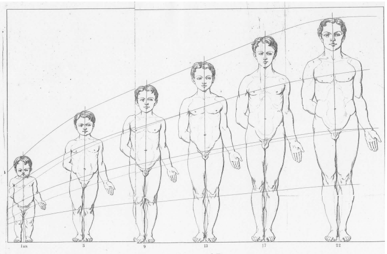
Nonparametric Quantile Regression

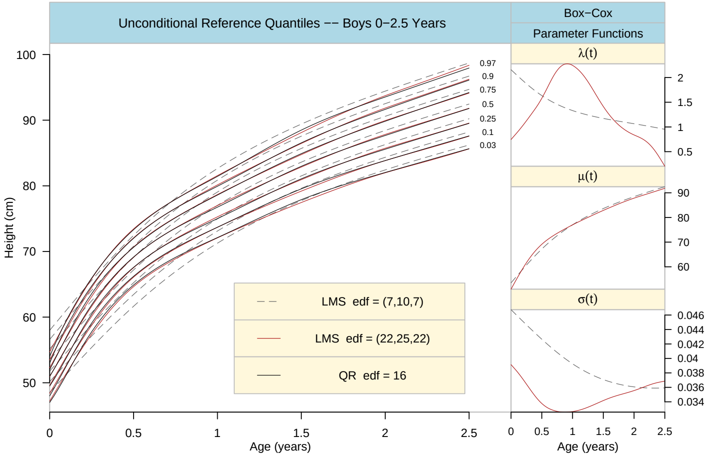
Roger Koenker

University of Illinois, Urbana-Champaign

University of Minho, 12-14 June 2017



In the Beginning, . . . were the Quantiles



Pere, Wei, He, and K *Stat. in Medicine* (2006)

Three Approaches to Nonparametric Quantile Regression

- Locally Polynomial (Kernel) Methods: `lprq(x, y, h`

$$\hat{\alpha}(\tau, x) = \text{argmin} \sum_{i=1}^{n} \rho_\tau(y_i - \alpha_0 - \alpha_1(x_i - x) - \ldots - \frac{1}{p!}\alpha_p(x_i - x)^p)$$
$$\hat{g}(\tau, x) = \hat{\alpha}_0(\tau, x)$$

- Series Methods `rq( y ~ bs(x,knots = k) + z)`

$$\hat{\alpha}(\tau) = \text{argmin}_\alpha \sum_{i=1}^{n} \rho_\tau(y_i - \sum_j \varphi_j(x_i)\alpha_j - z\beta)$$
$$\hat{g}(\tau, x) = \sum_{j=1}^{p} \varphi_j(x)\hat{\alpha}_j$$

- Penalty Methods `rqss(y ~ qss(x, lambda = 3.14) + z)`

$$\hat{g}(\tau, x) = \text{argmin}_g \sum_{i=1}^{n} \rho_\tau(y_i - g(x_i) - z\beta) + \lambda P(g)$$

Total Variation Regularization I

There are many possible penalties, ways to measure the roughness of fitted function, but total variation of the first derivative of g is particularly attractive:

$$P(g) = V(g') = \int |g''(x)|dx$$

As $\lambda \to \infty$ we constrain g to be closer to linear in x. Solutions of

$$\min_{g \in \mathcal{G}} \sum_{i=1}^{n} \rho_\tau(y_i - g(x_i)) + \lambda V(g')$$

are continuous and piecewise linear.

Example 1: Fish in a Bottle

Objective: to study metabolic activity of various fish species in an effort to better understand the nature of the feeding cycle. Metabolic rates based on oxygen consumption as measured by sensors mounted on the tubes.

Three primary aspects are of interest:

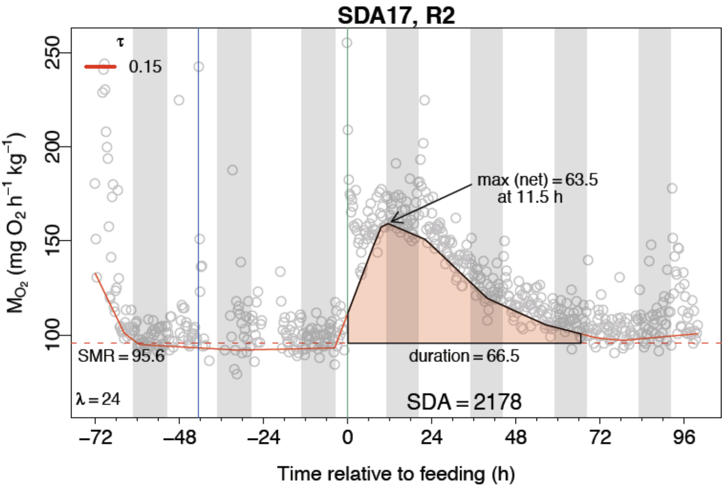
1. Basal (minimal) Metabolic Rate,
2. Duration and Shape of the Feeding Cycle, and
3. Diurnal Cycle.

Example 1: Some Experimental Details

Experimental data of Denis Chabot, Institut Maurice-Lamontagne, Quebec, Canada and his colleagues.

1. Basal (minimal) metabolic rate $\mathrm{M_{O_2}}$ (aka Standard Metabolic Rate SMR) is measured in $\mathrm{mg\ O_2\ h^{-1}\ kg^{-1}}$ for fish "at rest" after several days without feeding,
2. Fish are then fed and oxygen consumption monitored until $\mathrm{M_{O_2}}$ returns to its prior SMR level for several hours.
3. Elevation of $\mathrm{M_{O_2}}$ after feeding (aka Specific Dynamic Action SDA) ideally measures the energy required for digestion,
4. Procedure is repeated for several cycles, so each estimation of the cycle is based on a few hundred observations.

Example 1: Juvenile Codfish



Tuning Parameter Selection

There are two tuning parameters:

1. $\tau = 0.15$ the (low) quantile chosen to represent the SMR,
2. λ controls the smoothness of the SDA cycle.

One way to interpret the parameter λ is to note that it controls the number of effective parameters of the fitted model (Meyer and Woodroofe(2000):

$$p(\lambda) = \text{div } \hat{g}_{\lambda,\tau}(y_1, ..., y_n) = \sum_{i=1}^{n} \partial \hat{y}_i / \partial y_i$$

This is equivalent to the number of interpolated observations, the number of zero residuals. Selection of λ can be made by minimizing, e.g. Schwarz Criterion:

$$\text{SIC}(\lambda) = n \log(n^{-1} \sum \rho_\tau(y_i - \hat{g}_{\lambda,\tau}(x_i))) + \frac{1}{2} p(\lambda) \log n.$$

Total Variation Regularization II

For bivariate functions we consider the analogous problem:

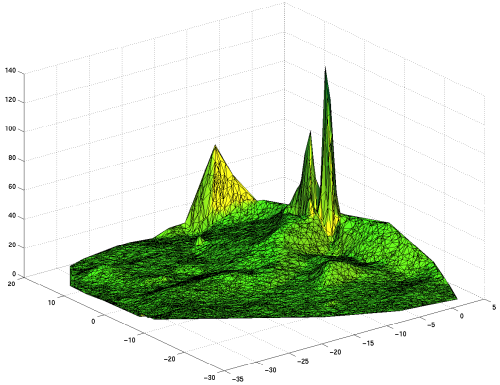
$$\min_{g \in \mathcal{G}} \sum_{i=1}^{n} \rho_\tau(y_i - g(x_{1i}, x_{2i})) + \lambda V(\nabla g)$$

where the total variation variation penalty is now:

$$V(\nabla g) = \int \|\nabla^2 g(x)\| dx$$

Solutions are again continuous, but now they are piecewise linear on a triangulation of the observed x observations. Again, as $\lambda \to \infty$ solutions are forced toward linearity.

Example 2: Chicago Land Values via TV Regularization



Chicago Land Values: Based on 1194 vacant land sales and 7505 "virtual" sales introduced to increase the flexibility of the triangulation. K and Mizera (2004).

Additive Models: Putting the pieces together

We can combine such models:

$$\min_{g \in \mathcal{G}} \sum_{i=1}^{n} \rho_\tau(y_i - \sum_j g_j(x_{ij})) + \sum_j \lambda_j V(\nabla g_j)$$

- Components g_j can be univariate, or bivariate.
- Additivity is intended to muffle the curse of dimensionality.
- Linear terms are easily allowed, or enforced.
- And shape restrictions like monotonicity and convexity/concavity as well as boundry conditions on g_j's can also be imposed.

Implementation in the R quantreg Package

- Problems are typically large, very sparse linear programs.
- Optimization via interior point methods are quite efficient,
- Provided sparsity of the linear algebra is exploited, quite large problems can be estimated.
- The nonparametric qss components can be either univariate, or bivariate
- Each qss component has its own λ specified
- Linear covariate terms enter formula in the usual way
- The qss components can be shape constrained.

```
fit <- rqss(y ~ qss(x1,3) + qss(x2,8) + x3, tau = .6)
```

Pointwise Confidence Bands

It is obviously crucial to have reliable confidence bands for nonparametric components. Following Wahba (1983) and Nychka(1983), conditioning on the λ selection, we can construct bands from the covariance matrix of the full model:

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

with

$$\tilde{X} = \begin{bmatrix} X & G_1 & \cdots & G_J \\ \lambda_0 H_K & 0 & \cdots & 0 \\ 0 & \lambda_1 P_1 & \cdots & 0 \\ \vdots & \dots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_j P_J \end{bmatrix} \quad \text{and} \quad \Psi = \text{diag}(\phi(\hat{u}_i/h_n)/h_n)$$

Pointwise bands can be constructed by extracting diagonal blocks of V.

Uniform Confidence Bands

Uniform bands are also important, but more challenging. We would like:

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

such that the true curve, g_0, is covered with specified probability $1 - \alpha$ over a given domain $\mathcal{X}$:

$$\mathcal{P}\{g_0(x) \in B_n(x) \mid x \in \mathcal{X}\} \geqslant 1 - \alpha.$$

We can follow the "Hotelling tube" approach based on Hotelling(1939) and Weyl (1939) as developed by Naiman (1986), Johansen and Johnstone (1990) Sun and Loader (1994) and others.

Uniform Confidence Bands

Hotelling's original formulation for parametric nonlinear regression has been extended to non-parametric regression. For series estimators

$$\hat{g}_n(x) = \sum_{j=1}^{p} \varphi_j(x)\hat{\theta}_j$$

with pointwise standard error $\sigma(x) = \sqrt{\varphi(x)^\top V^{-1}\varphi(x)}$ we would like to invert test statistics of the form:

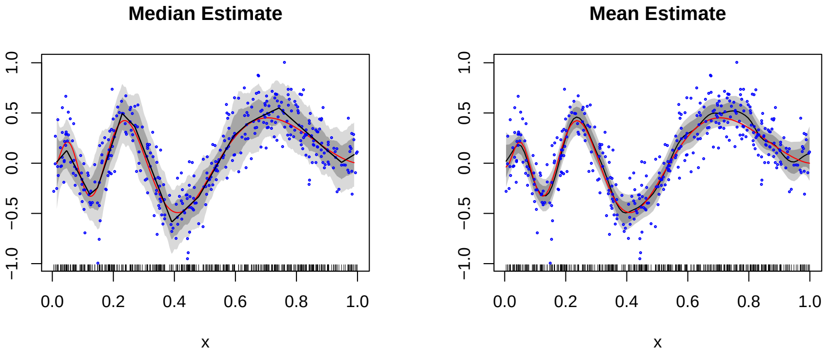
$$T_n = \sup_{x \in \mathcal{X}} \frac{\hat{g}_n(x) - g_0(x)}{\sigma(x)}.$$

This requires solving for the critical value, c_α in

$$\mathcal{P}(T_n > c) \leqslant \frac{\kappa}{2\pi}(1 + c^2/\nu)^{-\nu/2} + \mathcal{P}(t_\nu > c) = \alpha$$

where κ is the length of a "tube" determined by the basis expansion, t_ν is a Student random variable with degrees of freedom $\nu = n - p$.

Confidence Bands in Simulations



$$Y_i = \sqrt{x_i(1-x_i)} \sin\left(\frac{2\pi(1+2^{-7/5})}{x_i+2^{-7/5}}\right) + U_i, \quad i = 1, \cdots, 400, \quad U_i \sim \mathcal{N}(0, 0.04)$$

,

Simulation Performance

	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

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

Simulation Performance

	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

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

Example 3: Childhood Malnutrition in India

A larger scale problem illustrating the use of these methods is a model of risk factors for childhood malnutrition considered by Fenske, Kneib and Hothorn (2009).

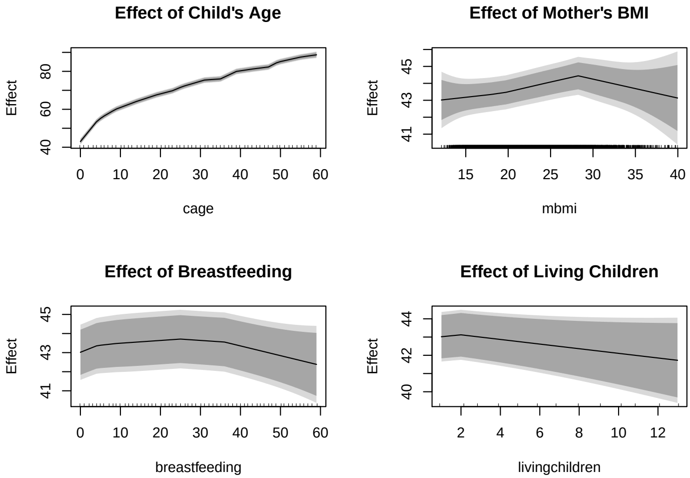
- They motivate the use of models for low conditional quantiles of height as a way to explore influences on malnutrition,
- They employ boosting as a model selection device,
- Their model includes six univariate nonparametric components and 15 other linear covariates.
- There are 37,623 observations on the height of children from India.

Example 3: R Formulation

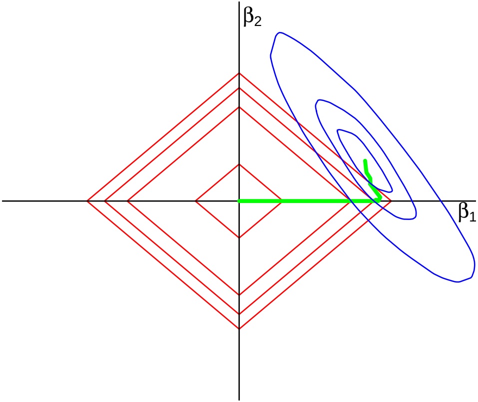
```
fit <- rqss(cheight ~ qss(cage, lambda = lam[1]) +
qss(bfed, lambda = lam[2]) + qss(mage, lambda = lam[3]) +
qss(mbmi, lambda = lam[4]) + qss(sibs, lambda = lam[5]) +
qss(medu, lambda = lam[6]) + qss(fedu, lambda = lam[7]) +
csex + ctwin + cbirthorder + munemployed + mreligion +
mresidence + deadchildren + wealth + electricity +
radio + television + frig + bicycle + motorcycle + car +
tau = 0.10, method = "lasso", lambda = lambda, data = india)
```

- The seven coordinates of `lam` control the smoothness of the nonparametric components,
- `lambda` controls the degree of shrinkage in the linear (lasso) coefficients.
- The estimated model has roughly 40,000 observations, including the penalty contribution, and has 2201 parameters.
- Fitting the model for a single choice of λ's takes approximately 5 seconds.

Example 3: Selected Smooth Components



Example 3: Lasso Shrinkage of Linear Components



Lasso λ Selection – Another Approach

Lasso shrinkage is a special form of the TV penalty:

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

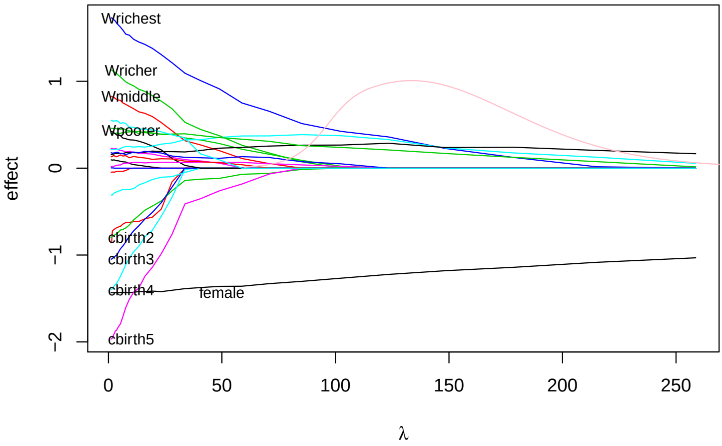
$$\begin{aligned}
\hat{\beta}_{\tau,\lambda} &= \text{argmin}\{R_\tau(b) + \lambda\|b\|_1\} \\
&\in \{b : \ 0 \in \partial R_\tau(b) + \lambda\partial\|b\|_1\}.
\end{aligned}$$

At the true parameter, $\beta_0(\tau)$, we have the pivotal statistic,

$$\begin{aligned}
\partial R_\tau(\beta_0(\tau)) &= \sum (\tau - I(F_{y_i}(y_i) \leqslant \tau))x_i \\
&\sim \sum (\tau - I(U_i \leqslant \tau))x_i
\end{aligned}$$

Proposal: (Belloni and Chernozhukov (2009)) Choose λ as the $1-\alpha$ quantile of the simulated distribution of $\|\sum(\tau - I(U_i \leqslant \tau))x_i\|_\infty$ with iid $U_i \sim U[0,1]$.

Example 3: Lasso Shrinkage of Linear Components



Conclusions

- Nonparametric specifications of $Q(\tau|x)$ improve flexibility.
- Additive models keep effective dimension in check.
- Total variation roughness penalties are natural.
- Schwarz model selection criteria are useful for λ selection
- Hotelling tubes are useful for uniform confidence bands
- Lasso Shrinkage is useful for parametric components.