The mesa distribution: an approximation likelihood for simultaneous nonlinear quantile regression

P. Richard Hahn
Booth School of Business, University of Chicago

Lane F. Burgette
RAND Corporation

April 13, 2012

Abstract
This paper proposes an approximation likelihood for Bayesian multiple quantile estimation. The simple form of the pseudo-density makes it possible to use in conjunction with nonparametric priors over quantile regression functions. Specifically, taking the conditional quantile functions to be sums of independent stochastic (log-Gaussian) processes yields a model for coherent multiple quantile regression which is at once simple, flexible and robust. A single-index version of the model is demonstrated, adapting the Gaussian process single-index model to the simultaneous multiple quantile regression setting.

keywords: asymmetric Laplace distribution, Bayesian, Gaussian process, multiple quantile regression, nonlinear regression, relevance detection, semiparametric, single-index model, quasi-likelihood

1 Introduction
Commonly, one is interested in how the conditional distribution of some response variable changes with respect to a set of predictor variables. Typical mean-regression models restrict this interest to the mean of the conditional distribution; heteroscedastic regression models further allow expected deviations from the mean to vary with the predictor variables. Potentially, arbitrary features of the response distribution may vary with the predictors. However, practitioners tend to find arbitrary distributions (even restricting to distributions with continuous or smooth density functions) difficult to reason with. Accordingly, it is often convenient to cast the density regression problem in terms of a finite vector of conditional quantile functions. How does location of the 10% tail vary with
a set of predictor variables, or the 90% tail? While such summaries do not fully characterize a
density – they remain entirely mute regarding what happens in between these landmark points –
they reduce the problem of comparing two distributions to the more practicable task of comparing
two vectors. This approach is called multiple quantile regression.

One may attempt to infer the full conditional density and from there rely on the implied quantile
vector as a parsimonious summary, but this approach has the immediate downside that specifying
a sensible prior over distributions requires dealing with the full densities from the outset, the
avoidance of which was a primary motivation of the quantile regression approach in the first place.
What one would want, rather, is to be able to specify the prior directly in terms of the conditional
quantile vector. In not allowing this, full density regression approaches are often unwieldy for
practical applications.

Accordingly, one might opt instead to use a parametric quasi-likelihood with parameters cor-
responding explicitly to the conditional quantiles of interest (plus, perhaps, a small number of
additional parameters). This approach includes methods based on the asymmetric Laplace density.
The patent shortcoming of asymmetric Laplace models, however, is that one must analyze one
quantile at a time, with the insensible result that the estimated quantile functions can intersect one
another so that, for example, in some regions of the predictor space the 10th percentile is inferred
to be higher than the 90th percentile. Thus, what the asymmetric Laplace model lacks is the ability
to do simultaneous multiple quantile regression.

This paper introduces a quasi-likelihood specifically designed for simultaneous multiple quantile
regression; it generalizes the asymmetric Laplace model and is easier to interpret than full density
regression approaches, especially in terms of prior specification. The simple form of the proposed
approximation density makes it possible to use in conjunction with nonparametric priors over
nonlinear quantile regression functions.
Additionally we consider a parsimonious model for multiple quantile regression via a single-index model where the predictor variables enter the conditional distribution of the response via a scalar linear combination. This approach generalizes the work of Hua et al. (2012), who develop a single-index asymmetric Laplace model. Moreover, we leverage the scalar nature of the single-index model to provide a more efficient sampling algorithm based on a convolution representation of Gaussian processes.

2 The “mesa” distribution for multiple quantile estimation

Before turning to the regression context, first consider the probability density function of some univariate response variable \(y \in \mathbb{R}\), denoted \(y \sim F\). Assume that \(F\) may be approximated to within a tolerable total variation distance by a parametric family \(F_{q,\theta}\) indexed by a set of \(k < \infty\) tuples \((q, \theta) = \{(q_1, \theta_1), \ldots, (q_k, \theta_k)\}\) such that \(q_j \in (0, 1)\) and \(\theta_j \in \mathbb{R}\) with \(q_j < q_{j'}\) and \(\theta_j < \theta_{j'}\) when \(j < j'\). These tuples map to probability distributions via a certain interpolation rule which generates the cumulative distribution function. Specifically, consider the rule that interpolates linearly between the tuples in the \((\theta-q)\)-plane and adjoins exponential tails with scale parameters chosen to preserve continuity of the corresponding density at both \(\theta_1\) and \(\theta_k\). Define \(\Delta_j = \theta_j - \theta_{j-1}\) and \(\pi_j = q_j - q_{j-1}\) where \(q_0 = 0\) and \(\pi_{k+1} = 1 - q_k\). Then

\[
f_q(y \mid \theta) = \begin{cases} 
\pi_1 \lambda_1 \exp(-\lambda_1 |y - \theta_1|) & \text{if } y \leq \theta_1 \\
\pi_j \Delta_j^{-1} & \text{if } \theta_{j-1} < y \leq \theta_j \\
\pi_{k+1} \lambda_{k+1} \exp(-\lambda_{k+1} |y - \theta_k|) & \text{if } y > \theta_k 
\end{cases}
\]
with $\lambda_1 = \frac{\pi_2}{\Delta_2 \pi_1}$ and $\lambda_{k+1} = \frac{\pi_k}{\Delta_k \pi_{k+1}}$. By construction

$$\int_{-\infty}^{\theta_j} f_{q,\theta}(y) \, dy = q_j.$$ \hfill (2)

Denote this family the “mesa” distribution owing to the butte-like profile of its densities, as illustrated in the top row of figure 1.

Note that different interpolation rules would work as well in principle, for example, cubic Hermite polynomials. The especially simple form of the mesa density facilitates ease of computation.

For any particular rule, increasing $k$ gives increasingly better approximations to any smooth density (Hjort and Walker, 2009). As those authors mention, in practical applications it is sensible to chose $k$ as a function of the number of replications $n$. This issue is not addressed further here, except to note that $k$ could reasonably be chosen via a stability approach as in Liu et al. (2010) and Meinshausen and Buhlmann (2010). By this we mean, intuitively, that one wants to select the smallest number $k$ such that inferences under the $k$-quantile and $(k+1)$-quantile models are
reasonably similar.

The mesa distribution has three main advantages over other approaches. First, because it is parametrized by multiple quantiles, it permits simultaneous estimation of conditional quantile functions; other popular methods require separate analyses for each quantile. Second, the mesa density allows direct specification of priors on multiple conditional quantiles; other approaches typically do this implicitly. Third, the simple form of the mesa density permits easy computation so that one may entertain complicated nonlinear functions of predictor variables. In the next section these advantages are demonstrated with a Gaussian process-based quantile regression model.

2.1 Summary of previous approaches

2.1.1 Asymmetric Laplace distribution

The likelihood-based literature on parametric quantile regression has focused on parametric pseudo-densities. The most widely used approach is based on the asymmetric Laplace distribution, with density function

\[ f(y \mid \theta, \phi, q) = \phi q (1 - q) \exp \{ -\phi (y - \theta) (q - 1) I_{y < \theta} \}, \]  

(3)

where \( \phi \) is a scale parameter and \( \theta \) is a scalar location parameter which coincides with the \( q \)th quantile. The form of the asymmetric Laplace density permits easy posterior computation (Yu and Moyeed, 2001) and yields consistent inference for the \( q \)th quantile, because the likelihood is maximized at the empirical quantiles. A detailed reference for various properties of the asymmetric Laplace distribution is Kotz, Kozubowski, and Podgórski (2001).

The asymmetric Laplace density is intimately related to well known classical approaches (Koenker, 2005) based on the check loss function:

\[ \rho_q(a) = (y - a) (q - 1) I_{y < a}. \]  

(4)
The usefulness of this loss function for estimating quantiles stems from the fact that expected loss over a distribution with cumulative distribution function $G$ is minimized by $a = G^{-1}(q)$. The asymmetric Laplace is obtained by exponentiating this (negative) loss and normalizing; introducing a prior over $\theta$ and applying Bayes rule yields a so-called Gibbs posterior (Jian and Tanner, 2008).

The major shortcoming of the asymmetric Laplace distribution is its unsuitability for simultaneous multiple quantile regression. Estimating quantile regressions independently is intuitively inefficient as well as logically incoherent: it can yield crossing regression estimates and the assumed response distribution depends on the quantile of interest. The mesa approximation density remedies these problems, but also reduces to the asymmetric Laplace in the case of a single quantile.

2.1.2 Jeffreys’ substitution likelihood and exponential tilted empirical likelihood

Semiparametric methods for quantile estimation include Jeffreys’ substitution likelihood (jsl) for quantiles (Jeffreys, 1961) and (Bayesian) exponential tilted empirical likelihood (betel). These two pseudo-likelihoods are very similar operationally; inference under the two methods, even for small sample sizes, are often indistinguishable (Lancaster and Jun, 2010). Defining $k(\theta) = \sum_i^n \mathbb{1}(y_i < \theta)$, Jeffreys substitution likelihood takes the form

$$\text{jsl}(\theta) = \binom{n}{k(\theta)} q^{k(\theta)} (1 - q)^{n-k(\theta)}, \quad (5)$$

while the tilted empirical likelihood takes the form

$$\text{betel}(y_i | \theta) = \mathbb{1}(y_i < \theta) \frac{q}{k(\theta)} + \mathbb{1}(y_i > \theta) \frac{1 - q}{n - k(\theta)}. \quad (6)$$

Posterior inference in either case follows proportionality analogous (but not identical, see Monahan and Boos (1992)) to an application of Bayes’ rule.
Jeffreys pseudo-likelihood was first proposed in Jeffreys (1961) for the single quantile case. Lavine (1995) demonstrates how to extend the method to a vector of quantiles and studies an asymptotic property (termed conservativeness) of its inferences. Dunson and Taylor (2005) extend the method to a linear regression.

The exponential tilted empirical likelihood has forerunners in the more general model of Schennach (2005) and the earlier Bayesian bootstrap (Rubin, 1981). It was first studied closely for quantile regression in Lancaster and Jun (2010) where it is derived as a maximum entropy density supported on the observed data. A similar argument, but over a finite interval, can be used to motivate the mesa approximation density.

Strengths of these semi-parametric approaches include computational simplicity and consistency, which again follows from a correspondence between empirical quantiles and the associated maximum likelihood estimates. Unlike the asymmetric Laplace approach, this approach is trivially applied to simultaneous multiple quantiles, for which it is still consistent.

However, extension to the regression setting proves unsatisfactory because regression lines are given posterior weight under jsl(θ) and betel(θ) based only on the number of data points above and below the line, irrespective of how tightly these lines trap the points, negating the benefit of a likelihood based approach.

Dunson and Taylor (2005) sidestep this problem by focusing on categorical predictors and employing a separate Jeffreys likelihood at each predictor value. For continuous predictors the method would need modification to incorporate some manner of binning of the observations, within which the Jeffreys likelihood (or exponential tilted empirical likelihood) could be applied.

The mesa likelihood remedies this problem naturally by defining a proper density with appropriate normalization.
2.1.3 Nonparametric Bayesian conditional density estimation

Fully nonparametric approaches are consistent for all (uncountably many) quantiles as long as the true density is within the support of the nonparametric prior and the prior places “enough” mass within a Kullback-Leibler neighborhood. As always, prior selection in nonparametric models is especially important for small sample performance; several options are described in the literature.

Taddy and Kottas (2010) specify and estimate a nonparametric joint predictor-response model from which the conditional density for any desired quantile may be deduced. This approach is most appropriate when the predictors are stochastic (as in observational data), but unsuitable for cases demanding a true design matrix, as in experimental settings for causal inference. This method essentially precludes using more than a handful of predictors because reliable joint density estimation for more than ten or so variables requires tremendous amounts of data. In the favorable case, when such vast sample sizes are available, the computational demands of Bayesian joint density estimation rapidly become prohibitive.

Alternatively, one can work directly with a conditional density specification as in Villani et al. (2009) and Dunson et al. (2007). Villani et al. (2009) use a mixture model where the parameters of the mixture components follow a Gaussian process. Dunson et al. (2007) use a generalization of
a Dirichlet process mixture model which leverages the stick-breaking representation of Sethuraman (1994) to induce predictor dependence in the mixture component probabilities. In general these approaches involve fewer parameters than the approach of Taddy and Kottas (2010) because they avoid working with a full joint distribution. In this setting, the primary practical difficulty is inducing the desired prior on quantile regression functions via the priors on $f(y \mid x)$. By using the mesa pseudo-density one has more direct prior control over the conditional quantile functions specified by the vector $q$.

Finally, Tokdar and Kadane (2011) consider simultaneous multiple linear quantile regression using a nonparametric approach. This is an important special case for interpretability reasons, but the mesa approximation with nonlinear conditional quantile functions may be preferred in more exploratory contexts.

3 Log-Gaussian process priors for nonlinear conditional quantiles

Applying the mesa approximation density for regression requires expressing the $k$ elements of $\theta$ as (unknown) functions of some $d$-dimensional predictor vector $x$:

$$\theta_j(x) = f_j(x), \quad j = 1, \ldots, k. \quad (7)$$

One must ensure the necessary inequalities of the quantile functions: $f_j(x) < f_{j'}(x)$ for $j < j'$ across all $x$. The approach taken here is to describe the various conditional quantile functions in terms of their offset from the median function, denoted $\theta_\mu$, so that

$$\theta_j(x) = \theta_\mu(x) + s_j \sum_{l \in J^*} \Delta_l(x). \quad (8)$$
Here, \( s_j = -1 \) if \( q_j < 0.5 \) and \( s_j = 1 \) if \( q_j > 0.5 \), \( J^* = \{ \min(j, j^*), \ldots, \max(j, j^*) \} \) and \( j^* \) is the value of \( l \) minimizing \( s_j(q_l - 0.5) \). This satisfies the order requirements as long as \( \Delta_j(x) \) a non-negative function, leaving one free to model \( \log(\Delta_j(x_i)) \) without any further restrictions. A convenient choice is to use a zero mean Gaussian process

\[
\log(\Delta_j) \sim \mathcal{GP}(0, \Sigma_x),
\]

where \( \Sigma_x \) is a covariance function depending on \( x \).

In summary, given covariates \( x_i \), each \( y_i \) is assumed to be drawn from the mesa distribution (1) with parameter \( \theta(x_i) \) defined as in (8); priors over \( \theta(x_i) \) are defined via Gaussian process priors over \( \theta_\mu(x_i) \) and \( \log(\Delta_j(x_i)) \) independently for each \( j \).

Computing posterior distributions under this model can be accomplished using a random-walk Metropolis-Hastings algorithm because the specification involves independent unrestricted components and evaluation of the likelihood function is straightforward. Careful indexing of the data partitions affected by \( \theta \) — the matrix of the \( k + 1 \) conditional quantile evaluations (including the median) over the \( n \) data points — can ease the computational burden, as the most intensive step is determining which region of the density to use to evaluate each point. While this issue is not explored here, efficient use of balanced search trees are key to scalability in \( n \), the number of observations. Predictor variables are rescaled to the unit cube. Note that hyperparameters of the Gaussian process covariance may be fixed to pre-determined values and the response variable rescaled to achieve plausible a priori fits (post-processing returns inferences to the original scale).

*Example: a single regressor.* In the single regressor setting, it is straightforward to plot mean
posterior quantile curves. Let

\[ y_i \mid x_i \sim N \left( -2.5 + 5x_i, \{5(x_i - 0.5)^2 + .1\}^2 \right), \]  \hspace{1cm} (10)

for \( x_i \) uniformly distributed on the unit interval. The mesa likelihood with Gaussian process priors can be fit to data from this distribution, using an exponential covariance function with range parameter set to 0.3 (recall that the predictor variable lives on the unit interval). Details of the sampling algorithm are given in Section 5, but note that because the predictors are scaled to the unit interval, the range parameter is straightforward to elicit qualitatively. The estimated quantile curves for \( q = (0.01, 0.1, 0.35, 0.5, 0.85, 0.95, 0.99) \) are shown in Figure 3.

![Figure 3: The left panel shows posterior estimates (mean) of quantile regression lines (median solid) superimposed over a scatterplot of the observed data (n = 500). Here q = (0.01, 0.1, 0.35, 0.5, 0.85, 0.95, 0.99). The right panel shows the corresponding true quantile curves.](image)

Example: two regressor case. Similarly, in two dimensions one can plot the posterior quantile surface. Let

\[ y_i \mid x_i \sim N \left( -2.5 + 5x_{1,i}, \{5(x_{2,i} - 0.5)^2 + .1\}^2 \right), \]  \hspace{1cm} (11)

for \( x_i = (x_{1,i}, x_{2,i}) \) uniformly distribute on the unit square. The estimated quantile curves for
$q = (0.01, 0.1, 0.35, 0.5, 0.85, 0.95, 0.99)$ were computed. As in the one-dimensional case, when the predictor variables are uniformly distributed, the likelihood favors surfaces which entrap the appropriate fraction of the observations as tightly as possible; the Gaussian process prior on the $\log(\Delta_j)$ functions favors smooth surfaces.

In dimensions greater than two, the quantile regression hyper-surfaces are impossible to visualize. Moreover, if irrelevant predictor variables are included in the model, the variance of the resulting Gaussian process can be difficult to calibrate. We present one approach to overcoming these difficulties in the next section.

4 Single-index multiple quantile regression

A Gaussian process single-index model, or GP-SIM, (Gramacy and Lian, 2011 to appear) has a covariance function that can be expressed in terms of a one-dimensional projection of the predictor variables so that $\Sigma_X \equiv \Sigma_\tau$ with $\tau = X^t w$, where $w$ is a $d$-element vector of weights. Here, we
advocate rescaling $\tau$ to span the unit interval:

$$
\tau_i = \frac{x_i^t w - \min(X^t w)}{\max(X^t w - \min(X^t w))}.
$$

(12)

This rescaling permits efficient computation, as described in Section 5. For identifiability, we set $w_1 = 1$; note that this strategy implies that the first predictor necessarily plays a non-zero role in defining the elements of $\tau$.

Regarding the prior, we note that it is often desirable to encourage sparsity of $w$, meaning that we favor zero elements a priori. Here we take a shrinkage approach by using a prior over the weights which has an infinite spike at zero and fatter-than-exponential tails, specifically $\pi(w_g) \propto \log (1 + 2w_g^{-2})$. This prior is motivated by the work in Carvalho et al. (2010), where it serves as an analytic upper bound. Regarding posterior summaries, note that the rescaling of $\tau$ in (12) makes the gross numbers arbitrary; to remedy this, we suggest looking at feature importance factors, which we define as

$$
\beta_g = \frac{|w_g|}{\sum_{l=1}^d |w_l|},
$$

(13)

for $g = 1, \ldots, d$. This metric measures the fractional contribution of each predictor towards defining the elements of $\tau$.

Example: single-index model with two predictors, one relevant and one irrelevant. Consider fitting a single index model to data generated from the model as in (10). Assume that the true $x_1$ was included, but that additionally a variable $x_2$ was included which is sampled uniformly from the unit interval, independently of $x_1$ and $y$. Scatterplots are shown in Figure 5 along with marginal posteriors (histograms) of $\beta_1$.

Perhaps more realistically, one can repeat this exercise using the transformed predictor variables
Figure 5: The first two scatterplots simply show the observed response variable against each of the two predictors included in the model. Notice that the second predictor is independent of the response. The third panel shows a histogram of posterior draws for $\beta_1$, reflecting the fact that $x_1$ alone determines the density of the response. Note that $w_2 = 1 - w_1$ in this case.

$$\tilde{x}_{1,j} = \frac{x_{1,j}}{\sqrt{2}} + \frac{x_{2,j}}{\sqrt{2}} \quad \text{and} \quad \tilde{x}_{2,j} = \frac{x_{1,j}}{\sqrt{2}} - \frac{x_{2,j}}{\sqrt{2}}.$$  

This reflects a situation in which each of the predictors plays some role in the conditional distribution of the response. Let $\tilde{w}$ denote the weight variable in this alternatively defined model; the posterior histogram of $\tilde{\beta}_1$ shown in figure 6 reflects the model successfully learning the correct convex combination to recover $x_1$.

Figure 6: The histogram of posterior draws for $\beta_1$, the importance factor of variable $\tilde{x}_1$ as defined above. The model accurately adjusts for the fact that the true $\tau$ involves both $\tilde{x}_1$ and $\tilde{x}_2$. 
The full model may be written compactly as

\[ y_i \mid \Delta_1, \ldots, \Delta_k, \theta_{\mu} \sim \text{mesa}(\Delta_1, \ldots, \Delta_k, \theta_{\mu}), \]

\[ \theta_{\mu} \mid \tau(X, w) \sim \mathcal{GP}(\Sigma_\tau), \]

\[ \log(\Delta_j) \sim \mathcal{GP}(\Sigma_\tau), \]

\[ \pi(w_g) \propto \log(1 + 2w_g^{-2}), \quad g = 2, \ldots, d, \]

where \( \Sigma_{i,i'} = \phi(\tau_i, \tau_{i'}) \) for a positive definite covariance function. Here we use the exponential covariance function \( \phi(\tau_i, \tau_{i'}) = v \exp(-\nu(\tau_i - \tau_{i'})^2) \) for fixed scale parameter \( v \) and bandwidth parameter \( \nu \).

\section{5 Computation}

Our computational approach offers leverages the uni-dimensionality of the single index to use a computationally efficient convolution approach to evaluating the Gaussian process prior. We work with a latent variable representation using a basis function expansion of \( \Sigma_\tau \). In particular for \( i = 1, \ldots, n \) we write

\[ \Delta_j = \exp(\Lambda_\tau \alpha_j), \quad j = 1, \ldots, k, \]

\[ \lambda_{ih} = v \exp(-\nu(\tau_i - t_h)^2), \quad h = 1, \ldots, m, \]

with \( \tau_i \) defined as in (12). In this representation \( \alpha_j \) is a \( m \)-by-1 matrix of latent regression coefficients for each \( j \) and \( t_h \) for \( t = 1, \ldots, m \) is a pre-fixed “knot” on the single-index scale where \( m \) is the number of knots used. By letting

\[ \alpha_{ih} \sim N(0, 1) \]
we recover (9) on marginalizing $\alpha_j$, up to approximation due to using $m < \infty$ knots (see chapter 4 of Ferreira and Lee (2007) for details). If it is determined that the approximation is acceptable with $m < n$, then evaluating $\Lambda$ can be done with fewer operations than evaluating $\Sigma_\tau$. The price one pays for this gain is having to handle the $m \times k$ latent regressors $\alpha_{hj}$. Crucially, however, this is a one-time fee, not growing with sample size.

We likewise use this latent regressor representation to express the Gaussian process median function

$$\theta_h = \Lambda_\tau \gamma,$$

$$\gamma_h \sim N(0,1).$$

Our overall sampling strategy is a random walk Metropolis-within-Gibbs approach, cycling through the $m$ rows of $\alpha$ and $\gamma$ given the other rows; to improve mixing we sample a new $w$ at each step:

sample $(w, \alpha^t_h, \gamma_h, | \alpha_{-h}, \gamma_{-h}, y, X)$ for each $h = 1, \ldots, m,$

where $\alpha^t_h$ is the $k$ dimensional column vector contain the $h$th element of $\alpha_j$ across all $j = 1, \ldots, k$ and $\alpha_{-h}$ and $\gamma_{-h}$ denote the corresponding matrix (resp. vector) with the $h$th column (resp. element) removed.

Resampling $w$ with each draw of $\alpha_{jh}$ is inefficient in the sense that we sample the vector $m$ times for each kept sample, but this redundancy allows us to explore the full joint distribution more readily, as the posterior distribution of $\alpha$ is apt to change substantially across different index weights. Thus at each step in the sampler we iterate through $h$, proposing new values via the
identities

\[ \tilde{\gamma}_h = \gamma_h + \sigma_{\gamma_h} \epsilon_{\gamma_h} \]
\[ \tilde{\alpha}_h = \alpha_h + L_h \epsilon_{\alpha_h}, \]
\[ \tilde{\eta}_g = \eta_g + \sigma_{\eta_g} \epsilon_{\eta_g} \quad \text{for } g = 1, \ldots, d, \]

where \( \epsilon_{\gamma_h}, \epsilon_{\alpha_h} \) and \( \epsilon_{\eta_g} \) are drawn as standard normal random variables. The \( k \times k \) matrices \( L_h \) are tuning parameters introduced to account for the fact that the elements of \( \alpha_h \) will likely exhibit negative correlation, due to the “nested” construction of the quantile functions in terms of the \( \Delta_j \); in practice we find that a single matrix \( L \) works well across all \( h \). Likewise the \( \sigma_{\gamma_h} \) and \( \sigma_{\eta_g} \) are tunable step-size parameters. As these relations are symmetric, the acceptance probability is given simply as

\[ \min \left( 1, \frac{f_q(y \mid X, \tilde{w}, \tilde{\alpha}, \tilde{\gamma})}{f_q(y \mid X, w, \alpha, \gamma)} \cdot \frac{\pi(\tilde{w})\pi(\tilde{\alpha})\pi(\tilde{\gamma})}{\pi(w)\pi(\alpha)\pi(\gamma)} \right), \]

where \( f_q(y) \) is as in equation 1, \( \pi(w) \) denotes the density of the logarithmic prior given in (14), and \( \pi(\gamma) \) and \( \pi(\alpha) \) denote independent standard normal densities.

6 Applied demonstration: moral hazard in management

In this section we demonstrate our methodology on the data of Yafeh and Yosha (2003), which is used to investigate the relationship between shareholder concentration and managerial expenditures with scope for private benefit. These data come from 185 Japanese industrial chemical firms listed on the Tokyo stock exchange. Following Taddy and Kottas (2010) we consider the model with response variable \( y \), the sales-deflated managerial and administrative expenses (\( \text{MH5} \) in the original paper), regressed upon predictor variables \textit{Leverage} (debt to total asset ratio), \( \log(\text{Assets}) \), the \textit{Age} of the firm, and \textit{TOPTEN}, percentage of ownership held by the ten largest shareholders.
Our posterior inference is neatly summarized by two illustrations. First, we look at the point-wise average of the $\tau$-$y$ scatterplot with the point-wise average conditional quantile curves superimposed. This plot provides an at-a-glance description of how the conditional quantile varies as one varies the composite index.

Figure 7: The point-wise average $\tau$-$y$ scatterplot with the point-wise average 10, 25, 50, 75 and 90th percentile conditional quantile curves overlayed. Compare to Figure 2 in Taddy and Kottas (2010).

The upshot of this plot is that the upper quantiles are much more variable with respect to the single index than are the lower quantiles and that the relationship is decreasing in $\tau$: companies with higher $\tau$ have less extreme (high) expenditures. When we turn to the interpretation of $\tau$, we find that this is consistent with companies that have more capacity for oversight (as measured by $TOPTEN$) and more incentive for oversight (as measured by $Leverage$) in fact exhibiting less extreme expenditure.

Second, to get a finer-grained look at the impact of the four company features we produce a pair-wise scatterplots of the feature importance factors and their marginal posterior histograms (Figure 8). This shows which variables appear to be driving the conditional quantile structure
relative to the others and provide an interpretation of the single index \( \tau \). Consistent with previous analyses we find that \textit{Leverage} and \textit{TOPTEN} are the most relevant variables, with posterior mean importance of 0.42 and 0.34 respectively, while \( \log(\text{Assets}) \) is three times less relevant and \textit{Age} scarcely plays any role at all (see Table 1).

Table 1: Posterior means and high probability density intervals of the feature importance factors.

<table>
<thead>
<tr>
<th>Feature</th>
<th>2.5%</th>
<th>Mean</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \log(\text{Assets}) )</td>
<td>0.12</td>
<td>0.15</td>
<td>0.22</td>
</tr>
<tr>
<td>\textit{Age}</td>
<td>0.05</td>
<td>0.09</td>
<td>0.13</td>
</tr>
<tr>
<td>\textit{Leverage}</td>
<td>0.34</td>
<td>0.42</td>
<td>0.48</td>
</tr>
<tr>
<td>\textit{TOPTEN}</td>
<td>0.29</td>
<td>0.34</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Figure 8: Posterior draws of the feature importance factors. We find that \textit{TOPTEN} and \textit{Leverage} have similar importance while \( \log(\text{Assets}) \) is less important and \textit{Age} is less important yet.
7 Conclusion

The mesa approximation likelihood generalizes the popular asymmetric Laplace model to allow coherent multiple quantile estimation and it retains the flavor of exponential tilted empirical likelihood while being suitable for regression. Meanwhile it is computationally and conceptually simpler than full density regression approaches while still permitting flexible nonlinear regression.

The mesa distribution provides a mechanism for efficiently inferring a scaffolding of nonlinear functions which parsimoniously characterize an observed data set. If full density regression is hamstrung due to a paucity of data relative to the dimensionality of the predictor space, an approximate density regression based on a handful of quantile functions may represent a good compromise, and the single-index approach combined with the mesa approximation likelihood represents an ideal first choice.

Finally, using the mesa density in conjunction with a of single-index log-Gaussian process priors permits the model to address the question ‘which predictor variables are driving conditional quantile functions?’ This question is very natural from a practical data analysis standpoint but is one for which earlier methods do not provide ready posterior summaries. In our model the posterior distribution of weights $w$ convey this information succinctly and, in combination with the marginal point-wise quantile curve plot, provide useful qualitative information not necessarily obvious from standard data summaries or model fits.

References


D. B. Dunson and J. Taylor. Approximate Bayesian inference for quantiles. *Journal of Nonpara-


