ADDITIVE MODELS FOR QUANTILE REGRESSION: MODEL SELECTION AND CONFIDENCE BANDAIDS

ROGER KOENKER

Abstract. Additive models for conditional quantile functions provide an attractive framework for non-parametric regression applications focused on features of the response beyond its central tendency. Total variation roughness penalities can be used to control the smoothness of the additive components much as squared Sobelev penalties are used for classical L_2 smoothing splines. We describe a general approach to estimation and inference for additive models of this type. We focus attention primarily on selection of smoothing parameters and on the construction of confidence bands for the nonparametric components. Both pointwise and uniform confidence bands are introduced; the uniform bands are based on the Hotelling (1939) tube approach. Some simulation evidence is presented to evaluate finite sample performance and the methods are also illustrated with an application to modeling childhood malnutrition in India.

1. Introduction

Models with additive nonparametric effects offer a valuable dimension reduction device throughout applied statistics. In this paper we describe some new estimation and inference methods for additive quantile regression models. The methods employ the total variation smoothing penalties introduced in Koenker, Ng, and Portnoy (1994) for univariate components and Koenker and Mizera (2004) for bivariate components. We focus on selection of smoothing parameters including lasso-type selection of parametric components, and on post selection inference methods, particularly confidence bands for nonparametric components of the model.

The motivation for these developments arose from an effort to compare the performance of additive modeling using penalty methods with the quantile regression boosting approach recently suggested by Fenske, Kneib, and Hothorn (2008) in a study of risk factors for early childhood malnutrition in India. Their study, based on the international Demographic and Health Survey (DHS) sponsored by USAID, explored models for children's height as an indicator of nutritional status. Interest naturally focuses on the lower conditional quantiles of height. The 2005-6 DHS survey of India involved roughly 37,000 children between the ages of zero to five. As detailed below in Table 4, there were a large number of potentially important discrete covariates: educational status of the parents, economic status of the household, mother's religion, ownership of various consumer durables as well as six or

Version: May 27, 2010. This research was partially supported by NSF grant SES-08-50060. I would like thank Torsten Hothorn and his colleagues for rekindling my interest in additive quantile regression models and for their help with the DHS data, thanks too to Victor Chernozhukov for very helpful suggestions on λ selection.

seven other covariates for which an additive nonparametric component was thought to be desirable. The application thus posed some serious computational and methodological challenges.

Rather than estimate models for mean height, or resort to binary response modeling for whether height exceeds an age-specific threshold, Fenske, Kneib, and Hothorn (2008) suggested estimating conditional quantile models for the first decile, 0.10, of heights. Boosting provides a natural approach to model selection in this context, and given the relatively large sample size also offers some computational advantages. So it became a challenge to see whether the additive modeling strategies described in Koenker (2005) could be adapted to problems of this scale and complexity. Initial forays into estimation of these models indicated that computational feasibility wasn't really an issue. In effect estimation of such models involves solving a fairly large but very sparse linear programming problem, a task for which modern interior point methods employing advances in sparse linear algebra are well-suited. Even though a typical model might have nearly 40,000 observations after augmentation by the penalty terms and 2200 parameters, estimation required only 5-10 seconds for (reasonable) fixed values of the smoothing penalty parameters. What loomed more ominously over the horizon was the problem of smoothing parameter selection, and ultimately the problem of evaluating the precision of estimated components after model selection.

We will begin by briefly describing the class of penalized additive models to be considered in the next section. Section 3 describes a general approach to selection of smoothing parameters. Sections 4 and 5 describe a general approach to constructing confidence bands, pointwise and uniform bands respectively, for the additive nonparametric components of the model. Section 6 reports some simulation evidence on model selection methods and confidence band performance. Section 7 returns to our motivating application and explores estimation and inference for malnutrition risk factors. Some comparisons with the boosting results of Fenske, Kneib, and Hothorn (2008) are made at the end of that section.

2. Additive Models for Quantile Regression

Additive models have received considerable attention since their introduction by Breiman and Friedman (1985) and Hastie and Tibshirani (1986, 1990). They provide a pragmatic approach to nonparametric regression modeling; by restricting nonparametric components to be composed of low-dimensional additive pieces we can circumvent some of the worst aspects of the notorious curse of dimensionality. It should be emphasized that we use the word "circumvent' advisedly, in full recognition that we have only swept difficulties under the rug by the assumption of additivity. When conditions for additivity are violated there will obviously be costs.

Our approach to additive models for quantile regression and especially our implementation of methods in `R` has been heavily influenced by Wood (2006, 2009) . In some fundamental respects the approaches are quite distinct: Gaussian likelihood is replaced by (Laplacean) quantile fidelity, squared $\mathcal{L}_2$ norms as measures of the roughness of fitted functions are replaced by corresponding $\mathcal{L}_1$ norms measuring total variation, and truncated basis expansions are supplanted by sparse algebra as a computational expedient. But in many other respects the structure of the models is quite similar. We will consider models for conditional quantiles indexed by $\tau \in (0, 1)$ of the general form:

$$Q_{Y_i|x_i,z_i}(\tau|x_i,z_i) = x_i^\top \theta_0 + \sum_{j=1}^{J} g_j(z_{ij}). \tag{1}$$

The nonparametric components g_j will be assumed to be continuous functions, either univariate, $\mathcal{R} \to \mathcal{R}$, or bivariate, $\mathcal{R}^2 \to \mathcal{R}$. We will denote the vector of these functions as $g = (g_1, \ldots, g_J)$. Our task is to estimate these functions together with the Euclidean parameter $\theta_0 \in \mathcal{R}^{p_0}$, by solving

$$\min_{(\theta_0,g)} \sum \rho_\tau(y_i - x_i^\top \theta_0 - \sum g_j(z_{ij})) + \lambda_0 \|\theta_0\|_1 + \sum_{j=1}^{J} \lambda_j \bigvee(\nabla g_j) \tag{2}$$

where $\rho_\tau(u) = u(\tau - I(u < 0)$ is the usual quantile objective function, $\|\theta_0\|_1 = \sum_{k=1}^{K} |\theta_{0k}|$ and $\bigvee(\nabla g_j)$ denotes the total variation of the derivative or gradient of the function g. Recall that for g with absolutely continuous derivative g' we can express the total variation of $g' : \mathcal{R} \to \mathcal{R}$ as

$$\bigvee(g'(z)) = \int |g''(z)| dz$$

while for $g : \mathcal{R}^2 \to \mathcal{R}$ with absolutely continuous gradient,

$$\bigvee(\nabla g) = \int \|\nabla^2 g(z)\| dz$$

where $\nabla^2 g(z)$ denotes the Hessian of g, and $\|\cdot\|$ will denote the usual Hilbert-Schmidt norm for matrices.

There is an extensive literature in image processing on the use of total variation smoothing penalties, initiated by Rudin, Osher, and Fatemi (1992). Edge-detection is an important consideration in imaging and total variation penalization permits sharp breaks in gradients that would be prohibited by conventional Sobolev penalties. Koenker and Mizera (2004) discuss the bivariate version of the total variation roughness penalty in greater detail and offer further motivation and references for it. In the univariate setting, $g : \mathbb{R} \to \mathbb{R}$, total variation penalties were suggested in Koenker, Ng, and Portnoy (1994) as computational convenient smoothing device for nonparametric quantile regression. Total variation penalties also underlie the taut-string methods of Davies and Kovac (2001), and the fused lasso methods of Tibshirani, Saunders, Rosset, Zhu, and Knight (2005), although both approaches focus primarily on penalization of the total variation of the function itself rather than its derivative.

Solutions to the variational problem (2) are piecewise linear with knots at the observed z_i in the univariate case, and piecewise linear on a triangulation of the observed z_i's in the bivariate case. This characterization greatly simplifies the computations required to solve (2), which can therefore be written as a linear program with (typically) a very sparse constraint matrix consisting mostly of zeros. This sparsity greatly facilitates efficient solution of the resulting problem, as described in Koenker and Ng (2005). Such problems are efficiently solved by modern interior point methods for linear programming. Backfitting is not required.

3. Model Selection

A challenging task for any regularization problem like (2) is the choice of the λ parameters. When, as in our application, there are several of these λ's then the problem is especially daunting. Following a proposal of Machado (1993) for parametric quantile regression, adapted to total variation penalized quantile regression by Koenker, Ng and Portnoy, we have relied upon the Schwarz (1978) like criterion

$$\mathrm{SIC}(\lambda) = n \log \hat{\sigma}(\lambda) + \tfrac{1}{2} p(\lambda) \log(n)$$

where $\hat{\sigma}(\lambda) = n^{-1} \sum_{i=1}^{n} \rho_\tau(y_i - \hat{g}(x,z))$, and $p(\lambda)$ is the effective dimension of the fitted model

$$\hat{g}(x,z) = x^\top \hat{\theta}_0 + \sum_{j=1}^{J} \hat{g}_j(z).$$

The quantity $p(\lambda)$ is usually defined for linear estimators in terms of the trace of a pseudo projection matrix, the matrix mapping observed response into fitted values. The situation is somewhat similar for quantile regression fitting except that we simply compute the number of zero residuals for the fitted model to obtain $p(\lambda)$. Recall that in unpenalized quantile regression fitting a p-parameter model yields precisely p zero residuals provided that the y_i's are in general position. This definition of $p(\lambda)$ can be viewed from a more unified perspective as consistent with the definition proposed by Meyer and Woodroofe (2000),

$$p(\lambda) = \mathrm{div}(\hat{g}) = \sum_{i=1}^{n} \frac{\partial \hat{g}(x_i, z_i)}{\partial y_i},$$

see Koenker (2005, p.243). A consequence of this approach to characterizing model dimension is that it is essential to avoid "tied" responses; we ensure this by "dithering" the response variable.

Optimizing $\mathrm{SIC}(\lambda)$ is still a difficult task made more challenging by the fact that the objective function is discontinuous. As any of the λ's increase so the regularization becomes more severe new constraints become binding and initially free parameters vanish from the model, thereby reducing the effective dimension of the model. When there are several λ's a prudent strategy would seem to be to explore informally, trying to narrow the region of optimization and then resort to some form of global optimizer to narrow the selection. In our applications we have relied on the R functions `optimize` for cases in which there is a single λ, and the simulated annealing option of `optim` when there are several λ's in play.

4. Pointwise Confidence Bands

Confidence bands for nonparametric regression introduce some new challenges. As with any shrinkage type estimation method there are immediate questions of bias. How do we ensure that the bands are centered properly? Bayesian interpretation of the bands as pioneered by Wahba (1983) and Nychka (1983) provides some shelter from these doubts. For our additive quantile regression models we have adopted a variant of the Nychka approach as implemented by Wood in the `mgcv` package.

As in any quantile regression inference problem we need to account for potential heterogeneity of the conditional density of the response. We do this by adopting Powell's (1991) proposal to estimate local conditional densities with a simple Gaussian kernel method.

The pseudo design matrix incorporating both the lasso and total variation smoothing penalties can be written as,

$$\tilde{X} = \begin{bmatrix} X_0 & X_1 & \cdots & X_J \\ \lambda_0 H_K & 0 & \cdots & 0 \\ 0 & \lambda_1 P_1 & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_J P_J \end{bmatrix}.$$

Here X_0 denotes the matrix representing the parametric covariate effects, the X_j's represent the basis expansion of the g_j functions, $H_K = [0 \vdots I_K]$ is the penalty contributions from the lasso excluding any penalty on the intercept and the P_j terms represent the contribution from the penalty terms on each of the smoothed components. The covariance matrix for the vector of estimates $\hat{\theta}$ of the full set of parameters, $\theta = (\theta_0^\top, \theta_1^\top, \cdots, \theta_J^\top)^\top$ is given by the sandwich formula,

$$V = \tau(1-\tau)(\tilde{X}^\top \Psi \tilde{X})^{-1}(\tilde{X}^\top \tilde{X})^{-1}(\tilde{X}^\top \Psi \tilde{X})^{-1} \tag{4.1}$$

where Ψ denotes a diagonal matrix with the first n elements given by the local density estimates,

$$\hat{f}_i = \phi(\hat{u}_i/h_n)/h_n$$

$\hat{u}_i$ is the ith residual from the fitted model, ϕ is the standard Gaussian density, and h is a bandwidth determined by one of the usual built-in rules. See Koenker (2005) Section 3.4 for further details. The remaining elements of the Ψ diagonal corresponding to the penalty terms are set to one.

Pointwise confidence bands can be easily constructed given this matrix V. A matrix G_j representing the prediction of g_j at some specified plotting points $z_{ij} : i = 1, \cdots, m$ is first made, so the m-vector with typical element, $\hat{g}_j(z_{ij})$, can be expressed as $G_j\hat{\theta}_j$ where $\hat{\theta}_j$ is the subvector of estimated coefficients of the fitted model pertaining to $\hat{g}_j$. We then extract the corresponding diagonal block, V_j of the matrix V, and compute the estimated covariance matrix, $V(G_j\hat{\theta}_j) = G_j V_j G_j^\top$, and finally we extract the square root of its diagonal. The only slight complication of this process is to recall that the intercept of the estimated model needs to be appended to each such prediction and properly accounted for in the extraction of the covariance matrix of the predictions. If this is not done then the variance at the lower support point of the fitted function degenerates to zero.

An obvious criticism of this pointwise approach to constructing confidence bands is that one may prefer to have uniform bands. This topic has received considerable attention in recent years; there are several possible approaches including resampling. Recent work by Krivobokova, Kneib, and Claeskens (2010) has shown how to adapt the early work of Hotelling (1939) to some GAM models. Similar methods can be adapted to additive quantile regression models, an approach that will be described in the next section.

5. Uniform Confidence Bands

Uniform confidence bands for nonparametric regression estimation impose a stronger probabilistic burden than the pointwise construction described above. We now require a band of the form,

$$B_n(x) = (\hat{g}_n(x) - c_\alpha \hat{\sigma}_n(x), \hat{g}_n(x) + c_\alpha \hat{\sigma}_n(x))$$

such that the random band B_n covers the true curve $\{g_0(x) : x \in \mathcal{X}\}$ with specified probability $1 - \alpha$, over a given domain, $\mathcal{X}$, for g

$$\mathbb{P}\{g_0(x) \in B_n(x) | x \in \mathcal{X}\} = 1 - \alpha.$$

Our construction will employ the same $\hat{\sigma}_n(x)$ local scale estimate described earlier, however c_α will need to change.

5.1. **Uniform Bands for Series Estimators.** For the sake of completeness we will begin by sketching some theoretical underpinnings of the Hotelling tube approach in the simplest Gaussian non-parametric setting for a series estimator, following the exposition of Johansen and Johnstone (1990). The key insight of Hotelling (1939) was the realization that the computation of the relevant rejection probability for band construction could be reduced to finding the volume of a tubular region embedded in a sphere. Subsequent work by Weyl (1939) and Naiman (1986) have generalized this approach to more general manifolds; initially we will focus on the classical Gaussian nonparametric settings.

Consider estimating the model,

$$y_i = g_0(x_i) + u_i, \tag{5.2}$$

with u_i iid $\mathcal{N}(0, \sigma^2)$. We adopt a series estimator of the form,

$$\hat{g}(x) = \text{argmin}_\theta \sum_{i=1}^n (y_i - \langle b(x_i), \theta \rangle)^2$$

that is we consider estimators from the set,

$$\mathcal{G} = \{g : g(x) = \langle b(x), \theta \rangle\},$$

where $b(x)$ denotes a vector, $(b_1(x), \cdots, b_p(x))^\top$ of basis functions for the series expansion. The likelihood ratio statistic for testing $H_0 : \theta = 0$ against a general alternative is based on the statistic

$$\mathcal{L} = \inf_{x \in \mathcal{X}} \sum_{i=1}^n (y_i - \langle b(x_i), \hat{\theta} \rangle)^2 / \sum_{i=1}^n Y_i^2.$$

Letting B denote the matrix with i row $(b_j(x_i))$, we have $\hat{\theta} = (B^\top B)^{-1} B^\top y$ and we will write,

$$\hat{g}(x) = \langle b(x)^\top (B^\top B)^{-1} B^\top, y \rangle \equiv \langle \ell(x), y \rangle.$$

The pointwise standard error of $\hat{g}(x)$ is,

$$\sigma(x) = \sqrt{\sigma^2 b(x)^\top (B^\top B)^{-1} b(x)}$$

so we want to consider test statistics of the form,

$$T_n = \sup_{x \in \mathcal{X}} \frac{\hat{g}(x) - g_0(x))}{\sigma(x)}.$$

Given the null distribution of T_n we can obtain a confidence set,

$$\mathcal{C} = \{g_0 \mid T_n < c_\alpha\}.$$

Following Johansen and Johnstone (1990), we can write $T_n = RW$ where

$$R^2 = (\hat{g}(x) - g_0(x))^2/\sigma^2(x) \sim \chi^2_p,$$

and letting $D = (B^\top B)^{-1}$,

$$\begin{aligned} W &= \sup_{x\in\mathcal{X}} \frac{\hat{g}(x) - g_0(x))}{\sigma(x)R} \\ &= \sup_{x\in\mathcal{X}} \frac{(D^{1/2}b(x))^\top D^{-1/2}(\hat{\theta} - \theta_0)}{\|D^{1/2}b(x)\|\ \|D^{-1/2}(\hat{\theta} - \theta_0)\|} \\ &\equiv \sup_{x\in\mathcal{X}} \gamma(x)\cdot U. \end{aligned}$$

Now, $\gamma = \{\gamma(x) : x \in \mathcal{X}\}$ is a curve on the sphere, $\mathcal{S}^{p-1}$, in p dimensions, and U is uniformly distributed on $\mathcal{S}^{p-1}$. The random variables R^2 and W are independent, with $R^2 \sim \chi^2_p$, so

$$\mathbb{P}(T_n > c) = \int_c^\infty \mathbb{P}(W > c/r)\mathbb{P}(R \in dr) \tag{5.3}$$

Hotelling (1939) showed that for non-closed, non-intersecting, curves, γ and w near 1,

$$\mathbb{P}(W > w) = \frac{|\gamma|}{2\pi}(1 - w^2)^{(p-2)/2} + \frac{1}{2}\mathbb{P}(B(1/2, (p-1)/2) \geq w^2) \equiv H_\gamma(w) \tag{5.4}$$

where $|\gamma| = \int_{\mathcal{X}} \|\dot{\gamma}(x)\|dx$ is the length of curve enclosed by the tube and $B(a, b)$ is a beta random variable. Naiman (1986) significantly weakened the conditions on γ showing that $H_\gamma(w)$ is an upper bound for the probability. Naiman bounds (5.3) by

$$\mathbb{P}(T_n > c) \leq \int_c^\infty \min\{H_\gamma(c/r), 1\}\mathbb{P}(R \in dr). \tag{5.5}$$

Relaxing the upper bound constraint of one, Knowles (1987) integrates the simplified version of (5.5) exactly to obtain the bound,

$$\mathbb{P}(T_n > c) \leq \frac{|\gamma|}{2\pi}e^{-c^2/2} + 1 - \Phi(c). \tag{5.6}$$

The foregoing assumes that σ^2 is known; if not, there is the corresponding formula that employs Student t bounds,

$$\mathbb{P}(T_n > c) \leq \frac{|\gamma|}{2\pi}(1 + c^2/\nu)^{-\nu/2} + \mathbb{P}(t_\nu > c), \tag{5.7}$$

where $\nu = n - p$ is the degrees of freedom of the estimated model.

It should be emphasized at this point that in the foregoing homoscedastic Gaussian setting the evaluation of these probability bounds is exact. More generally, we would have to rely on asymptotic approximations to justify the corresponding bands. For example, in settings where the u_i in (5.2) had heteroscedastic structure we might replace $\sigma^2(B^\top B)^{-1}$ in the earlier formulae with an appropriate Eicker-White sandwich. Sun, Loader, and McCormick (2000) consider Hotelling tube methods for generalized linear models employing Edgeworth expansion techniques to improve small sample performance; this does not appear to be a

practical approach for penalized estimators of the type considered here, so we must rely on large sample approximations in the sequel.

5.2. **Uniform Bands for Penalized Series Estimators.** Uniform confidence bands for penalized series estimators can be constructed in much the same way we have just described for the unpenalized case. Krivobokova, Kneib, and Claeskens (2010), drawing on earlier work by Sun (1993) and Sun and Loader (1994) consider generalized additive models with Gaussian penalties like those treated by Wood (2006). Maintaining our focus on the simple univariate model (5.2) we replace our fixed target function, g_0, with a random function $g_0(x) = \langle b(x), \theta \rangle$ with $\theta \sim \mathcal{N}(\theta_0, \Omega)$, yielding the hierarchical (mixed) model,

$$y \sim \mathcal{N}(B\theta_0, \sigma^2 I + B\Omega B^\top),$$

The optimal (BLUP) estimator for θ is now,

$$\hat{\theta} = (B^\top(\sigma^2 I + B\Omega B^\top)^{-1}B)^{-1}B^\top(\sigma^2 I + B\Omega B^\top)^{-1}y$$

and $\hat{g}(x) = \langle b(x), \hat{\theta} \rangle$ has variance,

$$\hat{\sigma}^2(x) = b(x)^\top(B^\top(\sigma^2 I + B\Omega B^\top)^{-1}B)^{-1}b(x).$$

Equivalently, we may consider the penalized estimator,

$$\begin{aligned} \tilde{\theta}(\lambda) &= \operatorname{argmin}\{\sum_{i=1}^{n}(y_i - \langle b(x_i), \theta \rangle)^2 + \lambda\theta^\top\Omega^{-1}\theta \\ &= (B^\top B + \lambda\Omega^{-1})^{-1}B^\top y. \end{aligned}$$

Here λ represents a free scaling parameter for the covariance matrix of θ. Typically, the choice of Ω imposes some form of smoothness on $\hat{g}$, and therefore λ controls the degree of smoothing.

An orthodox Bayesian would, at this point, assign prior distributions for σ^2 or λ, but lacking the courage of our convictions, one can fall back instead on asymptotic justifications for λ selection to rationalize the construction of bands as described above, modified to incorporate the new $\hat{\sigma}(x)$. This approach is closely tied to the uniform confidence band construction for local polynomial regression provided by Loader (2010) for the R package `locfit`. Krivobokova, Kneib, and Claeskens (2010) have recently implemented a version of this approach for a subclass of the GAM models encompassed by the `mgcv` package of Wood (2010). In Section 6 we will report some (limited) simulation experience with this approach and compare it with the bands constructed for total variation penalized quantile regression estimators.

5.3. **Uniform Bands for Penalized Quantile Regression Estimators.** The extension of the foregoing methods to the penalized quantile regression estimators described in Section 1, is quite straightforward. Pointwise confidence bands provide a local standard deviation estimate $\hat{\sigma}_j(z) : j = 1, \cdots, J$ for each of the additive components as described in Section 3. Given the construction of these local scale estimates, we can easily compute the Riemann approximation of the relevant tube length, and inversion of (5.7) yields a critical value, c_α, for the band,

$$\mathcal{C} = \{\hat{g}_j(z) - c_\alpha\hat{\sigma}_j(z), \hat{g}_j(z) + c_\alpha\hat{\sigma}_j(z) | z \in \mathcal{Z}\}.$$

More explicitly, the crucial quantity, $|\gamma|$, representing the length of the curve is computed as follows: let $\hat{g}_j = G_j\hat{\theta}_j$ denote a vector of plotting points, $\hat{g}_j(z_{ij}) : \ i = 1, \cdots, m$, for the estimated function and $\hat{V}_j$ denote the corresponding diagonal block of the estimated asymptotic covariance matrix of $\hat{\theta}$ given in (4.1). After Cholesky factorization of $\hat{V}_j$, we may write $\Xi_j = \hat{V}_j^{1/2} G_j$, a matrix with rows, $\xi_i : \ i = 1, \cdots, m$, and set $\gamma_i = \xi_i / \|\xi_i\|$ for $i = 1, \cdots, m$. Finally, we have the discrete approximation

$$|\gamma| = \int \|\dot{\gamma}(z)\| dz = \sum_{i=2}^{m} \|\gamma_i - \gamma_{i-1}\|.$$

Justification of the distributional properties of the analogues of the R^2 and W variables in this case follows from the asymptotic normality of the $\hat{\theta}_j$'s. This obviously requires conditions that control the selection of λ_j's; Krivobokova, Kneib, and Claeskens (2010) discuss the bias variance tradeoff implicit in this selection and give conditions for the validity of their bands for GAM estimators with estimated λ. The simulation results of the next section, provide some support for the asymptotic validity of this construction of the bands.

6. Some Simulation Evidence

To evaluate finite-sample performance of the confidence bands described above we have undertaken some simulation experiments. The experiments all employ some variant of the model,

$$g_0(x) = \sqrt{x(1-x)} \sin\left(\frac{2\pi(1+2^{-7/5})}{x+2^{-7/5}}\right),$$

introduced by Ruppert, Wand, and Carroll (2003), see Section 17.5.1. Design points, x_i, are generated as $U[0,1]$, and responses as,

$$y_i = g_0(x_i) + \sigma(x_i)u_i,$$

where the u_i's are iid Gaussian, t_3, t_1, or centered χ^2_3. The local scale factor $\sigma(x)$ is either constant, $\sigma(x) = \sigma_0$, or linearly increasing in x, $\sigma(x) = \sigma_0(1+x)$, with $\sigma_0 = 0.2$. All the experiments have sample size $n = 400$.

A typical realization of the experiment with iid Gaussian noise is shown in Figure 1. In the right panel we see the true function as the solid red curve, the sample observations as points, the estimated conditional mean model as the solid black curve. The latter curve is estimated by penalized Gaussian likelihood, using Wood's `mgcv` package, modified slightly to accommodate the uniform confidence band construction provided by Krivobokova, Kneib, and Claeskens (2010) in the R package `Confbands`. More explicitly we employ the command,

```
ghat <- gam(y ~ s(x,bs = "os", k=40))
```

The `"os"` option specifies the O-spline basis of Wand and Ormerod (2008) rather than Wood's default thin-plate basis, and setting $k = 40$ effectively increases the initial number of basis functions from its default value. The latter option improves the centering of the bands and thereby the coverage performance. The heavier grey band is the 0.95 *pointwise* band as implemented in the `mgcv` package, while the lighter grey band is the 0.95 *uniform* band as implemented in the `Confbands` package.

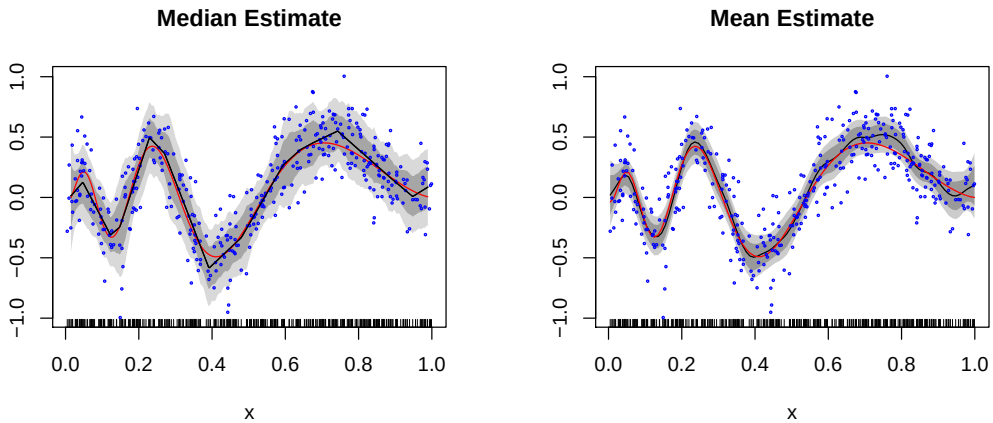

FIGURE 1. Confidence Bands for Penalized Estimators: Median and Mean estimated curves are shown in blue with the target curve in red. The heavier grey bands are 0.95 pointwise bands, while the lighter grey bands are the 0.95 uniform bands. Estimates are based on the same 400 Gaussian points.

In the left panel we illustrate the comparable fit for the median fit and the associated confidence bands. Now the solid black line is the piecewise linear estimate based on the total variation penalization described in Section 1. The selection of λ is done using the procedure described in Section 2; more specifically, we use the following R specification:

```
g <- function(lam,y,x) AIC(rqss(y ~ qss(x, lambda = lam)),k = -1)
lamstar <- optimize(g, interval = c(0.001, .5), x = x, y = y)
f <- rqss(y ~ qss(x, lambda = lamstar$min))
```

Bands are then constructed following the procedure described in Sections 3 and 4, conditional on the selection of λ. Again the heavier grey band is the pointwise 0.95 band, and the lighter grey band is the 0.95 uniform band. As is to be expected in this Gaussian setting, the bands for the mean are somewhat narrower than their median counterparts.

In Table 1 we report simulation results for the iid error models. The first three columns of the table are devoted to accuracy of the two estimation methods. Root mean integrated squared error (RMISE) and mean integrated absolute error (MIAE) give an overall impression of the precision of the two estimators. Mean effective degrees of freedom (MEDF), provides a measure of the average complexity (dimensionality) of the estimated models. In the Gaussian case, not surprisingly, the mean (gam) estimator performs somewhat better than the median (rqss) estimator, but is somewhat more profligate in terms of the effective degrees of freedom of the estimated models. For the t_3 error model, the performance comparison is reversed, now the rqss estimator is somewhat better than the gam estimator, and rqss is still considerably more parsimonious. For Cauchy errors, the gam estimator is poor, but the rqss estimator performance remains quite good, only slightly worse than its performance in the t_3 case. Finally, in the χ^2_3 case, our attempt to explore the consequences

	Accuracy			**Pointwise**		**Uniform**	
	RMISE	MIAE	MEDF	Pband	Uband	Pband	Uband
Gaussian							
rqss	0.063	0.046	12.936	0.960	0.999	0.323	0.920
gam	0.045	0.035	20.461	0.956	0.998	0.205	0.898
t_3							
rqss	0.071	0.052	11.379	0.955	0.998	0.274	0.929
gam	0.071	0.054	17.118	0.948	0.994	0.159	0.795
t_1							
rqss	0.099	0.070	9.004	0.930	0.996	0.161	0.867
gam	35.551	2.035	8.391	0.920	0.926	0.203	0.546
χ^2_3							
rqss	0.110	0.083	8.898	0.950	0.997	0.270	0.883
gam	0.096	0.074	14.760	0.947	0.987	0.218	0.683

TABLE 1. Performance of Penalized Estimators and Their Confidence Bands: IID Error Model

	Accuracy			**Pointwise**		**Uniform**	
	RMISE	MIAE	MEDF	Pband	Uband	Pband	Uband
Gaussian							
rqss	0.081	0.063	10.685	0.951	0.998	0.265	0.936
gam	0.064	0.050	17.905	0.957	0.999	0.234	0.940
t_3							
rqss	0.091	0.070	9.612	0.952	0.998	0.241	0.938
gam	0.103	0.078	14.656	0.949	0.992	0.232	0.804
t_1							
rqss	0.122	0.091	7.896	0.938	0.997	0.222	0.893
gam	78.693	4.459	7.801	0.927	0.958	0.251	0.695
χ^2_3							
rqss	0.145	0.114	7.593	0.947	0.998	0.307	0.921
gam	0.138	0.108	12.401	0.941	0.973	0.221	0.626

TABLE 2. Performance of Penalized Estimators and Their Confidence Bands: Linear Scale Model

of asymmetric error, we again see better accuracy for the gam estimates with somewhat more parsimonious estimation than in the Gaussian case due to the weaker signal to noise ratio of the model. Again, we should stress that the strict normal theory that underlies the finite sample justification of the Hotelling tube approach is obviously inapplicable for most of these simulations, but the asymptotic approximations appear to be adequate to obtain decent performance from the estimated bands.

Turning to the performance of the confidence bands, we consider two measures of performance. First, we compute for each realization of the experiment the proportion of gridded x values at which the band covers the value of $g_0(x)$. Averaging these proportions over the 1000 replications of the experiment gives coverage quite close the nominal coverage of 0.95 for the pointwise bands, while for the uniform bands this measure of coverage is almost unity. An alternative measure of performance for the bands is to simply compute the frequency with which the band covers the g_0 at *all* values of x. These frequencies are reported in the last two columns of the table for the pointwise and uniform bands. As expected, the pointwise bands uniform coverage is poor, on average they cover 0.95 of the curve, but they cover the entire curve only occasionally, achieving around .15 to .32 coverage. The uniform bands, as expected, perform much better achieving essential their nominal coverage probabilities of 0.95 except at the Cauchy and χ^2 where coverage is somewhat attenuated. Coverage for the rqss estimator is consistently better than for the gam estimator except in the iid Gaussian case.

Table 2 reports similar results for the linear scale model. In most respects the results are quite similar to those for the iid error table. Notably, the uniform coverage performance of the rqss bands is somewhat better for these models.

Our tentative conclusion from this exercise is that both the accuracy of the rqss estimator and its associated confidence bands are quite respectable, at least in comparison with the penalized least squares estimators represented by the gam estimator and its associated bands. In the final section of the paper we will illustrate our proposed methods on a more challenging empirical example involving several additive components in a model of risk factors for childhood malnutrition.

6.1. **Lasso selection of linear covariates.** We now complicate the foregoing simulation setup by introducing a group of additional covariates assumed to enter linearly, only a few of which are anticipated to be "significant." The usual "lasso" penalty is used to select these covariates and we explore the performance of various λ selection strategies for the lasso components and the validity of post selection inference. We maintain the same structure for the smooth nonparametric component, but augment the model with 24 new covariates that enter linearly. All 24 covariates are jointly Gaussian with unit variance. Six of the covariates have non-negligible impact on the response, the remaining 18 are irrelevant. To explore the impact of correlation among these covariates the first 12 covariates, including all the significant ones, are equicorrelated with correlation coefficient, $\rho = 0.5$. The remaining 12 covariates are independent.

After considerable exploration of λ selection for the lasso component using the SIC optimization methods described above, it was concluded that this approach often produced insufficient shrinkage. Given the very rough surface defined by the SIC criterion, simulated annealing was employed for the optimization, but this had the additional drawback that it was quite slow. Fortunately, an alternative λ selection for the lasso component has been recently suggested by Belloni and Chernozhukov (2009). This approach has the added advantage that it is computationally extremely simple and quick.

	Accuracy			Pointwise		Uniform		Covariates	
	RMISE	MIAE	MEDF	Pband	Uband	Pband	Uband	Positives	Negatives
Gaussian									
iid error	0.061	0.046	22.164	0.966	0.999	0.402	0.944	1.000	0.030
linear scale	0.073	0.053	20.641	0.955	0.998	0.311	0.920	0.999	0.038
t_3									
iid error	0.107	0.076	17.546	0.918	0.992	0.105	0.771	0.982	0.058
linear scale	0.116	0.086	17.691	0.940	0.993	0.268	0.822	0.914	0.107
t_1									
iid error	0.082	0.063	20.170	0.955	0.998	0.328	0.913	0.988	0.060
linear scale	0.094	0.072	18.663	0.949	0.996	0.282	0.888	0.970	0.073
χ^2_3									
iid error	0.127	0.097	16.273	0.926	0.993	0.159	0.783	0.908	0.090
linear scale	0.152	0.119	15.717	0.930	0.994	0.229	0.802	0.788	0.141

TABLE 3. Performance of Penalized Estimators and Their Confidence Bands: With Lasso Covariate Selection

To motivate this alternative approach to λ selection, let

$$R_\tau(b) = \sum_{i=1}^n \rho_\tau(y_i - x_i^\top b)$$

and consider minimizing,

$$R_\tau(b) + \lambda \|b\|_1.$$

At a solution, $\hat\beta$, we have the subgradient condition,

$$0 \in \partial R_\tau(\hat\beta) + \lambda \partial \|\hat\beta\|_1.$$

At $\beta = \beta_0(\tau)$, the true parameter vector, we have

$$\partial R_\tau(\beta_0(\tau)) = \sum_{i=1}^n (\tau - I(y_i \le x_i^\top \beta_0(\tau)))x_i = \sum_{i=1}^n (\tau - I(F_i(y_i) \le \tau))x_i$$

a random vector whose distribution is easily simulated by replacing $F_i(y_i)$ by random uniforms. The subgradient of the ℓ_1-norm, $\|\cdot\|_1$, is an element of the p-dimensional cube $[-1,1]^p$. Thus, simulating realizations of the random vector

$$S_n = \sum_{i=1}^n (\tau - I(U_i \le \tau))x_i$$

we can assert that the event $\|S_n\|_\infty \le \lambda$ should hold with high probability, provided of course that λ is chosen sensibly so that $\hat\beta$ is close to $\beta_0(\tau)$. Belloni and Chernozhukov (2009) propose choosing $\hat\lambda$ as a $(1-\alpha)$ quantile of the simulated distribution of $\|S_n\|_\infty$, or perhaps a constant multiple of such a quantile for some $c \in (1,2]$. In what follows we adopt a naïve version of this proposal with $\alpha = 0.05$ and $c = 1$. See Belloni and Chernozhukov (2009) for a much more thorough discussion of the rationale for this approach including

proofs of its asymptotic optimality. For the remaining simulations we maintain the one-dimensional SIC optimization for the choice of the λ for the smooth component g. This is done primarily to ensure comparability with the preceding results.

Table 3 reports results of a new experiment using the same models as in the earlier tables, but now augmented by the 24 linear covariates and the lasso penalty. For these simulations the coefficients on the first six covariates was set to 0.1. The first seven columns of the table provide comparable information to that found in the earlier tables for the performance of the estimates of the smooth component, g, and its confidence bands. We see some loss of efficiency in the RMISE and MAIE estimates, hardly surprising given the burden of estimating a considerably larger model. The effective degrees of freedom of the newly estimated models are roughly increased by six in the Gaussian and Cauchy cases, which would seem to bode well for the model selection strategy of the lasso. Performance of the confidence bands is also still quite good, although the uniform bands are somewhat less accurate for the t_3 and χ^2_3 cases.

The last two columns of the table report the observed frequency, in the 1000 trials of the experiment, that the six covariates with non-zero coefficients are selected ("positives"), and that ("negatives") the six correlated covariates were selected. The remaining 12 uncorrelated covariates were selected in less than 0.01 percent of cases. Except for the χ^2_3 setting the selection is quite good: all six important covariates are selected with high probability, and the covariates with zero coefficients are rarely selected. It should be noted that "selected" in the present context means that, given the chosen λ's, conventional inference employing standard errors from the estimated matrix, V of (4.1) yields p-values less than 0.05.

As in other applications of the lasso, there is a temptation to refit the model, once the covariate selection is done in the first phase. This yields some modest improvement in performance, but nothing terribly unexpected. We return to this point when we address lasso selection of parametric components of the model of malnutrition risk in the next section.

7. Risk Factors for Childhood Malnutrition

An application motivated by a recent paper by Fenske, Kneib, and Hothorn (2008) illustrates the range of the models described above. To investigate risk factors for childhood malnutrition we consider determinants of children's heights in India. The data comes originally from the Demographic and Health Surveys (DHS) conducted regularly in more than 75 countries; we employ a selected sample of 37,649 observations constructed similarly to the sample used by Fenske, Kneib, and Hothorn (2008) except that we have included the number of living siblings as an additional covariate. All children in the sample are between the ages of 0 and 5. We will consider six covariates entering as additive nonparametric effects in addition to the response variable height: the child's age, and months of breastfeeding, the mother's body mass index (bmi), her age and years of education, and the father's years of education. Summary statistics for these variables appear in Table 4 . There are also a large number of discrete covariates that enter the model as parametric effects; these variables are also summarized in Table 4. In the terminology of `R` categorical variables are entered as factors, so a variable like mother's religion that has five distinct levels accounts for 4 model parameters. For all the binary, consumer durable variables ownership is coded as one.

TABLE 4. Summary Statistics for the Response and Continuous Covariates

Variable	Units	Min	Q1	Q2	Q3	Max
Child's Height	cm	45.00	73.60	84.10	93.20	120.00
Child's Age	months	0.00	16.00	31.00	45.00	59.00
BreastFeeding	months	0.00	9.00	15.00	24.00	59.00
Mother's BMI	kg/m^2	12.13	17.97	19.71	22.02	39.97
Mother's Age	years	13.00	21.00	24.00	28.00	49.00
Mother's Ed	years	0.00	0.00	5.00	9.00	21.00
Father's Ed	years	0.00	2.00	8.00	10.00	22.00
Living Children	kids	1.00	2.00	2.00	3.00	13.00

Prior studies of malnutrition using data like the DHS have typically either focused on mean height or transformed the response to binary form and analyzed the probability that children fall below some conventional height cutoff. However, it seems more natural to try to estimate models for some low conditional quantile of the height distribution. This is the approach adopted by FKH, who employ boosting as a model selection device, and the one we will employ here. It is also conventional in prior studies including FKH, to replace the child's height as response variable by a standardized Z-score. This variable is called "stunting" in the DHS data and it is simply an age adjusted version of height with age-specific location and scale adjustments.

Variable	Counts	Percent
csex		
male	19591	52.0
female	18058	48.0
ctwin		
singlebirth	37196	98.8
twin	453	1.2
cbirthorder		
1	11491	30.5
2	10714	28.5
3	6304	16.7
4	3761	10.0
5	5379	14.3
munemployed		
unemployed	24002	63.8
employed	13647	36.2
mreligion		
hindu	26019	69.1
muslim	6051	16.1
christian	3807	10.1
sikh	697	1.9
other	1075	2.9
mresidence		
urban	13973	37.1
rural	23676	62.9

Variable	Counts	Percent
wealth		
poorest	6630	17.6
poorer	6858	18.2
middle	7814	20.8
richer	8454	22.5
richest	7893	21.0
electricity		
no	10433	27.7
yes	27216	72.3
radio		
no	25351	67.3
yes	12298	32.7
television		
no	19423	51.6
yes	18226	48.4
refrigerator		
no	31091	82.6
yes	6558	17.4
bicycle		
no	19924	52.9
yes	17725	47.1
motorcycle		
no	30223	80.3
yes	7426	19.7
car		
no	36285	96.4
yes	1364	3.6

In our experience this preliminary adjustment is detrimental to the estimation of the effects of interest so we have reverted to using height itself as a response variable. The construction of the Z-score seems to presuppose that none of the other covariates matters, and yet this is precisely the object of the subsequent analysis. Inclusion of age as a non-parametric effect after Z-score adjustment of the response is an admission that the original rescaling was inadequate and needs modification in view of other covariate effects. It seems preferable to estimate the age specific effect together with the other covariate effects in one step. Delbaere *et. al.* (2007) argue against using a similar Z-score adjustment of birthweights for gestational age.

The `R` specification of the model to be estimated is given by

```
f <- rqss(height~ qss(cage,lambda = lam[1]) + qss(mage, lambda = lam[2]) +
     qss(bfed,lambda = lam[3]) + qss(mbmi, lambda = lam[4]) +
     qss(medu, lambda = lam[5]) + qss(fedu, lambda = lam[6]) +
     qss(livingchildren, lambda = lam[7]) + csex + ctwin + cbirthorder +
     munemployed + mreligion + mresidence + wealth + electricity + radio +
     television + refrigerator + bicycle + motorcycle + car, tau = .10,
     method = "lasso", lambda = lam[8], data = india)
```

The formula given as the first argument specifies each of the seven non-parametric "smooth" terms. In the present instance each of these is univariate, each requires specification of a λ determining its degree of smoothness. The remaining terms in the formula are specified as is conventional in other `R` linear model fitting functions. The argument `tau` specifies the quantile of interest and `data` specifies the dataframe within which all of the formula variables are defined. The `method = "lasso"` indicates that a lasso penalty should be imposed on the linear covariate effects with `lam[8]` specified as the lasso value of λ.

Optimizing SIC(λ) over $\lambda \in \mathbb{R}^8_+$ is a difficult task. Since children's heights are reported only to the nearest millimeter, we begin by "dithering" the response, randomly perturbing values by adding an uniformly distributed half-millimeter "noise" $U[-0.05, 0.05]$. This ensures that fitted quantile regression models avoid degenerate solutions involving "tied" responses. Such solutions are dangerous from a model selection standpoint because they may misrepresent the model dimension when counting zero residuals. A prudent optimization strategy would seem to be to explore the space of λ's informally, trying to narrow the region of optimization and then resort to some form of global optimizer to further narrow the selection. Initial exploration was conducted by considering all of the continuous covariate effects excluding the child's age as a group, and examining one dimensional grids for λ's for this group, for the child's age, and the lasso λ individually. Preliminary experiments using simulated annealing yielded λ's for the $\hat{g}$ terms of the model of $\{16, 67, 78, 85, 78, 82, 80\}$. Age of the child, representing the usual growth curve, required considerably more flexibility than the other effects and therefore received a smaller λ. The choice of the λ parameter for the lasso contribution of the model posed a more difficult challenge. Preliminary SIC optimization for this parameter produced quite small values that failed to zero out more than one or two of the remaining 24 coefficients. Choosing the lasso λ as described in the previous section with $c = 1$ and $\alpha = 0.05$, according the proposal of Belloni and Chernozhukov (2009), yielded $\lambda = 237$. This choice had the effect of zeroing out all but three of the covariates. Given this feast-or-famine disparity it is tempting to consider an intermediate values as an alternative.

In Table 5 we report estimated coefficients and their standard errors for several models corresponding to different values of the lasso λ. The first column reports results for the unconstrained model with no lasso shrinkage. The last column is based on the $\lambda = 237$ value as selected in the simulations, the column headed $\lambda = 146$ corresponds to the median value of the simulated reference distribution for λ, and $\lambda = 60$ corresponds roughly the lower support point of the reference distribution. To compare the resulting four versions of the fitted model we fix the λ selections for the smooth g components, selected covariates are then identified for each value of the lasso λ, and finally, the model is reestimated without any lasso shrinkage with only the selected covariates.

Some of results are unsurprising: girls tend to be shorter than boys by a little more than a centimeter, but other results are puzzling. Covariates that are quite strongly significant in the unconstrained ($\lambda = 0$) version of the model like birth order of the child or the family wealth variables do not necessarily survive the lasso shrinkage. Birth order is particularly intriguing; having already conditioned on the number of children we see a strong monotone relationship indicating that children later in the birth order are shorter than their older siblings. Despite the highly significant nature of these unconstrained estimates these coefficient succumb early on to the pressure applied by the lasso. Curiously, the household durable

Covariate	$\lambda = 0$	$\lambda = 60$	$\lambda = 146$	$\lambda = 237$
Intercept	43.017 (0.605)	43.469 (0.570)	43.731 (0.522)	43.476 (0.520)
Female	−1.434 (0.085)	−1.427 (0.086)	−1.416 (0.087)	−1.401 (0.087)
Twin	−0.874 (0.360)	-	-	-
Birth2	−0.824 (0.124)	−0.230 (0.099)	-	-
Birth3	−1.085 (0.178)	-	-	-
Birth4	−1.460 (0.242)	-	-	-
Birth5	−2.037 (0.314)	−0.689 (0.187)	-	-
M-Unemployed	0.093 (0.094)	-	-	-
M-Muslim	−0.049 (0.128)	-	-	-
M-Christian	0.392 (0.156)	0.533 (0.152)	-	-
M-Sikh	0.020 (0.362)	-	-	-
M-other	−0.328 (0.223)	-	-	-
M-Rural	0.234 (0.105)	-	-	-
Poorer	0.436 (0.153)	-	-	-
Middle	0.840 (0.171)	0.518 (0.138)	-	-
Richer	1.145 (0.201)	0.707 (0.162)	-	-
Richest	1.752 (0.256)	1.295 (0.220)	0.846 (0.139)	-
Electricity	0.203 (0.132)	0.279 (0.132)	0.566 (0.122)	0.566 (0.119)
Radio	0.025 (0.098)	0.105 (0.097)	-	-
TV	0.153 (0.122)	0.193 (0.122)	0.350 (0.119)	0.462 (0.113)
Fridge	0.114 (0.152)	0.156 (0.154)	-	-
Bicycle	0.459 (0.090)	0.443 (0.091)	0.396 (0.089)	0.372 (0.089)
Motorcycle	0.163 (0.134)	0.246 (0.133)	-	-
Car	0.566 (0.232)	-	-	-

TABLE 5. Linear Covariate Estimates with Standard Errors for Several Lasso Selected Models

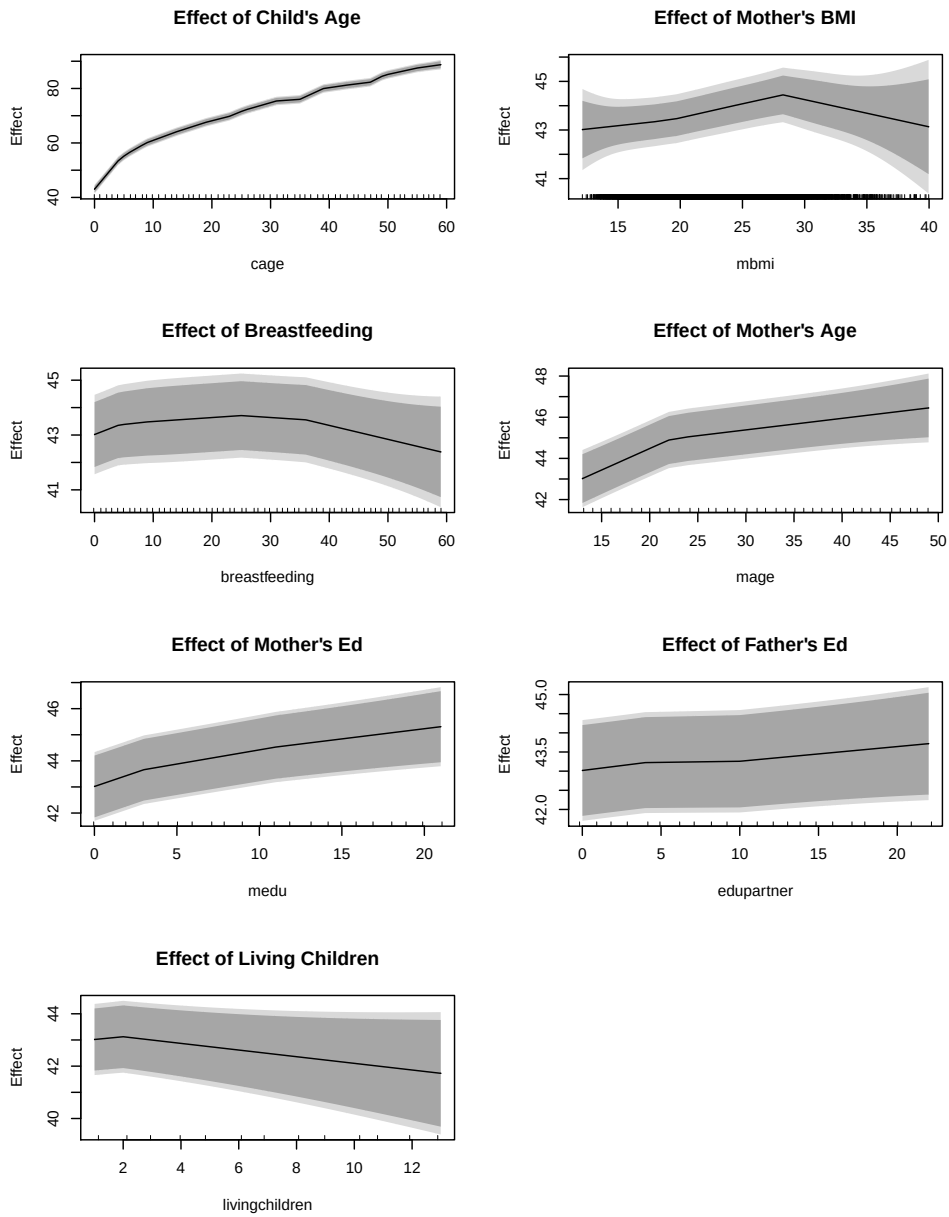

FIGURE 2. Estimated Smooth Components of the Malnutrition Model

ownership variables, none of which look particularly promising at $\lambda = 0$, are more successful withstanding the lasso shrinkage. From both a substantive and a methodological perspective, the lasso performance for this application is somewhat disappointing. Obviously, we are still entitled to some skepticism about the "automatic" nature of recent innovations in model selection.

Figure 2 shows the estimated effects for the seven continuous covariates and their associated confidence bands. As in the earlier plot the darker band is the pointwise 0.95 band, and the lighter band represents the 0.95 uniform band constructed with the Hotelling tube procedure. Clearly the effect of age and the associated growth curve is quite precisely estimated, but the remaining effects show considerably more uncertainty. Mother's BMI has a positive effect up to about 30 and declines after that, similarly breastfeeding appears advantageous up until about 28 months, and then declines somewhat. Breastfeeding beyond 28 months is apparently quite common in India; in our DHS sample roughly 30 percent of children older than 28 months were reported to have been breastfed beyond 28 months. The other effects are essentially linear given the selected λ's for these additive model components.

7.1. **Comparison with Results from Boosting.** Several difficulties inhibit a detailed comparison of the foregoing results with the boosting results for similar models by Fenske, Kneib, and Hothorn (2008) that originally motivated this research. There is a slight difference in the selected samples, attributable to our desire to include the number of living children in each household as an additional covariate. FKH's full sample is 37,623 while ours is 37,649. More significantly, FKH used a random two-thirds of the sample for estimation and the remaining one-third for determining the optimal stopping iteration of the boosting algorithm. Even more crucially, FKH used the age-adjusted Z-score for height as the response variable, while we used height directly. These differences make comparison of the smooth components particularly tricky since the scaling of covariate effects is strongly influenced by the Z-score rescaling. Thus, we confine our comparisons to qualitative features of the estimates of the smooth nonparametric components. The boosting estimates of the effects of mother's BMI, and months of breastfeeding are quite similar to those seen in our Figure 2, very mildly increasing and then decreasing about midway over the range of the covariate. Likewise, mother's years of education has an approximately linear effect, while father's education is essentially negligible. Mother's age has a monotone increasing effect in the the additive model estimates, with an initial slope slightly larger than the slope after age 23. In contrast the boosting estimate exhibits a much steaper slope in the ages 15 to 23.

Turning to the comparison of the effects of the discrete covariates, the boosting estimates exhibit the same monotone decreasing pattern for birth order that we remarked upon earlier for the additive model estimates. Children later in the birth order are shorter than their older siblings, at least if we focus on the first conditional decile. Without adjustment for the number of siblings the interpretation of the birth order effect is ambiguous, but results for the additive model specification have shown that the birth order effects persist even when the number of siblings is accounted for. Higher household economic status, not surprisingly yields taller children as for the additive model estimates, and girls are shorter than boys.

None of the durable ownership variables produce large boosting effects with the exception of bicycle ownership; this too is consistent with the additive model results.

Perhaps a more telling comparison of the boosting and the additive model approaches lies in their model selection results. Fenske, Kneib, and Hothorn (2008) report in their Table 5 the proportion of iterations in which each variable is selected by the boosting algorithm. For $\tau = 0.1$ none of these proportions are terribly impressive: age of the child is strongest achieving 0.272, but recall that this is after the Z-score adjustment. Curiously, the next strongest boosting effect is father's education at 0.137 even though the magnitude of the estimated effect is tiny. Gender of the child, which is the only really consistent effect after lasso shrinkage of the additive model is a weak performer in the boosting competition appearing in only 0.019 of the iterations. Mother's BMI, age and months of breastfeeding perform somewhat better with proportions 0.064, 0.092, and 0.091 respectively. Electricity and TV ownership almost never appear in the boosting results while on the contrary they are two of only four discrete effects left standing after the most severe lasso shrinkage.

An advantage of the additive model framework, as we have tried to argue above, lies in the ability to construct confidence bands and other inferential procedures. This appears to be a more difficult task for boosting and related approaches. Of course, the validity of post model selection inference always merits some healthy skepticism. Only through further comparison of methods in related empirical circumstances can we build confidence in their validity.

8. Conclusion

Post-selection model inference involves many delicate issues as recent work by Pötscher and Leeb (2009) has stressed. Reliable confidence bands for nonparametric additive components constitutes one important aspect of this general challenge. Hotelling's (1939) tubes seem to offer a viable approach to confidence bands for additive quantile regression models provided λ selection is done judiciously. Simulation evidence suggests that the asymptotic approximations required to justify use of the Hotelling tube approach are at least as successful in the rqss context as they were for earlier proposals designed for gam fitting. We have also seen that the ubiquitous ℓ_1 lasso penalty can be adapted for selection of parametric components in these models. Doubtless, further work will yield new refinements, but some progress can be recognized. All of the methods described above have been implemented for the R package `quantreg`, (Koenker (2010)), I hope that this will encourage others to explore these methods.

References

Belloni, A., and V. Chernozhukov (2009): "L_1-Penalized Quantile Regression in High-Dimensional Sparse Models," `http://arXiv.org/abs/1001.0188`.

Breiman, L., and J. Friedman (1985): "Estimating optimal transformations for multiple regression and correlation," *Journal of the American Statistical Association*, pp. 580–598.

Davies, P. L., and A. Kovac (2001): "Local extremes, runs, strings and multiresolution," *Ann. Statist.*, 29, 1–48.

Delbaere, I., S. Vansteelandt, D. D. Bacquer, H. Verstraelen, J. Gerris, P. D. Sutter, and M. Temmerman (2007): "Should we adjust for gestational age when analysing birth weights? The use of z-scores revisited," *Human Reproduction*, 22, 2080–2083.

Fenske, N., T. Kneib, and T. Hothorn (2008): "Identifying Risk Factors for Severe Childhood Malnutrition by Boosting Additive Quantile Regression," preprint.

Hastie, T., and R. Tibshirani (1986): "Generalized Additive Models," *Statistical Science*, 1, 297–310.

Hastie, T., and R. Tibshirani (1990): *Generalized Additive Models*. Chapman-Hall.

Hotelling, H. (1939): "Tubes and spheres in n-space and a class of statistical problems," *American J of Mathematics*, 61, 440–460.

Johansen, S., and I. M. Johnstone (1990): "Hotelling's Theorem on the Volume of Tubes: Some Illustrations in Simultaneous Inference and Data Analysis," *The Annals of Statistics*, 18, 652–684.

Knowles, M. (1987): "Simultaneous confidence bands for random functions," Ph.D. thesis, Stanford University.

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

——— (2010): "quantreg: Quantile Regression, v4.45," `http://www.r-project.org/package=quantreg`.

Koenker, R., and I. Mizera (2004): "Penalized triograms: total variation regularization for bivariate smoothing," *J. Royal Stat. Soc. (B)*, 66, 145–163.

Koenker, R., and P. Ng (2005): "A Frisch-Newton Algorithm for Sparse Quantile Regression," *Mathematicae Applicatae Sinica*, 21, 225–236.

Koenker, R., P. Ng, and S. Portnoy (1994): "Quantile smoothing splines," *Biometrika*, 81, 673–680.

Krivobokova, T., T. Kneib, and G. Claeskens (2010): "Simultaneous Confidence Bands for Penalized Spline Estimators," *J. of Am. Stat. Assoc.*, forthcoming.

Loader, C. (2010): "locfit: Local Regression, Likelihood and Density Estimation, v1.5-5," `http://www.r-project.org/package=locfit`.

Machado, J. (1993): "Robust model selection and M-estimation," *Econometric Theory*, pp. 478–493.

Meyer, M., and M. Woodroofe (2000): "On the degrees of freedom in shape-restricted regression," *Annals of Stat.*, 28, 1083–1104.

Naiman, D. (1986): "Conservative confidence bands in curvilinear reression," *Annals of Statistics*, 14, 896–906.

Nychka, D. (1983): "Bayesian Confidence Intervals for smoothing splines," *J. of Am. Stat. Assoc.*, 83, 1134–43.

Pötscher, B., and H. Leeb (2009): "On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD and thresholding," *J. Multivariate Analysis*, 100, 2065–2082.

Powell, J. L. (1991): "Estimation of monotonic regression models under quantile restrictions," in *Nonparametric and Semiparametric Methods in Econometrics*, ed. by W. Barnett, J. Powell, and G. Tauchen. Cambridge U. Press: Cambridge.

Rudin, L., S. Osher, and E. Fatemi (1992): "Nonlinear total variation based noise removal algorithms," *Physica D*, 60, 259–268.

Ruppert, D., M. Wand, and R. J. Carroll (2003): *Semiparametric Regression*. Cambridge U. Press.

Schwarz, G. (1978): "Estimating the dimension of a model," *The annals of statistics*, pp. 461–464.

Sun, J. (1993): "Tail Probabilities of the Maxima of Gaussian Random Fields," *Annals of Probability*, 21, 34–71.

Sun, J., and C. Loader (1994): "Simultaneous Confidence Bands for Linear Regression and Smoothing," *Annals of Statistics*, 22, 1328–1347.

Sun, J., C. Loader, and W. McCormick (2000): "Confidence bands in generalized linear models," *Annals of Statistics*, 28(2), 429–460.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005): "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society B*, pp. 91–108.

Wahba, G. (1983): "Bayesian "Confidence Intervals" for the cross-validated smoothing spline," *J. Royal Stat. Soc. (B)*, 45, 133–50.

Wand, M. P., and J. T. Ormerod (2008): "On Semiparametric Regression with O'Sullivan penalized Splines," *Aust. N.Z. J. of Statistics*, 50, 179–198.

Weyl, H. (1939): "On the Volume of Tubes," *Am J. Math*, 61, 461–472.

Wood, S. (2006): *Generalized Additive Models: An Introduction with R*. Chapman-Hall.

——— (2010): "mgcv: GAMs with GCV/AIC/REML smoothness estimation and GAMMs by PQL," `http://www.r-project.org/package=mgcv`.