ADDITIVE MODELS FOR QUANTILE REGRESSION: SOME NEW METHODS FOR R

ROGER KOENKER

Abstract. This brief report describes some recent developments of the R `quantreg` package to incorporate methods for additive models. The methods are illustrated with an application to modeling childhood malnutrition in India.

Models with additive nonparametric effects offer a valuable dimension reduction device throughout applied statistics. In this paper we describe some recent developments of additive models for quantile regression. These 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.

Additive models have received considerable attention since their introduction by 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 a price to pay.

1. Additive Models for Quantile Regression

Our approach to additive models for quantile regression and especially our implementation of methods in R is 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 other respects the structure of the models is quite similar. We will consider models for conditional quantiles of the general form:

Version: September 5, 2009. This research was partially supported by NSF grant SES-08-50060. I would like to express my appreciation to Ying Li for excellent research assistance. All of the methods described below have been implemented in version 4.42 of the `quantreg` package for R, Koenker (2009).

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

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 $\beta \in \mathcal{R}^K$, by solving

$$\min_{(\beta,g)} \sum \rho_\tau(y_i - x_i^T\beta + \sum g_j(z_{ij})) + \lambda_0 \parallel \beta \parallel_1 + \sum_{j=1}^{J} \lambda_j \bigvee(\nabla g_j) \tag{2}$$

where $\parallel \beta \parallel_1 = \sum_{k=1}^{K} |\beta_k|$ 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,

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

where $\nabla^2 g(z)$ denotes the Hessian of g, and $\parallel \cdot \parallel$ will denote the usual Hilbert-Schmidt norm for matrices. As it happens, solutions to (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 greatly simplifies the computations required to solve (2), which can now 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 like those implemented in the `quantreg` package.

2. A Model of Childhood Malnutrition in India

An application motivated by a recent paper by Fenske, Kneib, and Hothorn (2008) illustrates the full range of the models described above. As part of a larger investigation of malnutrition we are interested in determinants of children's heights in India. The data comes from Demographic and Health Surveys (DHS) conducted regularly in more than 75 countries. We have 37,623 observations on children between the ages of 0 and 6. 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), age and years of education, and the father's years of education. Summary statistics for these variables appear in Table 1 . There are also a large number of discrete covariates that enter the model as parametric effects; these variables are also summarized in Table 1. 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.

TABLE 1. Summary Statistics for the Response and Continuous Covariates

Ctab	Units	Min	Q1	Q2	Q3	Max
Chgt	cm	45.00	73.60	84.10	93.20	120.00
Cage	months	0.00	16.00	31.00	45.00	59.00
Bfed	months	0.00	9.00	15.00	24.00	59.00
Mbmi	kg/m^2	12.13	17.97	19.71	22.02	39.97
Mage	years	13.00	21.00	24.00	28.00	49.00
Medu	years	0.00	0.00	5.00	9.00	21.00
Fedu	years	0.00	2.00	8.00	10.00	22.00

Variable	Counts	Percent
csex		
male	19574	52.0
female	18049	48.0
ctwin		
singlebirth	37170	98.8
twin	453	1.2
cbirthorder		
1	11486	30.5
2	10702	28.4
3	6296	16.7
4	3760	10.0
5	5379	14.3
mreligion		
christian	3805	10.1
hindu	26003	69.1
muslim	6047	16.1
other	1071	2.8
sikh	697	1.9
mresidence		
urban	13965	37.1
rural	23658	62.9
deadchildren		
0	31236	83.0
1	4640	12.3
2	1196	3.2
3	551	1.5

Variable	Counts	Percent
wealth		
poorest	6625	17.6
poorer	6858	18.2
middle	7806	20.7
richer	8446	22.4
richest	7888	21.0
munemployed		
unemployed	24002	63.8
employed	13621	36.2
electricity		
no	10426	27.7
yes	27197	72.3
radio		
no	25333	67.3
yes	12290	32.7
television		
no	19414	51.6
yes	18209	48.4
refrigerator		
no	31070	82.6
yes	6553	17.4
bicycle		
no	19902	52.9
yes	17721	47.1
motorcycle		
no	30205	80.3
yes	7418	19.7
car		
no	36261	96.4
yes	1362	3.6

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 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 basically just an age adjusted version of height with age-specific location and scale adjustments. In our experience this preliminary adjustment is highly detrimental to the estimation of the effects of interest so we have reverted to using height itself as a response variable.

In R specification of the model to be estimated is given by

```
f <- rqss(Chgt~ qss(Cage,lambda = 20) + qss(Mage, lambda = 80) +
     qss(Bfed,lambda = 80) + qss(Mbmi, lambda = 80) +
     qss(Medu, lambda = 80) + qss(Fedu, lambda = 80) +
     munemployed + csex + ctwin + cbirthorder + mreligion +
     mresidence + deadchildren + wealth + electricity + radio +
     television + refrigerator + bicycle + motorcycle + car, tau = .10,
     method = "lasso", lambda = 40, data = india)
```

The formula given as the first argument specifies each of the six 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 like lm( ) and rq( ). The argument `tau` specifies the quantile of interest and `data` specifies the dataframe within which all of the formula variables are defined.

2.1. **λ-Selection.** A challenging task for any regularization problem like (2) is the choice of the λ parameters. Since we have 7 of these the problem is especially daunting. Following the suggestion originally appearing in Koenker, Ng and Portnoy we relied upon the SIC-type 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

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

The quantity $p(\lambda)$ is usually defined for linear least-squares estimators as the trace of a pseudo projection matrix. 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),

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

see Koenker (2005, p.243). A consequence of this approach to characterizing model dimension is that it is necessary to avoid "tied" responses; we ensure this by "dithering" the response variable. Heights measured to the nearest millimeter are replaced by randomly perturbed values by adding uniformly distributed "noise" $U[-0.05, 0.05]$.

Optimizing $SIC(\lambda)$ over $\lambda \in \mathbb{R}_+^7$ is still a difficult task made more challenging by the fact that the objective is discontinuous at points where new constraints become binding and previously free parameters vanish. The 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. 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. This procedure produced rough starting values for the following simulated annealing safari:

```
set.seed(1917)
malnu <- cbind(india, dChgt = dither(india$Chgt))

sic <- function(lam){
a <- AIC(rqss(dChgt~csex+qss(cage,lambda=lam[1])+
        qss(mbmi,lambda=lam[2])+ qss(Bfed,lambda=lam[3])+
        qss(Mage,lambda=lam[4])+ qss(Medu,lambda=lam[5])+ qss(Fedu,lambda=lam[6])+
        csex + ctwin+cbirthorder+ munemployed+mreligion+mresidence +
        wealth+electricity+radio+television+refrigerator+bicycle+motorcycle+car,
        tau=0.1, method="lasso", lambda=lam[7], data=malnu), k=-1)
print(c(lam,a))
a
}
g <- optim(c(20,80,80,80,80,80,20),sic,method="SANN",control=list(maxit=1000,
temp=5000, trace=10,REPORT=1))
```

Each function evaluation takes about 7 seconds, so 1000 steps of the simulated annealing algorithm required about two hours. The "solution" yielded:

```
$par
[1] 16.34189 67.92552 78.49549 85.05942 77.81752 82.51737 17.63161
$value
[1] 245034.0
```

Thus, the original starting values seem to be somewhat vindicated. We would not claim that the "solutions" produced by this procedure are anything but rough approximations. However, in our experience choosing λ's anywhere in a moderately large neighborhood of this solution obtained this way yield quite similar inferential results we will now describe.

2.2. **Confidence Bands and Post-Selection Inference.** 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) provide 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,

$$\tilde{X} = \begin{bmatrix} X & G_1 & \cdots & G_J \\ \lambda_0 H_K & 0 & \cdots & 0 \\ 0 & \lambda_1 P_1 & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_j P_J \end{bmatrix}.$$

Here X denotes the matrix representing the parametric covariate effects, the G_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 full set of parameters, $\theta = (\beta^\top, \gamma_1^\top, \cdots, \gamma_J^\top)^\top$ is given by the sandwich formula,

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

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

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

$\hat{u}_i$ is the ith residual from the fitted model, and h is a bandwidth determined by one the usual built-in rules. 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 D representing the prediction of g_j at some specified plotting points $z_{ij} : i = 1, \cdots, m$ is first made, then we extract the corresponding chunk of the matrix V, and compute the estimated covariance matrix of of the vector $D\hat{\theta}$. Finally, we extract the square root of the diagonal of this matrix. The only slight complication of this strategy is to remember that the intercept should be appended to each such prediction and properly accounted for in the extraction of the covariance matrix of the predictions.

To illustrate the use of these confidence bands, Figure 1 shows the six estimated smoothed covariate effects and the associated confidence bands. This plot is produced by refitting the model with the selected λ's, calling the fitted model object `fit` and then using the command

```
plot(fit, bands = TRUE, page = 1)
```

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 is advantageous up until about 30 months, and then declines somewhat. (Breastfeeding after 36 months is apparently quite common in India as revealed by the DHS survey.)

What about inference on the parametric components of the model? We would certainly like to have some way to evaluate the "significance" of the remaining parametric coefficients in the model. Again bias effects due to shrinkage create some serious doubts, from a strict frequentist viewpoint these doubts may be difficult to push aside. See for example the recent work of Pötscher and Leeb (2009). However, a Bayesian viewpoint may again rescue the naive application of the covariance matrix estimate discussed above. When we employ

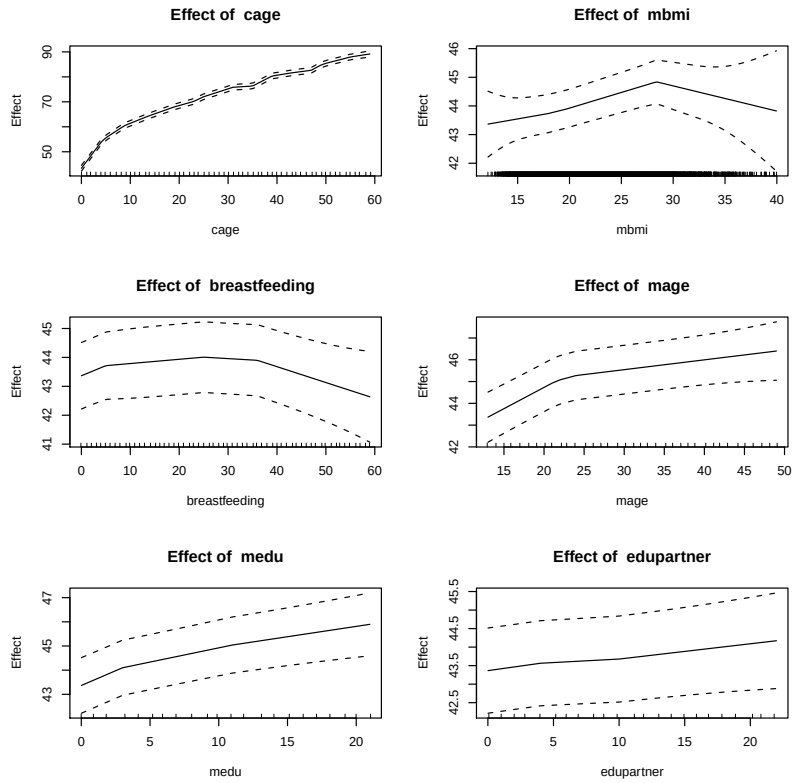

FIGURE 1. Smooth covariate effects on children's heights with pointwise confidence bands.

this covariance matrix to evaluate the parametric component of the model, we obtain the following table from R using the usual `summary(fit)` command.

```
Parametric coefficients:
                    Estimate Std. Error  t value Pr(>|t|)
(Intercept)         4.336e+01  5.753e-01   75.382  < 2e-16 ***
csexfemale         -1.405e+00  4.516e-02  -31.110  < 2e-16 ***
ctwintwin          -6.550e-01  2.504e-02  -26.157  < 2e-16 ***
cbirthorder2       -6.492e-01  4.411e-02  -14.719  < 2e-16 ***
cbirthorder3       -9.491e-01  4.246e-02  -22.355  < 2e-16 ***
cbirthorder4       -1.437e+00  4.013e-02  -35.807  < 2e-16 ***
cbirthorder5       -2.140e+00  3.975e-02  -53.837  < 2e-16 ***
munemployedemployed 9.753e-02  4.453e-02    2.190 0.028532 *
mreligionhindu     -2.111e-01  4.185e-02   -5.043 4.61e-07 ***
mreligionmuslim    -1.957e-01  3.991e-02   -4.904 9.42e-07 ***
mreligionother     -3.934e-01  3.005e-02  -13.090  < 2e-16 ***
mreligionsikh      -2.353e-13  2.766e-02 -8.5e-12 1.000000
mresidencerural     1.465e-01  4.357e-02    3.363 0.000773 ***
```

```
wealthpoorer          2.126e-01  4.374e-02    4.861 1.17e-06 ***
wealthmiddle          5.880e-01  4.230e-02   13.899  < 2e-16 ***
wealthricher          8.368e-01  3.999e-02   20.924  < 2e-16 ***
wealthrichest         1.358e+00  3.540e-02   38.367  < 2e-16 ***
electricityyes        2.414e-01  4.345e-02    5.556 2.78e-08 ***
radioyes              4.073e-02  4.530e-02    0.899 0.368547
televisionyes         1.793e-01  4.378e-02    4.096 4.21e-05 ***
refrigeratoryes       1.289e-01  3.969e-02    3.247 0.001168 **
bicycleyes            3.940e-01  4.489e-02    8.778  < 2e-16 ***
motorcycleyes         1.764e-01  4.193e-02    4.207 2.60e-05 ***
caryes                3.633e-01  3.214e-02   11.303  < 2e-16 ***
```

There are a number of peculiar aspects to this table. Somewhat surprisingly, our "optimal" choice of the lasso λ of 17.63 only zeros out one coefficient – the effect of the relatively small minority of sikhs. For all the remaining coefficients the effect of the lasso shrinkage is to push coefficients toward zero, *but also to reduce their standard errors.* The implicit prior represented by the lasso penalty acts as data augmentation that improves the apparent precision of the estimates. Whether this should be regarded as a Good Thing is really questionable. To contrast the conclusions drawn from this table with somewhat more conventional methods, we have reestimated the model maintaining the smoothing λ's at their "optimized" values, but setting the lasso λ to zero.

```
Parametric coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         43.51139    0.64391  67.574  < 2e-16 ***
csexfemale          -1.44232    0.08421 -17.128  < 2e-16 ***
ctwintwin           -0.86987    0.34680  -2.508  0.01214 *
cbirthorder2        -0.76125    0.10883  -6.995 2.70e-12 ***
cbirthorder3        -1.13288    0.14098  -8.036 8.88e-16 ***
cbirthorder4        -1.60645    0.18238  -8.808  < 2e-16 ***
cbirthorder5        -2.34391    0.20206 -11.600  < 2e-16 ***
munemployedemployed  0.09254    0.09348   0.990  0.32221
mreligionhindu      -0.42625    0.15390  -2.770  0.00561 **
mreligionmuslim     -0.50185    0.18902  -2.655  0.00793 **
mreligionother      -0.76162    0.25700  -2.963  0.00304 **
mreligionsikh       -0.39472    0.39786  -0.992  0.32114
mresidencerural      0.23299    0.10362   2.248  0.02456 *
wealthpoorer         0.45847    0.15372   2.982  0.00286 **
wealthmiddle         0.89591    0.17073   5.248 1.55e-07 ***
wealthricher         1.23945    0.20023   6.190 6.07e-10 ***
wealthrichest        1.83644    0.25340   7.247 4.33e-13 ***
electricityyes       0.14807    0.13215   1.120  0.26253
radioyes             0.01751    0.09701   0.180  0.85679
televisionyes        0.16862    0.12103   1.393  0.16359
refrigeratoryes      0.15100    0.14808   1.020  0.30787
bicycleyes           0.42391    0.08897   4.764 1.90e-06 ***
motorcycleyes        0.20167    0.13193   1.529  0.12637
caryes               0.49681    0.23161   2.145  0.03196 *
```

This table is obviously quite different: coefficients are somewhat larger in absolute value and more importantly standard errors are also somewhat larger. The net effect of removing the lasso "prior" is that many of the coefficients that looked "significant" in the previous version of the table are now of doubtful impact. Since we regard the lasso penalty more as an expedient model selection device rather than an accurate reflection of informed prior opinion, the latter table seems to offer a more prudent assessment of the effects of the parametric contribution to the model. A natural question would be: does the refitted model produce different plots of the smooth covariate effects? Fortunately, the answer is no, replotting Figure 1 with the unlasso'd parametric fit yields a figure that is almost indistinguishable from the orginal.

Most of the estimated parametric effects are unsurprising: girls are shorter than boys even at the 10th percentile of heights, children later in the birth order tend to be shorter, mother's who are employed and wealthier have taller children, religious differences are very small, and some household capital stock variables have a weak positive effect on heights, even after the categorical wealth variable is accounted for.

The `summary(fit)` command also produces F-tests of the joint significance of the nonparametric components, but we will defer the details of these calculations. A further issue regarding these nonparametric components would be the transition from the pointwise confidence bands that we have described above to uniform bands. This topic has received quite a lot of attention in recent years, although the early work of Hotelling (1939) has been crucial. Recent work by Krivobokova, Kneib, and Claeskens (2009) has shown how to adapt the Hotelling approach for the some GAM models in the Wood `mgcv` package. It appears that similar methods can be adapted to `rqss` fitting; I hope to report on this in future work.

References

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.

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

——— (2009): "quantreg: A Quantile Regression Package for R," `http://cran.r-project.org/src/contrib/PACKAGES.html#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 (2009): "Simultaneous Confidence Bands for Penalized Spline Estimators," preprint.

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

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*, forthcoming.

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.

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

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

——— (2009): "mgcv: GAMs with GCV/AIC/REML smoothness estimation and GAMMs by PQL," `http://cran.r-project.org/src/contrib/PACKAGES.html#mgcv`.