Emulating distributions

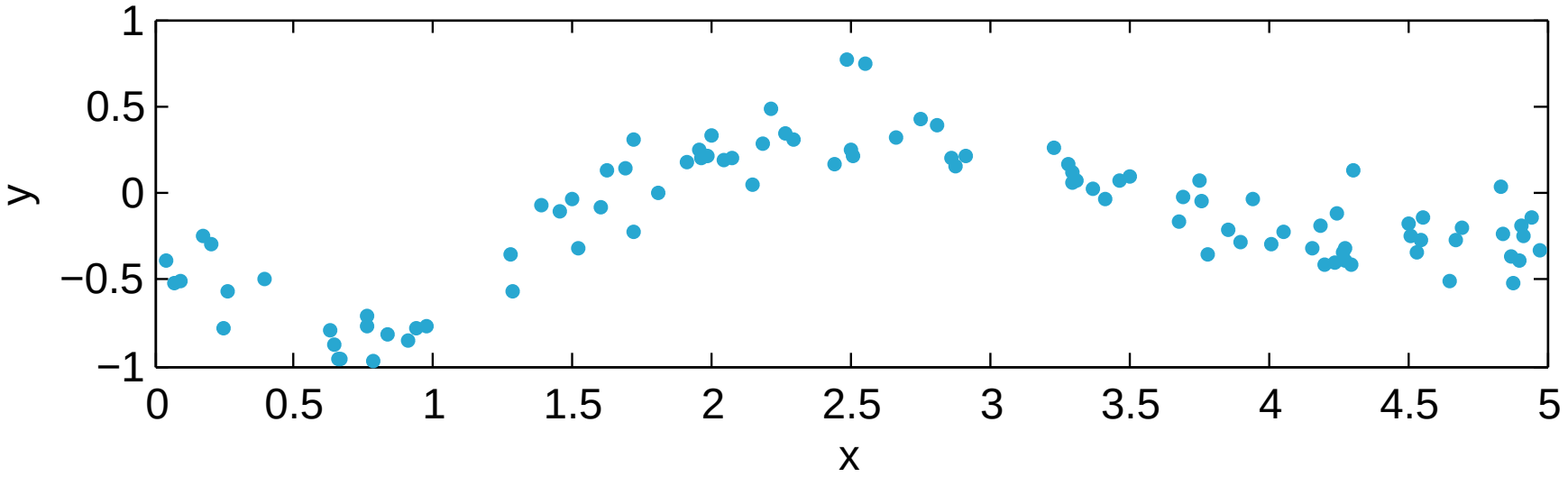
Remi Barillec, Alexis Boukouvalas, Dan Cornford

MUCM meeting, 21-22 September 2011

Outline

Overview of the problem

- We consider a stochastic simulator: $f(x) \sim p(y|x)$
- The simulator is run at a set of inputs $X = (x_1, \ldots, x_N)$ and produces a set of scalar outputs $Y = (y_1, \ldots, y_N)$ where y_i is a sample from $p(y_i|x_i)$



- Due to the variability of the simulator output, it is likely a large number of runs will be required for typical applications (e.g. calibration, ...) ➟ **build an emulator**

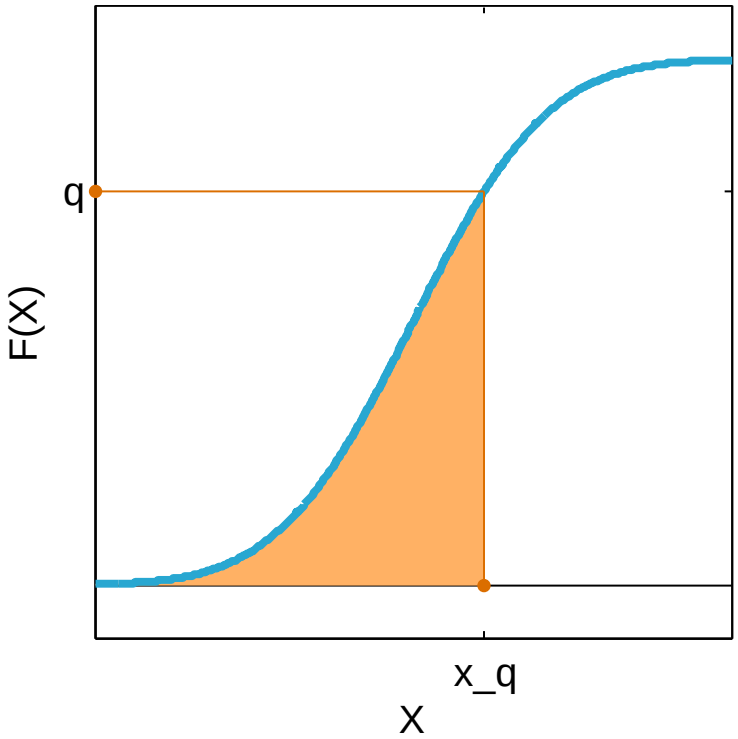
Emulating distributions

- $p(y|x)$ is usually unknown (but we can sample from it)
- We assume that $p(y|x)$ varies smoothly in x
- Ideally, we would like to emulate the full $p(y|x)$ ➟ we need some way of representing a distribution
- Possible options include:
  - Parameteric methods (e.g. we know the distribution shape)
  - Non-parameteric methods (mixture models, histogram approximation, quantile regression)
  - Sample-based methods: Monte-Carlo, ensemble methods, unscented methods

Do we really need the full distribution?

- Do we need to emulate the full $p(y|x)$? Decision making often only requires some summary of the distribution:
  - average value
  - measure of the spread
  - probability of exceeding some threshold
- The most common summaries are the **moments** and the **quantiles**
- Moments might not be informative enough if the distribution is skewed, non-Gaussian or multi-modal
- Quantiles are less restrictive, but harder to work with (if we want to use a GP model)

Quantiles



- The q-th quantile ($q \in [0, 1]$) of a scalar random variable $X \sim P(X)$ can be defined as:

$$x_q \equiv \sup_x \{P(X < x) \leq q\}$$

- Other slightly different definitions exist (infimum of $P(X < x) \geq q$, interpolation for a sample population...)

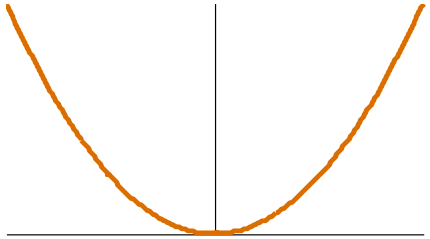
Quantiles

- Quantile regression is commonly used for[1]:
  - **Medical diagnostics:** percentiles can be used to identify unusual health conditions, especially when the reference range depends strongly on the patient's age (e.g. overweight/underweight)
  - **Survival analysis:** to quantify the chance of survival after medical operation (e.g. heart transplant), again as a function of patient age
  - **Finance:** risk assessment (tails of the distribution)
  - **Economics:** study of income distribution, household electricity consumption
  - **Environmental modelling:** distribution of rainfall (with lower quantiles linked to drought and higher quantiles to flooding), pollution...
  - **Detecting heteroscedasticity:** in the homoscedastic case, quantiles should be parallel (uniform variance). Non-parallel quantiles can help identify the heteroscedastic nature of a dataset (or simulator output)

[1] Yu et al. (2003)

How do we estimate quantiles?

- Consider a random variable $Y \sim p(Y)$
- Let's assume that we obtained a sample $(Y_1, ..., Y_N)$ from $p(Y)$
- We want to find a value $\hat{y}$ which minimises the loss $L(y, Y)$



- If we take the squared loss $L(y, Y) = (y - Y)^2$, we can easily show that $\hat{y}$ is the sample mean, i.e.

$$\hat{y} = \hat{E}[Y] = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

$\hat{y}$ for a normal sample

- More generally, we can show with a bit of calculus that the value $\hat{y}$ which minimises the expected loss:

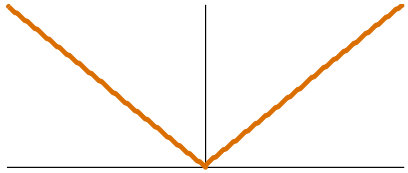
$$E[L] = \int L(y, Y)\, p(Y)\, dY$$

is $\hat{y} = E[Y]$.

How do we estimate quantiles?

- Similarly, if we take the absolute loss:

$$L(y, Y) = |y - Y|$$



it can be shown[2] that the value of y which minimises the expected loss $E[L]$ is the median of $p(Y)$

$\hat{y}$ for a normal sample

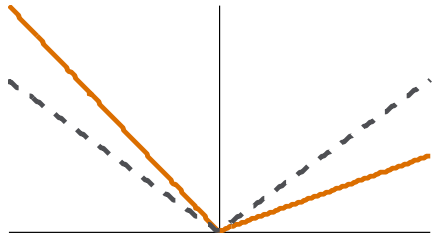
- Unlike in the quadratic case, the median doesn't have a closed form expression
- But we can easily find it by solving an optimisation problem ➟ **nice!**

[2] O'Hagan and Forster (2004)

How about other quantiles?

- To get to other quantiles, the trick is to use a "tilted" absolute error[3], i.e. for quantile q:

$$L_q(y, Y) = \begin{cases} q\,(y - Y) & \text{if } y \geq Y \\ -(1-q)\,(y - Y) & \text{if } y < Y \end{cases}$$
$$\equiv |y - Y|_q$$



$\hat{y}$ for $q = 0.25$

- Again, it can be shown[4] that the value of y which minimises the expected loss $E[L]$ is the q-th quantile
- In particular, for $q = 0.5$, we are back to the usual absolute loss

[3] Koenker and Hallock (2001); Yu et al. (2003)

[4] O'Hagan and Forster (2004)

Model inference for quantiles

- If we now consider that y is a function (our simulator output), we want to find a model $m(x, \beta)$ which approximates the quantiles of $p(y|x)$
- Initially, we look at a single quantile q
- We are looking for $\hat{\beta}_q$ which minimise the expected loss:

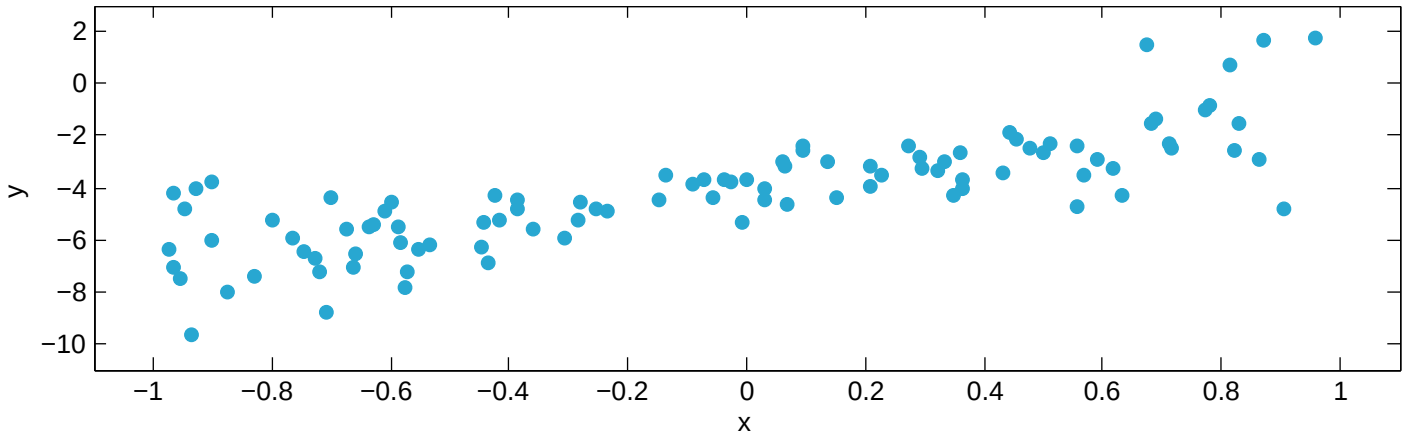
$$E[L_q] = \int |y - m(x, \beta_q)|_q \; p(y|x) \; dy$$

- For a set of simulator (input,output) pairs (x_i, y_i), this is equivalent to minimising:

$$\sum_{i=1}^{N} |y_i - m(x_i, \beta_q)|_q$$

with respect to β_q.

Example



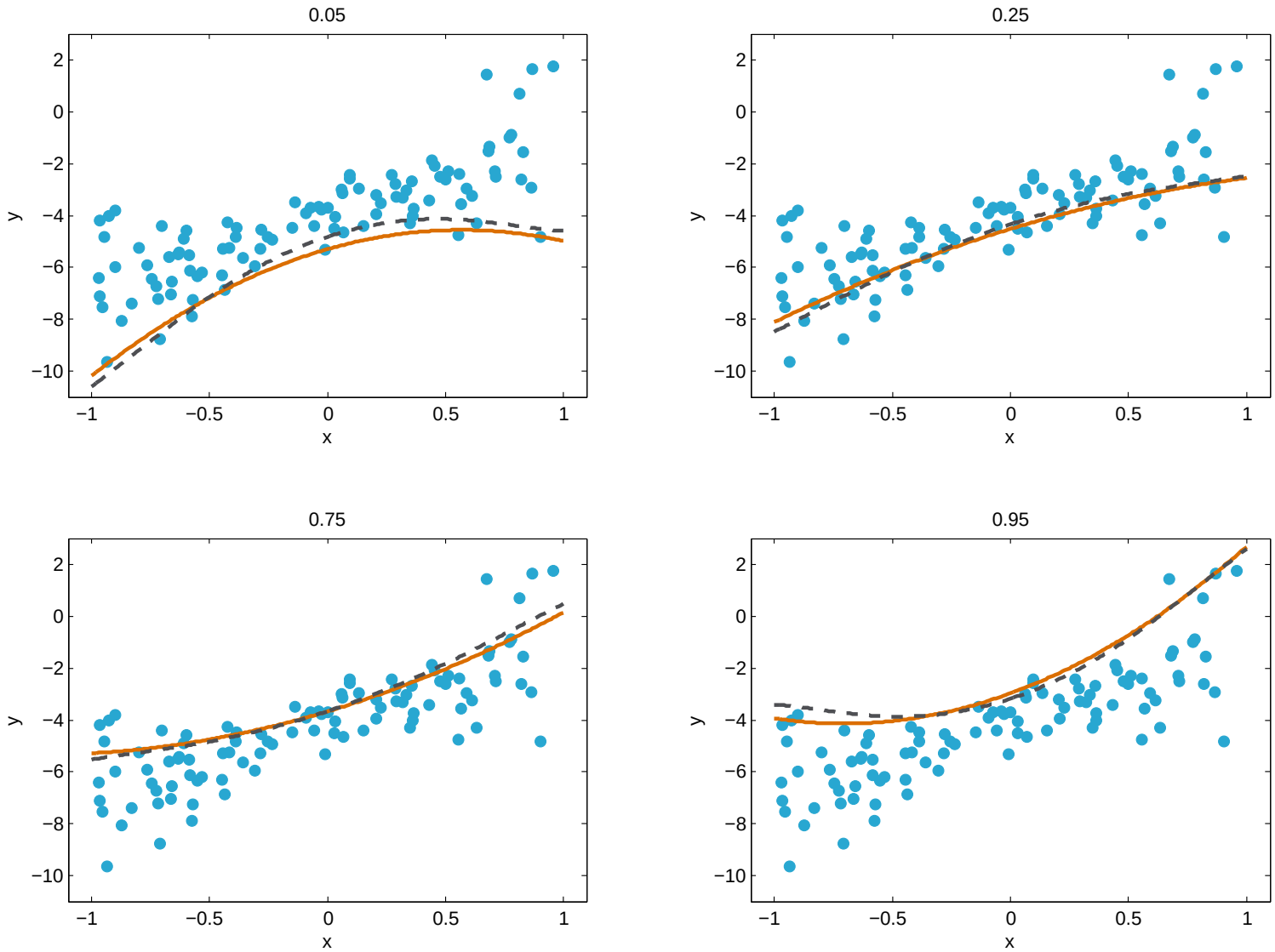
- We use a synthetic simulator

$$y(x) = 3x - 4 + \varepsilon$$

- ε is some Gaussian white noise with $\sigma(x) = 1 + \sin(x)^2$
- We fit a quadratic model to 4 different quantiles (independently):

$$m_q(x, \beta_q) = \beta_{q,2} x^2 + \beta_{q,1} x + \beta_{q,0}$$

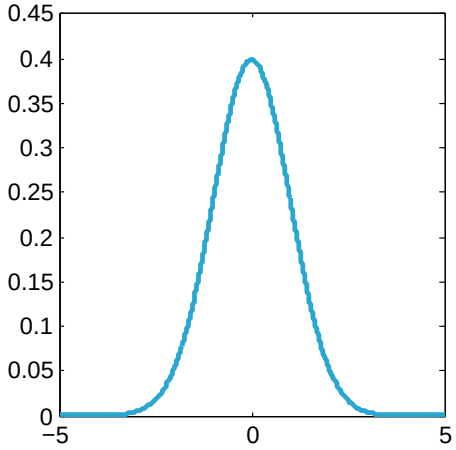
Example (cont.)



Wait! How about the uncertainty?

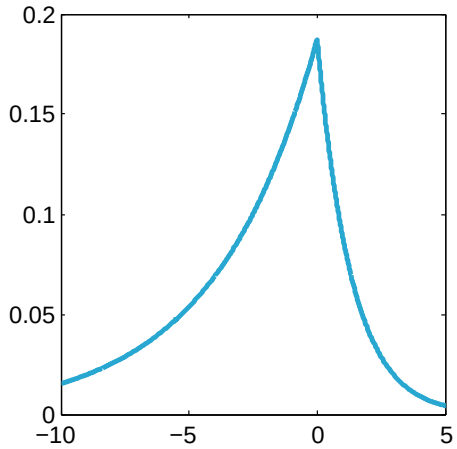
- If we want to be Bayesian about things, we need a likelihood function
- We know that there is an equivalence between the quadratic loss function and the Gaussian likelihood:

$$p(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} L(m(x), y)\right]$$



- Similarly, for the tilted absolute loss, the likelihood is the asymmetric (or skew) Laplace[5]:

$$p(y|x) = \frac{q(1-q)}{\sigma} \exp\left[-\frac{1}{\sigma} L_q(m(x), y)\right]$$



[5] Yu and Moyeed (2001); Kotz et al. (2001)

Wait! How about the uncertainty?

- The posterior over our model parameters is given by:

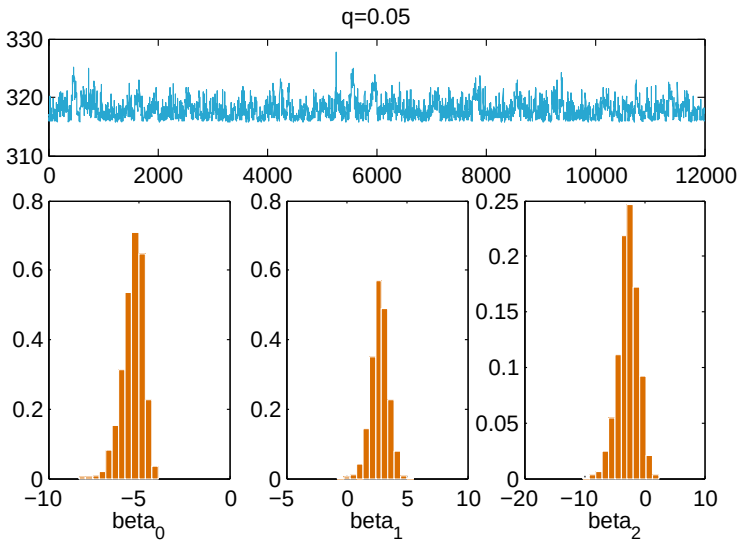
$$p(\beta|y) \propto p(y|x,\beta)\,p(\beta)$$
$$\propto \prod_{i=1}^{N} p(y_i|x_i,\beta)\,p(\beta)$$

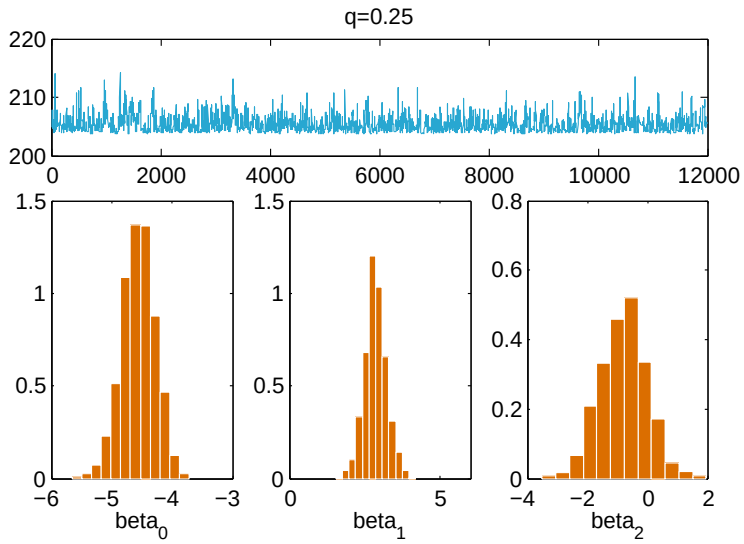
- We can look for a plugin estimate (MAP or ML)
- But ideally, we want to estimate the full posterior
- In the absence of a conjugate prior, we can use MCMC to sample from the posterior

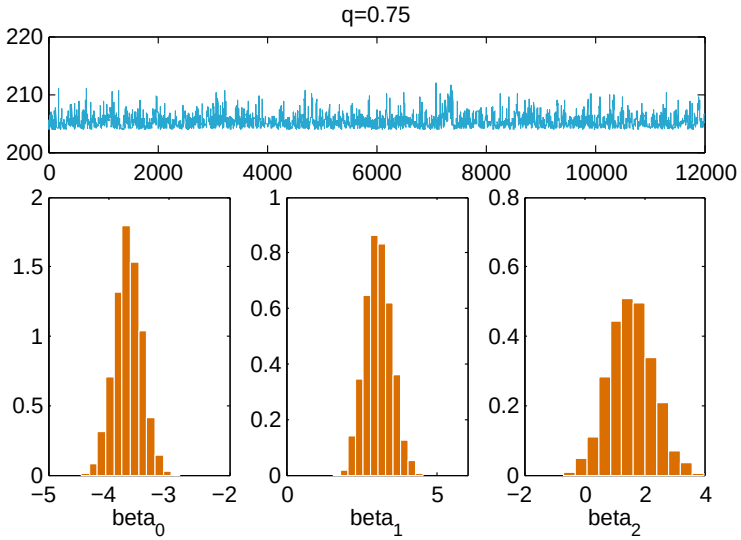
Wait! How about the uncertainty?

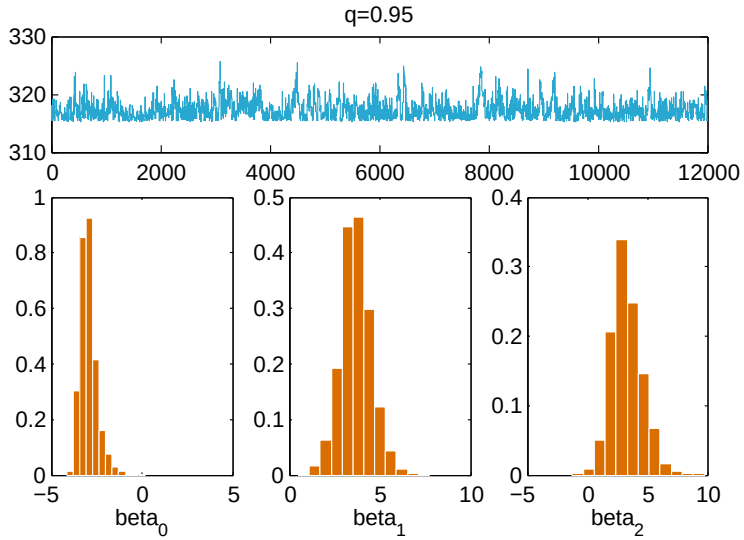
- We use an uninformative prior $p(\beta) \propto 1$
- A Metropolis algorithm is used to get the posterior samples (4 chains with different initial conditions, 3000 samples each, 500 samples discarded as burn-in)
- The posterior samples are plugged in to the model to get samples from the predicted quantile functions
- The quantiles of the predicted quantile process are estimated empirically from these samples

Example - posterior parameters for 4 different quantiles

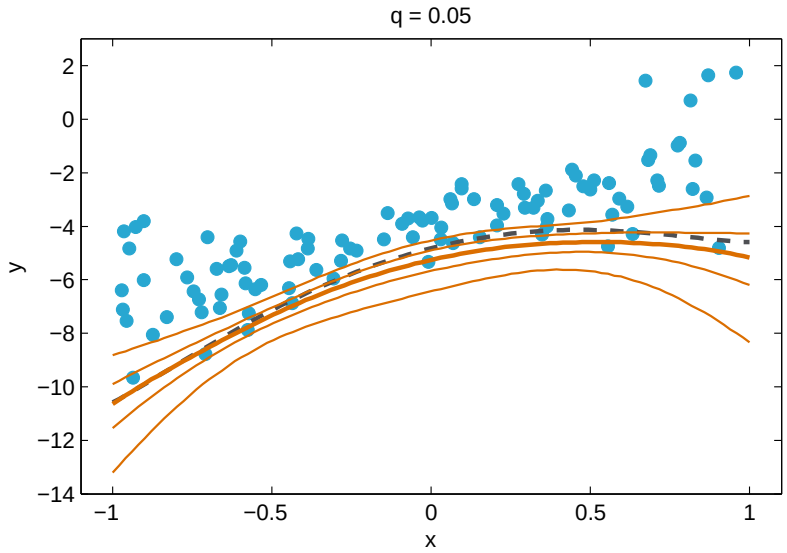


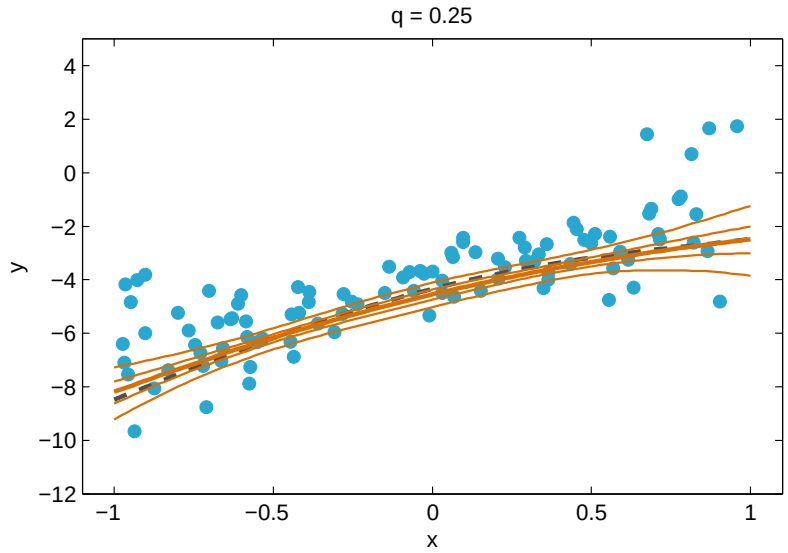


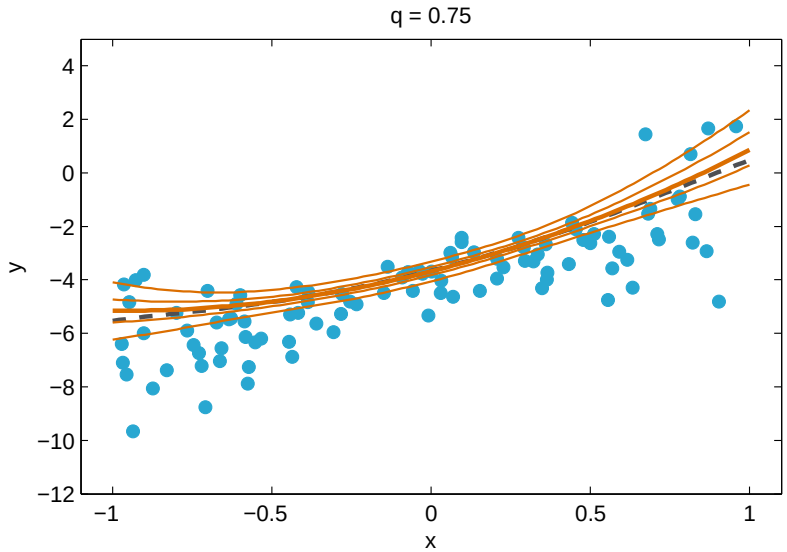


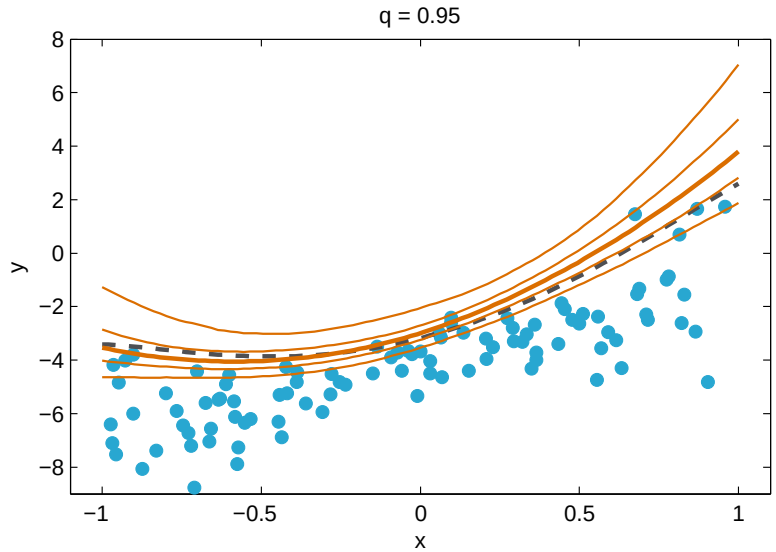


Example - quantile prediction (with own quantiles)

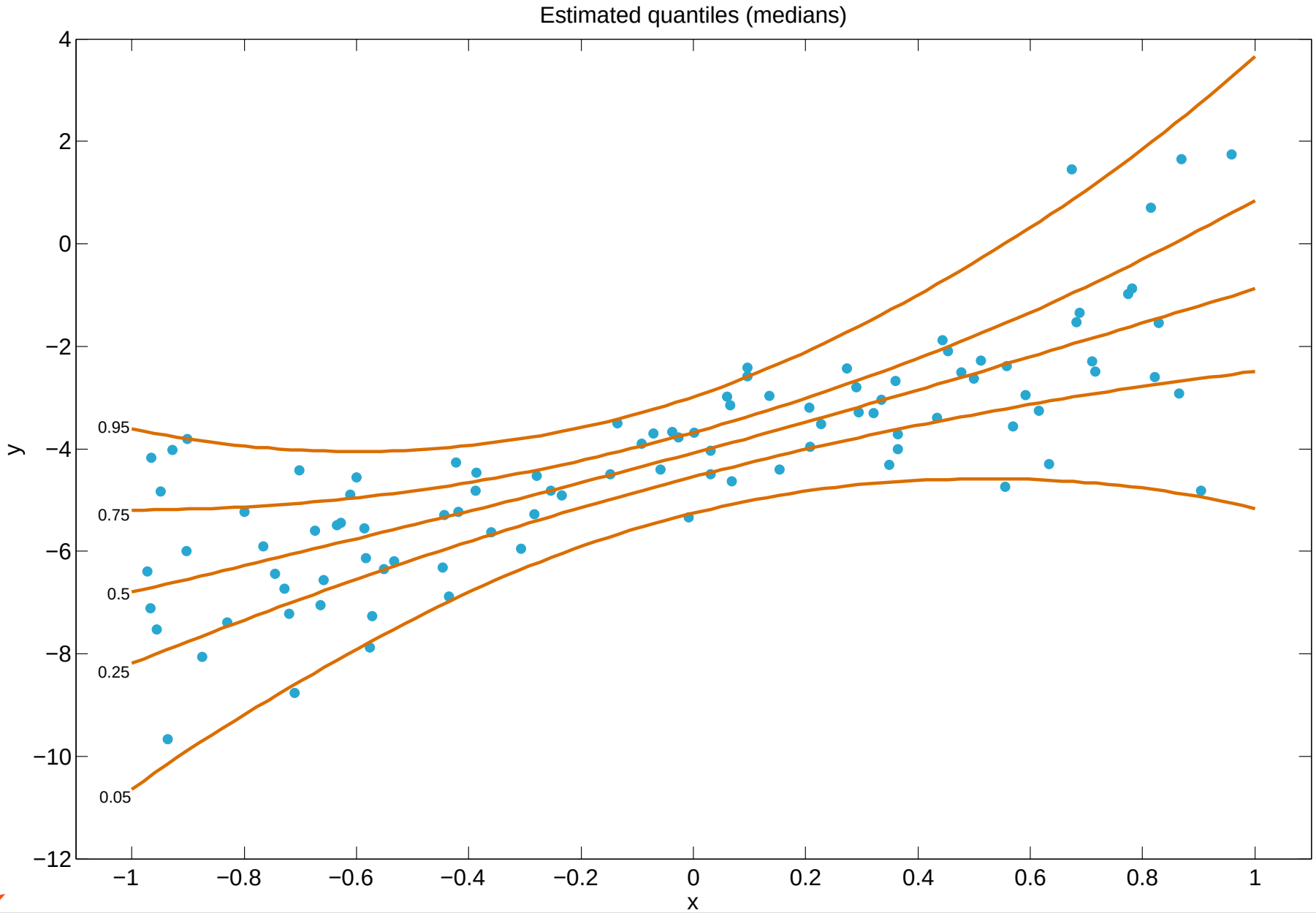




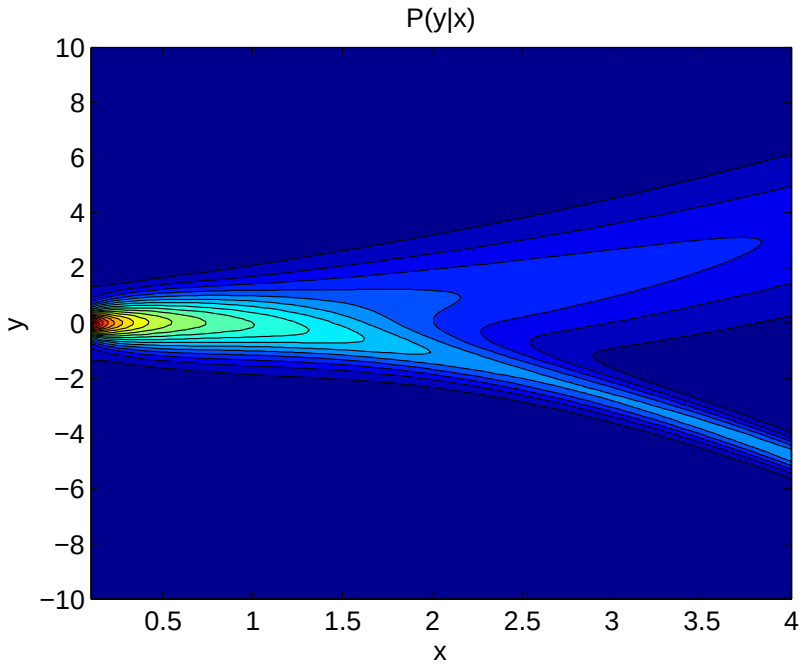


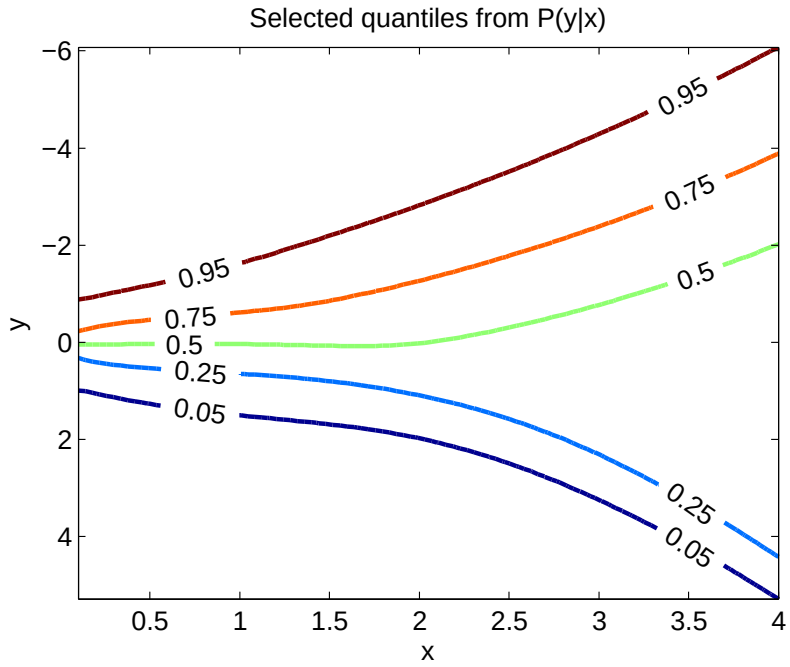


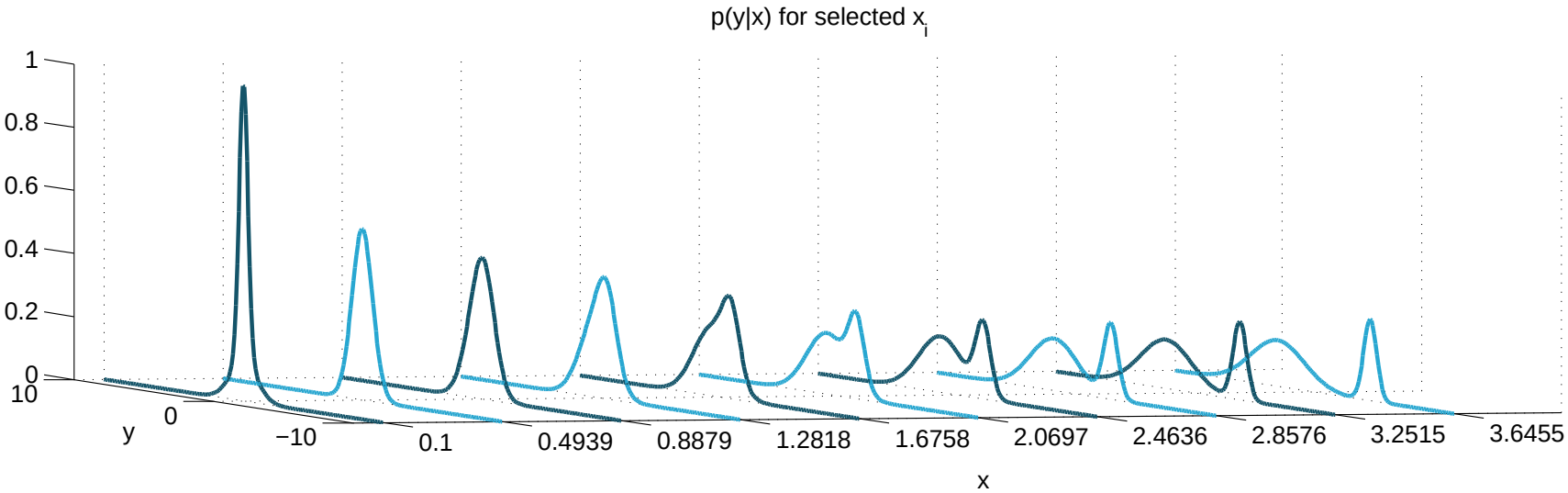
Example (cont.)



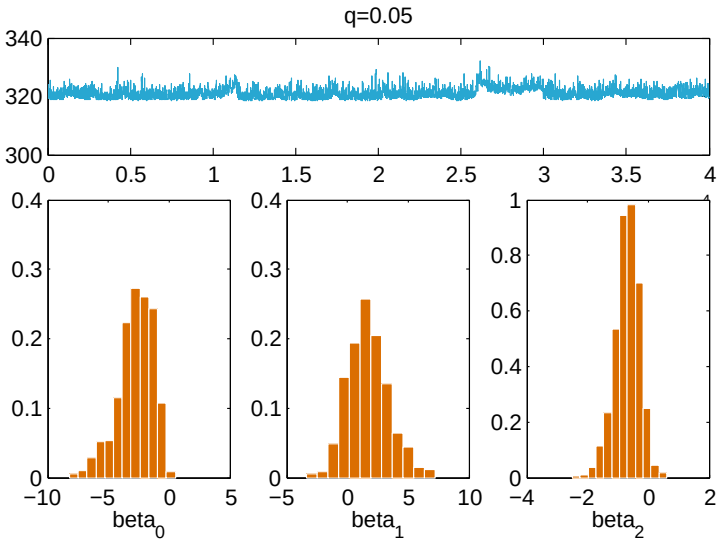
Bimodal example

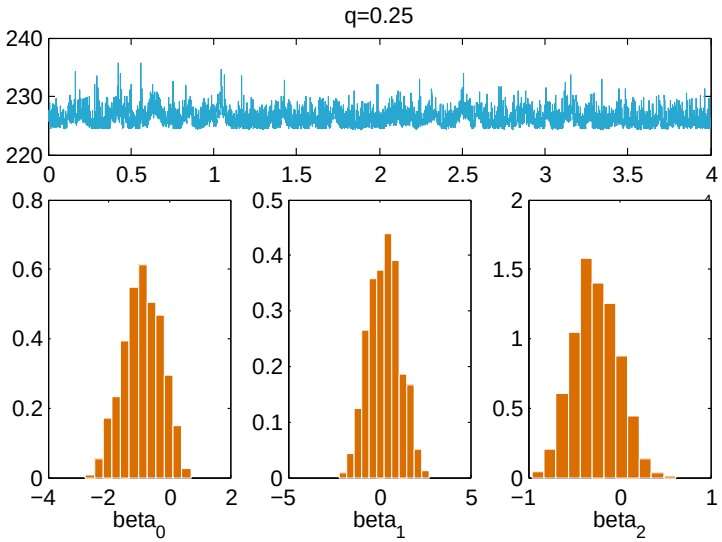


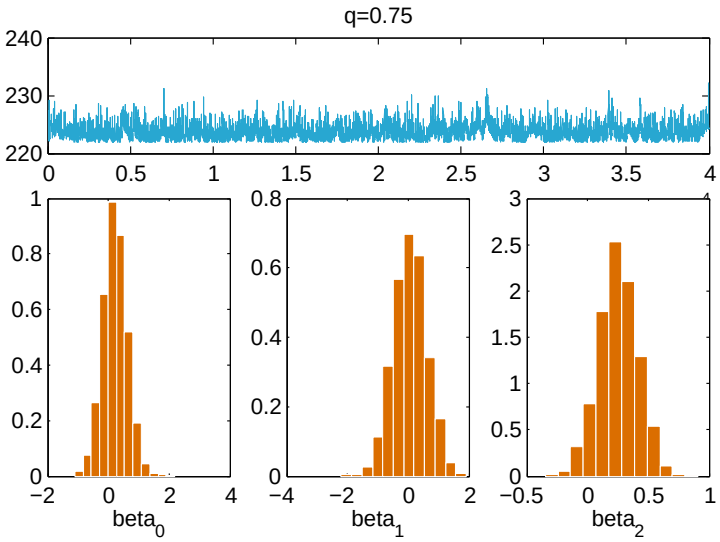


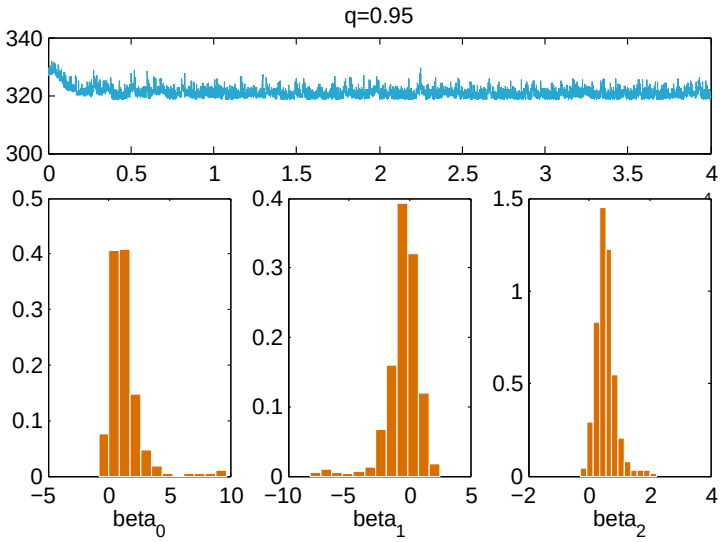


Bimodal example - Posterior parameters

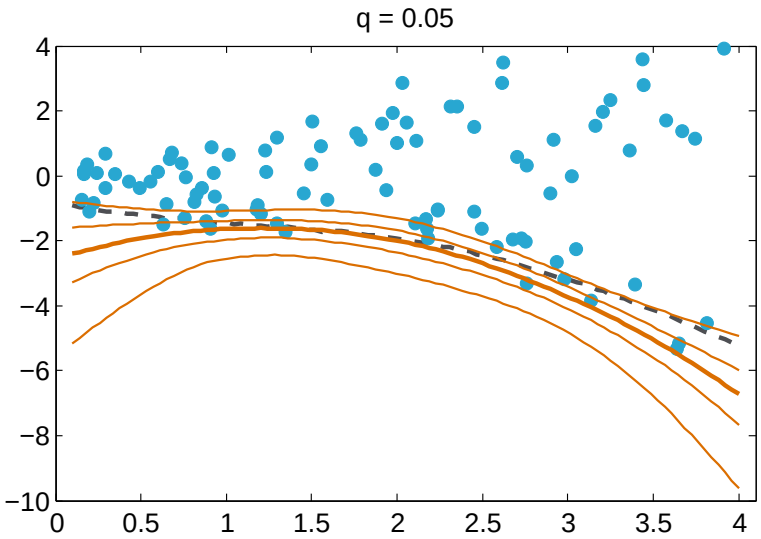


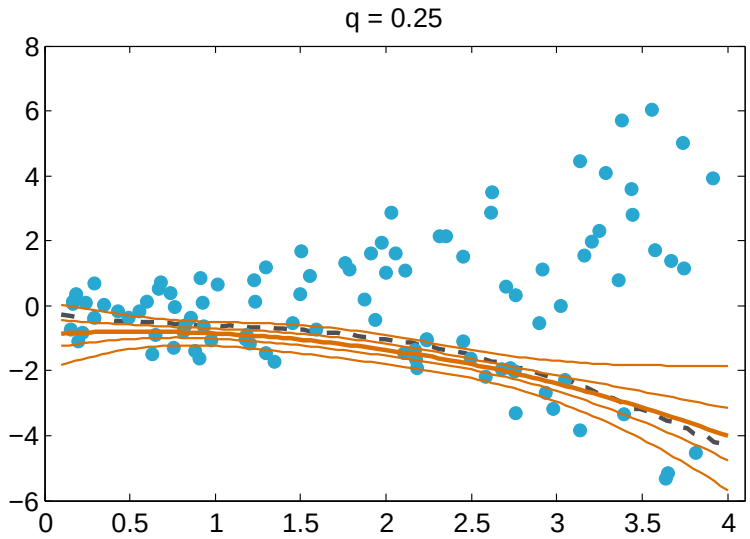


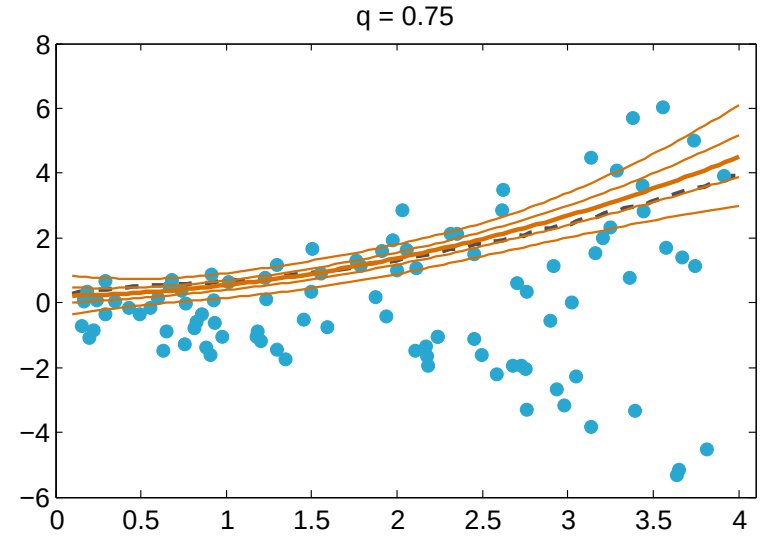


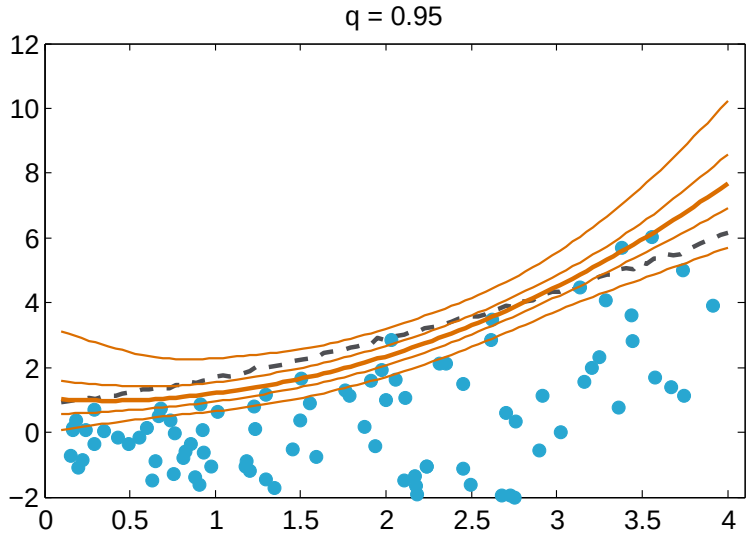


Bimodal example - Estimated quantiles

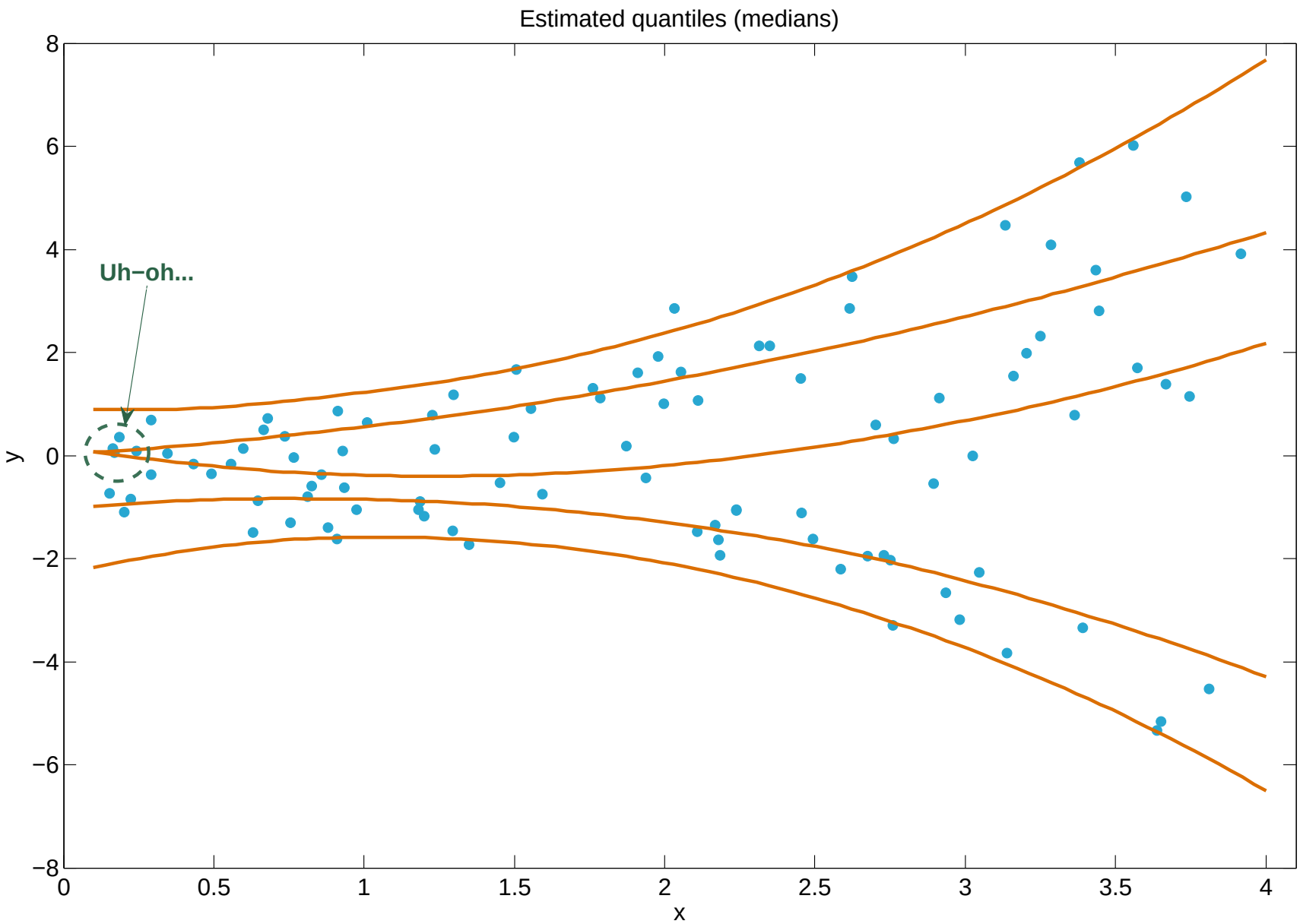








Example (cont.)



Limitations

- There can be order-violation issues when estimating the quantiles independently
- This can often be addressed with a post-processing operation
- Some more robust models for joint estimation of several quantiles have recently been proposed[6]

[6] Taddy and Kottas (2010)

What next? / Open questions

- The regression model needs not be linear: *Kottas et al.*[7] use a Gaussian process as the regression function (and a Dirichlet process over the error)
- The main issue with a GP comes from the intractable evidence:

$$p(y|x) = \int p(y|f)\, p(f|x)\, df$$

- There are ways around it: Laplace approximation, EP algorithm...
- Do we really need a GP...?
- What is a good design for quantile emulation (repeated observations vs. space filling)?
- How do we validate our quantile prediction (when we don't know the true quantile)?

[7] Kottas et al. (2007)

References

Roger Koenker and Kevin F. Hallock. Quantile regression. *The Journal of Economic Perspectives*, 15(4):143–156, 2001.

A. Kottas, M. Krnjajic, and M Taddy. Model-based approaches to nonparametric bayesian quantile regression. In American Statistical Association, editor, *ASA Proceedings of the Joint Statistical Meetings, Alexandria*, pages 1137–48, 2007.

S. Kotz, T.J. Kozubowski, and K. Podgorski. *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Number 183. Birkhauser, 2001.

A. O'Hagan and J. Forster. *Kendall's Advanced Theory of Statistics*, volume 2B Bayesian Inference (2nd ed.). Arnold Publishers, 2004.

M Taddy and A. Kottas. A bayesian nonparametric approach to inference for quantile regression. *Journal of Business and Economic Statistics*, (3):357–369, 2010.

Keming Yu and Rana A. Moyeed. Bayesian quantile regression. *Statistics & Probability Letters*, 2001.

Keming Yu, Zudi Lu, and Julian Stander. Quantile regression: applications and current research areas. *The Statistician*, 52(3):331–350, 2003.