Context models on sequences of covers

Non-parametric closed-form Bayesian estimation of (conditional) measures

Christos Dimitrakakis

FIAS, J.W. Goethe University

March 20, 2012

Setup

- Observations $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.
- Sequences $x^t \triangleq (x_k : k = 1, \ldots, t)$, $x_k \in \mathcal{X}$, $y^t \triangleq (y_k : k = 1, \ldots, t)$, $y_k \in \mathcal{Y}$.
- The set of all sequences $\mathcal{X}^* \triangleq \bigcup_k \mathcal{X}^k$.
- Problem: online estimation of $\mathbb{P}(y_{t+1} \mid x^{t+1}, y^t)$.

Main idea

Use contextual independence to break down problem.

Applications

- (Conditional) density estimation
- Regression
- Clustering
- Classification

Density estimation

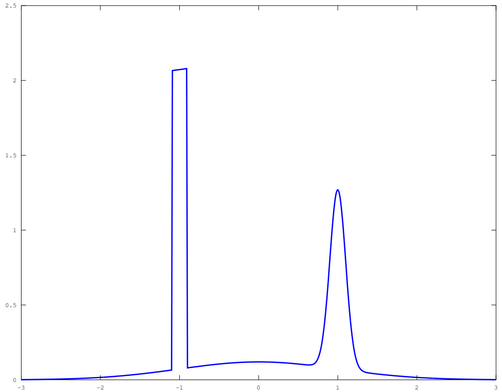


Figure: The generating density

Histogram-based methods

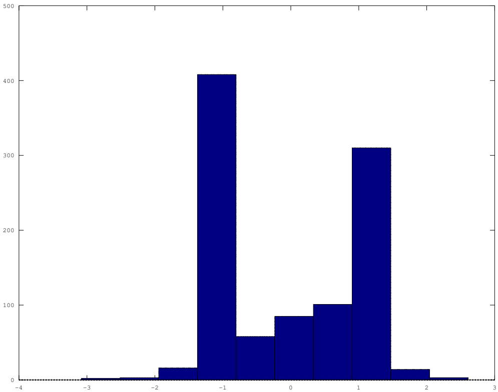


Figure: 10^3 samples

Problem: How to choose the number of bins.

Histogram-based methods

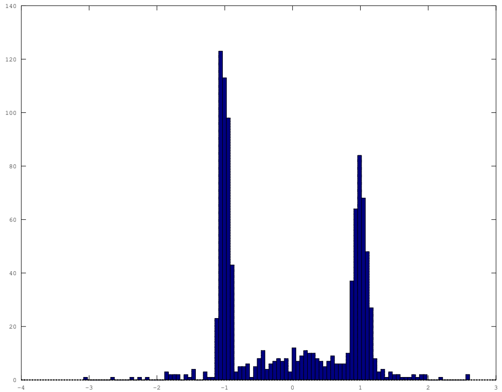


Figure: 10^3 samples

Problem: How to choose the number of bins.

Histogram-based methods

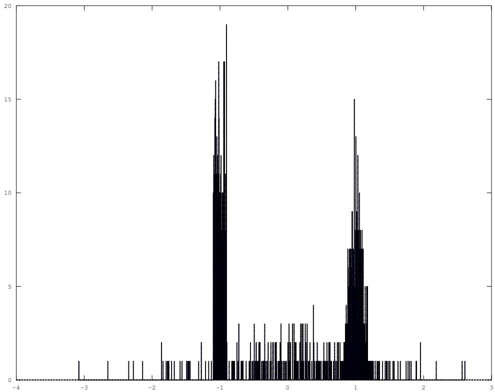


Figure: 10^3 samples

Problem: How to choose the number of bins.

Kernel-based methods

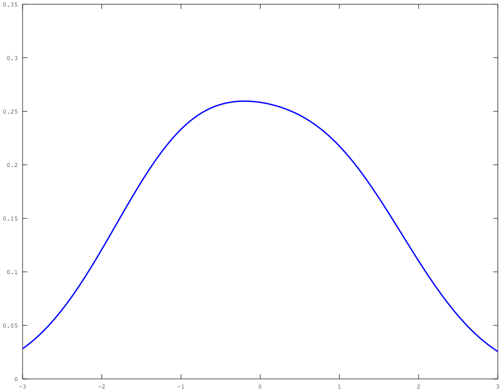


Figure: 10^3 samples

Problem: How to choose the bandwidth.

Kernel-based methods

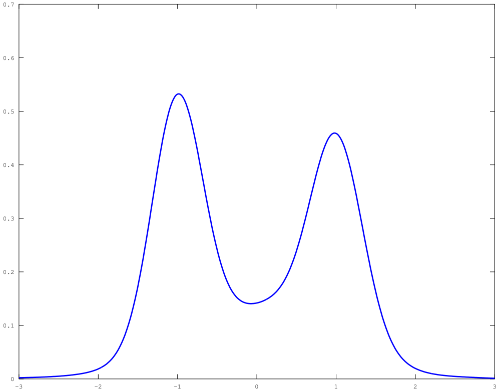


Figure: 10^3 samples

Problem: How to choose the bandwidth.

Kernel-based methods

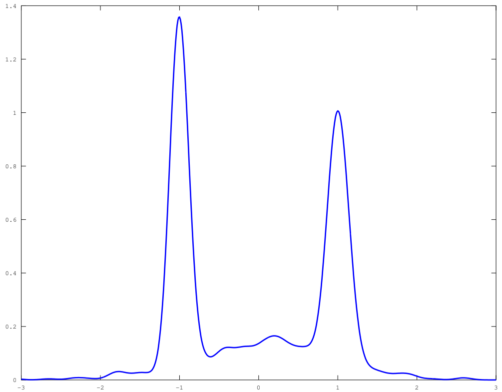


Figure: 10^3 samples

Problem: How to choose the bandwidth.

Kernel-based methods

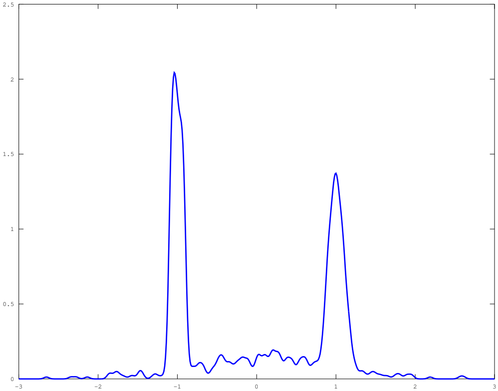


Figure: 10^3 samples

Problem: How to choose the bandwidth.

High-probability bounds

Lemma (Hoeffding's inequality)

If $X_i \in [0, 1]$, are independently distributed and $S_n = \frac{1}{n}\sum_{i=1}^n X_i$:

$$\mathbb{P}\left(|S_n - \mathbb{E}\, S_n| > \epsilon\right) < 2e^{-n\epsilon^2}. \tag{1}$$

High probability histograms

- k observations in $\mathcal{X}$, acceptable error probability δ.
- We partition $\mathcal{X}$ into $k^{1/3}$ sets each containing at least $k^{2/3}$.
- We use this partition as the basis for an empirical measure q.
- With probability at least $1 - \delta$, the error of the empirical measure is uniformly bounded by

$$\sqrt{\frac{\ln \frac{2}{\delta} k^{1/3}}{2k^{2/3}}} = \sqrt{\frac{\ln \frac{2}{\delta} + \frac{1}{3} \ln k}{2k^{2/3}}} = \tilde{O}(k^{1/3})$$

Histogram-based methods

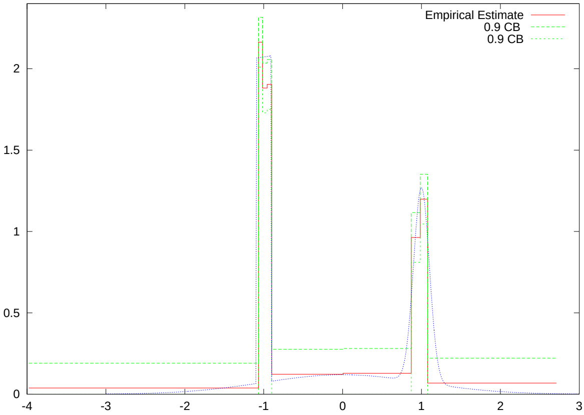


Figure: 10^3 samples

- Problem (minor): Requires sorting.
- Advantage: It is prior-free.

Histogram-based methods

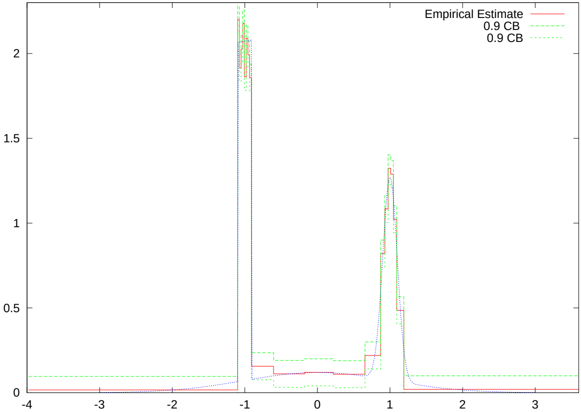


Figure: 10^4 samples

- Problem (minor): Requires sorting.
- Advantage: It is prior-free.

Histogram-based methods

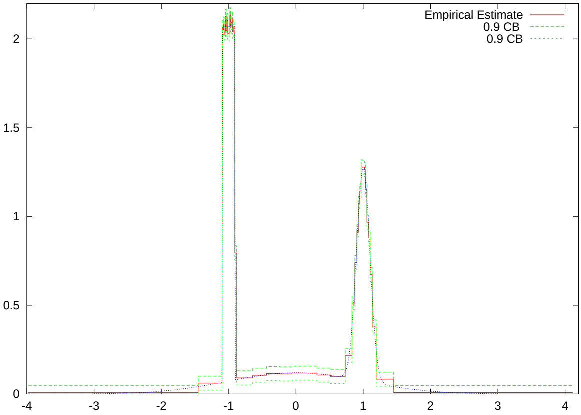


Figure: 10^5 samples

- Problem (minor): Requires sorting.
- Advantage: It is prior-free.

Histogram-based methods

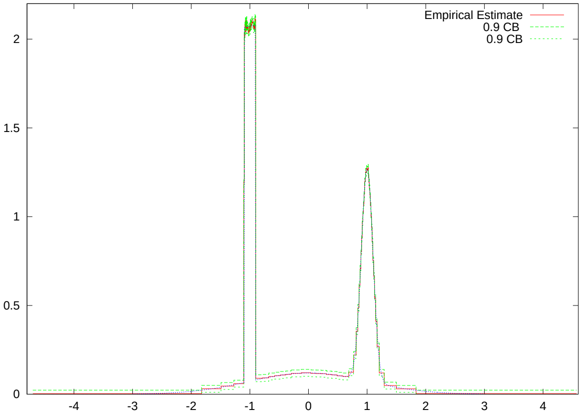


Figure: 10^6 samples

- Problem (minor): Requires sorting.
- Advantage: It is prior-free.

Context models

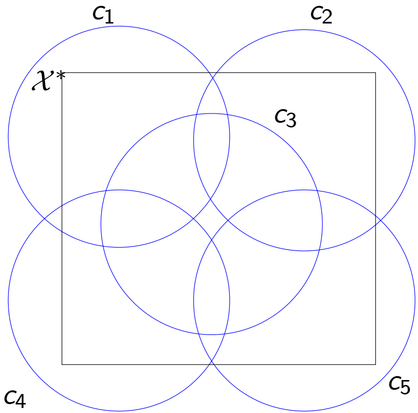


Figure: A cover set $S = \{c_1, c_2, c_3, c_4, c_5\}$.

- Let S be a cover of $\mathcal{X}^*$.

Context models

- Let S be a cover of $\mathcal{X}^*$.
- For each $c \in S$, define a model

$$\mathbb{P}(x_{t+1} \mid x^t, c).$$

- We may now estimate

$$\mathbb{P}(x_{t+1} \mid x^t) = \sum_c \mathbb{P}(x_{t+1} \mid x^t, c)\, \mathbb{P}(c \mid x^t).$$

Context models

- Let S be a cover of $\mathcal{X}^*$.
- For each $c \in S$, define a model

$$\mathbb{P}(x_{t+1} \mid x^t, c).$$

- We may now estimate

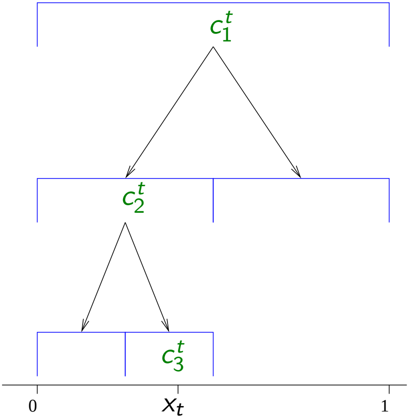
$$\mathbb{P}(x_{t+1} \mid x^t) = \sum_c \mathbb{P}(x_{t+1} \mid x^t, c)\, \mathbb{P}(c \mid x^t).$$

Issues

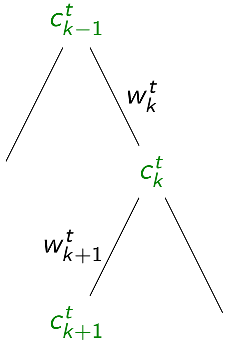
- What should the structure of S be?
- How can we estimate $\mathbb{P}(c \mid x^t)$?

Example: binary partition tree

$$\mathcal{X} = [1, 0]$$



A random walk on trees



Intution

- For every x^t, obtain a sequence $c^t_1, \ldots, c^t_D$, $D \leq t$.
- Mix the prediction of c^t_k with the predictions of $c^t_1, \ldots, c^t_{k-1}$.
- For each x^t, start from the deepest matching context c^t_D and walk up. When at level k, stop w.p. w^t_k.

The update

- Let B_k be the event that we stop at $1, \ldots, k$. Then define:

$$w_k^t \triangleq \mathbb{P}(c_k^t \mid B_k, x^t) \qquad \text{(p. of stopping at level } k\text{)}$$

$$\phi_k^t(x_{t+1}) \triangleq \mathbb{P}(x_{t+1} \mid x^t, c_k^t) \qquad \text{(prediction of } k\text{-th context)}$$

The update

- Let B_k be the event that we stop at $1, \ldots, k$. Then define:

$$w_k^t \triangleq \mathbb{P}(c_k^t \mid B_k, x^t) \qquad \text{(p. of stopping at level } k\text{)}$$

$$\phi_k^t(x_{t+1}) \triangleq \mathbb{P}(x_{t+1} \mid x^t, c_k^t) \qquad \text{(prediction of } k\text{-th context)}$$

- We obtain the following recursion:

$$\mathbb{P}(x_{t+1} \mid x^t, B_k) = \phi_k^t(x_{t+1}) w_k^t + \mathbb{P}(x_{t+1} \mid x^t, B_{k-1})(1 - w_k^t)$$

The update

- Let B_k be the event that we stop at $1, \ldots, k$. Then define:

$$w_k^t \triangleq \mathbb{P}(c_k^t \mid B_k, x^t) \qquad \text{(p. of stopping at level } k\text{)}$$

$$\phi_k^t(x_{t+1}) \triangleq \mathbb{P}(x_{t+1} \mid x^t, c_k^t) \qquad \text{(prediction of } k\text{-th context)}$$

- We obtain the following recursion:

$$\mathbb{P}(x_{t+1} \mid x^t, B_k) = \phi_k^t(x_{t+1}) w_k^t + \mathbb{P}(x_{t+1} \mid x^t, B_{k-1})(1 - w_k^t)$$

$$w_k^{t+1} = \frac{\phi_k^t(x_{t+1}) w_k^t}{\mathbb{P}(x_{t+1} \mid x^t, B_k)} \tag{2}$$

Application to density estimation

- First use of stopping: Hutter 2005, BayesTree
- Extension to sampling trees: Wong and Ma, 2010, Optional Pólya tree.

A context tree estimator

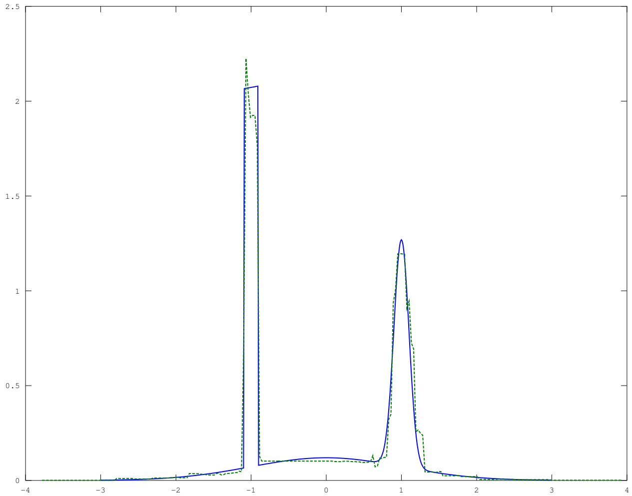


Figure: 10^3 samples

A context tree estimator

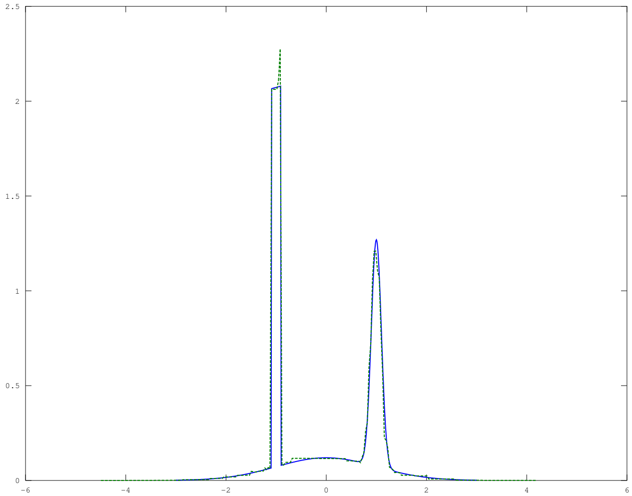


Figure: 10^4 samples

A context tree estimator

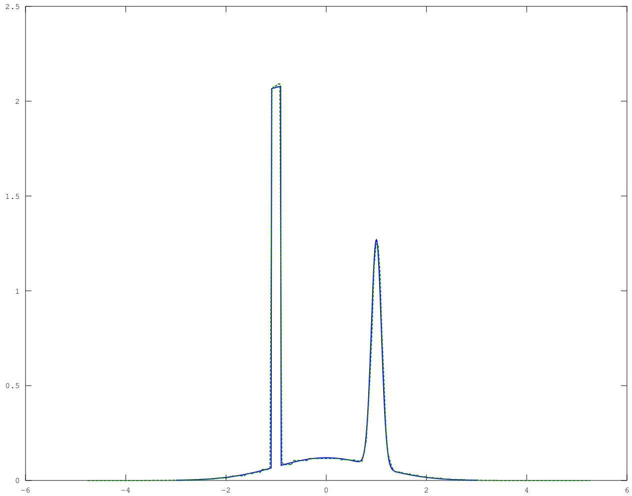


Figure: 10^5 samples

A context tree estimator

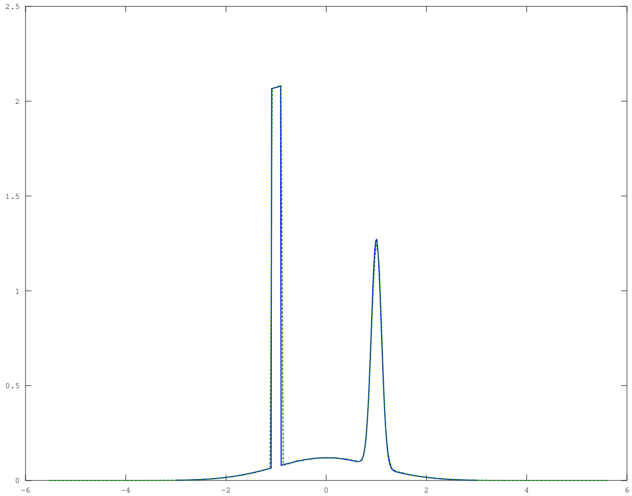


Figure: 10^6 samples

Inference on sequences of covers

Local mixtures

A collection C_k^t at each level k, s.t. $x^t \in c \; \forall c \in C_k^t$.

$$\mathbb{P}(x_{t+1} \mid B_k, x^t) = \psi_k^t(x_{t+1}) w_k^t + (1 - w_k^t)\, \mathbb{P}(x_{t+1} \mid B_{k-1}, x^t), \quad (3)$$

where

$$\psi_k^t(x_{t+1}) \triangleq \mathbb{P}(x_{t+1} \mid c \in C_k^t, x^t) \quad (4)$$

is the prediction at level k. If we stop, we select the i-th context from C_k^t, with probability:

$$v_{k,i}^t \triangleq \mathbb{P}(c = i \mid c \in C_k^t, x^t). \quad (5)$$

- Further generalisations possible at increased computation cost.
- Relaxes the requirement to define a partition tree.
- The problem of generating suitable covers remains.

Generating the covers

- It is better to use a data-driven process.
- In our case, $\mathcal{X} \subset \mathbb{R}^n$, so we used a KD-tree.
- Cover trees are also possible.

Indirect conditional density estimation

Naive approach

- Estimate the joint distribution $\mathbb{P}(x, y)$.
- Explicitly calculate the conditional

$$\mathbb{P}(y \mid x) = \frac{\mathbb{P}(x, y)}{\mathbb{P}(x)} = \frac{\mathbb{P}(x, y)}{\int_{\mathcal{Y}} \mathbb{P}(x, y) \, \mathrm{d}\mu(y)}$$

Context-based Conditional density estimation

Consider $x \in \mathcal{X}, y \in \mathcal{Y}$ and modelling the conditional density $f(y|x) = f(x, y)/f(x)$.

Modelling a density at each context

For any cover we $\mathcal{X}$, we use local models:

$$\phi_c(y \mid x) \triangleq f(y \mid c, x),$$

where the dependence on x may be dropped. We then have:

$$f(y \mid x) = \sum_c \mathbb{P}(c \mid x)\phi_c(y \mid x)$$

For example the density $\phi_c(y \mid x)$ can be modelled by a tree defined on the same covers, a Gaussian, or a mixture of both.

Density estimation

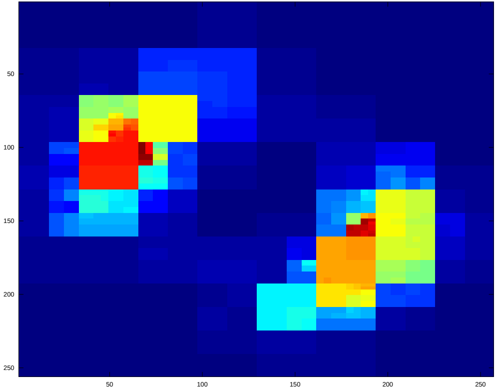


Figure: 10^3 samples

KD-tree partition on $\mathbb{R}^2$.

Density estimation

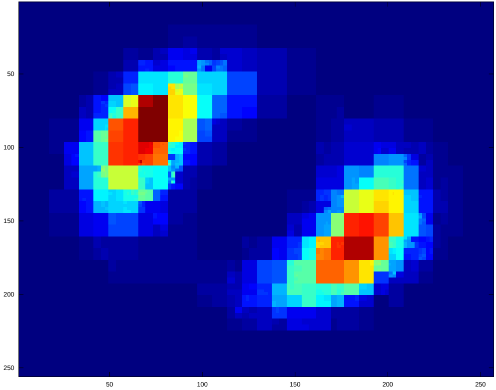


Figure: 10^4 samples

KD-tree partition on $\mathbb{R}^2$.

Density estimation

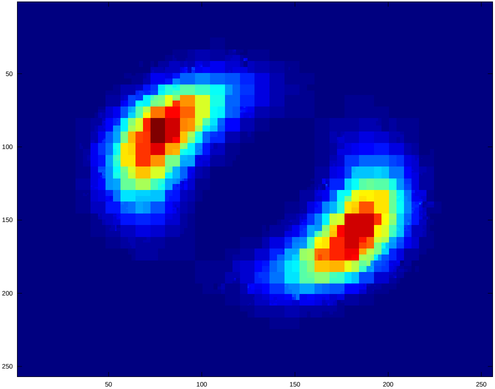


Figure: 10^5 samples

KD-tree partition on $\mathbb{R}^2$.

Density estimation

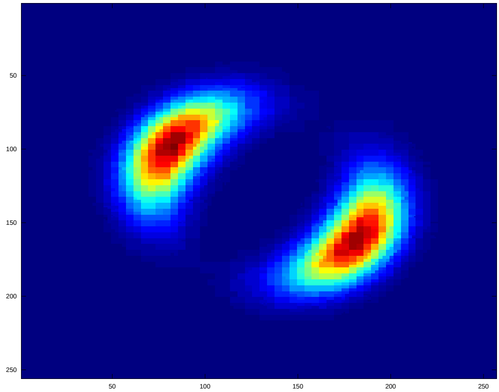


Figure: 10^6 samples

KD-tree partition on $\mathbb{R}^2$.

Conditional density estimation

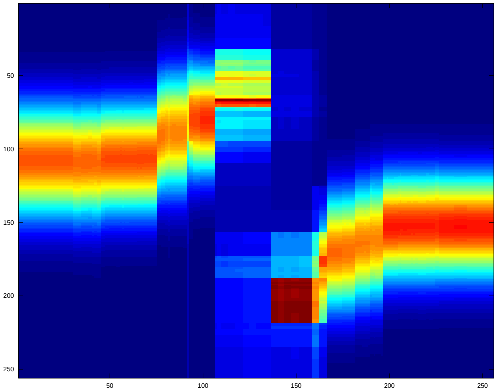


Figure: 10^3 samples

KD-tree partition on $\mathcal{X}$, mixture of KD-tree density estimates and Normal-Wishart estimates for $\mathcal{Y}$.

Conditional density estimation

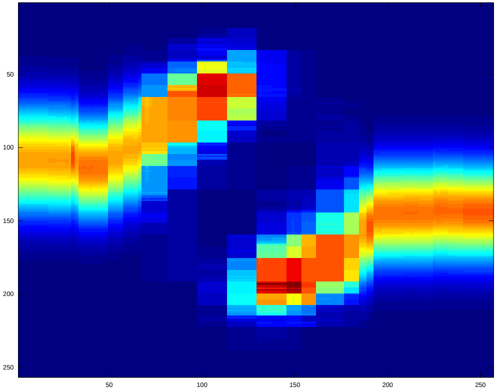


Figure: 10^4 samples

KD-tree partition on $\mathcal{X}$, mixture of KD-tree density estimates and Normal-Wishart estimates for $\mathcal{Y}$.

Conditional density estimation

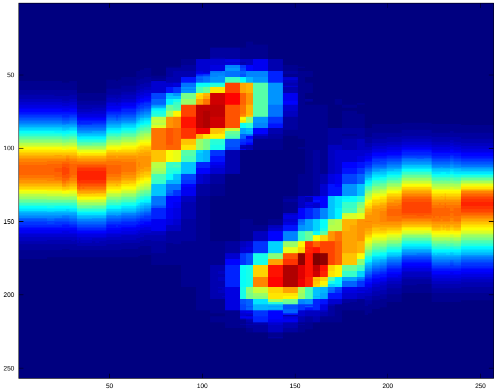


Figure: 10^5 samples

KD-tree partition on $\mathcal{X}$, mixture of KD-tree density estimates and Normal-Wishart estimates for $\mathcal{Y}$.

Conditional density estimation

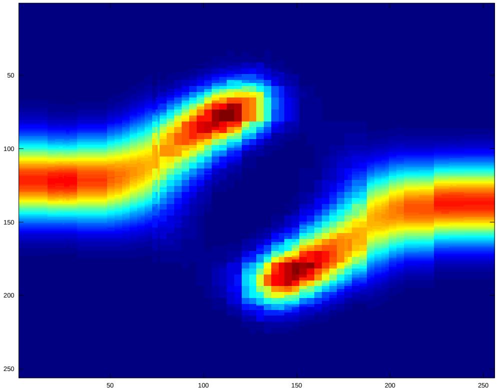


Figure: 10^6 samples

KD-tree partition on $\mathcal{X}$, mixture of KD-tree density estimates and Normal-Wishart estimates for $\mathcal{Y}$.

Classification

Consider observations $x \in \mathcal{X}$ and class labels $y \in \mathcal{Y}$.

Using a classifier at each context

For any cover we $\mathcal{X}$, we use a local classifier:

$$\phi_c(y \mid x) \triangleq f(y \mid c, x),$$

where the dependence on x may be dropped. We then have:

$$f(y \mid x) = \sum_c \mathbb{P}(c \mid x)\phi_c(y \mid x).$$

The classifier can be a linear, nearest-neighbour, a mixture ...

Conclusion

Results

- Incremental, fast, closed-form Bayesian inference.
- Automatically adjusts to amount of available data.
- Has close to state-of-the-art performance.
- Very general setting.

Extensions

- Application to reinforcement learning.

Open problems

- Finite-sample bounds.
- Smoothing.