UPPSALA
UNIVERSITET

Distribution-Free Prediction

Perspectives from a recovering Bayesian

Dave Zachariah

Talk @ ERNSI 2017

dave.zachariah@it.uu.se

UPPSALA UNIVERSITET

Background

- Wågberg, J., Zachariah, D., Schön, T.B., Stoica, P.: Prediction performance after learning in Gaussian process regression, AISTATS, 2017.
- Zachariah, D., Stoica, P., and Schön, T.B.: Online Learning for Distribution-Free Prediction, submitted, 2017.
- Zachariah, D. and Stoica, P.: Cramér-Rao Bounds in Statistical Inference, in preparation.

dave.zachariah@it.uu.se

UPPSALA UNIVERSITET

Aims

1. Subjectivist-Bayesian vs. Frequentist interpretations
2. Bayesian paradigm is powerful but sometimes misleading
3. Distribution-free approach to tackle challenges

dave.zachariah@it.uu.se

UPPSALA
UNIVERSITET

Prediction problem

dave.zachariah@it.uu.se

Data-generating process

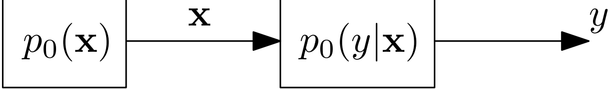


Figure : Process with inputs $\mathbf{x}$ and outputs y

Dataset $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$

dave.zachariah@it.uu.se

Data-generating process

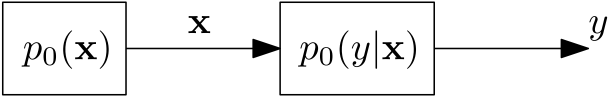


Figure : Process with inputs $\mathbf{x}$ and outputs y

Example #1:

- $\mathbf{x}$ spatial coordinates
- y ozone density
- $n = 17\,340$ samples

dave.zachariah@it.uu.se

Data-generating process

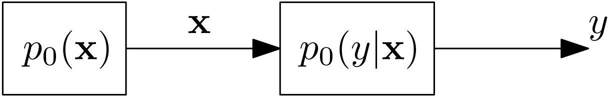


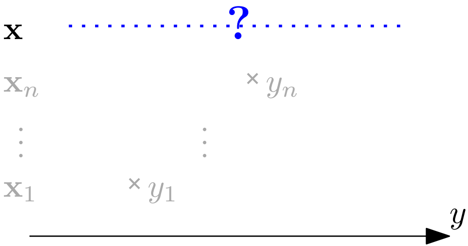
Figure : Process with inputs $\mathbf{x}$ and outputs y

Example #2:

- $\mathbf{x}$ individual background covariates
- y weekly wage
- $n = 219\ 673$ samples

dave.zachariah@it.uu.se

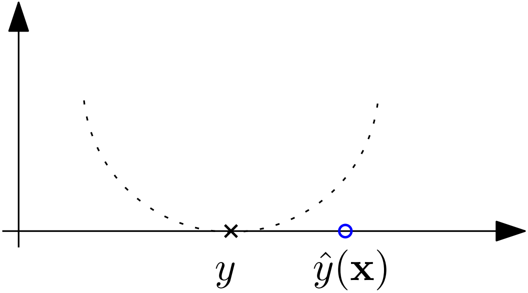
Prediction method as a decision rule



- New sample $(\mathbf{x}, y) \sim p_0(\mathbf{x}, y)$
- Given $\mathbf{x}$ predict y using $\mathcal{D}$

dave.zachariah@it.uu.se

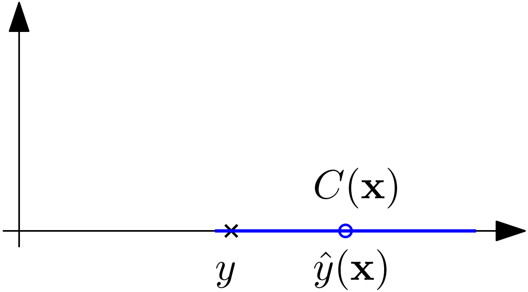
Prediction method as a decision rule



Decision-rule $\widehat{y}(\mathbf{x})$ has risk $\mathcal{R} = \mathrm{E}\Big[|y - \widehat{y}(\mathbf{x})|^2\Big]$

dave.zachariah@it.uu.se

Uncertainty region



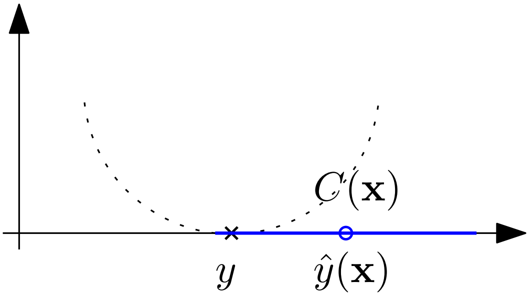
Uncertainty region

$$C(\mathbf{x}) = \left\{ y' \; : \; d\left(y', \widehat{y}(\mathbf{x})\right) \leq \kappa \right\}$$

constructed using $\mathcal{D}$

dave.zachariah@it.uu.se

Decisions under uncertainty

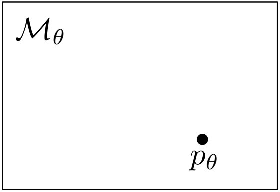


To proceed we specify a model p_θ of unknown p_0

dave.zachariah@it.uu.se

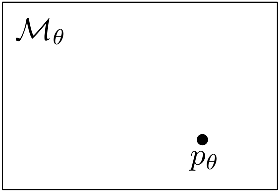
UPPSALA UNIVERSITET

Specifying a model class



dave.zachariah@it.uu.se

Model class



Data-generating process:

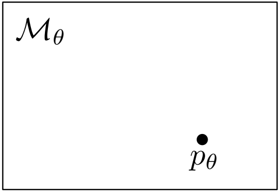
$$p_0(\mathbf{x}, y, \mathcal{D})$$

Model:

$$p_{\theta}(\mathbf{x}, y, \mathcal{D})$$

dave.zachariah@it.uu.se

Model class



Data-generating process:

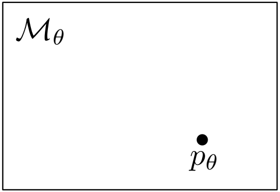
$$p_0(\mathbf{x}, y, \mathcal{D})$$

Model (w/ marginalized latent variables):

$$p_{\theta}(\mathbf{x}, y, \mathcal{D}) = \int p_{\theta}(\mathbf{x}, y, \mathbf{z}, \mathcal{D}) \, d\mathbf{z}$$

dave.zachariah@it.uu.se

Model class



Data-generating process:

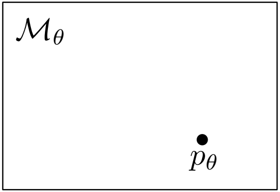
$$p_0(\mathbf{x}, y, \mathcal{D})$$

Model (factorized form):

$$p_\theta(\mathbf{x}, y, \mathcal{D}) = p_\theta(\mathbf{x}, \mathcal{D})\, p_\theta(y|\mathbf{x}, \mathcal{D})$$

dave.zachariah@it.uu.se

Model class



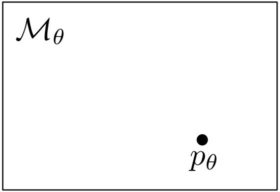
Example: i.i.d. samples

$$\mathcal{M}_\theta = \left\{ p_\theta(\mathbf{x}, \mathcal{D})\, p_\theta(y|\mathbf{x}, \mathcal{D}) \;:\; \mu_\theta(\mathbf{x}) = \boldsymbol{\phi}^\top(\mathbf{x})\boldsymbol{\theta} \right\}$$

with feature vector $\phi(\mathbf{x})$

dave.zachariah@it.uu.se

Model class

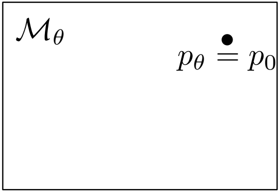


Example: Gaussian process

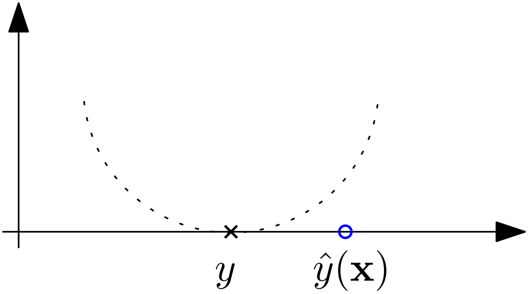
$$\mathcal{M}_\theta = \left\{ p_\theta(\mathbf{x}, \mathcal{D}) \; p_\theta(y|\mathbf{x}, \mathcal{D}) \text{ Gaussian } : \; \mu_\theta(\mathbf{x}), k_\theta(\mathbf{x}, \mathbf{x}') \right\}$$

with posterior mean and covariance functions using $\mathcal{D}$

Ideal case



Optimal decision-rule



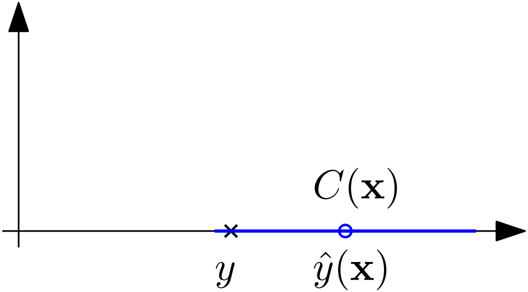
Bound on risk

$$\mathcal{R} \geq \mathrm{E}_{x,\mathcal{D}}\Big[\mathrm{Var}[y|\mathbf{x},\mathcal{D}] \Big]$$

Attained by decision rule $\widehat{y}(\mathbf{x}) = \mu_{\theta}(\mathbf{x})$

dave.zachariah@it.uu.se

Uncertainty region



Since $\sigma^2_\theta(\mathbf{x}) = \mathrm{Var}[y|\mathbf{x}, \mathcal{D}]$, construct uncertainty region

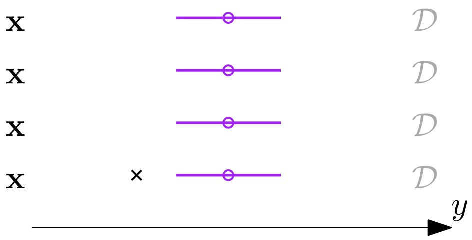
$$C_\theta(\mathbf{x}) = \left\{ y' \ : |y' - \mu_\theta(\mathbf{x})| \leq \kappa \sigma_\theta(\mathbf{x}) \right\}$$

dave.zachariah@it.uu.se

Uncertainty region: Credibility

Credibility

$$\Pr\{y \in C_\theta(\mathbf{x}) \mid \mathbf{x}, \mathcal{D}\} \;\geq\; 1 - \kappa^{-2}$$



Individualized decision-making and beliefs

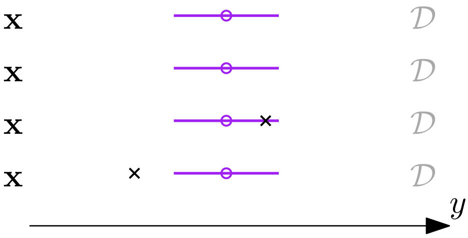
dave.zachariah@it.uu.se

UPPSALA UNIVERSITET

Uncertainty region: Credibility

Credibility

$$\Pr\{y \in C_\theta(\mathbf{x}) \mid \mathbf{x}, \mathcal{D}\} \;\geq\; 1 - \kappa^{-2}$$



Individualized decision-making and beliefs

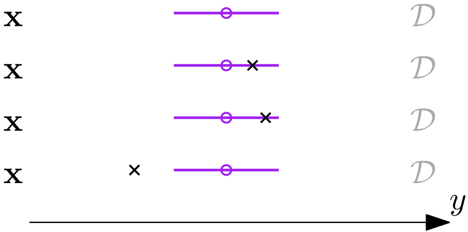
dave.zachariah@it.uu.se

UPPSALA UNIVERSITET

Uncertainty region: Credibility

Credibility

$$\Pr\{y \in C_\theta(\mathbf{x}) \mid \mathbf{x}, \mathcal{D}\} \geq 1 - \kappa^{-2}$$



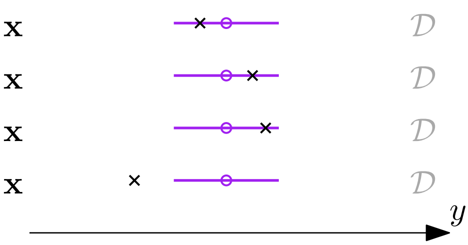
Individualized decision-making and beliefs

dave.zachariah@it.uu.se

Uncertainty region: Credibility

Credibility

$$\Pr\{y \in C_\theta(\mathbf{x}) \mid \mathbf{x}, \mathcal{D}\} \;\geq\; 1 - \kappa^{-2}$$



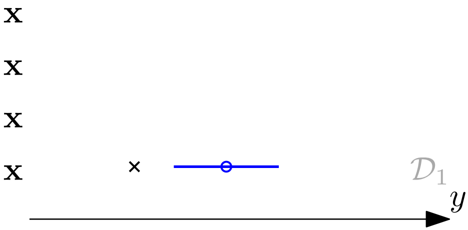
Individualized decision-making and beliefs

dave.zachariah@it.uu.se

Uncertainty region: Confidence

Confidence

$$\Pr\{y \in C_\theta(\mathbf{x}) \mid \mathbf{x}\} \;\geq\; 1 - \kappa^{-2}$$



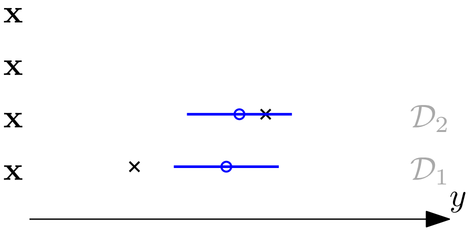
Reproducible decision-making and coverage

dave.zachariah@it.uu.se

Uncertainty region: Confidence

Confidence

$$\Pr\{y \in C_\theta(\mathbf{x}) \mid \mathbf{x}\} \;\geq\; 1 - \kappa^{-2}$$



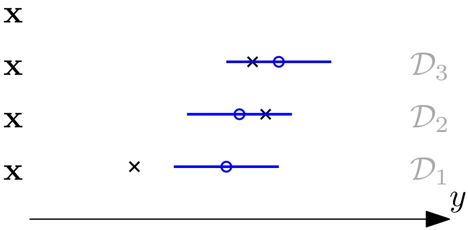
Reproducible decision-making and coverage

dave.zachariah@it.uu.se

Uncertainty region: Confidence

Confidence

$$\Pr\{y \in C_\theta(\mathbf{x}) \mid \mathbf{x}\} \geq 1 - \kappa^{-2}$$



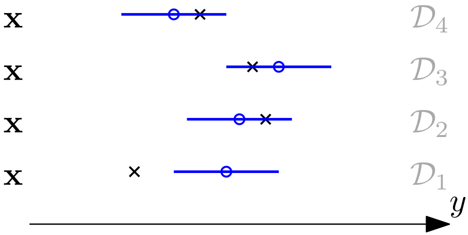
Reproducible decision-making and coverage

dave.zachariah@it.uu.se

Uncertainty region: Confidence

Confidence

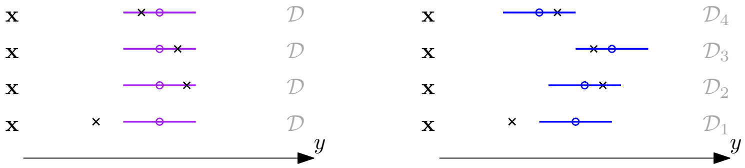
$$\Pr\{y \in C_\theta(\mathbf{x}) \mid \mathbf{x}\} \;\geq\; 1 - \kappa^{-2}$$



Reproducible decision-making and coverage

dave.zachariah@it.uu.se

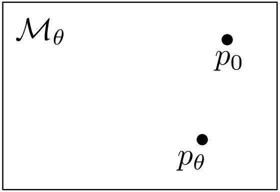
Subjectivist vs. Frequentist interpretations



Subjectivist:	Frequentist:
Belief	Coverage
Credibility	Confidence
Individualized	Reproducible

dave.zachariah@it.uu.se

Well-specified case



dave.zachariah@it.uu.se

UPPSALA UNIVERSITET

Defining the best model

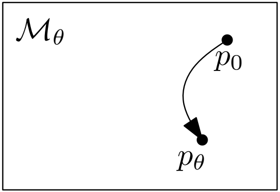


Figure : Divergence Δ of p_θ from p_0

dave.zachariah@it.uu.se

Defining the best model

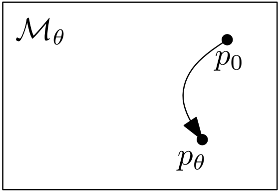


Figure : Divergence Δ of p_θ from p_0

Best model:

$$\boldsymbol{\theta}_\star = \arg\min_{\boldsymbol{\theta}} \; \Delta(\boldsymbol{\theta}) \qquad \text{e.g.} \quad \Delta(\boldsymbol{\theta}) = \widehat{\mathsf{E}}\left[|y_i - \mu_{\boldsymbol{\theta}}(\mathbf{x}_i)|^2 \right]$$

dave.zachariah@it.uu.se

Defining the best model

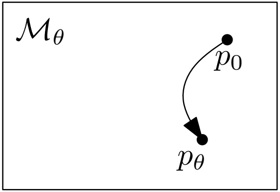


Figure : Divergence Δ of p_θ from p_0

Best model:

$$\boldsymbol{\theta}_\star = \arg\min_{\boldsymbol{\theta}} \; \Delta(\boldsymbol{\theta}) \qquad \text{e.g.} \quad \Delta(\boldsymbol{\theta}) = \mathrm{E}_{y|X}\left[\ln \frac{p_0(\mathbf{y}|\mathbf{X})}{p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{X})} \right]$$

dave.zachariah@it.uu.se

Defining the best model

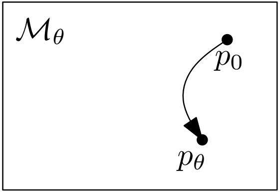


Figure : Divergence Δ of p_θ from p_0

Well-specified case: $p_0 \in \mathcal{M}_\theta \quad \Rightarrow \quad p_{\theta_\star} = p_0$

dave.zachariah@it.uu.se

Defining the best model

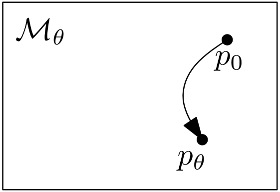
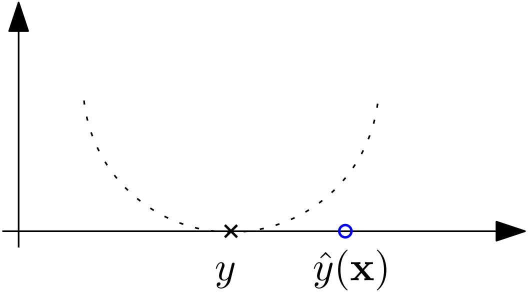


Figure : Divergence Δ of p_θ from p_0

Learned model $\widehat{\boldsymbol{\theta}}$ that approaches $\boldsymbol{\theta}_\star$ as n grows

dave.zachariah@it.uu.se

Bound on risk when best model is unknown



Bound in Gaussian case

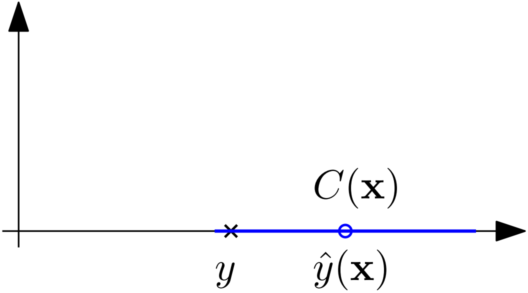
If bias is invariant w.r.t. $\boldsymbol{\theta}_\star$:

$$\mathcal{R} \geq \mathrm{E}_{x,X}\Big[\,\mathrm{Var}[y|\mathbf{x},\mathbf{X}] + \mathbf{g}^\top \mathbf{J}^{-1}\mathbf{g}\,\Big],$$

where $\mathbf{J}$ is a Fisher information matrix.

dave.zachariah@it.uu.se

Uncertainty region: unknown properties

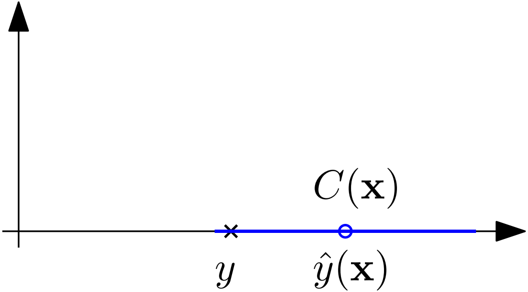


Constructed as before but evaluated at $\widehat{\boldsymbol{\theta}}$:

$$C_{\widehat{\theta}}(\mathbf{x}) = \left\{ y' : |y' - \mu_{\widehat{\theta}}(\mathbf{x})| < \kappa \sigma_{\widehat{\theta}}(\mathbf{x}) \right\}$$

dave.zachariah@it.uu.se

Uncertainty region: unknown properties



Credibility:

$$\Pr\left\{y \in C_{\widehat{\theta}}(\mathbf{x}) \mid \mathbf{x}, \mathcal{D}\right\} \quad ?$$

Confidence:

$$\Pr\left\{y \in C_{\widehat{\theta}}(\mathbf{x}) \mid \mathbf{x}\right\} \quad ?$$

Approximations possible when $n \to \infty$

dave.zachariah@it.uu.se

Uncertainty region: underperformance

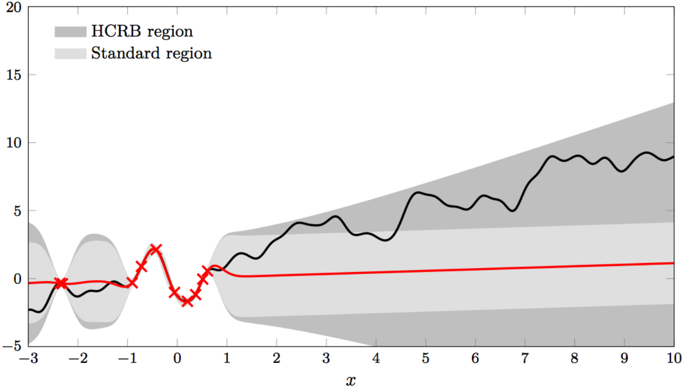
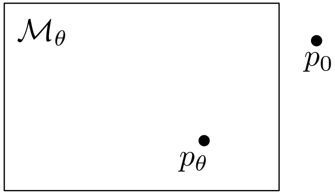


Figure : Standard region $C_\theta(x)$ with $\kappa = 3$ around prediction $\widehat{y}(x)$

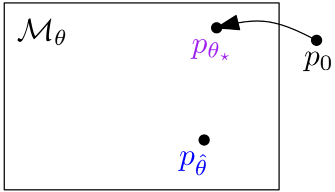
dave.zachariah@it.uu.se

Misspecified case



dave.zachariah@it.uu.se

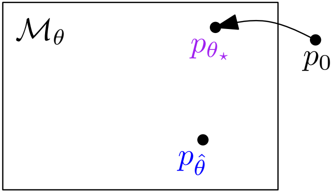
Seeking the best model



$$\boldsymbol{\theta}_\star = \arg\min_{\boldsymbol{\theta}} \; \Delta(\boldsymbol{\theta})$$

dave.zachariah@it.uu.se

Seeking the best model



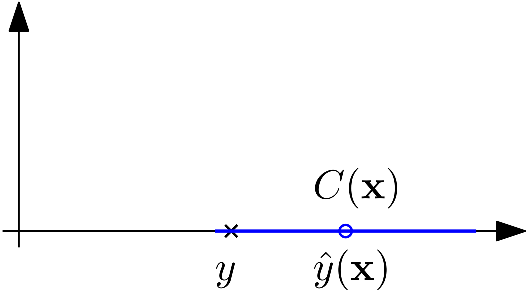
$$\boldsymbol{\theta}_\star = \arg\min_{\boldsymbol{\theta}} \; \Delta(\boldsymbol{\theta})$$

- How good is best model $\boldsymbol{\theta}_\star$?
- How far away is learned model $\widehat{\boldsymbol{\theta}}$?

dave.zachariah@it.uu.se

UPPSALA UNIVERSITET

Uncertainty region: unknown properties



Credibility:

$$\Pr\left\{y \in C_{\widehat{\theta}}(\mathbf{x}) \mid \mathbf{x}, \mathcal{D}\right\} \quad ???$$

Confidence:

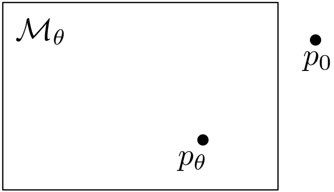
$$\Pr\left\{y \in C_{\widehat{\theta}}(\mathbf{x}) \mid \mathbf{x}\right\} \quad ???$$

dave.zachariah@it.uu.se

Distribution-free approach

dave.zachariah@it.uu.se

Simple model class



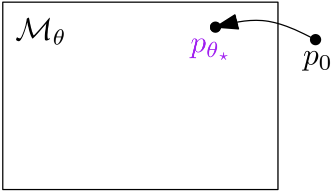
Here: i.i.d. samples

$$\mathcal{M}_\theta = \left\{ p_\theta(\mathbf{x}, \mathcal{D}) \, p_\theta(y|\mathbf{x}, \mathcal{D}) \; : \; \mu_\theta(\mathbf{x}) = \boldsymbol{\phi}^\top(\mathbf{x})\boldsymbol{\theta} \right\}$$

with p-dimensional feature vector $\boldsymbol{\phi}(\mathbf{x})$

dave.zachariah@it.uu.se

Defining the best model



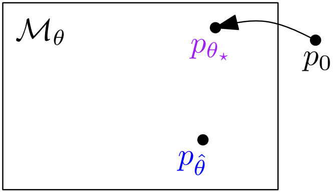
Best model:

$$\boldsymbol{\theta}_\star = \underset{\boldsymbol{\theta}\,:\,\|\boldsymbol{\theta}\|_0 \leq k}{\arg\min}\ \Delta(\boldsymbol{\theta}) \quad \text{where} \quad \Delta(\boldsymbol{\theta}) = \widehat{\mathsf{E}}\left[\left|y_i - \mu_{\boldsymbol{\theta}}(\mathbf{x}_i)\right|^2\right]$$

UPPSALA UNIVERSITET

dave.zachariah@it.uu.se

Learning a model



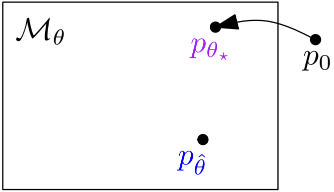
Learned model:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sqrt{\Delta(\boldsymbol{\theta})} + \frac{1}{\sqrt{n}} \|\boldsymbol{\varphi} \odot \boldsymbol{\theta}\|_1,$$

with weight $\varphi_j = \|\widetilde{\boldsymbol{\phi}}_j\|_2/\sqrt{n}$ from feature j

dave.zachariah@it.uu.se

Learning a model



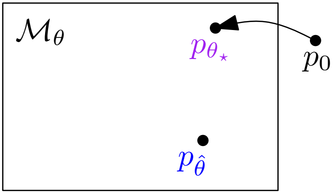
Guarantees

When φ_j above a prediction error level of best model:

$$\widehat{\mathsf{E}}\left[|\mu_{\theta_\star}(\mathbf{x}_i) - \mu_{\widehat{\theta}}(\mathbf{x}_i)|^2\right] \leq \frac{2}{n}\|\boldsymbol{\varphi} \odot \boldsymbol{\theta}_\star\|_1^2 + 4\sqrt{\frac{\Delta(\boldsymbol{\theta}_\star)}{n}}\|\boldsymbol{\varphi} \odot \boldsymbol{\theta}_\star\|_1$$

dave.zachariah@it.uu.se

Learning a model

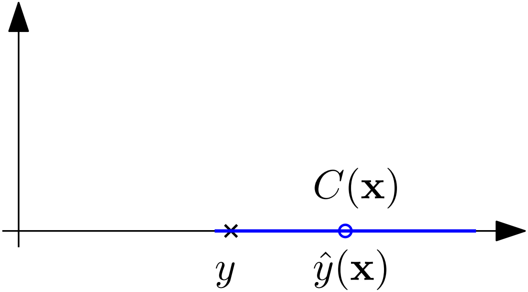


Computational requirements

Recursive computation of $\widehat{\boldsymbol{\theta}}$ via matrix-inversion free updates with runtime $\mathcal{O}(p^2 n)$ and memory $\mathcal{O}(p^2)$

dave.zachariah@it.uu.se

Uncertainty region under misspecification

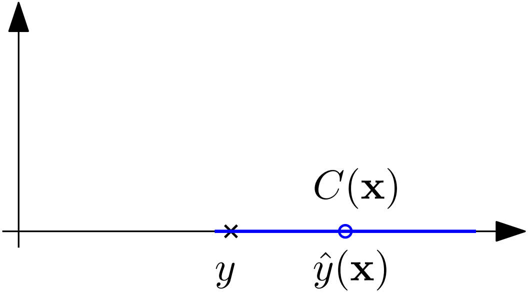


Constructing uncertainty region:

1. Randomly split $\mathcal{D}$ into $\mathcal{D}'$ and $\mathcal{D}''$
2. Learn $\widehat{\boldsymbol{\theta}}$ using $\mathcal{D}'$ and sort residuals $r_i = |y_i - \mu_{\widehat{\theta}}(\mathbf{x}_i)|$ from $\mathcal{D}''$
3. Let $\overline{r}_\kappa$ denote the $\lceil (n/2+1)\kappa \rceil$th smallest residual

dave.zachariah@it.uu.se

Uncertainty region under misspecification

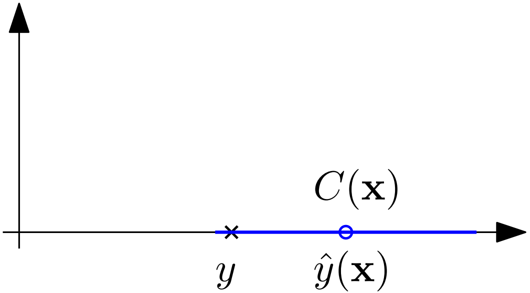


Construct uncertainty region

$$C_{\widehat{\theta}}(\mathbf{x}) = \left\{ y' \; : \; |y' - \mu_{\widehat{\theta}}| \leq \overline{r}_{\kappa} \right\}$$

dave.zachariah@it.uu.se

Uncertainty region under misspecification



Marginal confidence

When $(\mathbf{x}_i, y_i) \sim p_0(\mathbf{x}, y)$ independent,

$$\Pr\{\ y \in C_{\widehat{\theta}}(\mathbf{x})\ \}\ \geq\ \kappa$$

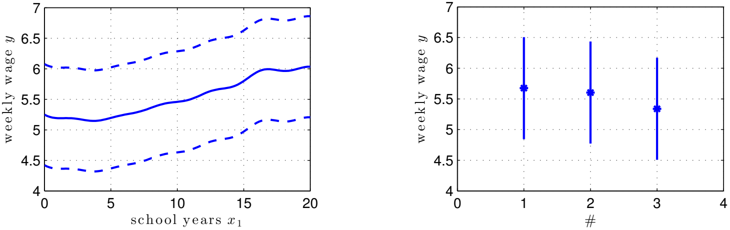
dave.zachariah@it.uu.se

UPPSALA
UNIVERSITET

Empirical illustrations

dave.zachariah@it.uu.se

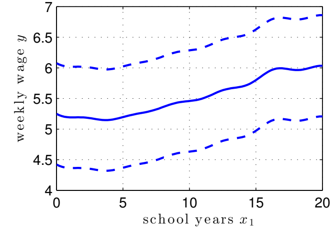
Example: US income data, 1930s cohort

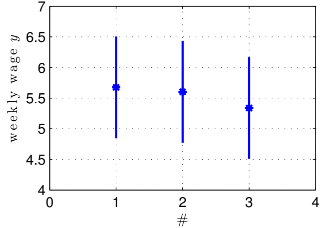


- $\mathbf{x}$ years in school, marital status, ethnicity, region, etc.
- y weekly wage [log-units]
- $n = 219\ 673$ samples
- $\mathcal{M}_\theta$ where $\phi(\mathbf{x})$ is linear and wavelet-based

dave.zachariah@it.uu.se

Example: US income data, 1930s cohort

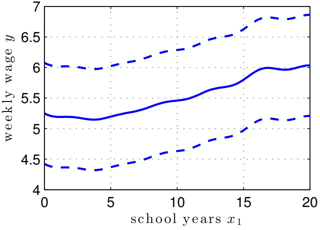


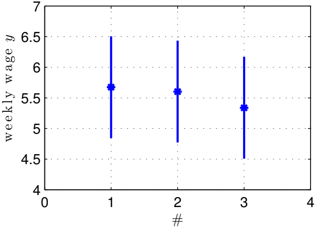


Left: Predicted wage vs. years in school.
Right: Predicted wage for individuals with 12 years of schooling, but differing backgrounds.

dave.zachariah@it.uu.se

Example: US income data, 1930s cohort





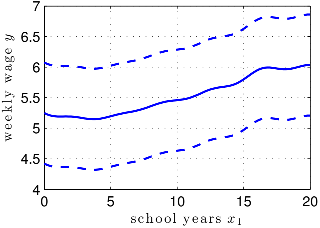
Output dynamic range $y \in [-2.34,\ 10.53]$

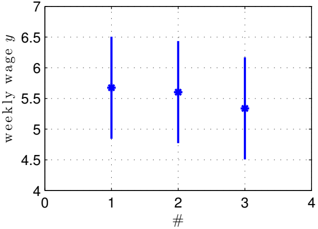
$$\Rightarrow \widehat{\text{RMSE}} = 0.62$$

using $\bar{n} \sim 110 \times 10^3$ test individuals

dave.zachariah@it.uu.se

Example: US income data, 1930s cohort





Output dynamic range $y \in [-2.34,\ 10.53]$

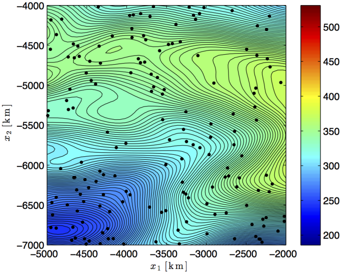
$$\kappa = 0.90 \quad \Rightarrow \quad \Pr\widehat{\{y \in C_{\widehat{\theta}}\}} = 89.9\%$$

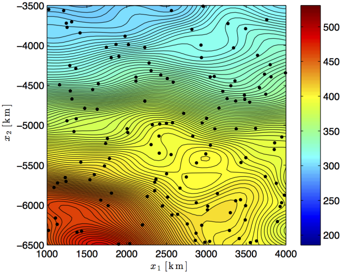
using $\bar{n} \sim 110 \times 10^3$ test individuals

dave.zachariah@it.uu.se

UPPSALA UNIVERSITET

Example: Global ozone data

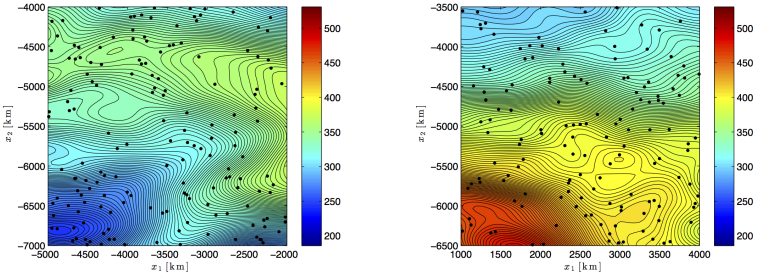




- $\mathbf{x}$ spatial coordinates [km]
- y ozone density [DU]
- $n = 17\,340$ samples
- $\mathcal{M}_\theta$ where $\phi(\mathbf{x})$ is wavelet-based

dave.zachariah@it.uu.se

Example: Global ozone data



Output dynamic range $y \in [179,\ 542]$

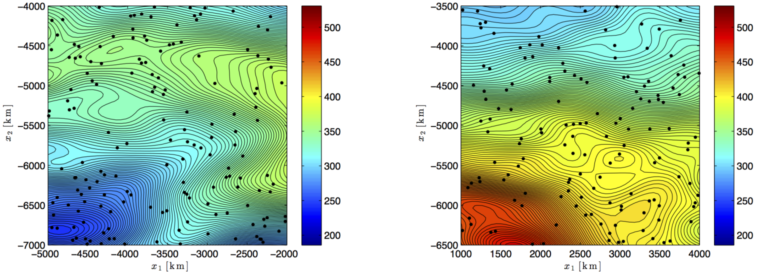
$$\Rightarrow \widehat{\text{RMSE}} = 6.74$$

using $\bar{n} \sim 156 \times 10^3$ test samples

dave.zachariah@it.uu.se

UPPSALA
UNIVERSITET

Example: Global ozone data



Output dynamic range $y \in [179,\ 542]$

$$\kappa = 0.90 \quad \Rightarrow \quad \widehat{\Pr\{y \in C_{\hat{\theta}}\}} = 90.0\%$$

using $\bar{n} \sim 156 \times 10^3$ test samples

dave.zachariah@it.uu.se

UPPSALA
UNIVERSITET

Conclusions

dave.zachariah@it.uu.se

UPPSALA UNIVERSITET

Conclusions

1. Subjectivist-Bayesian vs. Frequentist interpretations

dave.zachariah@it.uu.se

UPPSALA UNIVERSITET

Conclusions

1. Subjectivist-Bayesian vs. Frequentist interpretations
   - Neither Bayes' rule nor prior are distinguishing features ...
   - ... rather belief vs. coverage

dave.zachariah@it.uu.se

Conclusions

1. Subjectivist-Bayesian vs. Frequentist interpretations
   - Neither Bayes' rule nor prior are distinguishing features ...
   - ... rather belief vs. coverage
2. Bayesian paradigm is powerful but possibly intoxicating

dave.zachariah@it.uu.se

Conclusions

1. Subjectivist-Bayesian vs. Frequentist interpretations
   - Neither Bayes' rule nor prior are distinguishing features ...
   - ... rather belief vs. coverage
2. Bayesian paradigm is powerful but possibly intoxicating
   - Construct excellent decision rules ...
   - ... but also seriously misleading uncertainty regions

dave.zachariah@it.uu.se

UPPSALA UNIVERSITET

Conclusions

1. Subjectivist-Bayesian vs. Frequentist interpretations
    - Neither Bayes' rule nor prior are distinguishing features ...
    - ... rather belief vs. coverage
2. Bayesian paradigm is powerful but possibly intoxicating
    - Construct excellent decision rules ...
    - ... but also seriously misleading uncertainty regions
3. Distribution-free approach can tackle misspecification

dave.zachariah@it.uu.se

UPPSALA UNIVERSITET

Conclusions

1. Subjectivist-Bayesian vs. Frequentist interpretations
   - Neither Bayes' rule nor prior are distinguishing features ...
   - ... rather belief vs. coverage
2. Bayesian paradigm is powerful but possibly intoxicating
   - Construct excellent decision rules ...
   - ... but also seriously misleading uncertainty regions
3. Distribution-free approach can tackle misspecification
   - Simple model class with meaningful best model
   - Performance guarantees and scalable in n ...
   - ... uncertainty regions with valid coverage

dave.zachariah@it.uu.se