Distribution-free Prediction Bands for Non-parametric Regression

Yang Yu

Department of Statistics
Purdue University

April 6, 2016

(Work by Jing Lei and Larry Wasserman)

Problem:

- Given observations $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}^1$ for $i = 1, \ldots, n$, where $\mathcal{X} \subset \mathbb{R}^d$,
- we want to predict Y_{n+1} given a future predictor X_{n+1}.

Goal:

- To construct a prediction set $C_n(X_{n+1})$ that contains Y_{n+1} with probability at least $1 - \alpha$.
- The collection of prediction sets $C_n = \{C_n(x) : x \in \mathbb{R}^d\}$ forms a prediction band.

Introduction
Prior Work

Table 1: Parametric vs. Non-parametric Methods

	Parametric methods	Non-parametric methods
Assumptions?	Linear assumption, Gaussian assumption (relaxed by quantile regression)	Any smooth distribution
Coverage guarantee (Validity)? ($\mathbb{P}\{Y_{n+1} \in C_n(X_{n+1})\} \geq 1-\alpha$)	Finite sample ✓	Asymptotic ✓
	Linear **?**	Finite sample **?**

Introduction

Prior Work

Table 2: Two Important Classes of Non-parametric Methods

	Usual non-parametric prediction set	Quantile regression prediction set
Form	$\hat{m}(x) \pm z_{\alpha/2}\sqrt{(\hat{\sigma}^2 + s^2)}$	$[\hat{f}_{\alpha/2}(x), \hat{f}_{1-\alpha/2}(x)]$
Drawbacks	Finite sample validity ✗ ($\mathbb{P}\{Y_{n+1} \in C_n(X_{n+1})\} \geq 1-\alpha$)	
	Optimal ✗ (in the form of an interval)	

Introduction

Prior Work

The work by Vovk *et al.* (2009):

- provides finite sample marginal validity.
- However, they focused on linear predictors and did not address efficiency or conditional validity.

In this work:

- The results in Vovk *et al.* (2009) are extended and conditional coverage as well as efficiency are studied.

Outline

Validity and Efficiency

Marginal Validity

Without covariates:

- We observe $Z_1, \ldots, Z_n \overset{iid}{\sim} P$, $Z_i \in \mathbb{R}^d$ for $i = 1, \ldots, n$.
- We want a set $T_n = T_n(Z_1, \ldots, Z_n) \subseteq \mathbb{R}^d$ such that

$$\mathbb{P}(Z_{n+1} \in T_n) \geq 1 - \alpha, \text{ for all } P.$$

With covariates:

- Let $Z_i = (X_i, Y_i)$.
-

$$\mathbb{P}\{(X_{n+1}, Y_{n+1}) \in C_n\} \geq 1 - \alpha, \text{ for all } P,$$

or

$$\mathbb{P}\{Y_{n+1} \in C_n(X_{n+1})\} \geq 1 - \alpha, \text{ for all } P,$$

is the definition of a prediction set for the joint distribution (X, Y).

Validity and Efficiency

Marginal Validity

Definition 1 (Marginal Validity or Joint Validity, Shafer and Vovk (2008))

$$\mathbb{P}\{Y_{n+1} \in C_n(X_{n+1})\} \geq 1 - \alpha, \text{ for all } P, \tag{1}$$

where $\mathbb{P} = P^{n+1}$ is the joint measure of $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$.

The optimal joint prediction set at level $1 - \alpha$ is an upper level set of the joint density

$$C^{(\alpha)} = \{(x, y) : p(x, y) \geq t^{(\alpha)}\}, \tag{2}$$

where $t^{(\alpha)}$ is chosen such that $P(C^{(\alpha)}) = 1 - \alpha$.

- It is defined when the joint distribution of (X, Y) is known.
- Optimality refers to minimizing the Lebesgue measure maintaining the probability coverage at the nominal level.
- It can lead to an unsatisfactory prediction band.

Validity and Efficiency

Optimal Joint Prediction Band

X and Y are independent standard normal distributions.

- According to equation (2), the optimal prediction set for any α is a circle centered at the origin as described by the grey area.
- But intuitively, the best prediction band at level α should be $C(x) = [-z_{\alpha/2}, z_{\alpha/2}]$, for all x, as described by the area between the two broken lines, since observing X provides no information about Y.
- Only requiring marginal validity for prediction bands is not enough.

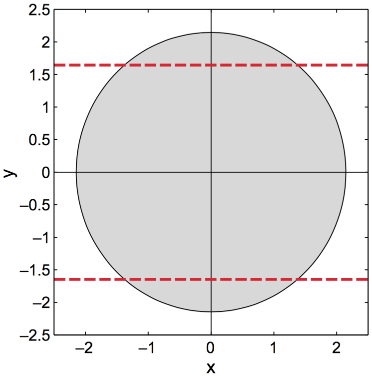

Figure 1: Joint prediction set for bivariate independent Gaussian distributions, $\alpha = 0.1$

Validity and Efficiency

Optimal Joint Prediction Band

Pointwise conditional coverage:
$P\{Y \in C(x) | X = x\}$

- The 'optimal' joint prediction set (the chain curve) tends to overestimate the set when x is in the high density area and to underestimate for low density x.
- It may be tempting to insist on a more stringent probability guarantee such as the conditional validity.

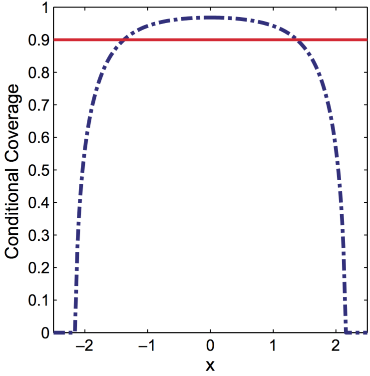

Figure 2: Pointwise conditional coverage for joint prediction set, $\alpha = 0.1$

Validity and Efficiency

Conditional Validity

Definition 2 (Conditional Validity)

$$\mathbb{P}\{Y_{n+1} \in C_n(X_{n+1}) | X_{n+1} = x\} \geq 1 - \alpha, \text{ for all } P \text{ and almost all } x. \quad (3)$$

However, it is shown that finite sample conditional validity is impossible to achieve for continuous distributions.

Definition 3 (Asymptotic Conditional Validity)

$$\sup_x [\mathbb{P}\{Y_{n+1} \notin C_n(x) | X_{n+1} = x\} - \alpha]_+ \xrightarrow{\text{P}} 0 \tag{4}$$

as $n \to \infty$. Here, the supremum is taken over the support of P_X, the marginal distribution of X under P.

Validity and Efficiency

Conditional Oracle Band

Define an oracle band as the counterpart of expression (2) for conditionally valid bands:

$$C_P(x) = \{y : p(y|x) \geq t^{(\alpha)}(x)\}, \tag{5}$$

where $t^{(\alpha)}(x) \equiv t_x^{(\alpha)}$ satisfies

$$\int \mathbb{1}\{p(y|x) \geq t^{(\alpha)}(x)\} p(y|x) \mathrm{d}y = 1 - \alpha.$$

$C_P = \{C_P(x) : x \in \mathbb{R}^d\}$ is called the conditional oracle band.

- It is defined when the joint distribution of (X, Y) is known.
- C_P minimizes $\mu\{C(x)\}$ for all x among all bands satisfying $\inf_x P\{Y \in C(x)|X = x\} \geq 1 - \alpha$.

Definition 3 (Asymptotic Efficiency)

For an estimator C_n, $C_n(x)$ is close to $C_P(x)$ uniformly over all x:

$$\sup_x \mu\{C_n(x) \triangle C_P(x)\} \xrightarrow{\mathrm{P}} 0 \tag{6}$$

where $\triangle$ denotes symmetric set difference.

Validity and Efficiency

Local Validity

A new notion is developed, called local validity, which interpolates between marginal and conditional validity, and is achievable with a finite sample.

Definition 4 (Local Validity)

Let $\mathcal{A} = \{A_j : j \geq 1\}$ be a partition of $\mathrm{supp}(P_X)$. A prediction band C_n is locally valid with respect to $\mathcal{A}$ if

$$\mathbb{P}\{Y_{n+1} \in C_n(X_{n+1}) | X_{n+1} \in A_j\} \geq 1 - \alpha, \text{ for all } j \text{ and all } P. \quad (7)$$

- In the limiting case $\mathcal{A} = \{\mathrm{supp}(P_X)\}$, and local validity becomes marginal validity.
- In the extremal case, A_j shrinks to a single point $x \in \mathbb{R}^d$, and local validity approximates conditional validity.

Validity and Efficiency

Local Validity

Relationship between different types of validity:

- Local validity is stronger than marginal validity but weaker than conditional validity.

Proposition 1

If C is conditionally valid, then it is also locally valid for any partition $\mathcal{A}$. If C is locally valid for some partition $\mathcal{A}$, then it is also marginally valid.

We will construct a finite sample locally valid prediction band, which satisfies

(a) finite sample marginal validity,
(b) asymptotic conditional validity and
(c) asymptotic efficiency.

Marginally Valid Prediction Band

Construction

Extend the idea of conformal prediction (Shafer and Vovk (2008) and Vovk *et al*. (2005, 2009)) to construct joint prediction sets by using kernel density estimators, as described in Lei *et al*. (2011).

A simple scenario without covariates:

- Suppose that we observe $Z_1, \ldots, Z_n \sim P$ and we want a prediction set for Z_{n+1}.
- The idea is to test $H_0 : Z_{n+1} = z$ for each z and then to invert the test.

Marginally Valid Prediction Band

Construction

- For any z let $\hat{p}_n^z(\cdot)$ be a kernel density estimator with bandwidth h_n based on the augmented data $\mathrm{aug}(\mathbf{Z}; z) = (Z_1, \ldots, Z_n, z)$.
- Define

$$C_n \equiv C_n(Z_1, \ldots, Z_n) = \{z : \pi_n(z) \geq \alpha\}$$

where

$$\pi_n(z) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{1}\{\sigma_i(z) \leq \sigma_{n+1}(z)\}$$

is the p-value for the test, $\sigma_i(z) = \hat{p}_n^z(Z_i)$ for $i = 1, \ldots, n$ and $\sigma_{n+1}(z) = \hat{p}_n^z(z)$.

Marginally Valid Prediction Band
Construction

- C_n is finite sample marginally valid:

$$\mathbb{P}(Z_{n+1} \in C_n) \geq 1 - \alpha \text{ for all } P.$$

- It can be shown that C_n is close to $C^{(\alpha)}$ with high probability where $C^{(\alpha)}$ is the optimal prediction set as defined in expression (2).

- The statistic σ_i is an example of a conformity measure.
- More generally, a conformity measure $\sigma_i(z) = \sigma\{\text{aug}(\mathbf{Z}, z), Z_i\}$ indicates how well a data point Z_i agrees with the augmented data set $\text{aug}(\mathbf{Z}, z)$.

Marginally Valid Prediction Band

Construction

Let $Z = (X, Y)$.

- For any $(x, y) \in \mathbb{R}^d \times \mathbb{R}^1$, let $(\mathbf{X}, \mathbf{Y}) = (X_1, Y_1, \ldots, X_n, Y_n)$ be the data set and $\operatorname{aug}\{\mathbf{X}, \mathbf{Y}; (x, y)\}$ be the augmented data with $X_{n+1} = x$ and $Y_{n+1} = y$.
- Define $\hat{p}_n^{(x,y)}$ as the kernel density estimator with bandwidth h_n from the augmented data.
- Define the conformity measure

$$\sigma_i(x, y) := \hat{p}_n^{(x,y)}(X_i, Y_i), \text{ for } i = 1, \ldots, n,$$

and

$$\sigma_{n+1}(x, y) := \hat{p}_n^{(x,y)}(x, y),$$

and p-value

$$\pi_n(x, y) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{1}\{\sigma_i(x, y) \leq \sigma_{n+1}(x, y)\}.$$

Marginally Valid Prediction Band

Construction

- Define

$$\hat{C}^{(\alpha)}(x) = \{y : \pi_n(x, y) \geq \alpha\}. \tag{8}$$

Lemma 1

$\hat{C}^{(\alpha)}(x)$ is finite sample marginally valid:

$$\mathbb{P}\{Y_{n+1} \in \hat{C}^{(\alpha)}(X_{n+1})\} \geq 1 - \alpha \text{ for all } P.$$

- Computing $\hat{C}^{(\alpha)}$ is expensive since we need to find the p-value $\pi_n(x, y)$ for every (x, y).
- The sandwich approximation, an accurate approximation C_n^+ to $\hat{C}^{(\alpha)}$, avoids the augmentation step altogether but preserves finite sample validity.

Marginally Valid Prediction Band

Sandwich Approximation

Algorithm 1 (Sandwich Slicer Algorithm)

(a) Let $\hat{p}(x, y)$ be the joint density estimator.

(b) Let $Z_i = (X_i, Y_i)$ and let $Z_{(1)}$, $Z_{(2)}$,..., denote the sample ordered increasingly by $\hat{p}(X_i, Y_i)$.

(c) Let $j = \lfloor n\alpha \rfloor$ and define

$$C_n^+(x) = \left\{ y : \hat{p}(x, y) \geq \hat{p}(X_{(j)}, Y_{(j)}) - \frac{K_x(0)K_y(0)}{nh^{d+1}} \right\}. \qquad (9)$$

- It can be shown that $\hat{C}^{(\alpha)} \subseteq C_n^+$ and hence C_n^+ also has finite sample marginal validity.
- Moreover, C_n^+ has the same asymptotic properties as $\hat{C}^{(\alpha)}$ if h_n is chosen appropriately.
- But C_n^+ is not asymptotically efficient nor does it satisfy asymptotic conditional validity.

Locally Valid Prediction Band

Construction

Extend the idea of conformal prediction to construct prediction bands with local validity. These bands will also be asymptotically efficient and have asymptotic conditional validity.

- We consider partitions $\mathcal{A} = \{A_k, k \geq 1\}$ in the form of equilateral cubes with sides of length w_n.
- Let $n_k = \sum_{i=1}^{n} \mathbb{1}(X_i \in A_k)$ be the histogram count.
- Define $\hat{p}_n^{(x,y)}$ as yhe kernel density estimator with bandwidth h_n from the augmented data for the local marginal density of Y is, for any $(x, y) \in A_k \times \mathbb{R}^1$,

$$\hat{p}_n^{(x,y)}(v|A_k) = \frac{1}{(n_k+1)h_n}\left(\sum_{i=1}^{n} \mathbb{1}(X_i \in A_k) K\left(\frac{v - Y_i}{h_n}\right) + K\left(\frac{v-y}{h_n}\right)\right).$$

Locally Valid Prediction Band

Construction

- For any $(x, y) \in A_k \times \mathbb{R}^1$, define the p-value

$$\pi_{n,k}(x, y) = \frac{1}{n_k + 1} \sum_{i=1}^{n+1} \mathbb{1}(X_i \in A_k) \mathbb{1}\{\hat{p}_n^{(x,y)}(Y_i | A_k) \leq \hat{p}_n^{(x,y)}(Y_{n+1} | A_k)\}.$$

- The band

$$\hat{C}^{(\alpha)}_{\text{loc}}(x) = \{y : \pi_{n,k}(x, y) \geq \alpha\} \tag{10}$$

for $x \in A_k$ has finite sample local validity.

Proposition 2

$\hat{C}^{(\alpha)}_{\text{loc}}(x)$ is finite sample locally valid and hence finite sample marginally valid.

Locally Valid Prediction Band

Construction

$\hat{C}_{\text{loc}}^{(\alpha)}$: Conformal Optimized Prediction Set estimator (COPS)

- 'Optimized' denotes the effort of minimizing the average set length $\mathbb{E}[\mu\{C_n(X_{n+1})\}]$.

Locally Valid Prediction Band

Sandwich Approximation

Algorithm 2 is a fast approximation algorithm that is analogous to algorithm 1. The resulting approximation also satisfies finite sample local validity as well as asymptotic efficiency and asymptotic conditional validity.

Algorithm 2 (Local Sandwich Slicer Algorithm)

(a) Divide $\mathcal{X}$ into bins $A_1, \ldots, A_m$.

(b) Apply algorithm 1 separately on all Y_is within each A_k.

(c) Output $C_n^+(x)$: the resulting set of A_k for all $x \in A_k$.

Asymptotic Properties

Assumptions

The following assumption put boundedness and smoothness conditions on the marginal density p_X, conditional density $p(y|x)$ and its derivatives.

Assumption 1 (Regularity of Marginal and Conditional Densities)

(a) The marginal density of X satisfies $0 < b_1 \le p_X(x) \le b_2 < \infty$ for all x in $[0,1]^d$.

(b) For all x, $p(\cdot|x)$ is in Hölder class $\Sigma(\beta, L)$. Correspondingly, the kernel K is a valid kernel of order β.

(c) For any $0 \le s \le \lfloor \beta \rfloor$, $p^{(s)}(y|x)$ is continuous and uniformly bounded by L for all x and y.

(d) The conditional density is Lipschitz in x: $\|p(\cdot|x) - p(\cdot|x')\|_\infty \le L\,\|x - x'\|$.

The next assumption gives a sufficient regularity condition, 'γ-exponent' condition (Polonik (1995)), on the upper level sets $L_x(t) \equiv \{p(y|x) \geq t\}$.

Assumption 2 (Regularity of Conditional Density Level Set)

There are positive constants ε_0, γ, c_1 and c_2 such that, for all $x \in [0,1]^d$,

$$c_1 \varepsilon^\gamma \leq \mathbb{P}[y : |p(y|x) - t_x^{(\alpha)}| < \varepsilon | X = x] \leq c_2 \varepsilon^\gamma$$

for all $\varepsilon \leq \varepsilon_0$, where $t_x^{(\alpha)}$ is the cut-off value such that $P_x\{L_x(t_x^{(\alpha)})\} = \mathbb{P}[\{y : p(y|x) \geq t_x^{(\alpha)}\} | X = x] = 1 - \alpha$. Moreover, $\inf_x t_x^{(\alpha)} \geq t_0 > 0$.

Theorem 1 (Convergence Rate on Asymptotic Efficiency)

Let $C_{\mathrm{loc}}^{(\alpha)}$ be the prediction band given by the local conformity procedure as described in equation (21). Choose $w_n \asymp r_n$ and $h_n \asymp r_n^{1/\beta}$. Under assumptions 1 and 2, for any $\lambda > 0$, there is a constant A_λ, such that

$$\mathbb{P}[\sup_{x \in \mathcal{X}} \mu\{C_{\mathrm{loc}}^{(\alpha)}(x) \triangle C_P(x)\} \geq A_\lambda {r_n}^{\gamma_1}] = O(n^{-\lambda}),$$

where $\gamma_1 = \min(1, \gamma)$ and

$$r_n = \left\{\frac{\log(n)}{n}\right\}^{\beta/\{\beta(d+2)+1\}}. \tag{11}$$

Thus, in the common case $\gamma = 1$, the convergence rate on the asymptotic efficiency of the locally valid prediction band $C_{\mathrm{loc}}^{(\alpha)}$ is r_n.

Asymptotic Properties

Rate of Convergence

Lemma 2

Under assumptions 1 and 2, the local band $C_{\mathrm{loc}}^{(\alpha)}$ is asymptotically conditionally valid.

The approximation in algorithm 2, C_n^+, also satisfies the same asymptotic efficiency and conditional validity results.

Asymptotic Properties

Minimax Bound

Theorem 2 (Lower Bound on Estimation Error)

Let $\mathcal{P}(\beta, L)$ be the class of distributions satisfying assumptions 1 and 2 with $\gamma = 1$. Fix an $\alpha \in (0, 1)$; there is a constant $c = c(\alpha, \beta, L, d) > 0$ such that, for all large n,

$$\inf_{\hat{C}_n} \sup_{P \in \mathcal{P}(\beta, L)} \mathbb{E}_P[\mu\{C_{\text{loc}}^{(\alpha)}(x) \triangle C_P(x)\}] \geq c r_n$$

where the infimum is over all estimators $\hat{C}_n$ based on a sample size of n.

In the most common case $\gamma = 1$, r_n, the convergence rate on the asymptotic efficiency of the locally valid prediction band $C_{\text{loc}}^{(\alpha)}$, is indeed minimax rate optimal.

Tuning Parameter Selection

Idea

Two parameters to choose:

- w_n: width of the cubic partition $\mathcal{A}$.
- $h_{n,k}$: kernel bandwidth of the estimated local marginal density $\hat{p}(\cdot|A_k)$.

Goal:

- Choose $(w_n, h_{n,k})$ such that the resulting conformal set has smallest Lebesgue measure $\mu(\hat{C})$.

Idea: (Completely data driven)

- Split the sample into two equal-sized subsamples.
- Apply the tuning algorithm on one subsample.
- Use the output bandwidth on the other subsample to obtain the prediction band.

(Two-stage procedure, to preserve finite sample marginal validity)

Tuning Parameter Selection

Procedure

Algorithm 3 (Bandwidth Tuning for COPS)

Input data $\mathcal{Z}$, level α and candidate sets $\mathcal{W}$ and $\mathcal{H}$.

(a) Split the data set into two equal-sized subsamples $\mathcal{Z}_1$ and $\mathcal{Z}_2$.

(b) For each $w \in \mathcal{W}$:

  (i) construct partition $\mathcal{A}^w$;

  (ii) for each bin A_k and candidate kernel bandwidth h construct local conformal prediction set $\hat{C}^1_{h,k}$, each at level $1-\alpha$, using data $\mathcal{Z}_1$;

  (iii) let $h^*_{h,k} = \arg\min_{h\in\mathcal{H}} \mu(\hat{C}^1_{h,k})$, for all k;

  (iv) let $Q(w) = (1/n)\sum_k n_k \mu(\hat{C}^1_{h^*_{w,k},k})$

(c) Choose $\hat{w} = \arg\min Q(w)$; $\hat{h}_{\hat{w},k} = h^*_{\hat{w},k}$.

(d) Construct partition $\mathcal{A}_{\hat{w}}$. For $x \in A_k$, output prediction band $\hat{C}(x) = \hat{C}^2_{\hat{h}_{\hat{w},k},k}$, where $\hat{C}^2_{h,k}$ is the local conformal prediction set estimated from data $\mathcal{Z}_2$ in local set A_k.

- The band $\hat{C}$ is locally valid and marginally valid.
- How about its asymptotic efficiency? (An open question)

Data Examples
Synthetic Example

Data:

- $d = 1$
- $X \sim \mathrm{Unif}[-1.5, 1.5]$
- $(Y|X = x) \sim 0.5N\{f(x) - g(x), \sigma^2(x)\} + 0.5N\{f(x) + g(x), \sigma^2(x)\}$, where $f(x) = (x-1)^2(x+1), g(x) = 2\sqrt{(x+0.5)}\mathbb{1}(x \geq -0.5)$, $\sigma^2(x) = \frac{1}{4} + |x|$
- $n = 1000$

Data Examples

Synthetic Example

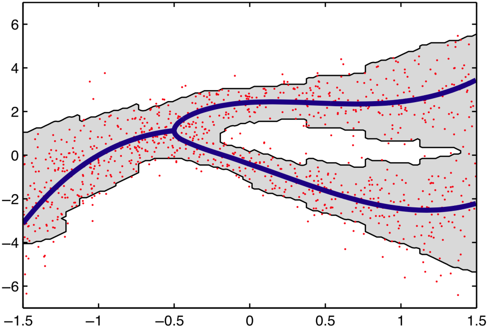

Figure 3: Generated data (red dots)

Features of data:

- For $x \leq -0.5$, $(Y|X = x)$ is a Gaussian distribution centered at $f(x)$ with varying variance $\sigma^2(x)$.
- For $x \geq -0.5$, $(Y|X = x)$ is a two-component Gaussian mixture and, for large values of x, the two components have little overlap.

Data Examples

Synthetic Example

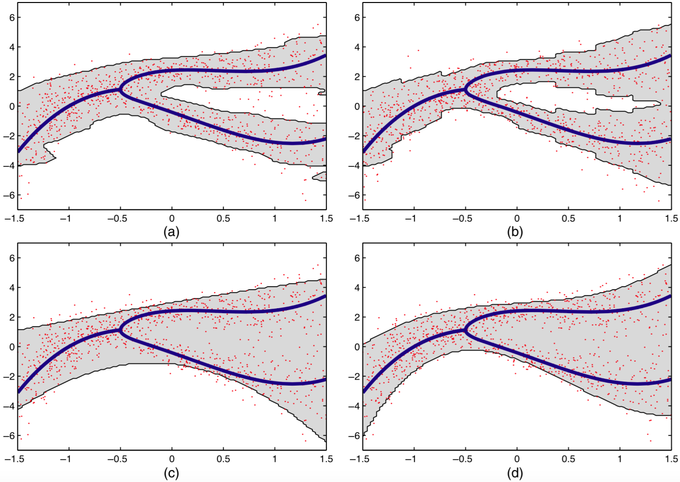

Figure 4: (a) Marginal conformal bands, (b) local conformal bands, (c) quadratic quantile regression, (d) cubic quantile regression, $\alpha = 0.1$

Marginal (a) vs. Conditional (b)

- The locally valid band gives the desired coverage for all values x.
- The marginally valid band overcovers for smaller values of x, and undercovers for larger values of x.

Conformal (a, b) vs. Quantile regression (c, d)

- The conformal regions correctly capture the bifurcated structure,
- but the quantile regression methods do not.

Data Examples

Synthetic Example

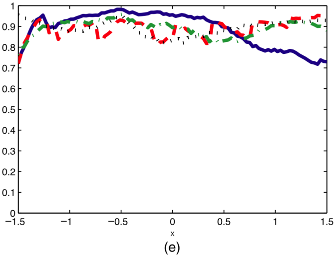

Figure 5: (e) Conditional coverage as a function of x (blue: marginal, red: local, green: quadratic, black: cubic), (f) integrated Lebesgue measure of the prediction regions over 100 repetitions

- The conformal method has correct finite sample coverage.
- The conformal regions have smaller average Lebesgue measure.

Data Examples

Car Data

- We want to predict the miles per gallon by the horsepower.
- The relationship between miles per gallon and horsepower is far from linear, so some transformation (the inverse of miles per gallon) must be applied before linear model fitting.

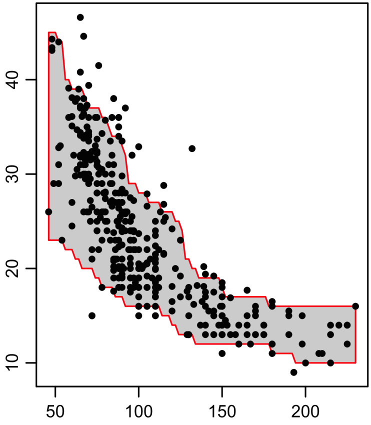

Figure 6: Car data (black dots)

Data Examples

Car Data

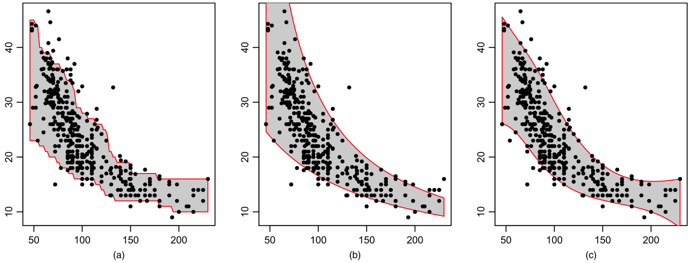

Figure 7: (a) Local conformal bands, (b) linear regression with variable transformation, (c) spline-based quantile regression, $\alpha = 0.1$

- The linear regression prediction band is too wide for small values of horse power and too narrow for large values, owing to the non-uniform noise level.
- The spline-based quantile regression is similar to the conformal band albeit a little smoother.
- The local conformal method does not involve choosing the variable transformation or specifying a model.

Final Remarks
Summary

$C_{\text{loc}}^{(\alpha)}$ or C_n^+: The first prediction band with the following properties:

(a) Finite sample (marginal and local), distribution-free validity
  - Distribution-free: no assumptions on P are required.

(b) Asymptotic conditional validity

(c) An explicit rate for asymptotic efficiency, achieving the minimax bound

(d) Completely data-driven tuning parameters selection

Final Remarks

Future Work

Future Work:

(a) Establish a rigorous result on the asymptotic efficiency for the data-driven bandwidth.

(b) The bands in this work are not suitable for high dimensional problems.
  - Develop methods for constructing prediction bands that exploit sparsity assumptions.
  - Yield valid prediction and variable selection simultaneously.

References

- Lei, J., Robins, J. and Wasserman, L. (2013) Distribution free prediction sets. *J. Am. Statist. Ass*, **108**, 278–287.
- Lei, J. and Wasserman, L. (2014) Distribution free prediction bands for nonparametric regression. *J. R. Statist. Soc.* B, **76**, 71–96.
- Polonik, W. (1995) Measuring mass concentrations and estimating density contour clusters–an excess mass approach. *Ann. Statist.*, **23**, 855–881.
- Shafer, G. and Vovk, V. (2008) A tutorial on conformal prediction. *J. Mach. Learn. Res.*, **9**, 371–421.
- Vovk, V., Gammerman, A. and Shafer, G. (2005) *Algorithmic Learning in a Random World*. New York: Springer.
- Vovk, V., Nouretdinov, I. and Gammerman, A. (2009) On-line predictive linear regression. *Ann. Statist.*, **37**, 1566–1590.