NONLINEAR QUANTILE REGRESSION DESIGN

Linglong Kong[1] and Douglas P. Wiens[1]

- **Model Robustness**. We consider experiments in which the investigator chooses inputs X, and observes an output Z; these and the resulting predictions are related through a, possibly imperfectly specified, parametric model:

  Input (X) $\longrightarrow$ [Parametric model] $\longrightarrow$ Output (Z) $\longrightarrow$ [Predict $Z_{|newX,model}$]

  - Choose design points x_i at which to observe Z; aim for efficiency (small variance when the model is right) and accurate predictions (small biases if the parametric model is wrong).
  - The 'best' design for a slightly wrong model can be much more than slightly sub-optimal. (Box and Draper 1959 etc.)

- **Robust estimation: Quantile Regression**
  - Assume that the τ-quantile of the output Z at input $\boldsymbol{x}$ is a possibly nonlinear function $F(\boldsymbol{x};\boldsymbol{\beta})$:

$$\tau = P_{Z_{|\boldsymbol{x}}}\left(Z \leq F\left(\boldsymbol{x};\boldsymbol{\beta}_\tau\right)\right).$$

  - Estimation is by quantile regression; inherently resistant to y-outliers.
  - More efficient than LSE under non-normal distributions; no moment assumptions made (e.g. Cauchy errors are possible).
  - Provides a satisfying picture of the manner in which the response is affected by the covariates.

- **Example** Dette and Trampisch (('DT') JASA 2012) report an experiment carried out by Cressie and Keightley ('CK') 1979):
  Response Z = amount of estrogen bound to a receptor, x = amount of hormone; Michaelis-Menten response:

$$z = F(x;\boldsymbol{\beta}) = \frac{\beta_1 x}{\beta_2 + x}.$$

- Linear approximation (expand around initial estimate $\boldsymbol{\beta}_0$):

$$(Z - F(\boldsymbol{x};\boldsymbol{\beta}_0) =)\, Y = \boldsymbol{f}_0'(\boldsymbol{x})\,\boldsymbol{\theta} + \text{random error} \qquad (1)$$

  for $\boldsymbol{\theta} = \boldsymbol{\beta} - \boldsymbol{\beta}_0$ and

$$\boldsymbol{f}_0'(x) = \left(\frac{x}{\beta_2 + x}, -\frac{\beta_1 x}{(\beta_2 + x)^2}\right)_{|\boldsymbol{\beta}=\boldsymbol{\beta}_0}. \qquad (2)$$

[1]Department of Mathematical and Statistical Sciences; University of Alberta, Edmonton, Alberta; Canada T6G 2G1. e-mail: lkong@ualberta.ca, doug.wiens@ualberta.ca

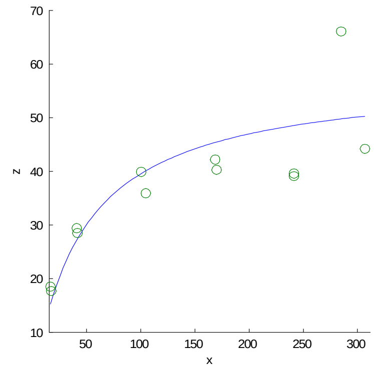

Figure 1: Data gathered by Cressie and Keightley (1979) with least squares response curve $F(x;\boldsymbol{\beta}_0)$ using initial estimate $\boldsymbol{\beta}_0 = (57.98, 46.43)'$.

As 'design space' we take a grid χ of $N = 100$ equally spaced points spanning $[1, 400]$; we will choose $n = 20$, not necessarily distinct, points $x \in \chi$, at which to observe Y.

- See Figure 1. The poor fit suggests a need for robustness of some form.
- The experimenter, acting as though the model is correct and the errors are homoscedastic, computes the quantile regression estimate (see Figure 2)

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{t}} \sum_{i=1}^{n} \rho_\tau \left(Y_i - \boldsymbol{f}_0'(\boldsymbol{x}_i)\, \boldsymbol{t} \right).$$

- The $(Y, \boldsymbol{x}, \boldsymbol{\theta})$ formulation (1) is only an approximation, partly because of the linearizing, and also possibly because the original Michaelis-Menten model may itself have been misspecified, either with respect to the local parameter, or the functional form of the assumed MM-response $F(\boldsymbol{x};\boldsymbol{\beta})$. We suppose that in fact the model is

$$Y_\tau = \boldsymbol{f}_0'(\boldsymbol{x})\, \boldsymbol{\theta}_\tau + \delta_n(\boldsymbol{x}) + \sigma(\boldsymbol{x})\, \varepsilon, \tag{3}$$

for some 'small' model error δ_n. We define the 'true' parameter by

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{t}} \frac{1}{N} \sum_{i=1}^{N} E_{Y|\boldsymbol{x}} \left[\rho_\tau \left(Y - \boldsymbol{f}_0'(\boldsymbol{x}_i)\, \boldsymbol{t} \right) \right]; \tag{4}$$

carrying out this minimization and taking a first order approximation results in the orthogonality of the 'model residuals' $\delta_n(\boldsymbol{x}_i)$ and the regressors:

$$\left(n^{-1/2} g_\varepsilon(0) + O(1) \right) \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{f}_0(\boldsymbol{x}_i) \sqrt{n}\, \delta_n(\boldsymbol{x}_i) = \mathbf{0}. \tag{5}$$

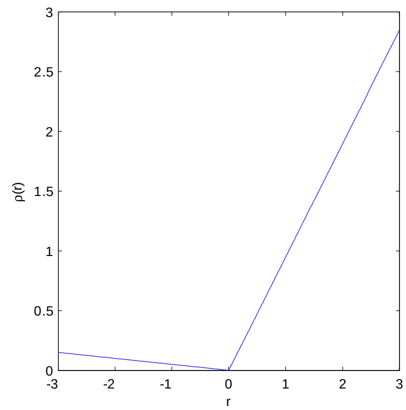

Figure 2: Check function $\rho_\tau(r) = r(\tau - I(r < 0))$; $\tau = .95$.

- We seek designs for (3) which are robust against increased mean squared errors of the predicted conditional quantiles $\hat{Y}_\tau = \boldsymbol{f}'_0(\boldsymbol{x})\hat{\boldsymbol{\theta}}_\tau$:

$$MSE = E\left[\{\text{predicted value} - \text{true value}\}^2\right]$$
$$= E\left[\left\{\hat{Y}_\tau(\boldsymbol{x}_i) - Y_\tau(\boldsymbol{x}_i)\right\}^2\right]$$

engendered by δ_n or by nonconstant $\sigma(\cdot)$.

- For the asymptotics, the effect of δ_n must drop at the same rate as standard error (*Reason*: mse = s.e.^{2}+ bias^2), and so we assume the existence of a bounded limit:

$$\delta_0(\boldsymbol{x}) = \lim_{n\to\infty} \sqrt{n}\delta_n(\boldsymbol{x}), \text{ with } N^{-1}\sum_{i=1}^{N} \delta_0^2(\boldsymbol{x}_i) \le \eta^2, \tag{6}$$

for given η^2. We also impose a bound $N^{-1}\sum_{i=1}^{N}\sigma^2(\boldsymbol{x}_i) \le \sigma_0^2$ for a given σ_0^2 ($= 1$ w.l.o.g.).

- **Optimality and variational mathematics**: In KW we establish asymptotic normality of the estimate $\hat{\boldsymbol{\theta}}_n$, from which we obtain the MSE matrix $\textsc{mse}(\delta_0, \boldsymbol{\sigma})$ of $\hat{\boldsymbol{\theta}}_n$. Our loss function is to be asymptotic, average MSE when the conditional quantile $Y_\tau(\boldsymbol{x}) = \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\theta} + \delta_n(\boldsymbol{x})$, for $\boldsymbol{x} \in \chi$, is incorrectly estimated by $\hat{Y}_n(\boldsymbol{x}) = \boldsymbol{f}'(\boldsymbol{x})\hat{\boldsymbol{\theta}}_n$, i.e.

$$\textsc{amse} = \lim_n \frac{1}{N}\sum_{i=1}^{N} E\left[\left\{\sqrt{n}\left(\hat{Y}_n(\boldsymbol{x}_i) - Y_\tau(\boldsymbol{x}_i)\right)\right\}^2\right].$$

This is evaluated, and then maximized over $\delta_0 \in \Delta_0$ (a class defined by (5) and (6)) using variational methods. In terms of the design measure

$$\xi_i = \text{fraction of observations made at } \boldsymbol{x}_i,$$

and

$$\boldsymbol{A} = N^{-1}\sum_{i=1}^{N}\boldsymbol{f}_0\left(\boldsymbol{x}_i\right)\boldsymbol{f}_0'\left(\boldsymbol{x}_i\right),$$
$$\boldsymbol{B} = \sum_{\xi_i>0}\boldsymbol{f}_0\left(\boldsymbol{x}_i\right)\boldsymbol{f}_0'\left(\boldsymbol{x}_i\right)\left(\frac{\xi_i}{\sigma(\boldsymbol{x}_i)}\right),$$
$$\boldsymbol{S} = \boldsymbol{B}^{-1}\left[\sum_{\xi_i>0}\boldsymbol{f}_0\left(\boldsymbol{x}_i\right)\boldsymbol{f}_0'\left(\boldsymbol{x}_i\right)\xi_i\right]\boldsymbol{B}^{-1},$$
$$\boldsymbol{T} = \boldsymbol{B}^{-1}\left[\sum_{\xi_i>0}\boldsymbol{f}_0\left(\boldsymbol{x}_i\right)\boldsymbol{f}_0'\left(\boldsymbol{x}_i\right)\left(\frac{\xi_i}{\sigma(\boldsymbol{x}_i)}\right)^2\right]\boldsymbol{B}^{-1},$$

we obtain that $\max_{\Delta_0}\text{AMSE}$ is $\frac{\tau(1-\tau)}{g_\varepsilon^2(0)}+\eta^2$ times

$$\mathcal{L}_\nu\left(\xi|\sigma\right) = \left(1-\nu\right)tr\left(\boldsymbol{AS}\right)+\nu ch_{\max}\left(\boldsymbol{AT}\right), \tag{7}$$

where $\nu = \eta^2\Big/\left\{\frac{\tau(1-\tau)}{g_\varepsilon^2(0)}+\eta^2\right\}$.

- The first component $(tr\left(\boldsymbol{AS}\right))$ of $\mathcal{L}_\nu\left(\xi|\sigma\right)$ arises solely from variation – $\boldsymbol{f}_0'\left(\boldsymbol{x}\right)\boldsymbol{S}\boldsymbol{f}_0\left(\boldsymbol{x}\right)$ is the asymptotic variance of $\sqrt{n}\boldsymbol{f}_0'\left(\boldsymbol{x}\right)\hat{\boldsymbol{\theta}}_n = \sqrt{n}\hat{Y}_n\left(\boldsymbol{x}\right)$. A 'classical' (non-robust) design aims to minimize $\mathcal{L}_0$; this is appropriate if one has absolute faith in one's model.
- The second $(ch_{\max}\left(\boldsymbol{AT}\right))$ arises from bias – the asymptotic bias of $\sqrt{n}\boldsymbol{f}'\left(\boldsymbol{x}\right)\hat{\boldsymbol{\theta}}_n$ is

$$\boldsymbol{f}'\left(\boldsymbol{x}\right)\boldsymbol{B}^{-1}\left[\sum_{\xi_i>0}\boldsymbol{f}_0\left(\boldsymbol{x}_i\right)\delta_0\left(\boldsymbol{x}_i\right)\xi_i\right]\left(=\boldsymbol{c}'\left(\boldsymbol{x}\right)\boldsymbol{d},\ \text{say}\right);$$

  this is squared, averaged over χ and maximized over $\boldsymbol{d} = \left(\delta_0\left(\boldsymbol{x}_1\right),...,\delta_0\left(\boldsymbol{x}_N\right)\right)'$. This amounts to maximizing a quadratic form $\boldsymbol{d}'\left[N^{-1}\sum_{i=1}^{N}\boldsymbol{c}\left(\boldsymbol{x}_i\right)\boldsymbol{c}'\left(\boldsymbol{x}_i\right)\right]\boldsymbol{d}$ subject to a bound (from (6)) $\boldsymbol{d}'\boldsymbol{d}\leq N\eta^2$ and a linear constraint

$$\sum_{i=1}^{N}\boldsymbol{f}_0\left(\boldsymbol{x}_i\right)\sqrt{n}\delta_n\left(\boldsymbol{x}_i\right)\sim\left(\boldsymbol{f}_0\left(\boldsymbol{x}_1\right),...,\boldsymbol{f}_0\left(\boldsymbol{x}_N\right)\right)\boldsymbol{d}=\boldsymbol{0}.$$

  This leads to $ch_{\max}\left(\boldsymbol{AT}\right)$.
- We parameterize the designs by $\nu\in[0,1]$, which may be chosen by the experimenter, representing his relative concern for errors due to bias rather than to variation. Once ν is chosen, the designs do not depend upon τ.

• See Figure 3 for a comparative plot of the regressors (2) and the least favourable model error function $\delta_n\left(\boldsymbol{x}\right)=\delta_0\left(\boldsymbol{x}\right)/\sqrt{n}$. These model errors are essentially constant and slightly negative except at the design points.

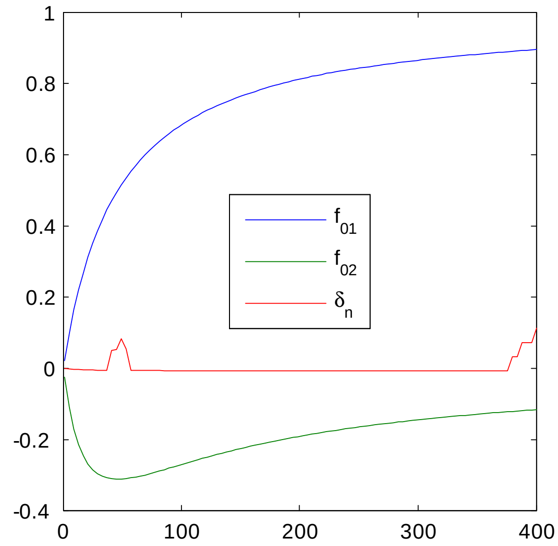

Figure 3: Regressors $\boldsymbol{f}_0(\boldsymbol{x})$ and least favourable model error $\delta_n(\boldsymbol{x})$; $\nu = .1$, $\eta = 1$.

- **Design construction**. We compare five designs – ES (n equally spaced points spanning $\chi = [1, 400]$) and:

KW1 These attain minimax AMSE, i.e. minimize (7), for a particular value of ν; each is assessed for $0 \leq \nu \leq 1$. When $\nu = 0$ the loss is the average variance of the predicted values. The minimization is carried out via a genetic algorithm. Computationally rather intensive.

KW2 We have found designs minimizing the maximum AMSE, with the maximum evaluated not only over δ but also over variance functions $\sigma^2(x_i) \propto \xi_i^r$ for $r \in (-\infty, \infty)$. It turns out that $r = 1$ is least favourable, and that the minimizing design must be supported on n distinct points. These points are found very quickly and simply via an exchange algorithm.

DT1 These 'D-optimal' designs minimize the determinant of the asymptotic covariance matrix of the parameter estimates, assuming homoscedastic errors. They place equal weight on two points, derived explicitly in DT.

DT2 As for DT1, but derived assuming heteroscedastic errors $\sigma(x) \propto 1/F(\boldsymbol{x}; \boldsymbol{\beta}_0)$.

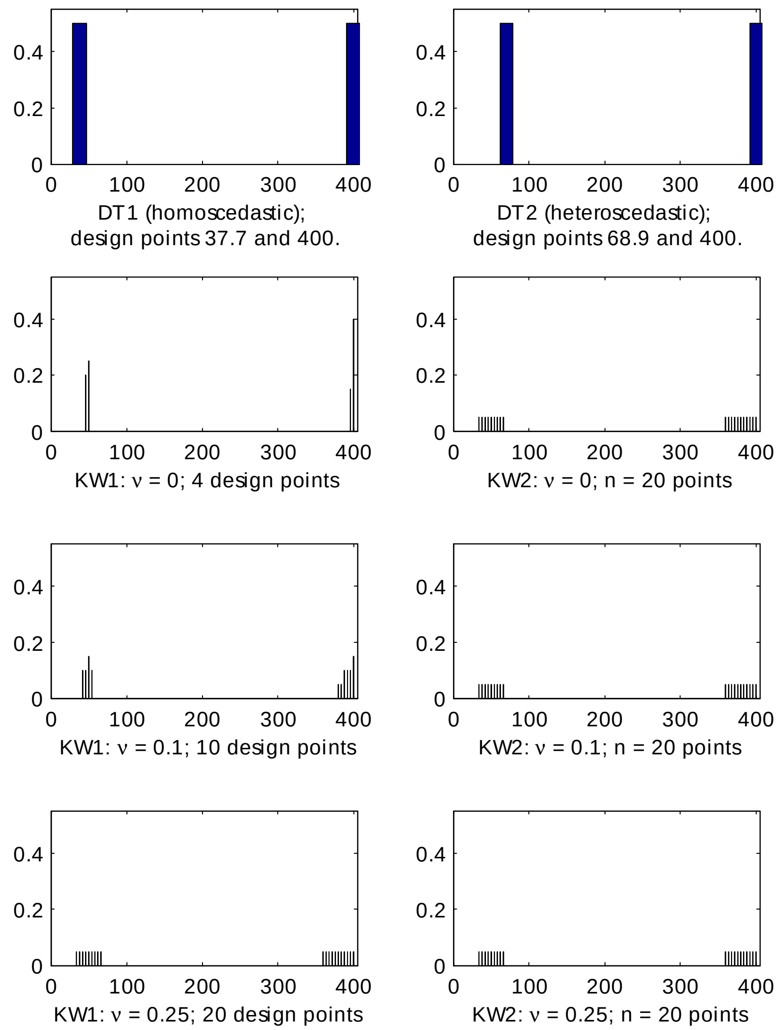

Figure 4: Designs; $n = 20$. Top plots are the replicated designs of DT, constructed assuming homoscedasticity (left) or heteroscedasticity (right). Those of KW are constructed for optimality at the indicated values of ν. Those in the left panel minimize the maximum loss (7), either for constant σ or heteroscedasticity of the specified form $\sigma(x) \propto 1/F(\boldsymbol{x}; \boldsymbol{\beta}_0)$; those in the right panel are the 'no replicate' designs, minimax against heteroscedasticity as well.

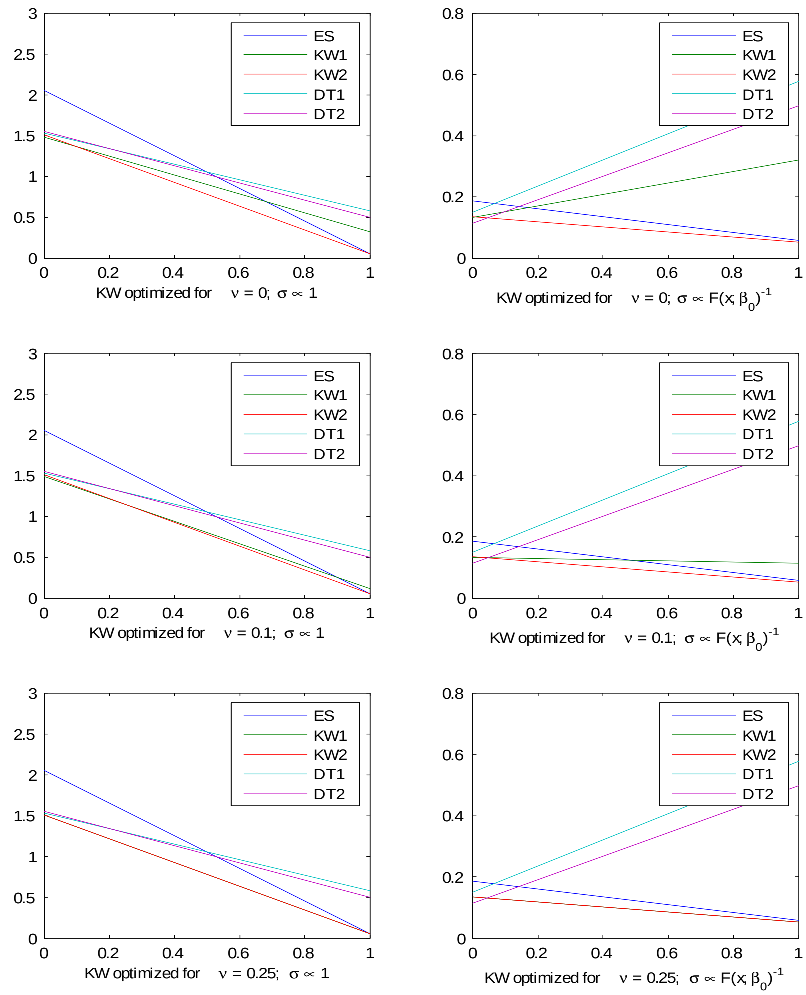

Figure 5: Maximum AMSE vs. ν. Our aim: small loss in efficiency when $\nu = 0$, large gain in robustness when $\nu > 0$. In the left panel the designs are assessed under homoscedasticity, in the right panel they are assessed under heteroscedastcity.

Conclusions and recommendations

- If the model is in doubt, then substantial reductions in MSE can be attained by employing notions of robustness.
- If the 'classically' optimal design is available, then an easy robustification comes about by spreading its replicates into clusters of design points at distinct but nearby locations.
- The very easily constructed n-point designs KW2 always performed at least as well as the more computationally intensive KW1. They were very nearly fully efficient (at $\nu = 0$) and uniformly more robust (when $\nu > 0$). Thus <u>the robustness is obtained at almost no cost in efficiency</u>.

References

Box, G. E. P., and Draper, N. R. (1959), "A Basis for the Selection of a Response Surface Design," *Journal of the American Statistical Association*, 54, 622-654.

Cressie, N. A., and Keightley, D. D. (1979), "The Underlying Structure of the Direct Linear Plot With Application to the Analysis of Hormone-Receptor Interactions", *Journal of Steroid Biochemistry*, 11, 1173-1180.

Dette, H., and Trampisch, M. (2012), "Optimal Designs for Quantile Regression Methods," *Journal of the American Statistical Association*, 107, 1140-1151.

Kong, L., and Wiens, D. P. (2015), " Model-Robust Designs for Quantile Regression," *Journal of the American Statistical Association*, 110, 233-245.