Additive Models for Quantile Regression: Model Selection and Confidence Bands

Roger Koenker

University of Illinois, Urbana-Champaign

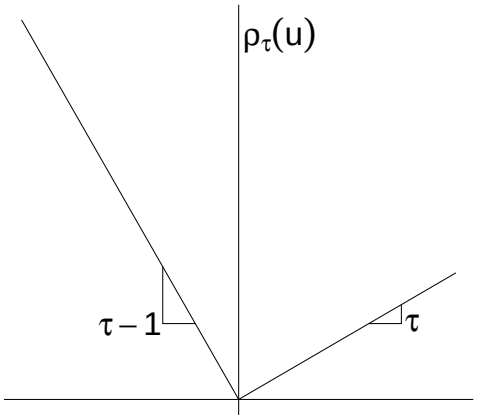
Einaudi Institute
Rome, 10 September 2012

Sample Quantiles via Optimization

Ordinary sample quantiles can be easily computed (without sorting) by optimizing:

$$\min_{\xi \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - \xi)$$

where $\rho_\tau(u) = u \cdot (\tau - I(u < 0))$



Linear Quantile Regression

Linear (in parameters) conditional quantile functions can be estimated by:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top \beta)$$

where $\rho_\tau(u) = u \cdot (\tau - I(u < 0))$ denotes the same "check" function.

- Median solutions minimize sums of absolute errors.
- Some inherent robustness since only signs of residuals matter.
- Solutions efficiently computed via linear programming.
- Solutions interpolate p points when there are p parameters.

Nonparametric Quantile Regression

There are several approaches to estimating conditional quantile functions nonparametrically:

- Inversion of some form of local conditional distribution function estimators: Peracchi (2002), Matzkin (2003), Komunjer and Vuong (2006), Imbens and Newey (2009)
- Locally polynomial weighting: Chaudhuri (1991), Welsh (1996), Horowitz and Lee (2005), Spokoiny, Wang and Härdle (2012), ...
- Series/Sieve estimation: Shen, Shi and Wong (1999), Wei and He (2006), Chen (2007)
- Penalty methods: K, Ng and Portnoy (1994), Bosch, Ye and Woodworth (1995), K and Mizera (2004)

Penalized Quantile Regression

Non-parametric conditional quantile functions can be estimated by solving:

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

where P denotes a penalty term designed to control the roughness of the fitted function $\hat{g}$.

Penalized Quantile Regression

Non-parametric conditional quantile functions can be estimated by solving:

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

where P denotes a penalty term designed to control the roughness of the fitted function $\hat{g}$.
Typically, in fitting conditional mean models, Wahba, Reinsch, etc.,

$$P(g) = \int (g''(x))^2 dx$$

or some more exotic Sobolev form, as e.g. Ramsay and Silverman (2005).

Total Variation Regularization I

There are many possible penalties, ways to measure the roughness of fitted functions, but total variation of the first derivative of g is particularly convenient in the context of quantile regression:

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

As $\lambda \to \infty$ we force $\hat{g}$ to be more nearly 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 (K, Ng and Portnoy (Biometrika, 1994)). This is a natural analogue of the classical (Wahba) $\mathcal{L}_2$ smoothing spline, and a Lasso penalty *avant la lettre*.

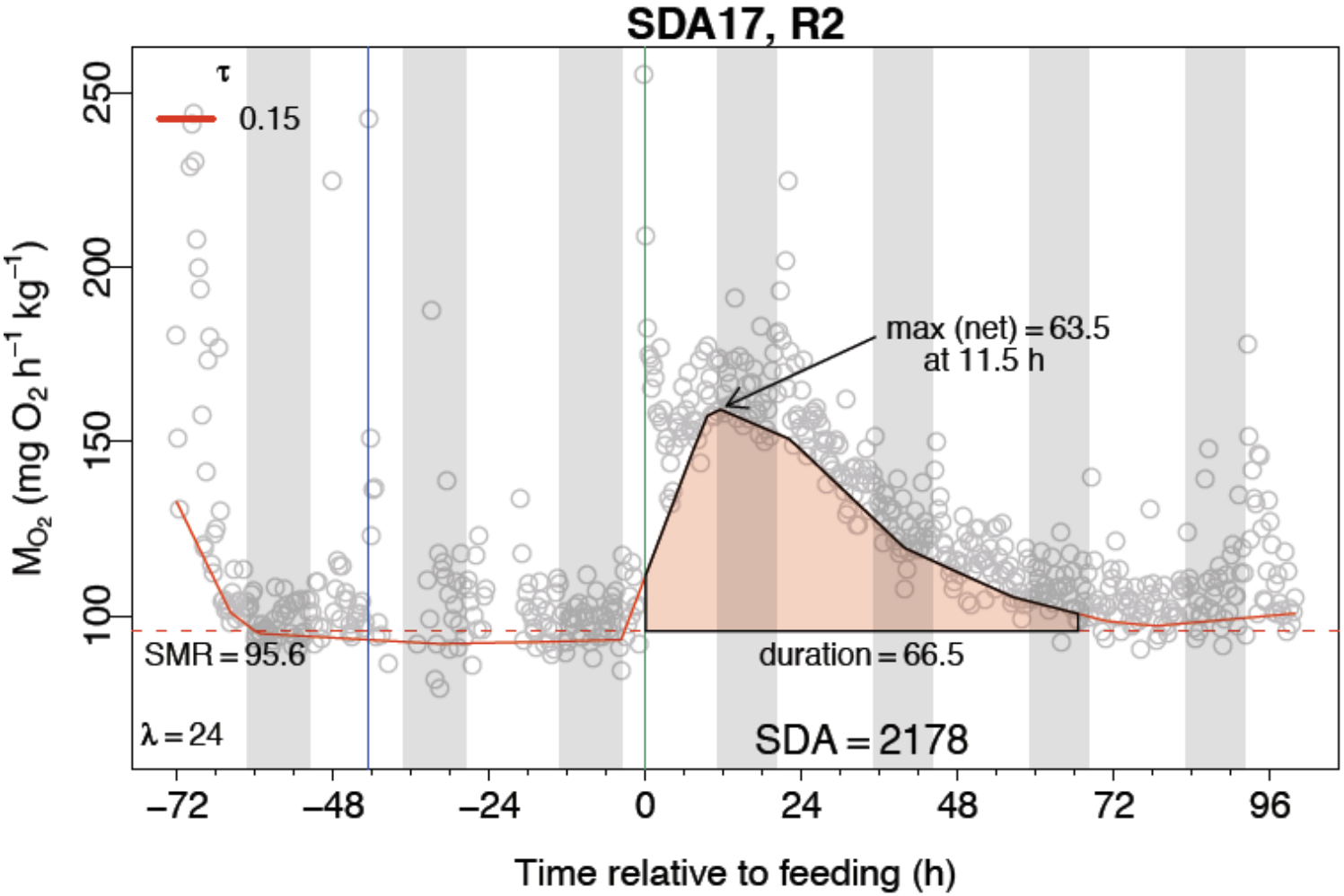
Fish in a Bottle

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

Three primary aspects are of interest:

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

Juvenile Codfish



Experimental data from Denis Chabot, Institut Maurice-Lamontagne, Quebec, Canada.

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:

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

See e.g. Machado (ET, 1993), Li and Zhu (JGCS, 2008), Xu and Ying (AISM, 2010).

Total Variation Regularization II

For bivariate functions we consider the analogous problem:

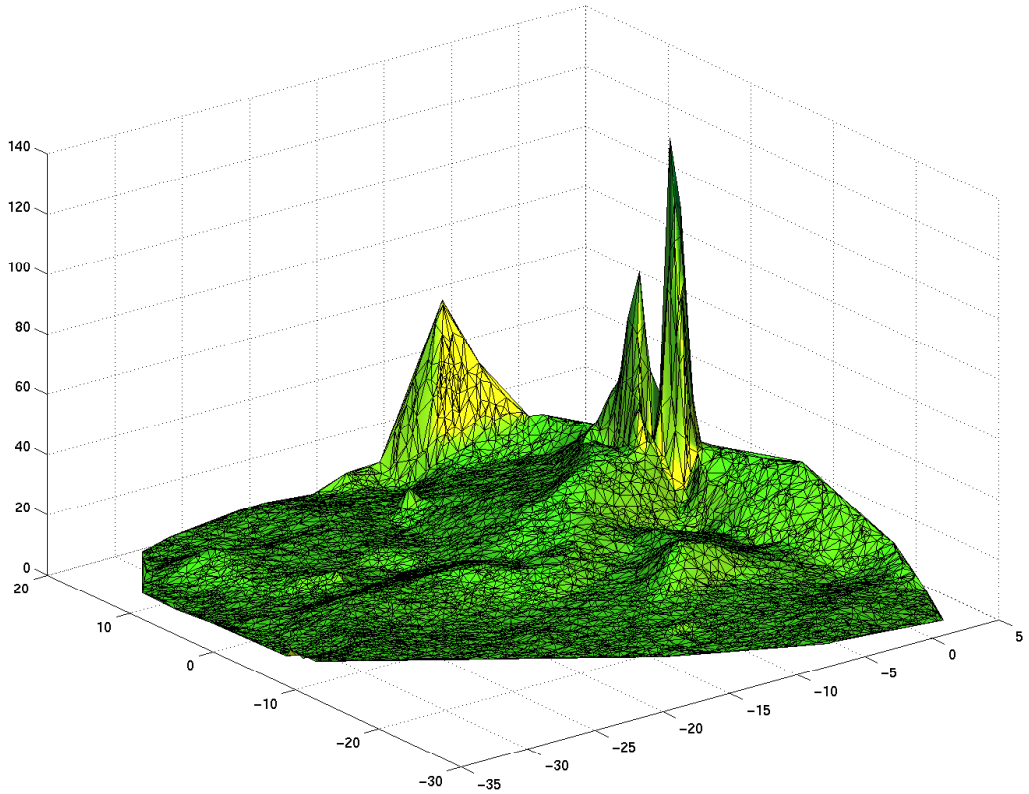
$$\min_{g \in \mathcal{G}} \sum_{i=1}^{n} \rho_\tau(y_i - g(x_{1i}, x_{2i})) + \lambda V(\nabla g)$$

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 x observations, for $\|\cdot\|$, the Hilbert-Schmidt norm. Again, as $\lambda \to \infty$ solutions are forced toward linearity.

Chicago Land Values via TV Regularization



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

Additive Models: Putting the pieces together

We can now 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 penalized by Lasso
- Shape restrictions like monotonicity and convexity/concavity as well as boundry conditions on g_j's can also be easily imposed.

Implementation in the R quantreg Package

- Problems typically yield large, very sparse linear programs.
- Optimization via interior point methods are quite efficient,
- Exploiting sparsity of the linear algebra problems scale well,
- Nonparametric qss components can be either univariate, or bivariate
- Each qss component has its own λ,
- Linear covariate terms enter formula in the usual way,
- The qss components can be shape constrained, monotone, convex,

```
fit <- rqss(y ~ qss(x1,lambda = 3, constraint = "I") +
            qss(x2,lambda = 8) + x3, tau = .15)
```

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 pointwise bands from the covariance matrix of the full model:

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

with $\Psi = \text{diag}(f_{Y_i|x_i}(F^{-1}_{Y_i|x_i}))$,

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

Bands for the nonparametric additive components 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 follow the "Hotelling tube" approach initiated by Hotelling(1939) and Weyl (1939) and developed by Naiman (1986), Siegmund and Knowles (1988), Johansen and Johnstone (1990) Sun and Loader (1994), Krivobokova, Kneib and Claeskens (2010), and others.

Uniform Confidence Bands

As in Krivobokova, Kneib and Claeskens (2010) we have a fitted component,

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

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

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

The Hotelling approach requires a critical value, c_α solving

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

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

Digression on Hotelling Tubes Construction

Suppose that we have an "partially linear model" of the form:

$$Y_i = z_i^\top \alpha + \lambda_i(\tau)\beta + \varepsilon_i$$

with parameters, α, β, τ. The functions, $\lambda_i(\tau)$ are something like Box-Cox transformations of an observable covariate, e.g. $\lambda_i(\tau) = (x_i^\tau - 1)/\tau$.

Digression on Hotelling Tubes Construction

Suppose that we have an "partially linear model" of the form:

$$Y_i = z_i^\top \alpha + \lambda_i(\tau)\beta + \varepsilon_i$$

with parameters, α, β, τ. The functions, $\lambda_i(\tau)$ are something like Box-Cox transformations of an observable covariate, e.g. $\lambda_i(\tau) = (x_i^\tau - 1)/\tau$.
Our task is to test the hypothesis:

$$H_0 : \beta = 0$$

based on the likelihood ratio statistic,

$$L = \inf_\tau \sum (Y_i - \hat{\beta}_\tau \lambda_i(\tau))^2 / \sum Y_i^2$$

after possible preliminary projection to remove α and an abuse of notation.

Hotelling Tubes and the Glorified Cosine

Given τ, we have $\hat{\beta}_\tau = Y^\top \lambda(\tau) / \parallel \lambda(\tau) \parallel^2$ so the likelihood ratio is,

$$\begin{aligned} L &= \inf_\tau \parallel Y - \lambda(\tau)\hat{\beta}_\tau \parallel^2 / \parallel Y \parallel^2 \\ &= \inf_\tau \parallel Y \parallel^{-2} (\parallel Y \parallel^2 - 2(Y^\top \lambda)^2 / \parallel \lambda \parallel^2 + (Y^\top \lambda)^2 / \parallel \lambda \parallel^2) \\ &= 1 - \sup_\tau \left(\frac{\lambda(\tau)^\top Y}{\parallel \lambda(\tau) \parallel \parallel Y \parallel} \right)^2 \\ &\equiv 1 - \sup_\tau (\gamma(\tau)^\top U)^2 \end{aligned}$$

Under the null $U = Y / \parallel Y \parallel$ is uniformly distributed on the sphere S^{n-1} and $\gamma(\tau) = \lambda(\tau) / \parallel \lambda(\tau) \parallel$ is a curve in S^{n-1}.

Hotelling Tubes as Rejection Regions

The test rejects when $W = \sup_\tau \gamma(\tau)^\top U$ exceeds some critical value $w = \cos\theta$ which is equivalent to

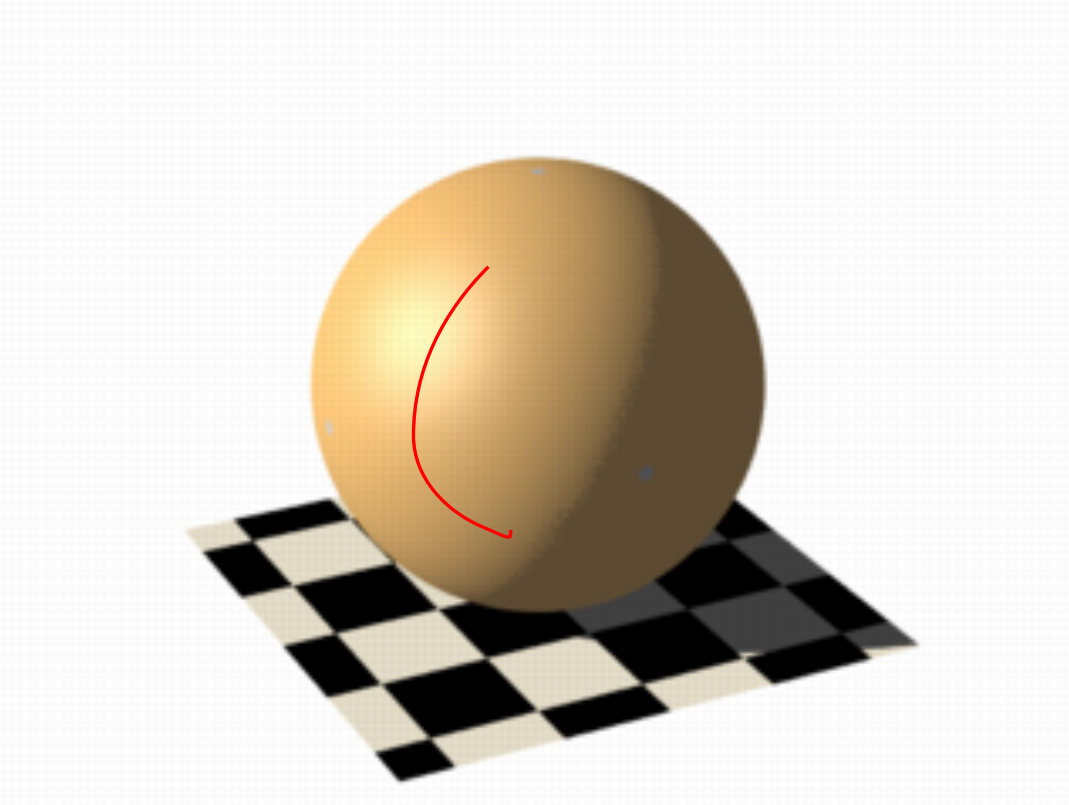
$$\begin{aligned} \{U \in \gamma^\theta\} &= \{u \in S^{n-1} : \sup_t u^\top \gamma(t) \geqslant \cos\theta\} \\ &= \{u \in S^{n-1} : d(u,\gamma) \leqslant (2(1-w))^{1/2}\} \end{aligned}$$

The distance $d(u,\gamma)$ is called the "angular or geodesic radius θ about γ:"

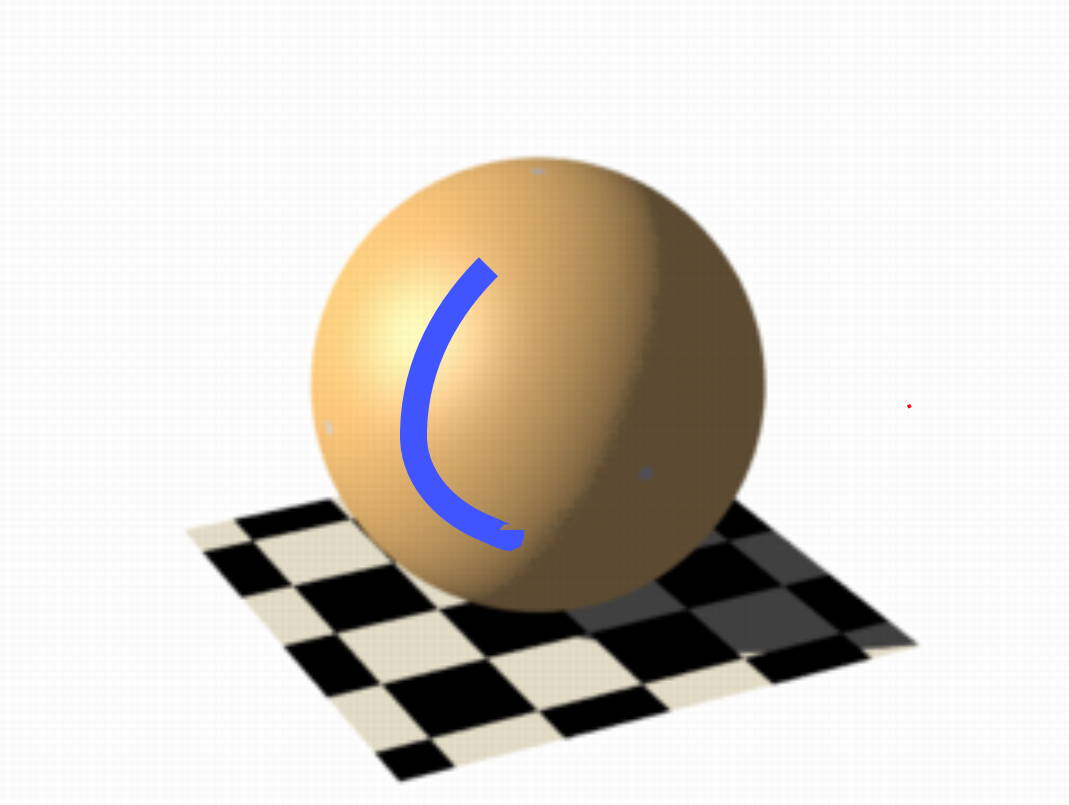
$$\begin{aligned} d^2(u,\gamma) &= \sin^2(\theta) + (1-\cos(\theta))^2 \\ &= 2(1-\cos\theta). \end{aligned}$$

So when the distance is small, U falls inside the tube, and we reject. Given the uniformity of U on the sphere it is (relatively) easy to chose the radius of the tube.

The curve $\gamma(t)$ on $\mathcal{S}^2$.



The tube γ^θ on $\mathcal{S}^2$.



Tubes for Sieves

Now suppose that we have the nonparametric sieve model,

$$Y_i = \sum_{j=1}^{d} \beta_j a_j(t_i) + \varepsilon_i$$

with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $t \in \mathcal{T} \subset \mathbb{R}$. Our objective is to find a positive c such that

$$P_{\beta,\sigma,\Sigma}(|\beta^\top a(t) - \hat{\beta}^\top a(t)| \leqslant c\sigma(a(t)^\top \Sigma a(t))^{1/2} \text{ for all } t \in \mathcal{T}) \approx 1 - \alpha$$

uniformly in β, σ. That is, we have the test statistic,

$$T = \sup_{a \in C} \frac{a^\top(\hat{\beta} - \beta)}{\sqrt{a^\top \Sigma a}} \equiv T(X, \xi) = \sup_{a \in C} \frac{a^\top(X - \xi)}{\sqrt{a^\top \Sigma a}}$$

where $X \sim \mathcal{N}(\xi, \Sigma)$. We'd like to make a confidence statement about $\{a^\top \xi | a \in C\}$ when C is some sort 1-dimensional "curve."

Tubes for Sieves (2)

Write $T = RW$ where $R^2 = (X - \xi)^\top \Sigma^{-1}(X - \xi) \sim \chi^2_d$ and

$$\begin{aligned} W &= \sup_{a \in C} \frac{a^\top (X - \xi)}{\sqrt{a^\top \Sigma a}\sqrt{(X - \xi)\Sigma^{-1}(X - \xi)}} \\ &= \sup_{a \in C} \frac{(\Sigma^{1/2} a)^\top \Sigma^{-1/2}(X - \xi)}{\| \Sigma^{1/2} a \| \| \Sigma^{-1/2}(X - \xi) \|} \\ &\equiv U^\top \gamma(a) \end{aligned}$$

Tubes for Sieves (2)

Write $T = RW$ where $R^2 = (X-\xi)^\top \Sigma^{-1}(X-\xi) \sim \chi^2_d$ and

$$\begin{aligned} W &= \sup_{a \in C} \frac{a^\top (X-\xi)}{\sqrt{a^\top \Sigma a}\sqrt{(X-\xi)\Sigma^{-1}(X-\xi)}} \\ &= \sup_{a \in C} \frac{(\Sigma^{1/2}a)^\top \Sigma^{-1/2}(X-\xi)}{\| \Sigma^{1/2}a \| \| \Sigma^{-1/2}(X-\xi) \|} \\ &\equiv U^\top \gamma(a) \end{aligned}$$

So, as before, $\gamma = \gamma(C) \subset S^{d-1}$ and U is uniform on S^{d-1}. R and W don't depend on ξ, Σ or they do, but only via γ. $R^2 \perp\!\!\!\perp W$ and $R^2 \sim \chi^2_d$ so

$$\mathcal{P}(T > c) = \int_c^\infty \mathcal{P}(W > c/r)\mathcal{P}(R \in dr)$$

Tubes for Sieves (3)

The random variable W has the same form as in the previous example so

$$\mathcal{P}(W > w) = \frac{\| \gamma \|}{2\pi}(1 - w^2)^{(d-2)/2} + \frac{1}{2}\mathcal{P}(B \geqslant w^2) \equiv b_\gamma(w)$$

Naiman (1986) suggests the bound

$$\mathcal{P}(T > c) \leqslant \int_c^\infty \min\{b_\gamma(c/r), 1\}\mathcal{P}(R \in dr)$$

and Knowles (1987) suggests ignoring the $b_\gamma < 1$ constraint and integrates explicitly the bound to get,

$$\mathcal{P}(T > c) \leqslant \frac{\| \gamma \|}{2\pi} e^{-c^2/2} + 1 - \Phi(c)$$

where again $\| \gamma \|$ is the length of γ. Finally, we invert to get the critical value c_α.

Simulation Design

All the simulations employ the Wand, Ruppert and Carroll (2003) test function:

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

Three model flavors:

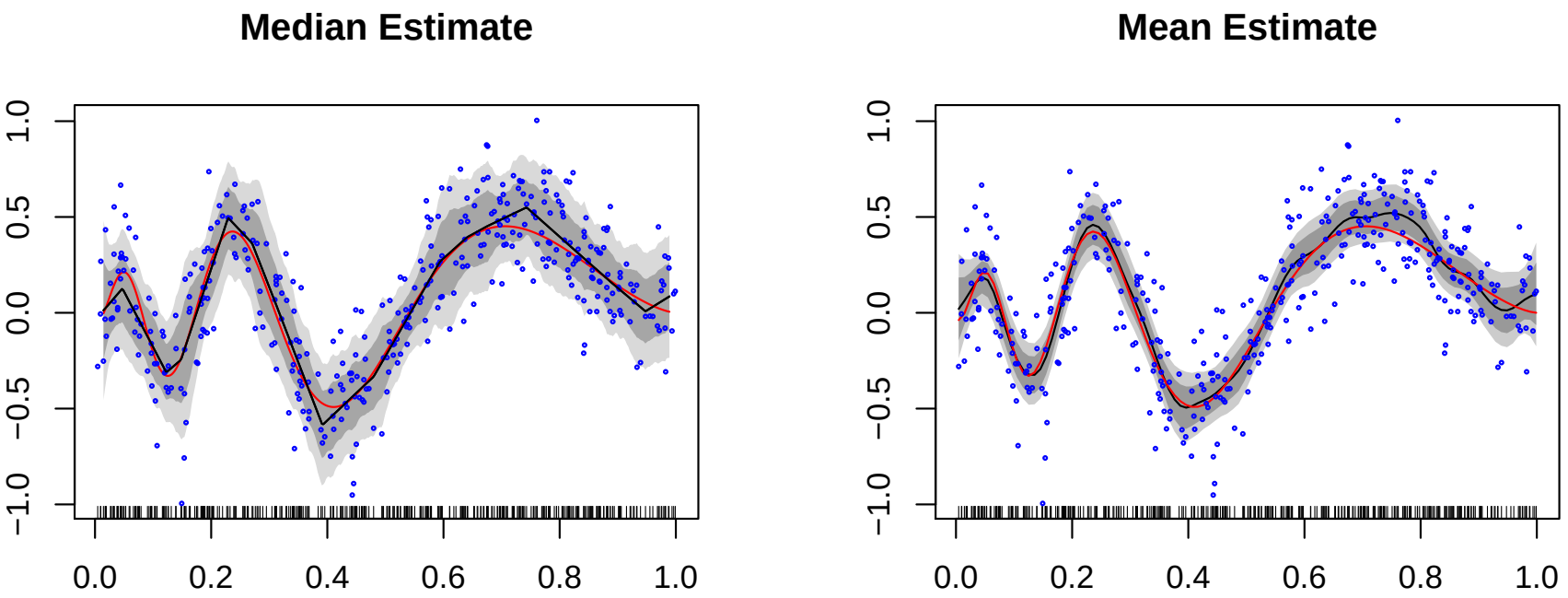
$$\text{iid error} \quad Y_i = g_0(x_i) + \sigma_0 U_i$$

$$\text{linear scale} \quad Y_i = g_0(x_i) + \sigma_0(1+x_i)U_i$$

$$\text{nuisance covariates} \quad Y_i = g_0(x_i) + z_i^\top \gamma + \sigma_0 U_i$$

Sample size: $n = 400$, replications $R = 1000$, $U_i \sim F$, and four choices of the error distribution, $F \in \{\Phi, t_3, t_1, \chi^2_3\}$.

Confidence Bands in Simulations



Mean bands are based on Krivobokova, Kneib and Claeskens (2010), for Simon Wood's `mgcv` fits, median bands based on `quantreg rqss` estimates and SIC λ-selection.

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

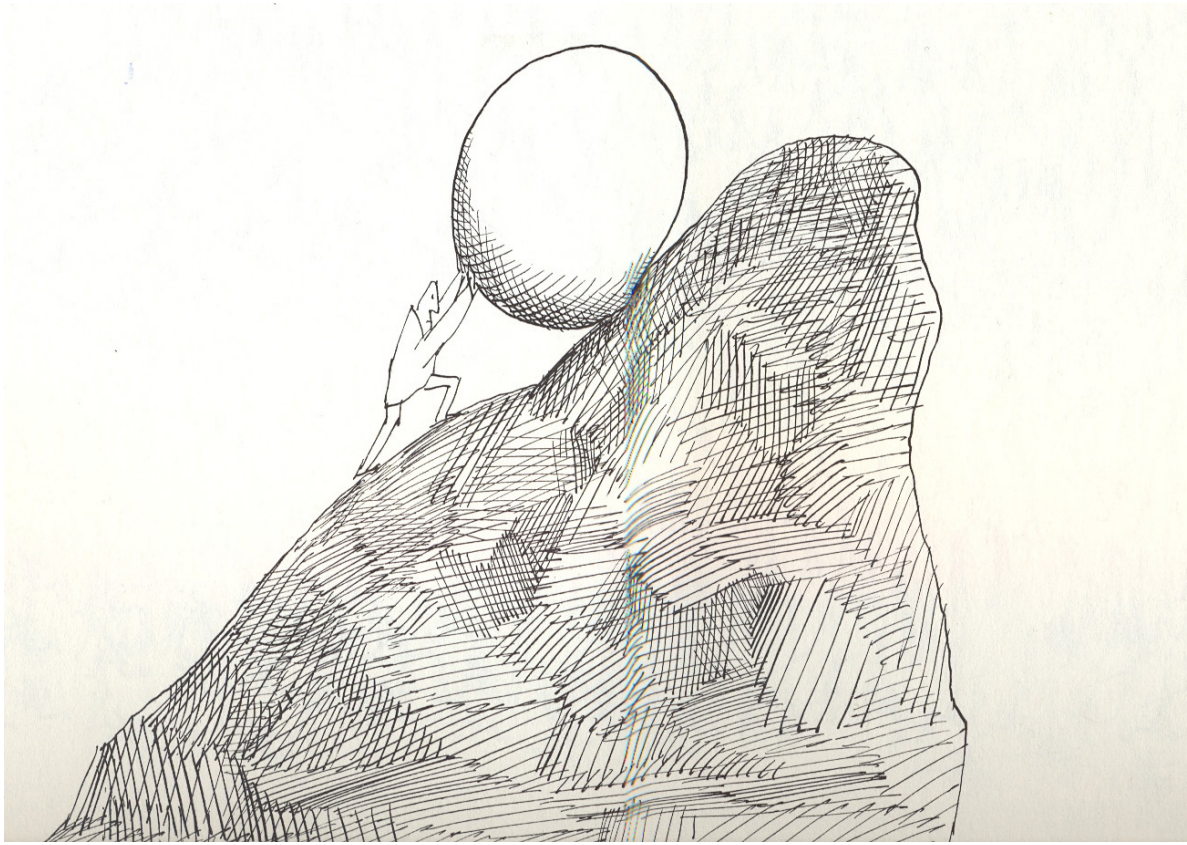
	Accuracy			Pointwise		Uniform	
	RMISE	MIAE	MEDF	Pband	Uband	Pband	Uband
Gaussian							
rqss	0.081	0.063	10.685	0.951	0.998	0.265	0.936
gam	0.064	0.050	17.905	0.957	0.999	0.234	0.940
t_3							
rqss	0.091	0.070	9.612	0.952	0.998	0.241	0.938
gam	0.103	0.078	14.656	0.949	0.992	0.232	0.804
t_1							
rqss	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

Conclusions

- Flexible nonparametric specifications of conditional quantiles,
- Total variation roughness penalties are convenient and natural,
- Additive models keep effective dimension in check,
- Schwarz model selection criteria are useful for λ selection,
- Hotelling tubes are useful for uniform confidence bands,
- Lasso Shrinkage is useful for parametric components.

Sisyphus and the $\mathcal{L}_2$ Ball – Statistics in the 20th Century



Sisyphus and the $\mathcal{L}_1$ Ball – Statistics in the 21th Century