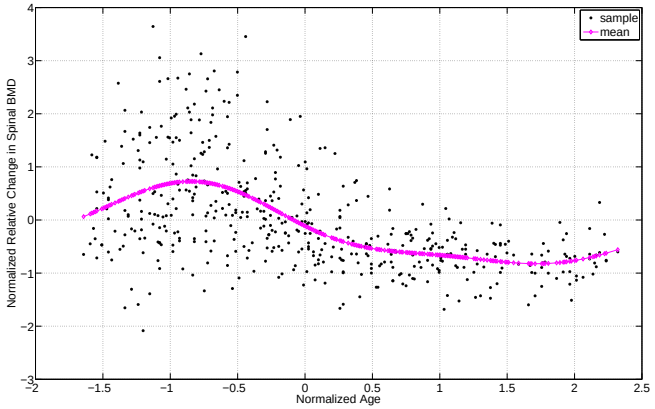
Kernel Conditional Quantile Estimation via Reduction Revisited

Novi Quadrianto
Novi.Quad@gmail.com

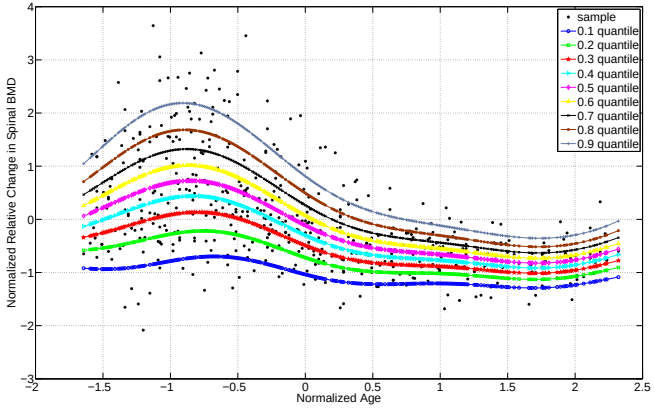
The Australian National University, Australia
NICTA, Statistical Machine Learning Program, Australia

Joint work with
Kristian Kersting, Mark Reid,
Tiberio Caetano and Wray Buntine

The Problem



Mean Regression



Quantile Regression

Mean regression:

computing a regression curve corresponding to the mean of a (conditional) distribution.

Quantile regression:

computing regression curves corresponding to various percentage points of a (conditional) distribution.

Quantile Regression

(Formal) Definition of Quantile Regression

Consider a random variable $y \in \mathbb{R}$ and let $\tau \in (0, 1)$. The conditional quantile $q_\tau(x)$ for a pair of random variables $(x, y) \in \mathcal{X} \times \mathbb{R}$ is defined as the function $q_\tau : \mathcal{X} \to \mathbb{R}$ for which pointwise $q_\tau(x)$ is the infimum over q for which $\Pr(y \leq q_\tau | x) = \tau$.

Applications

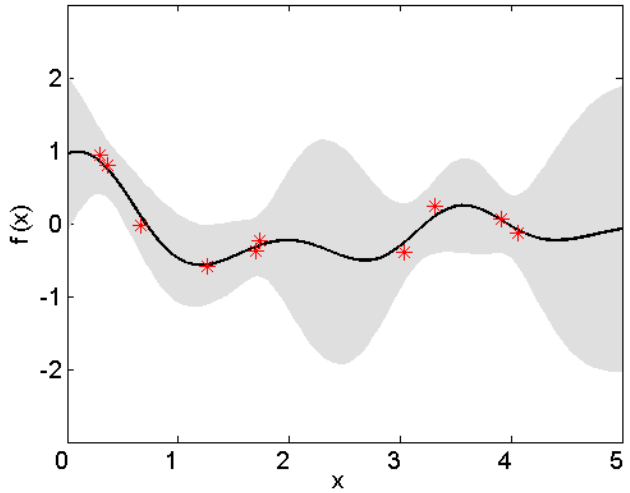
- data mining
- econometrics
- social sciences
- ecology
- bioinformatics
- ...

Gaussian Processes

(Not-so-Formal) Definition of Gaussian Processes

- It is a generalization of multivariate Gaussian distributions over finite dimensional vectors to infinite dimensionality .
- Each draw from a Gaussian process is a function .

One of Applications: mean regression



Quantile Regression via Reduction

Observations :

- If conditional distribution, $p(y|x)$, is known , quantile regression becomes an easy problem.
- We reduce a hard quantile problem to yet another hard intermediate problem, i.e. distribution modeling of $p(y|x)$.
- However, we can harness Gaussian processes in tackling the distribution modeling.

Our Approach:

Estimate conditional distribution via Gaussian processes and subsequently slice the distribution at desired quantile levels.

Quantile Regression via Reduction

Given m observed data points $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{m}$, where $y_i \in \mathbb{R}$ (the set of outputs) and $x_i \in \mathbb{R}^d$ (the set of inputs), infer a conditional quantile function $q_\tau(x)$ from observed data points.

The Model:

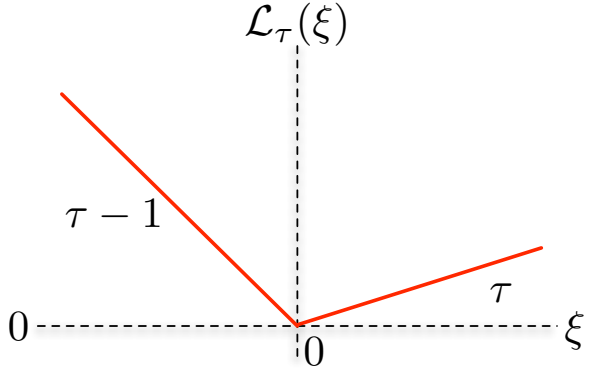
- Prior distribution: $q \sim \mathcal{GP}(m(x), k(x, x'))$
- Likelihood function: $y|q, x \sim \mathcal{N}(q, \sigma_n^2)$
- Predictive distribution (in the form of the standard GP mean regression): for a new input, x_*
  $q^*|x_*, X, Y \sim \mathcal{N}(\mu_*, \sigma_*'^2)$ with the moments as follows

$$\begin{aligned} \mu_* &= k^{*T}(\sigma_n^2 I + K)^{-1} Y \\ \sigma_*'^2 &= k(x_*, x_*) - k^{*T}(\sigma_n^2 I + K)^{-1} k^*. \end{aligned}$$

Quantile Regression via Reduction

The Model:

- Point Prediction:
  - We need a notion of a loss function, i.e.

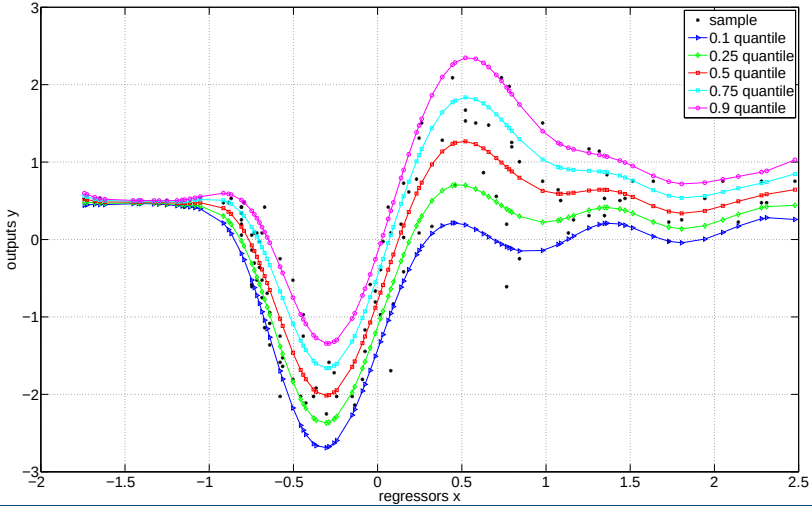


  - Pinball Loss Function
  - We need to minimize a Bayes risk w.r.t the loss function, i.e.

$$
\begin{aligned}
q_\tau^{(\text{opt})} &= \underset{q_\tau}{\operatorname{argmin}}\, \mathbf{E}_{p(y|x)}[\mathcal{L}_\tau(y - q_\tau)] \\
&= \underset{q_\tau}{\operatorname{argmin}} \{(\mu - q)\left[\tau - \Phi_{\mu,\sigma^2}(q)\right] + \sigma\phi_{\mu,\sigma^2}(q)\}
\end{aligned}
$$

Noise Dependent Case

- (Sad) Reality: in real world problems, the noise rate is dependent on the input variables.
- Solution: model distribution via heteroscedastic Gaussian processes.
- Our contribution to heteroscedastic model: joint learning of the free parameters of the latent and observed processes



Performance Guaranteed!

Our quantile estimator has bounded regret .

Theorem 1 *Suppose* $p_* = \mathcal{N}(\mu_*, \sigma_*^2)$ *is a predictive distribution at the point* x_* *and the true point distribution is* $p = \mathcal{N}(\mu, \sigma^2)$. *Then, if* $KL(p_*||p) \leq \epsilon$, *the regret of the corresponding* τ*-th quantile estimator* q_τ^* *satisfies*

$$\Delta\mathcal{R}_\tau(q) \leq \sqrt{2\epsilon}(\tau\sigma + 1)(|\Phi^{-1}(\tau)| + 1).$$

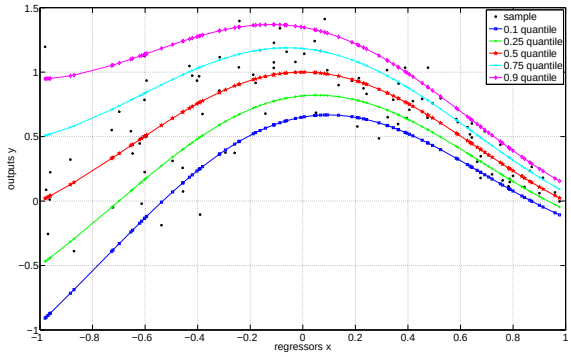
Proof Please refer to the paper. ■

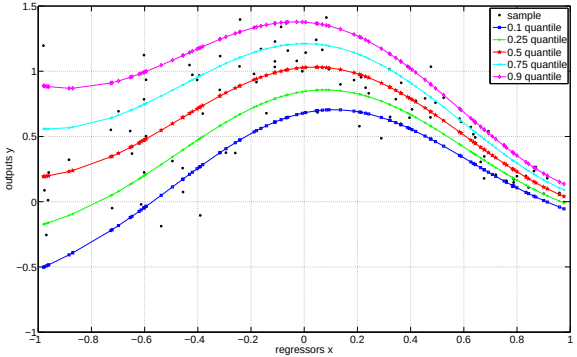
Related Work

- Linear method (Koenker & Bassett 1978): the quantile function is a linear function of inputs, i.e. $q_\tau(x) := \langle x, \beta(\tau) \rangle$, where $\beta(\tau)$ is obtained by minimizing the pinball loss function via linear programming.
- Kernel method (Takeuchi et al. 2006): the dual of regularized pinball loss optimization is minimized via quadratic programming.
- Reduction method (Langford et al. 2006): solving a series of importance weighted binary classification problems.
- Bayesian method (Yu & Moyeed 2001): modeling an asymmetric Laplace likelihood and improper uniform prior with MCMC to infer posterior distribution.

Experiments

Toy Data





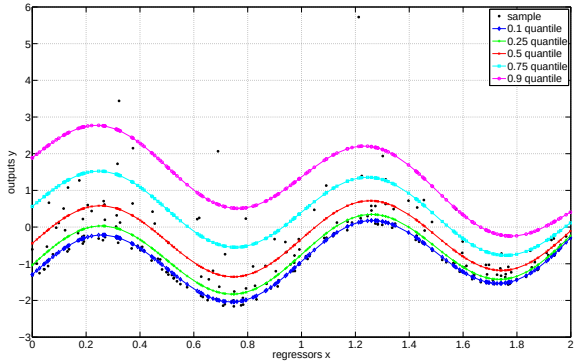
Ground Truth Ours

Data: $x \sim U(-1,1)$ and $y = \mu(x) + \sigma(x)\xi$ with $\mu(x) = \mathrm{sinc}(x), \sigma(x) = 0.1\exp(1-x)$, and $\xi \sim \mathcal{N}(0,1)$.

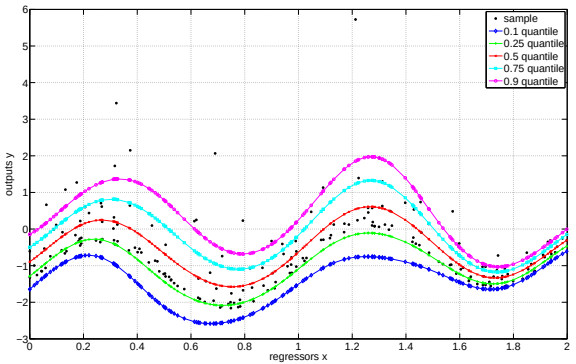
τ	0.1	0.25	0.5	0.75	0.9
QSVM	0.0822	0.0641	0.0274	0.0238	0.0937
HQGP	0.0621	0.0410	0.0306	0.0379	0.0563

Experiments

Toy Data



Ground Truth



Ours

Data: $x \sim U(0,2)$ and $y = \mu(x) + \sigma(x)\xi$ with $\mu(x) = \sin(2\pi x), \sigma(x) = \sqrt{(2.1-x)/4}$, and $\xi \sim \chi^2_{(1)} - 2$.

	τ				
	0.1	0.25	0.5	0.75	0.9
QSVM	0.0987	0.2465	0.5090	0.8044	0.9393
HQGP	0.5286	0.2938	0.2398	0.4793	0.9927

Experiments

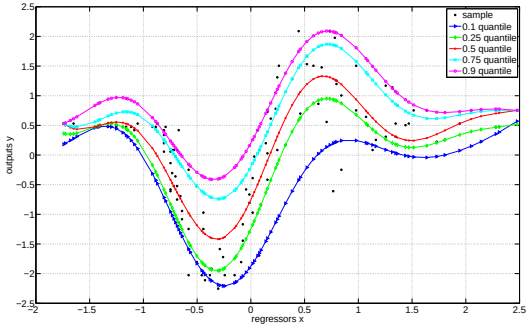
Real Data

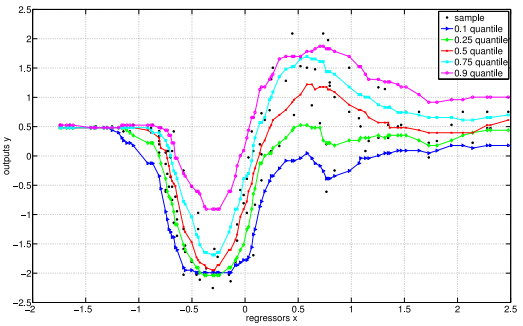
Dataset	m	*Method*	τ 0.1	0.5	0.9
Antigen	97	Linear	0.2927±0.1049	0.2637±0.0502	0.2921±0.0873
		QSVM	0.1229±0.0328	0.2494±0.0326	0.1281±0.0180
		Reduction	0.1218±0.0314	0.2664±0.0213	0.1311±0.0152
		QGP	0.1158±0.0208	0.2553±0.0283	0.1268±0.0148
Weather	238	Linear	0.2911±0.0341	0.2931±0.0238	0.3011±0.0445
		QSVM	0.0672±0.0149	0.2177±0.0335	0.1181±0.0131
		Reduction	0.0749±0.0114	0.1757±0.0275	0.1226±0.0109
		QGP	0.0573±0.0099	0.0971±0.0154	0.0682±0.0168
Motorcycle	133	Linear	0.3963±0.0803	0.3897±0.0186	0.3870±0.0560
		QSVM	0.090±0.0115	0.2022±0.019	0.0852±0.0076
		Reduction	0.0944±0.0109	0.1903±0.0149	0.0830±0.0099
		QGP	0.092±0.0245	0.1859±0.0175	0.0889±0.0099
		HQGP	0.0790±0.0194	0.1872±0.0210	0.0697±0.0162
BMD	485	Linear	0.3283±0.0280	0.3246±0.0398	0.3242±0.0726
		QSVM	0.1220±0.0174	0.3065±0.0387	0.1522±0.0247
		Reduction	0.1205±0.0197	0.3113±0.0414	0.1536±0.0265
		QGP	0.1349±0.0139	0.3097±0.0448	0.1684±0.0296
		HQGP	0.1228±0.0171	0.3086±0.0448	0.1527±0.0267
California Housing	20640	Linear	0.2826±0.0383	0.2252±0.0093	0.2539±0.0678
		QSVM	†	†	†
		Reduction	0.1079±0.0143	0.2630±0.0229	0.1673±0.0061
		Sparse QGP	0.1039±0.0161	0.2721±0.0184	0.1758±0.0242

Experiments

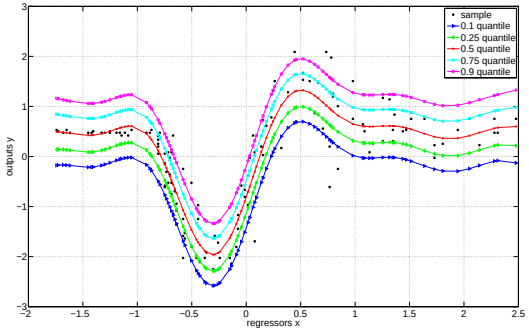
Real Data

Visualization on Silverman's Motorcyle Benchmark.

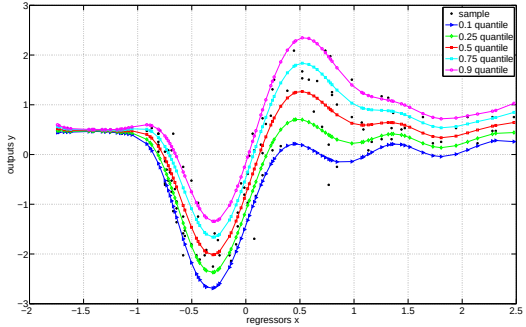




Quantile SVM



Reduction



Ours

Ours (Heteroscedastic)

Summary

Take home messages

- We propose a quantile estimator which is simple to implement;
- enjoys non-parametric and probabilistic model;
- principles learning of free parameters;
- sparse approximation;
- enforced non-crossing constraint properties;
- and has performance guaranteed .