}$.

[3] If $\mathcal{D}=\mathcal{S}$ then the empirical quantile is $\widehat{\beta}=\underset{b}{\operatorname{argmin}}\ \sum_{j=1}^{J}\rho_\tau(s_j-b)$, which is non-unique if τJ is an integer (e.g. if $\tau=0.5$, then if J is even).

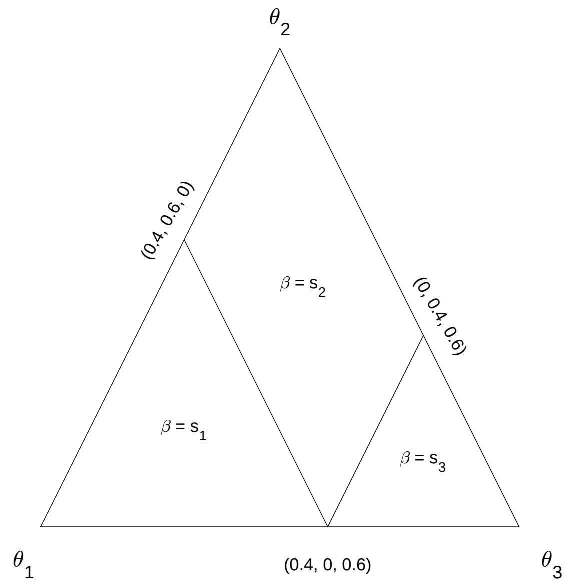

Figure 2: Ternary plots of θ_1, θ_2 and $\theta_3 = 1 - \theta_1 - \theta_2$ and the implied quantiles β at level $\tau = 0.4$. Here $\beta \in \{s_1, s_2, s_3\}$. $\mathcal{A}_k$ is the set of probabilities θ_1, θ_2 where $\beta = s_k$.

For the quantile problem, with probability one $\beta = t(\theta)$, so the "area formula" of Federer (1969) (see also Diaconis et al. (2013) and Bornn et al. (2016)) implies the marginal density for the probabilities is induced as

$$p(\theta) = p(\beta, \theta), \quad \beta = t(\theta),$$

as Proposition 1 shows that $\partial\beta/\partial\theta' = 0$. Here the right hand side is the density of the prior with respect to Hausdorff measure defined on $\Theta_{\beta,\theta}$, while the left hand side is the implied density of the prior distribution of θ with respect to Lebesgue measure defined on the simplex Δ.

The model's likelihood is,

$$\prod_{j=1}^{J} \theta_j^{n_j},$$

where $n_j = \sum_{i=1}^{n} 1(z_i = s_j)$. Then the posterior distribution of β, θ will be,

$$p(\theta|\mathcal{D}) = p(\beta = s_{k(\theta)}, \theta|\mathcal{D}) \propto p(\beta = s_{k(\theta)})p(\theta|\beta = s_{k(\theta)}) \prod_{j=1}^{J} \theta_j^{n_j}, \quad \beta = t(\theta). \tag{3}$$

This means that

$$p(\beta = s_k|\mathcal{D}) = \int_{\mathcal{A}_k} p(\theta|\mathcal{D}) \mathrm{d}\theta \propto p(\beta = s_k) \int_{\mathcal{A}_k} \left\{ p(\theta|\beta = s_k) \prod_{j=1}^{J} \theta_j^{n_j} \right\} \mathrm{d}\theta.$$

2.3 A class of $p(\theta|\beta)$ models

Assume $f_\Delta(\theta)$ is the density function of a continuous distribution on Δ, and define,

$$c_k = \Pr_{f_\Delta}(\beta = s_k) = \int_{\mathcal{A}_k} f_\Delta(\theta)\mathrm{d}\theta,$$

Then one way to build an explicit prior for $p(\theta|\beta)$ is to decide to set

$$p(\theta|\beta = s_k) = \frac{f_\Delta(\theta)}{c_k} 1_{\mathcal{A}_k}(\theta), \quad \theta \in \mathcal{A}_k.$$

Proposition 2 shows how to compute $\{c_k\}$.

Proposition 2 *Here* $c_1 = 1 - \Pr(\theta_1 < \tau)$, $c_J = \Pr\left(\sum_{j=1} \theta_j < \tau\right)$, *and* $c_k = \Pr\left(\sum_{j=1}^{k-1} \theta_j < \tau\right) - \Pr\left(\sum_{j=1}^{k} \theta_j < \tau\right)$, *for* $k = 2, ..., J-1$.

This conditional distribution can be combined with a fully flexible prior $\Pr(\beta = s_k) = b_k$, where $b_k > 0$, for $1 \leq k \leq J$, and $\sum_{k=1}^{J} b_k = 1$. Returning to the general case, this implies the joint

$$p(\theta) = p(\beta = s_{k(\theta)}, \theta) = \frac{b_{k(\theta)}}{c_{k(\theta)}} f_\Delta(\theta), \tag{4}$$

which means $\Pr(\beta = s_k) = \int_{\mathcal{A}_k} p(\beta, \theta; \alpha)\mathrm{d}\theta = b_k$, in which $p(\beta, \theta; \alpha)$ is the density of the parameters indexed by the parameters α. Note that $p(\theta)$ is discontinuous at the set boundaries (that is the zero Lebesgue measure set $\mathcal{N}$), and $p(\theta) \neq f_\Delta(\theta)$ unless $b_k = c_k$ for all k.

From (3) the posterior distribution of β, θ will be,

$$p(\beta = s_{k(\theta)}, \theta|\mathcal{D}) \propto \frac{b_{k(\theta)}}{c_{k(\theta)}} f_\Delta(\theta) \prod_{j=1}^{J} \theta_j^{n_j}, \quad \text{and} \quad p(\beta = s_k|\mathcal{D}) \propto \frac{b_k}{c_k} \int_{\mathcal{A}_k} \left\{ f_\Delta(\theta) \prod_{j=1}^{J} \theta_j^{n_j} \right\} \mathrm{d}\theta.$$

The Dirichlet case is particularly convenient.

2.4 Dirichlet special case

Let f_Δ be the Dirichlet density, $f_D(\theta; \alpha) = B(\alpha)^{-1} \prod_{j=1}^{J} \theta_j^{\alpha_j - 1}$, where $\alpha = (\alpha_1, ..., \alpha_J)$ is the vector of positive parameters, and $B(\alpha)$ is the beta function. Then c_k can be computed via Proposition 2 using the distribution function[4] of

$$\theta_k^+ \sim Be\left(\alpha_k^+, \alpha_J^+ - \alpha_k^+\right), \quad \text{where generically} \quad \alpha_k^+ = \sum_{j=1}^{k} \alpha_j.$$

[4] $\Pr\left(\theta_k^+ < \tau\right) = I_\tau(\alpha_k^+, \alpha_J^+ - \alpha_k^+) = B_k$, in which $I_\tau(\alpha, \beta) = B(\tau, \alpha, \beta)/B(\alpha, \beta)$ is the regularized incomplete beta function, $B(\tau, \alpha, \beta) = \int_0^\tau x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x$ is the incomplete beta function. When α_k^+ and $\alpha_J^+ - \alpha_k^+$ are large some care has to be taken in computing c_k. We have written $c_k = B_{k-1} - B_k = B_k\left\{\frac{B_{k-1}}{B_k} - 1\right\} = B_k\left\{\exp\left(\log B_{k-1} - \log B_k\right) - 1\right\}$ so $\log c_k = \log B_k + \log\left\{\exp\left(\log B_{k-1} - \log B_k\right) - 1\right\}$. Now $B(x, a, b) = {}_2F_1(a + b, 1, a+1, x)\frac{1}{a}x^a(1-x)^b$ where ${}_2F_1$ is the Gauss hypergeometric function. Hence we can compute $\log c_k$ accurately.

To mark their dependence on α, in the Dirichlet case we write $c_k = c_k(\alpha)$. We will refer to

$$p(\theta|\beta = s_k) = \frac{f_D(\theta;\alpha)}{c_k} 1_{\mathcal{A}_k}(\theta), \quad \theta \in \mathcal{A}_k, \tag{5}$$

as the density of $D_J(\alpha, k)$, the Dirichlet distribution on Δ^{J-1} with parameter α truncated to $\mathcal{A}_k$. This can be used in the following simple prior to posterior calculation.

Proposition 3 *When $f_\Delta(\theta) = f_D(\theta;\alpha)$, then*

$$\Pr(\beta = s_k|\mathcal{D}) = \frac{1}{C(\alpha, \mathbf{n})} \frac{c_k(\alpha + \mathbf{n})}{c_k(\alpha)} b_k,$$

where $\mathbf{n} = (n_1, ..., n_J)$. Here $C(\alpha, \mathbf{n})$ is the normalizing constant, which is computed via enumeration, $C(\alpha, \mathbf{n}) = \sum_{k=1}^J \frac{c_k(\alpha+\mathbf{n})}{c_k(\alpha)} b_k$. Further,

$$p(\theta|\mathcal{D}) = \frac{1}{C(\alpha, \mathbf{n})} \frac{\Pr(\beta = s_{k(\theta)})}{c_{k(\theta)}(\alpha)} f_D(\theta; \alpha + \mathbf{n}).$$

The Bayesian posterior mean or quantiles of the posterior can be computed by enumeration.

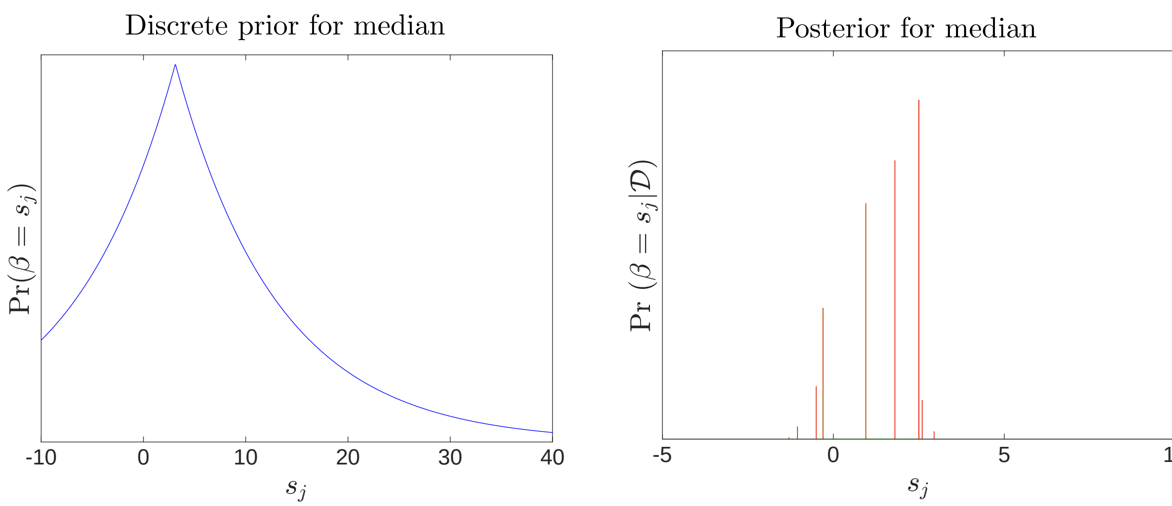

Figure 3: The left hand side shows the prior distribution of β for the discrete method, and the right shows the corresponding posterior for the first replication. Notice the posterior has many small atoms marked in short green lines. These points originate from the prior and represent around 1 data points.

2.5 Example: quantile of daily return of equities

Here we apply the methodology to functions of the log returns of a financial asset, $r_t = \log(P_t/P_{t-1})$, where P_t is the price at time $t = 1, 2, ..., n$. This is a challenging, especially when the available sample is too short, for instance for a newly introduced or illiquid assets, or for a hedge fund performance reported on a monthly basis. Here we illustrate this by looking at daily returns from 20 stocks equities. Our basic database covers 14 years, from January 2003 to December 2016. In our analysis we treat the sample quantile of the whole sample for each stock as the "*true*" value of that specific stock. Then for shorter subperiods, ranging from 6 months to 2 years, we compare two estimation methodologies: the sample quantiles as the benchmark versus the proposed Bayesian approach.

Before analysis we filter the returns as $r_t/\sqrt{h_t}$, where h_t is computed by an exponentially weighted moving average (EWMA) model on square returns, with the decay factor fixed to $\lambda = 0.94$ (following RiskMetrics$^{\text{TM}}$ daily methodology). We will write the subsample of these filtered returns as z_j, $j = 1, 2, ..., J$, and ignore their time series dependence. We then apply our methodology to these filtered returns aiming to estimate the corresponding filtered $\tau = 0.01$-quantile.

In the Bayesian model we use the posterior mean as the point estimator. The prior of θ is a symmetric Dirichlet distribution with $\alpha = 50/J$, and $b_k \propto e^{-\frac{1}{2}(s_k+3.14)^2}$ (that is a normal density function centered at $\tau = 0.01$-quantile of the Student's t distribution with 6 degrees of freedom and unit variance). Note that the symmetric 95% credible region of the prior's distribution is $(-5.10, -1.18)$, that covers both the Normal distribution quantile (-2.33) and the Student's t distribution with 3 degrees of freedom (-4.54) as two potentially extreme cases. The analysis has been performed for each stock and for non-overlapping time intervals of different lengths. For each stock, the estimate's error has been computed by comparing it to the sample quantile obtained from the whole sample for that stock.

In Table 1 we have reported the ratio of the mean absolute error (MAE) and root mean square error (RMSE) of the sample quantiles to the proposed Bayesian estimators. In almost all the cases the Bayesian estimators have significantly reduced the estimation errors. Although the prior distribution of the quantile over the relevant part of the real axis is relatively vague, it has been able to penalized the unreasonably small or large estimates, resulting in less variability and more efficiency.

2.6 Monte Carlo experiment

Here $\mathcal{D}$ is simulated from the long right hand tailed $z_i \overset{iid}{\sim} -\log \chi_1^2$, so the τ-quantile is $\beta_\tau = -\log\left\{F_{\chi_1^2}^{-1}(1-\tau)\right\}$. The empirical quantile $\widehat{\beta}_\tau$ will be used to benchmark the Bayesian procedures. The distribution of $\widehat{\beta}_\tau$ will be computed using its limiting distribution $\sqrt{n}\left(\widehat{\beta}_\tau - \beta_\tau\right) \xrightarrow{d} N(0, \tau(1-\tau)/f_z(\beta_\tau)^2)$, and by bootstrapping. In the limiting distribution case, the density of data, $f_z(\beta_\tau)$, has been estimated by a normal kernel and Silverman's optimal bandwidth.

	2 Years		1 Year		6 Months	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
MSFT	1.20	1.11	1.95	1.77	1.29	1.20
PFE	1.01	0.98	1.20	1.20	1.22	1.34
BAC	1.18	1.12	1.40	1.47	1.63	1.58
AAPL	1.12	1.25	1.29	1.38	1.28	1.28
MRK	1.31	1.37	1.55	1.61	1.53	1.47
MMM	1.12	1.18	1.11	1.13	1.46	1.65
AXP	1.23	1.28	1.55	1.53	1.22	1.22
BA	0.97	1.02	1.16	1.23	1.41	1.41
CAT	1.16	1.23	1.45	1.41	1.55	1.48
CVX	0.95	1.01	1.08	1.08	1.20	1.36
KO	1.21	1.58	1.44	1.59	1.54	1.64
DD	0.94	1.05	1.28	1.26	1.46	1.71
XOM	0.96	0.85	1.31	1.40	1.39	1.54
GE	1.07	1.06	1.60	1.94	1.07	1.24
HD	0.88	0.92	1.32	1.35	1.24	1.27
IBM	1.25	1.26	1.23	1.35	1.53	1.66
INTC	0.68	0.68	1.47	1.92	1.42	1.48
JNJ	1.75	1.70	1.53	1.45	1.38	1.33
JPM	1.27	1.31	1.37	1.57	1.43	1.45
MCD	1.30	1.24	1.58	1.44	1.51	1.33
mean	1.13	1.16	1.39	1.45	1.39	1.43

Table 1: Ratio of MAE and RMSE of the sample 0.01-quantiles to the Bayesian estimates for 20 stocks. The analysis have been performed on the normalized daily returns on intervals of different lengths (2 years, 1 years, and 6 months) of the whole sample (from 2003 to 2016). The whole sample quantile is considered to be the "*true*" value.

	$\tau = 0.5$				$\tau = 0.9$			
	Sample quantile		Posterior		Sample quantile		Posterior	
	CLT	Boot	Discrete	Data	CLT	Boot	Discrete	Data
	$n = 10$							
Bias	-0.157	0.152	0.345	0.206	1.245	-0.174	1.083	-0.167
$n^{1/2}$ SE	2.221	2.119	2.155	2.136	7.909	5.087	4.764	5.206
RMSE	0.720	0.687	0.764	0.706	2.794	1.618	1.856	1.655
Coverage	0.913	0.943	0.936	0.897	0.805	0.645	0.932	0.638
	$n = 40$							
Bias	-0.039	0.038	0.085	0.054	-0.224	0.050	0.438	0.098
$n^{1/2}$ SE	2.309	2.184	2.214	2.203	5.639	5.403	5.731	5.560
RMSE	0.367	0.347	0.360	0.353	0.919	0.856	1.007	0.885
Coverage	0.945	0.945	0.937	0.940	0.810	0.912	0.944	0.910
	$n = 160$							
Bias	0.020	0.010	0.021	0.014	0.072	0.014	0.102	0.025
$n^{1/2}$ SE	2.358	2.258	2.266	2.262	6.130	5.650	5.767	5.700
RMSE	0.187	0.179	0.180	0.179	0.490	0.447	0.467	0.451
Coverage	0.955	0.948	0.946	0.953	0.922	0.944	0.952	0.947
	$n = 320$							
Bias	-0.003	0.002	0.005	0.003	-0.015	0.002	0.023	0.004
$n^{1/2}$ SE	2.343	2.296	2.297	2.297	6.029	5.856	5.885	5.867
RMSE	0.093	0.091	0.091	0.091	0.239	0.231	0.234	0.232
Coverage	0.952	0.950	0.948	0.947	0.930	0.947	0.940	0.951

Table 2: Monte Carlo experiment using 25,000 replications and a highly biased prior. Coverage probability is based on a nominal 95% confidence or credible interval. Bayesian estimators are the posterior mean. RMSE denotes root mean square error. Boot denotes bootstrap, CLT implemented using a kernel for the asymptotic standard error.

We build two Bayesian estimators:

1. **Discrete.** Let $s_j = -10 + 50(j-1)/(J-1)$, where $J = 1,000$, and assume a prior

$$\Pr(\beta_\tau = s_k) \propto \exp\{-\lambda |s_k - \delta_\tau|\}, \tag{6}$$

where $\lambda = 0.1$, and $\delta_\tau = \beta_\tau + \gamma_\tau$, where $\gamma_\tau > 0$, and we let γ_τ increases when τ deviates from 0.5. This prior is not centered at the true value of the quantile and is more contaminated for the tail quantiles. In particular in our simulations we use $\gamma_{0.5} = 2.33$ and $\gamma_{0.9} = 6.03$. The data is binned using the support, and $\alpha = \frac{1}{J}$. Equation (6), for $\tau = 0.5$, is shown in Figure 3 together with the associated posterior for one replication of simulated data.

2. **Data.** The support $\mathcal{S}$ is the data (therefore $J = n$), $\alpha = \frac{1}{J}$, and the prior height (6) sits on those J points of support (so the prior changes in each replication).

Table 2 reports the results from 25,000 replications, comparing the four modes of inference. The asymptotic distribution of the empirical quantile provides a poor guide when working within a thin tail even when the n is quite large. In the center of the distribution it is satisfactory by the time n hits 40. The bootstrap performs poorly in the tail when n is tiny, but is solid when n is large.

Not surprisingly the bootstrap of the empirical quantile $\widehat{\beta}$ and the Bayesian method using support from the data are very similar. Assuming no ties, straightforward computations leads to,

$$\Pr(\widehat{\beta} = s_j) = F_{\mathrm{B}}\left(\lceil \tau J \rceil - 1; J, (j-1)/J\right) - F_{\mathrm{B}}\left(\lceil \tau J \rceil - 1; J, j/J\right)$$

where $F_{\mathrm{B}}(\cdot; n, p)$ is the binomial cumulative distribution function with size parameter n, and probability of success p. Interestingly, for large J, this is a close approximation to $c_j(\mathbf{1})$. This connection will become more explicit in the next subsection.

The discrete Bayesian procedure is by far the most reliable, performing quite well for all n. It has a large bias for small n, caused by the poor prior, but the coverage is encouraging. Overall, there is some evidence that for small samples the Bayesian estimators perform well. The two Bayesian procedures have roughly the same properties.

2.7 Comparison with Jeffrey's substitution likelihood

Some interesting connections can be established by thinking of α as being small.

Proposition 4 *Conditioning on the data and finite J and n, if $\alpha_k \downarrow 0$, and $\frac{\alpha_k}{\alpha_l} \to 1$, then $c_k(\alpha) \to \frac{1}{J}$, and,*

$$c_k(\alpha + \mathbf{n}) \quad \to \quad \sum_{j=n_{k-1}^+}^{n_k^+ - 1} f_B(j; n-1, \tau), \quad k = 1, 2, ..., J,$$

where, for $k = 0, 1, ..., n$, $f_B(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$, is the binomial probability function with the size parameter n, and the probability of success p, and $n_0^+ = 0$.

The reason why $n-1$, not n, appears in the limit of $c_k(\alpha + \mathbf{n})$ is that $\mathcal{S}$ only has n elements so j runs from 0 to $n-1$. The proposition means that if there are no ties and $\mathcal{D} = \mathcal{S}$, then

$$\begin{array}{lcl} c_k(\alpha + \mathbf{n}) & \to & f_B(k-1; J-1, \tau), \quad k = 1, 2, ..., J, \\ \Pr(\beta = s_k | \mathcal{D}) & \to & C(\mathbf{n})^{-1} f_B(k-1; J-1, \tau) b_k. \end{array}$$

Here $C(\mathbf{n})$ is the normalizing constant, computed via enumeration, $C(\mathbf{n}) = \sum_{k=1}^J f(k-1; J-1, \tau) b_k$.

The result in Proposition 4 is close to, but different from, Jeffrey's substitution likelihood $s(\beta) = f_B(k; J, \tau)$, for $s_k \leq \beta < s_{k+1}$ where $s_0 = -\infty$ and $s_{J+1} = \infty$ (Jeffrey has $n+1$ categories to choose from, not n, as he allows data outside $\mathcal{S}$). $s(\beta)$ is a piecewise constant, non-integrable function (which means it needs proper priors to make sense) in $\beta \in R$, while for us $\beta \in \mathcal{S}$ (and the posterior is always proper).

2.8 Comparison with Bayesian bootstrap

The prior and posterior distribution of β in the Bayesian bootstrap are $\Pr(\beta = s_k) = c_k(\alpha)$ and $\Pr(\beta = s_k|\mathcal{D}) = c_k(\alpha + \mathbf{n})$, respectively. Therefore, Proposition 3 demonstrates that the choice of $b_k = c_k(\alpha)$ delivers the Bayesian bootstrap (here the results are computed analytically rather than via simulation). If a Bayesian bootstrap was run, each draw would be weighed by $w_k = b_k/c_k(\alpha)$ to produce a Bayesian analysis using a proper prior; w_k is the ratio of the priors and does not depend upon the data. Finally, Proposition 4 implies that as $\alpha \downarrow 0$, so $c_k(\alpha) \to J^{-1}$. This demonstrates that, in the Bayesian bootstrap, the implied prior of β is the uniform discrete distribution on the support of the data. In many applications this is an inappropriate prior.

Remark 1 *To simulate from*

$$p(\theta|\mathcal{D}) = \frac{1}{C(\alpha, \mathbf{n})} \frac{b_{k(\theta)}}{c_{k(\theta)}(\alpha)} f_D(\theta; \alpha + \mathbf{n}), \quad b_k = \Pr(\beta = s_k),$$

write $m_k = \frac{b_k}{c_k}/C(\alpha, \mathbf{n})$, $m'_k = b_k/c_k$, $M = \max(m_1, ..., m_J)$ *and* $M' = \max(m'_1, ..., m'_J)$. *Now* $p(\theta|\mathcal{D}) \leq M f_D(\theta; \alpha + \mathbf{n})$, *for any* θ. *We can sample from* $p(\theta|\mathcal{D})$ *by drawing from Dirichlet*$(\alpha + \mathbf{n})$ *and accepting with probability* $m_{k(\theta)}/M = m'_{k(\theta)}/M'$. *The overall acceptance rate is* $1/M$. *If the prior on* β *is weakly informative then* $m'_k \simeq 1$ *for each* k, *and so the acceptance rate* $m'_{k(\theta)}/M' \simeq 1$.

2.9 A cheap approximation

If J is large, $\alpha \downarrow 0$ and no ties, then a central limit theory for binomial random variables implies

$$\frac{1}{J} \log f_{\mathrm{B}}(k-1; J-1, \tau) \simeq -\frac{\left(\frac{k-1}{J-1} - \tau\right)^2}{2\tau(1-\tau)},$$

which should be a good approximation unless τ is in the tails, or J is small. So the resulting approximations to the main posterior quantities are

$$\begin{aligned}
\widehat{\mathrm{E}}(\beta|\mathcal{D}) &= \sum_{j=1}^{J} w_j^* s_j, \quad w_k^* = w_k b_k / \sum_{j=1}^{J} w_j b_j, \quad w_k = \exp\left\{ -\frac{J\left(\frac{k-1}{J-1} - \tau\right)^2}{2\tau(1-\tau)} \right\}, \\
\widehat{\mathrm{Var}}(\beta|\mathcal{D}) &= \sum_{j=1}^{J} w_j^* \left\{ s_j - \widehat{\mathrm{E}}(\beta|\mathcal{D}) \right\}^2, \quad \widehat{F}_{\beta|\mathcal{D}}(\beta) = \sum_{j=1}^{J} w_j^* 1_{s_j \leq \beta}.
\end{aligned}$$

When the prior is flat, this is a kernel weighted average of the data where the weights are determined by the ordering of the data. So large weights are placed on data with ranks $(k-1)/(J-1)$ which are close to τ. This is very close to the literature on kernel quantiles, e.g. Parzen (1979), Azzalini (1981), Yang (1985) and Sheather and Marron (1990).

3 Hierarchical quantile models

3.1 Model structure

Assume a population is indexed by $i = 1, 2, ..., I$ subpopulations, and that our random variable Z again has known discrete support, $\mathcal{S} = \{s_1, ..., s_J\}$. Then we assume within the i-th subpopulation

$$\Pr(Z = s_j|\theta, i) = \theta_j^{(i)}, \tag{7}$$

thus allowing the distribution to change across the subpopulations. Here $\theta^{(i)} = (\theta_1^{(i)}, ..., \theta_{J-1}^{(i)})$, $\theta_J^{(i)} = 1 - \iota'\theta^{(i)}$, and $\theta = (\theta^{(1)}, ..., \theta^{(I)})$. We assume the data $\mathcal{D} = \{Z_1, ..., Z_n\}$ are conditionally independent draws from (7). We assume that each time we see datapoints we also see which subpopulation the datapoint comes from. The data from the i-th population will be written as $\mathcal{D}_i$.

For the i-th subpopulation, the Bayesian nonparametric τ quantile is defined as

$$\beta_i = \underset{b}{\operatorname{argmin}} \sum_{j=1}^{J} \theta_j^{(i)} \rho_\tau(s_j - b).$$

Collecting terms $\beta = (\beta_1, \beta_2, ..., \beta_I)'$, the crucial assumption in our model is that

$$f(\theta|\beta) = \prod_{i=1}^{I} f(\theta^{(i)}|\beta_i).$$

This says the distributions across subpopulations are conditionally independent given the quantile. That is, the single quantiles are the only feature which is shared across subpopulations.

We assume $\beta_i \in \mathcal{S}$, and the $\{\beta_i\}$ are i.i.d. across i, but from the shared distribution $\Pr(\beta_i = s_j|i, \pi) = \pi_j$, $i = 1, 2, ..., I$, where $\pi = (\pi_1, ..., \pi_{J-1})$, and $\pi_J = 1 - \iota'\pi$. We write a prior on π as $p(\pi)$. Then the prior on the hierarchical parameters is

$$f(\beta, \pi) = f(\pi)f(\beta|\pi) = f(\pi)\prod_{i=1}^{I} f(\beta_i|\pi).$$

This structured distribution will allow us to pool quantile information across subpopulations.

Our task is to make inference on $(\beta_1, \beta_2, ..., \beta_I)'$ from $\mathcal{D}$. When taken together, we call this a "nonparametric hierarchical quantile model". This can also be thought of as related to the Robbins (1956) empirical Bayes method, but here each step is nonparametric. By Bayes theorem,

$$f(\beta, \pi|\mathcal{D}) \propto f(\beta, \pi)f(\mathcal{D}|\beta, \pi). \tag{8}$$

We will access this joint density using simulation.

- **Algorithm 1: $\beta, \pi|\mathcal{D}$ Gibbs sampler**

1. Sample from $\Pr(\beta|\mathcal{D},\pi) = \prod_{i=1}^{I} \Pr(\beta_i|\mathcal{D}_i,\pi)$.
2. Sample from $f(\pi|\mathcal{D},\beta) = f(\pi|\beta)$.

In the Dirichlet case, we can sample from $\Pr(\beta_i|\mathcal{D}_i,\pi)$ using Proposition 3. If $f(\pi)$ is Dirichlet, then $\pi|\beta$ =Dirichlet$(\lambda+\nu)$, where $\nu = (\nu_1,...,\nu_J)$, and $\nu_j = \sum_{i=1}^{I} 1(\beta^{(i)} = s_j)$.

3.2 Example: a hierarchical model for quantiles of financial assets' returns

Here we apply the proposed hierarchical quantile model to the estimation of the extreme left quantile of financial assets' daily return. The model could provides an estimate for quantiles of assets for which little historical data is available, borrowing strength from the data on other assets. We use the filtered returns for $I = 20$ stocks from January 2013 to December 2016, described earlier in Example 2.5. In this model we assume the innovations are i.i.d. draws from a discrete distribution,

$$\Pr(Z = s_j|\theta, i) = \theta_j^{(i)}.$$

Similar to the previous example, we use a common support for all assets that is the union of the observed innovations for all assets ($J = 19,985$).

The prior distribution of $\theta^{(i)}$ is a Dirichlet distribution with $\alpha = (\alpha_1,...,\alpha_J)$, in which $\alpha_j = 4\tilde{\alpha}_j + \frac{1}{J}$ with $\tilde{\alpha}_j \propto \left(1+\frac{s_j^2}{6}\right)^{-\frac{7}{2}}$ (that is proportional to the density function of a Student's t-distribution with 6 degrees of freedom) and $\sum_{j=1}^{J} \tilde{\alpha}_j = 1$. Moreover we let $\pi \sim$ Dirichlet(λ), where $\lambda_j = \tilde{\lambda}_j + \frac{1}{J}$, in which $\tilde{\lambda}_j \propto e^{-\frac{1}{2}(s_j+3.14)^2}$ for $j = 1,...,J$, and $\sum_{j=1}^{J} \tilde{\lambda}_j = 5$ (that is a normal density function centered at the $\tau = 0.01$ quantile of a Student's t-distribution with 6 degrees of freedom and unit variance). In Table 3 the empirical and the Bayesian estimate of the quantiles have been reported. In Figure 4, we have depicted $\mathrm{E}(\pi)$ and $\mathrm{E}(\pi|\mathcal{D})$. Basically, $\mathrm{E}(\pi|\mathcal{D})$ is the probability distribution of the 0.01-quantile of an asset with no available data.

3.3 Example: batting records in cricket

We illustrate the hierarchical model using a dataset of the number of runs (which is a non-negative integer) scored in each innings by the most recent (by debut) $I = 300$ English test players. "Tests" are international cricket matches, typically played over 5 days. Here we look at only games involving the English national team. This team plays matches against Australia, Bangladesh, India, New Zealand, Pakistan, South Africa, Sri Lanka, West Indies and Zimbabwe. Batsmen can bat up to twice in each test, but some players fail to get to bat in an individual game due to the weather or

Ticker	Sample Quantile	Hierarchical Bayesian Quantile
MSFT	-2.90	-2.87
PFE	-2.73	-2.73
BAC	-2.67	-2.71
AAPL	-3.16	-3.13
MRK	-2.69	-2.76
MMM	-3.14	-3.18
AXP	-2.86	-3.07
BA	-3.14	-3.14
CAT	-3.18	-3.29
CVX	-2.83	-2.90
KO	-3.30	-3.23
DD	-2.68	-2.71
XOM	-2.73	-2.80
GE	-2.56	-2.59
HD	-2.77	-2.93
IBM	-3.22	-3.32
INTC	-2.97	-2.97
JNJ	-2.81	-2.87
JPM	-2.86	-2.88
MCD	-2.62	-2.84

Table 3: Empirical and Bayesian estimates of 0.01-quantile of the normalized daily returns of 20 stocks estimated by the hierarchical model.

due to the match situation. Some players are elite batsmen and score many runs, others specialize in other aspects of the game and have poor batting records without any runs.

The database starts on 14th December 1951 and ends on 22nd January 2016. Some of these players never bat, others have long careers, the largest of which we see in our database is 235 innings, covering well over 100 test matches. In test matches batsmen can continue their innings for potentially a very long time and so can accumulate very high scores. An inning can be left incomplete for a number of reasons, so the score is right-censored — such innings are marked as being "not out". By the rules of cricket at least 9% of the data must be right-censored. The database is quite large, and batting records are full of heterogeneity, highly skewed, partially censored and heavy tailed data. It is therefore a good test case for our methods.

Academic papers on the statistics of batting include Kimber and Hansford (1993), which is a sustained statistical analysis of estimating the average performance of batsmen just using their own scores. More recent papers include Philipson and Boys (2015) and Brewer (2013).

Our initial aim will be to make inference on the 0.5 quantile for every batsmen, even if they have never batted. To start we will ignore the "not out" indicator. The player-by-player empirical median ranges from 0 and 46, and is itself heavily negatively skewed.

The common support of data for all the players is $\mathcal{S} = \{0, 1, ..., 350\}$, therefore $J = 351$. The

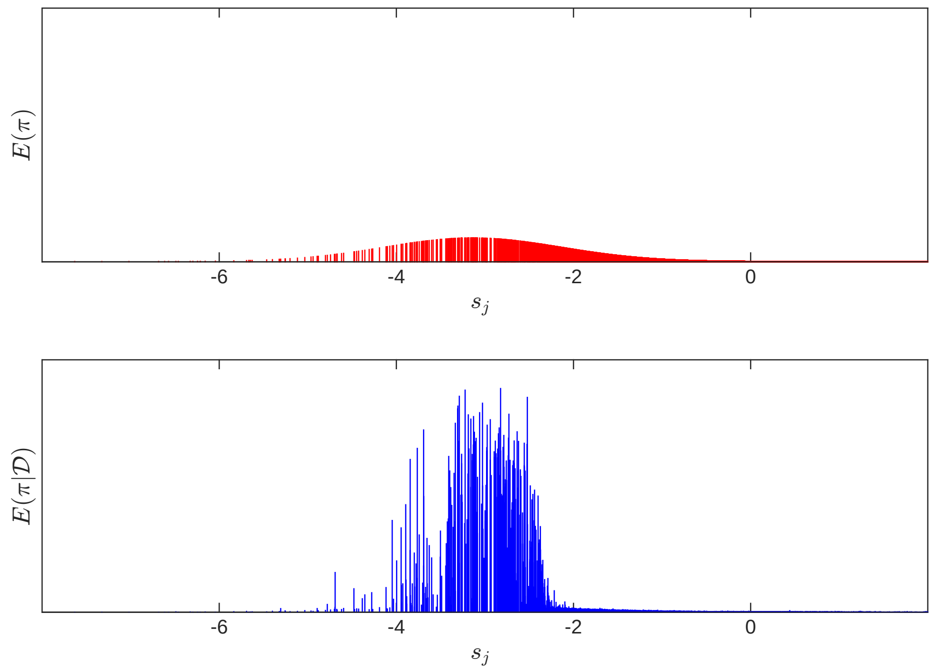

Figure 4: Extreme left quantile estimation for standardized financial asset returns: E(π) and E($\pi|\mathcal{D}$). The initial prior is proportional to a unit variance Normal distribution with mean equal to $\tau = 0.01$-quantile of a Student's t-distribution with 6 degrees of freedom.

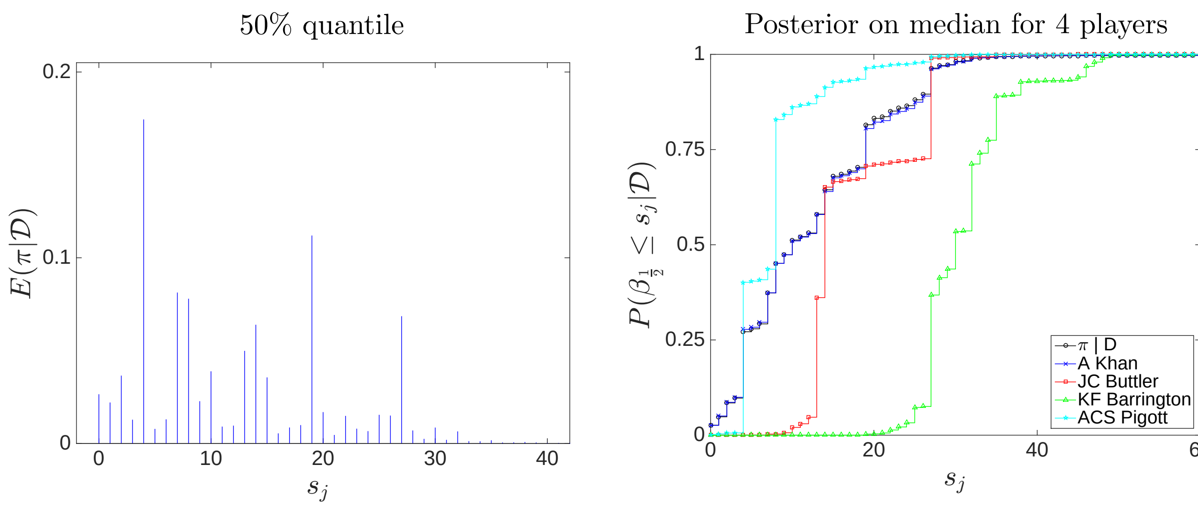

Figure 5: The left hand side shows the posterior distribution of the population probabilities of the quantiles $\pi_j = P(\beta = s_j)$, $\tau = 0.5$. The right hand side shows the posterior distribution of median of several players along with the posterior distribution of π. Notice in the case of Barrington there is only one innings which finished in the range 36 to 44 inclusive, which makes estimating the median unexpectedly hard (given how large a sample we have) and encourages the Bayesian method to aggressively shrink the estimator of the median.

prior distribution of $\theta^{(i)}$ is a Dirichlet distribution with $\alpha = (\alpha_1, ..., \alpha_J)$, where $\alpha_j = 4\tilde{\alpha}_j + \frac{1}{J}$ with $\tilde{\alpha}_j \propto e^{-0.03 s_j}$ and $\sum_{j=1}^{J} \tilde{\alpha}_j = 1$ (The empirical probability mass function of batting scores of all

English players in the matches started between 1930 and 1949, $p_i = \Pr(Z = s_j)$, is approximately proportional to $e^{-0.03 s_j}$. Therefore our Dirichlet prior for θ is approximately centered around this empirical probability mass function with large variability). We assume $\pi \sim \text{Dirichlet}(\lambda)$, where $\lambda_j = \tilde{\lambda}_j + \frac{1}{J}$, in which $\tilde{\lambda}_j \propto e^{-\frac{1}{2}\left(\frac{s_j-15}{15}\right)^2}$ for $j = 1, ..., J$, and $\sum_{j=1}^{J} \tilde{\lambda}_j = 5$. In the left hand side of Figure 5 we have depicted $\mathrm{E}(\pi|\mathcal{D})$ for the $\tau = 0.5$ median case. Figure 7 shows the results for the $\tau = 0.3$ and $\tau = 0.9$ cases. We will return to the non-median cases in the next subsection.

	Posterior						Posterior				
Batsman	$\widetilde{\beta}_{1/2}$	Q$_5$	Q$_{95}$	$\widehat{\beta}_{1/2}$	n_i	Batsman	$\widetilde{\beta}_{1/2}$	Q$_5$	Q$_{95}$	$\widehat{\beta}_{1/2}$	n_i
A Khan	12.8	1	27	–	0	CJ Tavare	17.4	13	25	19.5	56
ACS Pigott	7.8	4	19	6	2	PCR Tufnell	1.2	0	2	1	59
A McGrath	13.1	4	27	34	5	MS Panesar	1.6	0	4	1	64
AJ Hollioake	4.9	2	12	3	6	CM Old	8.4	7	11	9	66
JB Mortimore	16.7	9	19	11.5	12	JA Snow	5.5	4	8	6	71
DS Steele	18.8	7	38	43	16	DW Randall	14.6	9	19	15	79
PJW Allott	6.6	4	14	6.5	18	RC Russell	14.6	10	20	15	86
JC Buttler	17.4	13	27	13.5	20	MR Ramprakash	18.7	14	21	19	92
W Larkins	12.8	7	25	11	25	PD Collingwood	23.2	19	28	25	115
NG Cowans	4.1	3	7	3	29	RGD Willis	4.1	4	5	5	128
JK Lever	5.4	4	10	6	31	KF Barrington	31	25	46	46	131
M Hendrick	3.4	1	4	2	35	APE Knott	17.6	13	24	19	149
DR Pringle	7.8	4	9	8	50	IT Botham	20.7	15	27	21	161
C White	11.5	7	19	10.5	50	DI Gower	26.9	25	28	27	204
GO Jones	15.9	10	22	14	53	AJ Stewart	25.6	19	28	27	235

Table 4: Estimated median batting scores, treating not outs as if they were completed innings (i.e. ignoring right censoring). The batsman are ordered by sample size (i.e. the number of innings the batsman had). Table shows, for each batsmen, the mean of the Bayesian posterior of the median given the data, $\widetilde{\beta}_{1/2} = E(\beta_{1/2}|\mathcal{D})$, the sample median $\widehat{\beta}_{1/2}$ and the sample size n_i. Q_5 and Q_{95} are the estimates of the Bayesian 5% and 95% quantiles of the posterior distribution of the median, so indicates how uncertain we are about the Bayesian estimator of the mean. All the Bayesian quantities are estimated by simulation.

In the right hand side plot in Figure 5, the posterior distribution function of the median of scores for several players have been compared with the posterior distribution function of π (the black curve). For the first player, A. Khan (the blue curve), no data are available as he never batted, and the distribution is indistinguishable from that for $\mathrm{E}(\pi|\mathcal{D})$. A.C.S. Pigott played two innings for England, scoring 4 and 8 not out. The light blue curve shows that even with just two data points a lot of the posterior mass on the median has moved to the left, but the median is very imprecisely estimated (the estimate of median is 7.8 with 90% credible region $[4, 19]$). The red curve corresponds to J. C. Buttler, whose sample median (13) is close to $\mathbb{E}_{\pi|\mathcal{D}}(\beta) = 12.9$. His 20 actual scores were 85, 70, 45, 0, 59*, 13, 3*, 35*, 67, 14, 10, 73, 27, 7, 13, 11, 9, 12, 1, 42 (an asterisk in superscript denotes a not out). His scores are not particularly heavy-tailed and so the median is reasonably well determined (the estimate of median is 17.4 with 90% credible region

[13, 27]). The green line shows the results for K. F. Barrington who batted 131 times and one of the highest averages of any English batsman. His median is relatively high (31) but surprisingly not well determined (with 90% credible region [25, 46]). Remarkably he has only once scored between 36 and 44 (inclusive), so there is a whole range of possible scores where there is no data. This stretches the Bayesian nonparametric interval. The right hand side of Figure 5 shows this clearly. Of course a 90% interval would be much shorter as it would not include this blank range.

Table 4 shows estimated posterior mean of the median for 30 players, together with sample sizes, 90% intervals, putting 5% of the posterior probability in each tail. Also given is the empirical median. The players are sorted by sample size. It shows that when the sample size is small there is a great deal of borrowing across the subpopulations. However, when the subpopulation is large then the hierarchy does not make much difference. McGrath's scores are 69, 81, 34, 4, 13 (with sample median 34), so he has very little data in the middle (he either fails or scores highly), and therefore the procedure shrinks the median a great deal towards a typical median result (the Bayesian estimate is 13.1). Steele's sample median (43) is very high (it is very similar to Barrington's) and the sample size is low (16). The resulting Bayes estimate is still a high number (18.8), but is less than half of his sample median. Hence we think the evidence is that Steele was a very good batsmen, but there is not the evidence to rank him as a great batsman like Barrington. His record is more in line with Botham and Ramprakash.

Figure 6 highlights the shrinkage of the sample median by the hierarchical model. We plot the batsman's sample median $\widehat{\beta}_{1/2}$ against the batsman's sample size n_i. Blue arrows show that the Bayesian posterior mean of the median is below the sample median, that is, it is shrunk down. Red arrows are the opposite, the Bayesian estimator is above the sample median, so is moved upwards. The picture shows there is typically more shrinkage for small sample sizes. But also, high sample medians are typically shrunk more than low sample medians, but there are more medians which are moved up than down. All this makes sense: the data are highly skewed, so high scores can occur due to high medians or by chance. Hence until we have seen a lot of high scores, we should shrink a high median down towards a more common value. Analyzing a similar dataset, Efron and Morris (1977) use James-Stein estimators to find the baseball players' batting average, by shrinking toward the average of averages (see also Efron and Morris (1975)).

3.4 Estimating the quantile function

Of interest is β_τ, the τ-th quantile, as a function of τ. Here we estimate that relationship pointwise, building a separate hierarchical model for each value of τ. The only change we will employ is to

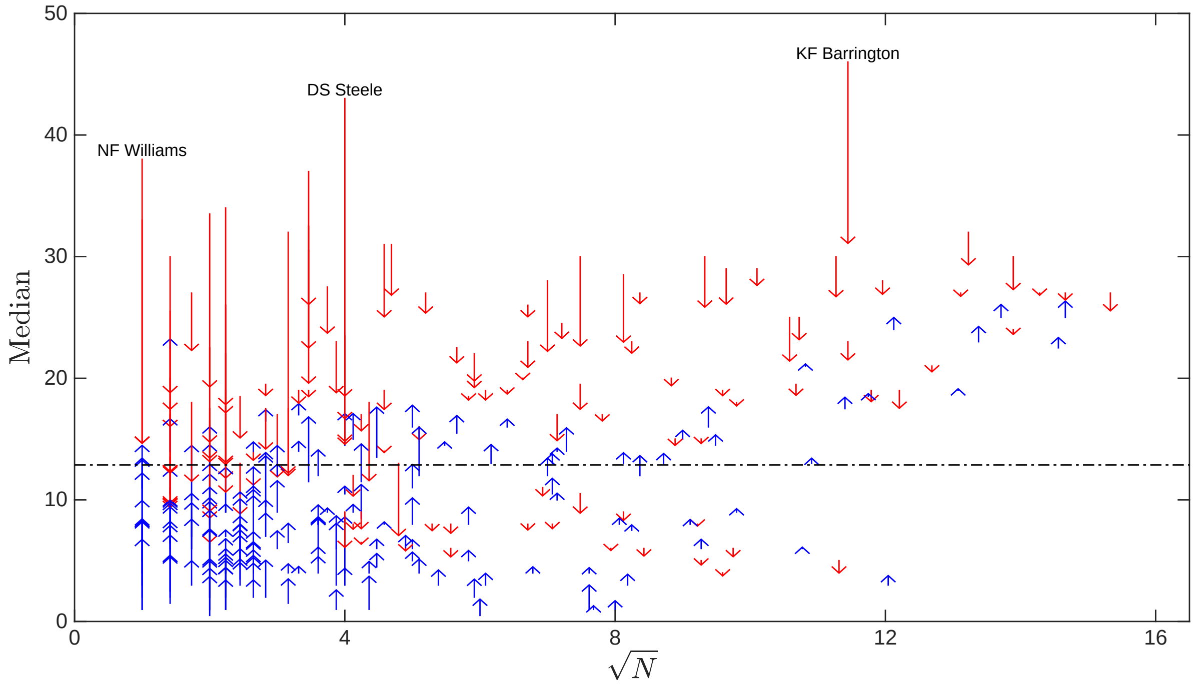

Figure 6: Sample median (arrow nocks) and mean of posterior distribution of medians (arrow heads) against the sample size for all players. The blue arrows indicate the estimates which were moved upwards, and the estimators which were moved down demonstrated by the red arrows. The dashed line is the expected value of β under $E(\pi|D)$.

set $\tilde{\lambda}_j \propto \exp\left\{-\left(s_j-\mu_\tau\right)^2/\sigma_\tau^2\right\}$, allowing $\mu_\tau = 15 + 15\Phi^{-1}(\tau)$ and $\sigma_\tau = 15$.

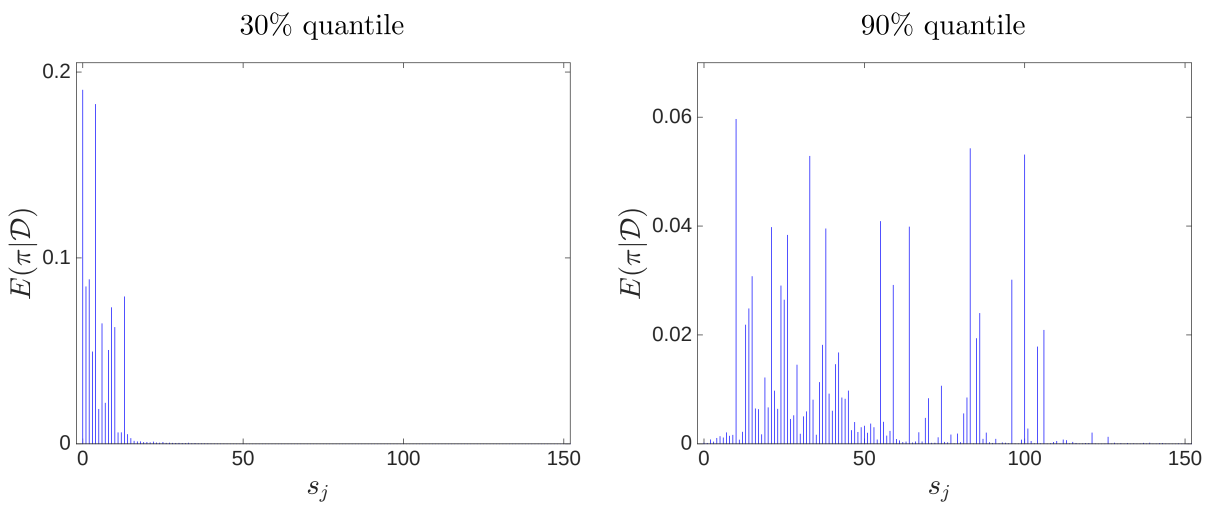

Figure 7: The left hand side shows the posterior distribution of the population probabilities of the quantiles $\pi_j = P(\beta = s_j)$, $\tau = 0.30$. The right hand side shows the corresponding result for $\tau = 0.90$.

Figure 7 shows the common mixing distribution $\mathrm{E}\left(\pi|\mathcal{D}\right)$ for two quantile levels $\tau = 0.30$ and

$\tau = 0.90$. Notice, of course, how different they are, with a great deal of mass on low scores when $\tau = 0.30$ and vastly more scatter for $\tau = 0.90$. This is because even the very best batsmen fail with a substantial probability, frequently recording very low scores. In the right hand tail, the difference between the skill levels of the players is much more stark, with enormous scatter.

We now turn to individual players. In Figure 8, the dashed blue line shows the empirical quantile function for P.J.W. Allott, while also plotted using a blue full line is the associated Bayesian quantile function $\mathrm{E}(\beta_\tau|\mathcal{D})$. The results are computed for $\tau \in \{0.01, 0.2, ..., 0.99\}$. The Bayesian function also shows a central 90% interval around the estimate.

In the same figure, the red curve shows the same object but for K.F. Barrington, who tended to score very highly, and also played a great deal (his n_i is around 8 times larger than Allott's). In both players' cases the lower quantiles are very precisely estimated and not very different, but at higher quantile levels the uncertainty is material and the differences stretch out. Further, at these higher levels the 90% intervals are typically skewed, with a longer right hand tail.

The Bayesian quantile functions seem shrunk more for Barrington, which looks odd as Allott has a smaller sample size. But Barrington has typically much higher scores (and so more variable) and so his quantiles are intrinsically harder to estimate and so are more strongly shrunk. His exceptionalism is reduced by the shrinkage.

For a moment we now leave the cricket example. We should note that we have ignored the fact some innings were not completed and marked "not out", a form of censoring. We now develop methods to overcome this deficiency.

4 Truncated data

4.1 Censored data

Here this methodology is extended to models with truncated data. The probabilistic aspect of the model is unaltered. Assume the support is sorted and known to be $\mathcal{S}$, and $\Pr(Z = s_j|\theta) = \theta_j$. However, in addition to some fully observed data, $\mathcal{D}_1 = \{z_1, ..., z_N\}$, there exist N' additional data, $\mathcal{D}_2 = \{s_{l_i}, ..., s_{l_{N'}}\}$, which we know has been right truncated. We assume the non-truncated versions of the data are independent over i, such that $U_i \geq s_{l_i}$, $1 \leq i \leq N'$, $U_i \in \mathcal{S}$, $\Pr(U = s_j|\theta) = \theta_j$. Write $U = (U_1, ..., U_{N'})$. Thus the data is $\mathcal{D} = \mathcal{D}_1 \bigcup \mathcal{D}_2$.

Inference on (β, π) is carried out by employing a Gibbs sampler in order to draw from $p(\beta, \pi, U|\mathcal{D})$. This adds a first step to Algorithm 1:

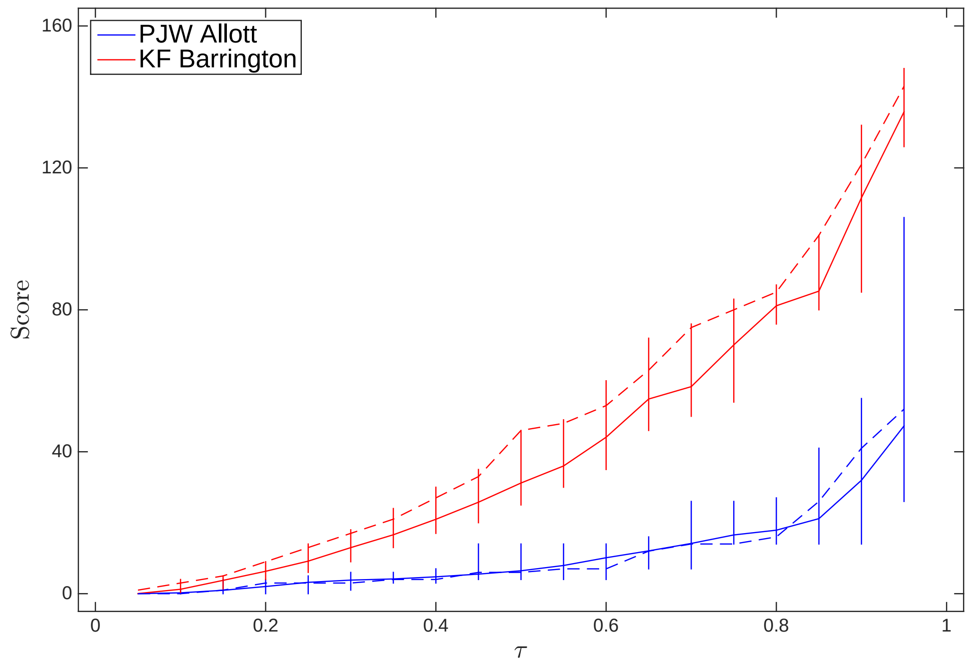

Figure 8: The pointwise estimated quantile function for two cricketers: P.J.W. Allott and K.F. Barrington. These calculations ignore the impact of censoring. Horizonal lines denote 90% posterior intervals with 5% in each tail. The curve for Allott uses his 18 innings, Barrington had 131 innings.

- **Algorithm 2: $\beta, \pi, U|\mathcal{D}$ Gibbs sampler**

1. Sample $\Pr(U|\beta, \mathcal{D}, \pi)$.

2. Sample $\Pr(\beta|\mathcal{D}, U, \pi)$.

3. Sample $f(\pi|\beta)$, returning to 1.

Sampling from $U|\left(\beta = s_k, \mathcal{D}\right)$ is carried out through data augmentation:

1. Sampling from $\Pr(U|\beta, \mathcal{D}, \pi)$ by,

   (a) Sample $\theta|\left(\beta = s_k, \mathcal{D}\right) \sim D_J(\alpha + \mathbf{n}, k)$.

   (b) Sample $U|\left(\beta = s_k, \theta, \mathcal{D}\right)$.

Step 1(b) is straightforward, while 1(a) is a truncated Dirichlet defined in (5). The online Appendix F shows how to simulate from $D_J(\alpha, k)$ exactly. Although it is tempting to sample $\theta|U, \mathcal{D}$ and $U|\theta, \mathcal{D}$, but this fails in practice; the reasons for this are described in detail in online Appendix E.

4.2 A Bayesian bootstrap for the censored data

A Bayesian bootstrap algorithm can be developed to deal with the censored data. Independent draws from the Bayesian bootstrap posterior distribution can be obtained by the following algorithm.

- **Algorithm 3: Bayesian bootstrap with censored data**

1. Draw $\theta^* \sim \text{Dirichlet}(\alpha + \mathbf{n})$.
2. For $1 \leq i \leq N'$, draw U_i from $\{s_{l_i}, ..., s_J\}$, with probability $\Pr(U_i = s_j) = \frac{\theta_j^*}{\sum_{k=l_i}^J \theta_k^*}$, and set $n_j' = \sum_1^{N'} 1(U_i = s_j)$, and $\mathbf{n}' = (n_1', ..., n_J')$.
3. Draw $\theta \sim \text{Dirichlet}(\alpha + \mathbf{n} + \mathbf{n}')$. Set $\beta = t(\theta)$. Go to 1.

4.3 Returning to cricket: the impact of not outs

In cricket scores at least 9% of scores in each innings must be not out, so right censoring is important. Not outs are particularly common for weaker batsmen who are often left not out at the end of the team's innings. In Section 3.3 we ignored this feature of batting. Here we return to it to correct the results.

Figure 9 shows the estimated pointwise quantile function for Barrington and Allott, taking into account the not outs. Both are shifted upwards, particularly Allott in the right hand tail. However, Allott's right hand tail is not precisely estimated.

Table 5 shows the Bayesian results for our selected 30 players, updating Table 4 to reflect the role of right censoring. Here n_i' denotes the number of not out, that is right censored innings, the player had. In many cases this is between 10% and 20% of the innings, but for some players it is far higher. R.G.S. Willis is the leading example, who had 55 not outs of 128 innings. A leading bowler, he usually batted towards the end of innings and was often left not out. His posterior mean of the median is inflated greatly by the statistical treatment of censoring. Further, the interval between Q_5 and Q_{95} is widened substantially. Other players are hardly affected, e.g. M.R. Ramprakash, who had 6 not outs in 92 innings.

Table 6 shows a ranking of players by the mean of the posteriors of the quantiles, at three different levels of quantiles. This shows how the rankings change greatly with the quantile level. For small levels, we can think of this as being about consistency. For the median it is about typical performance. For the 90% quantile this is about upside potential to bat long. A remarkable result

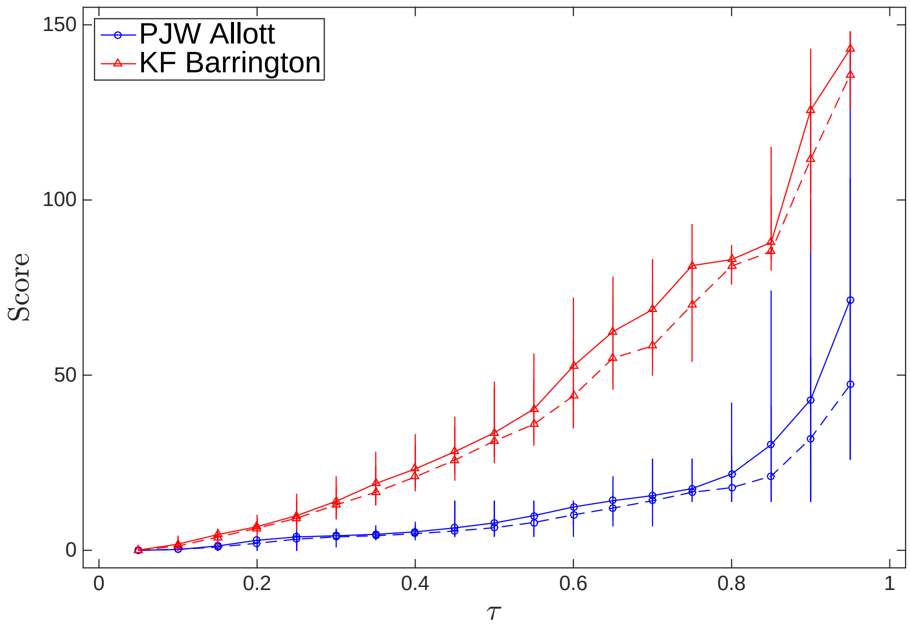

Figure 9: The censored-adjusted pointwise estimated quantile function for two cricketers: P.J.W. Allott and K.F. Barrington. The solid lines are the the estimates with the censored observations, and the dashed lines are obtained by ignoring that they are censored data. Horizonal lines denote 90% posterior intervals with 5% in each tail. The curve for Allott uses his 18 innings, Barrington had 131 innings.

is J.B. Bolus who has a very high $\beta_{0.30}$ quantile. His career innings were the following: 14, 43, 33, 15, 88, 22, 25, 57, 39, 35, 58, 67. He only played for a single year, but never really failed in a single inning. However, he never managed to put together very long memorable innings and this meant his Test career was cut short by the team selectors, who seem to not so highly value reliability.

Again K.F. Barrington is the standout batsman. He is very strong at all the different quantiles. Notice though he still had a 30% chance of scoring 14 or less. But once his innings was established his record was remarkably strong, typically playing long innings.

5 Conclusions

We provide a Bayesian analysis of quantiles by embedding the quantile problem in a larger inference challenge. This delivers quite simple ways of performing inference on a single quantile. The frequentist performance of our methods are similar to that of the bootstrap.

We extend the framework to introduce a hierarchical quantile model, where each subpopulation's distribution is modeled nonparametrically but linked through a nonparametric mixing distribution placed on the quantile. This allows non-linear shrinkage, adjusting to skewed and sparse data in

				Ignoring censoring in analysis					
	Bayesian			Bayesian			Empirical		
Batsman	$\widetilde{\beta}_{1/2}$	Q_5	Q_{95}	$\widetilde{\beta}_{1/2}$	Q_5	Q_{95}	n_i	n'_i	$\widehat{\beta}_{1/2}$
A Khan	14.7	4	30	12.7	2	27	0	0	–
ACS Pigott	9.7	4	27	7.8	4	19	2	1	4
A McGrath	14.3	4	31	13	4	27	5	0	34
AJ Hollioake	5.5	4	14	4.7	2	12	6	0	2
JB Mortimore	16.4	9	20	16.7	9	19	12	2	11
DS Steele	19.4	7	37	18.8	7	35	16	0	42
PJW Allott	7.8	4	14	6.7	4	14	18	3	6
JC Buttler	17.7	13	27	17.1	13	27	20	3	13
W Larkins	12.6	7	27	13.3	7	25	25	1	11
NG Cowans	5.9	4	10	4.1	3	7	29	7	3
JK Lever	6.7	4	11	5.3	4	8.5	31	5	6
M Hendrick	6.3	4	10	3.4	1	5	35	15	2
DR Pringle	8.3	7	10	7.7	4	9	50	4	8
C White	13.1	8	19	11.7	7	19	50	7	10
GO Jones	17.1	10	22	15.9	10	19	53	4	14
CJ Tavare	17.4	12.5	25	17.3	13	25	56	2	22
PCR Tufnell	4.8	1	9	1.2	0	2	59	29	1
MS Panesar	4	4	4	1.6	0	4	64	21	1
CM Old	8.9	7	13	8.4	7	11	66	9	9
JA Snow	7.9	4	9	5.5	4	8	71	14	6
DW Randall	14.6	9	19	14.5	10	19	79	5	15
RC Russell	16.2	9.5	24	14.6	12	20	86	16	15
MR Ramprakash	18.8	14	21	18.6	14	21	92	6	19
PD Collingwood	25	19	30	23.3	19	28	115	10	25
RGD Willis	8.5	7	10	4.1	4	5	128	55	5
KF Barrington	33.7	27	48	31	25	46	131	15	46
APE Knott	18.9	14	27	17.5	13	24	149	15	19
IT Botham	20.6	15	27	20.7	15	27	161	6	21
DI Gower	27.5	26	32	27	25	28	204	18	27
AJ Stewart	26.3	19	29.5	25.6	19	28	235	21	27

Table 5: Estimated median batting scores. Sample median is compared with two Bayesian estimators, where $\widetilde{\beta}_{1/2} = E(\beta_{1/2}|D)$. n_i is the number of innings, n'_i denotes the number of not outs which are treated as right censored data and $\widehat{\beta}_{1/2}$ is the empirical median. In the first model the not outs are assumed to be right censored observations. In the second model they are treated as if they were completed innings. Q_5 denotes the estimated 5% point on the relevant posterior distribution.

an automatic manner.

This approach is illustrated by the analysis of a large database from sports statistics of 300 Test cricketers. Each person's batting performance is modeled nonparametrically and separately, but linked through a quantile which is drawn from a common distribution. This allows us to shrink each cricketer's performance – a particular advantage in cases where the careers are very short.

The modeling approach is extended to allow for truncated data. This is implemented by using simulation based inference. This is illustrated in practice by looking at not outs in batting innings, where we think of the data as right censored.

	0.3 quantile				0.5 quantile				0.9 quantile			
rank	Batsman	$\widetilde{\beta}_{0.3}$	Q_5	Q_{95}	Batsman	$\widetilde{\beta}_{0.5}$	Q_5	Q_{95}	Batsman	$\widetilde{\beta}_{0.9}$	Q_5	Q_{95}
1	JB Bolus	16.6	4	33	KF Barrington	33.7	27	48	KF Barrington	121.3	101	143
2	KF Barrington	14.0	9	21	KP Pietersen	30.5	26	34	IR Bell	116.8	109	121
3	DI Gower	13.1	11	16	JH Edrich	29.5	22	35	GP Thorpe	115.8	94	119
4	AN Cook	12.9	11	13	G Boycott	29.1	23	35	PH Parfitt	115.6	86	121
5	ER Dexter	12.8	10	16	ER Dexter	28.5	27	32	IJL Trott	111.2	64	121
6	G Boycott	12.6	10	13	ME Trescothick	28.4	24	32	MC Cowdrey	111.1	96	119
7	GA Gooch	12.6	10	13	BL D'Oliveira	28.2	23	32	G Boycott	109.7	106	116
8	KP Pietersen	12.6	9	14	AJ Strauss	28.1	25	32	DL Amiss	106.9	64	119
9	RW Barber	12.5	6	13	R Subba Row	27.9	22	32	AN Cook	106.7	96	118
10	AJ Strauss	12.5	9	14	DI Gower	27.5	26	32	MP Vaughan	106.3	100	115
11	G Pullar	12.4	9	14	MC Cowdrey	27.3	23	32	ME Trescothick	105.4	90	113
12	ME Trescothick	12.4	9	14	AW Greig	27.2	19	32	KP Pietersen	105.2	96	119
13	MP Vaughan	12.4	9	13	AN Cook	27.2	22	32	AJ Strauss	105.0	83	112
14	MC Cowdrey	12.3	9	13	GA Gooch	27.1	22	30	AW Greig	102.8	96	110
15	JE Root	12.2	6	13	JB Bolus	27.1	15	36	N Hussain	102.5	85	109
16	R Subba Row	12.1	8	13	GP Thorpe	27.1	19	32	CT Radley	102.4	59	106
17	RA Smith	12.1	8	13	IJL Trott	26.7	19	35	JE Root	102.1	83	130
18	JM Parks	12.0	7	14	AJ Stewart	26.3	19	29	DI Gower	101.8	85	106
19	JG Binks	12.0	6	13	PH Parfitt	25.8	18	32	AJ Lamb	101.4	83	119
20	GP Thorpe	11.8	9	13	MP Vaughan	25.7	19	32	DS Steele	100.8	64	106

Table 6: Best 20 players ranked based on the mean of the posteriors of the quantiles, at three different levels of quantiles.

References

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika 68*, 326–328.

Boos, D. and J. F. Monahan (1986). Bootstrap methods using prior information. *Biometrika 73*, 77–83.

Bornn, L., N. Shephard, and R. Solgi (2016). Moment conditions and Bayesian nonparametrics. Unpublished paper: arXiv:1507.08645.

Brewer, B. J. (2013). Getting your eye in: A Bayesian analysis of early dismissals in cricket. Unpublished paper: School of Mathematics and Statistics, The University of New South Wales.

Butucea, C. and F. Comte (2009). Adaptive estimation of linear functionals in the convolution model and applications. *Bernoulli 15*, 69–98.

Carlin, B. P. and T. A. Louis (2008). *Bayes and Empirical Bayes Methods for Data Analysis* (3 ed.). Chapman and Hall.

Cavalier, L. and N. W. Hengartner (2009). Estimating linear functionals in Poisson mixture models. *Journal of Nonparametric Statistics 21*, 713–728.

Chamberlain, G. and G. Imbens (2003). Nonparametric applications of Bayesian inference. *Journal of Business and Economic Statistics 21*, 12–18.

Chernozhukov, V. and H. Hong (2003). An MCMC approach to classical inference. *Journal of Econometrics 115*, 293–346.

Diaconis, P., S. Holmes, and M. Shahshahani (2013). Sampling from a manifold. In G. Jones and X. Shen (Eds.), *Advances in Modern Statistical Theory and Applications*. Institute of Mathematical Statistics.

Dunson, D. and J. Taylor (2005). Approximate Bayesian inference for quantiles. *Journal of Nonparametric Statistics 17*, 385–400.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press.

Efron, B. (2013). Empirical Bayes modeling, computation, and accuracy. Unpublished paper, Department of Statistics, Stanford University.

Efron, B. and C. Morris (1975). Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association 70*, 311–319.

Efron, B. and C. Morris (1977). Stein's paradox in statistics. *Scientific American 236*, 119–127.

Federer, H. (1969). *Geometric Measure Theory*. New York: Springer–Verlag.

Feng, Y., Y. Chen, and X. He (2015). Bayesian quantile regression with approximate likelihood. *Bernoulli 21*, 832–850.

Hjort, N. and S. Petrone (2007). Nonparametric quantile inference with Dirichlet processes. In V. Nair (Ed.), *Advances in Statistical Modeling and Inference. Essays in Honor of Kjell A. Doksum*, pp. 463–492. World Scientific.

Hjort, N. L. and S. G. Walker (2009). Quantile pyramids for Bayesian nonparametrics. *The Annals of Statistics 37*, 105–131.

Jeffreys, H. (1961). *Theory of Probability*. Oxford: Oxford University Press.

Kimber, A. C. and A. R. Hansford (1993). A statistical analysis of batting in cricket. *Journal of the Royal Statistical Society, Series B 156*, 443–455.

Koenker, R. (2005). *Quantile Regression*. Cambridge: Cambridge University Press.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica 46*, 33–50.

Koenker, R. and J. Machado (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association 94*, 1296–1309.

Kottas, A. and M. Krnjajic (2009). Bayesian semiparametric modelling in quantile regression. *Scandinavian Journal of Statistics 36*, 297–319.

Kozumi, H. and G. Kobayashi (2011). Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation 81*, 1565–1578.

Lancaster, T. and S. J. Jun (2010). Bayesian quantile regression methods. *Journal of Applied Econometrics 25*, 287–307.

Lavine, M. (1995). On an approximate likelihood for quantiles. *Biometrika 82*, 220–222.

Lazar, N. A. (2003). Bayesian empirical likelihood. *Biometrika 90*, 319–326.

Li, Q., R. Xi, and N. Lin (2010). Bayesian regularized quantile regression. *Bayesian Analysis 5*, 1–24.

Lindley, D. V. and A. F. M. Smith (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, 1–41.

McAuliffe, J. D., D. M. Blei, and M. I. Jordan (2006). Nonparametric empirical Bayes for the Dirichlet process mixture model. *Statistical Computing 16*, 5–14.

Morris, C. N. and M. Lysy (2012). Shrinkage estimation in multilevel normal models. *Statistical Science 27*, 115–134.

Muller, U. (2013). Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica 81*, 1805–1849.

Parzen, E. (1979). Nonparametric statistical data modeling. *Journal of the American Statistical Association 74*, 105–121.

Parzen, E. (2004). Quantile probability and statistical data modeling. *Statistical Science 19*, 652–662.

Philipson, P. and R. Boys (2015). Who is the greatest? A Bayesian analysis of test match cricketers. Unpublished paper: New England Symposium on Statistics in Sports.

Robbins, H. (1956). An empirical Bayesian approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 157–163. University of California Press.

Rubin, D. B. (1981). The Bayesian bootstrap. *Annals of Statistics 9*, 130–134.

Sheather, S. J. and J. S. Marron (1990). Kernel quantile estimators. *Journal of the American Statistical Association 85*, 410–416.

Stein, C. M. (1966). An approach to recovery of interblock information in balanced incomplete block designs. In *Research Papers in Statistics (Festchrift J. Neyman)*, pp. 351–366. London: Wiley.

Tsionas, E. G. (2003). Bayesian quantile regression. *Journal of Statistical Computation and Simulation 73*, 659–674.

Yang, S. S. (1985). A smooth nonparametric estimator of a quantile function. *Journal of the American Statistical Association 80*, 1004–1011.

Yang, Y. and X. He (2012). Bayesian empirical likelihood for quantile regression. *The Annals of Statistics 40*, 1102–1131.

Yang, Y., H. J. Wang, and X. He (2016). Posterior inference in Bayesian quantile regression with asymmetric Laplace likelihood. *International Statistical Review 84*, 327–344.

Yu, K. and R. A. Moyeed (2001). Bayesian quantile regression. *Statistics and Probability Letters 54*, 437–447.