# Statistical Regular Pavings for Nonparametric Density Estimation: Emphasizing Tree Arithmetic

**Raaz**esh Sainudiin

http://lamastex.org

joint work with: Gloria Teng, Jennifer Harlow and Warwick Tucker

Laboratory for Mathematical Statistical Experiments, Christchurch Centre, and
Centre for Computable and Constructive Mathematics,
School of Mathematics and Statistics, University of Canterbury,
Christchurch, New Zealand

November 17, 2016, Oaxaca, Mexico

CMO Workshop in Constructive Mathematics and Interval Analysis, Oaxaca, Mexico

Real-world Motivations
- Air Traffic Co-trajectories. Teng, Kuhn & S, J. Aerospace Comput. Inf. & Com., 2012.
- A Hume Phenomenon. An Enquiry Conerning Human Undestanding, 1777.

Non-parametric Density Estimation

Arithmetic and Algebra of Plane Binary Trees - 2014 Primer

Randomized Priority Queue Markov chain

Minimum Distance Estimation
- Setting up MDE
- An Example

Conclusions and References

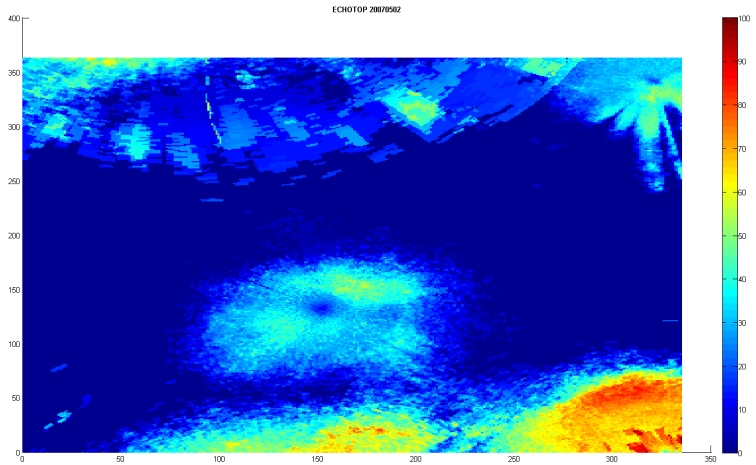
Section 1

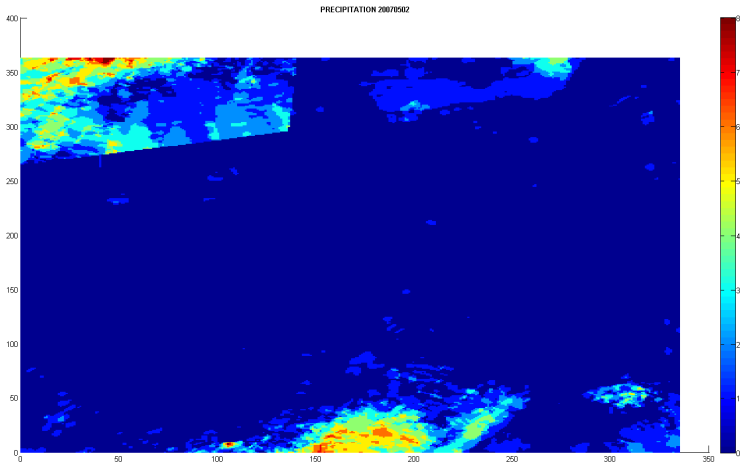
# Real-world Motivations

# 12 GBz Data for 59-days of Weather & Air Traffic

## Weather Data over Atlanta, GA – Cloud height & Precipitation

Measured every 1/2-hour and predicted every 5 minutes




There are $45 \times 59 = 2655$ half-an-hour blocks of weather data.

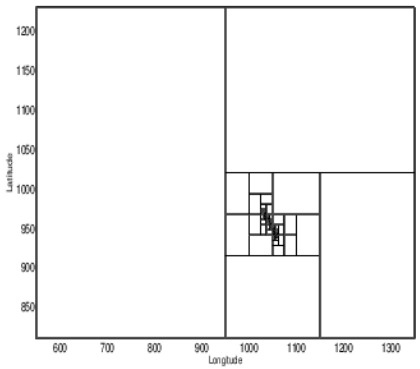
# 12 GBz Data for 59-days of Weather & Air Traffic

## Weather Tree – Leaves are Time-blocks over 65 days

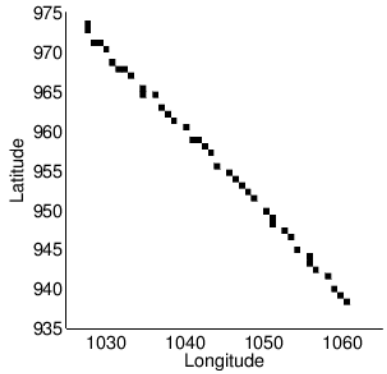
Neighbor-Joining Tree from Pairwise L1 Distances between Cloud height & Precipitation

# Dynamic $\mathbb{Z}$-MRPs for 12GB-z (59-days of Air Traffic)

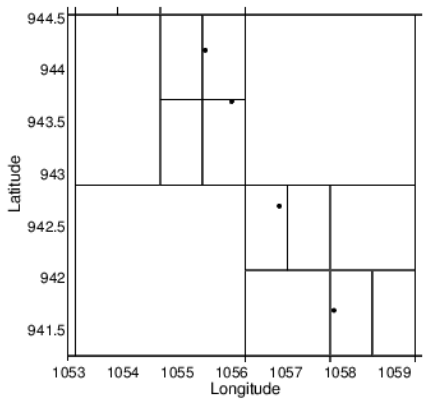
$\mathbb{Z}$-MRP of an aircraft trajectory and its tree
(every 4-6 sec. position data from radar sweep)



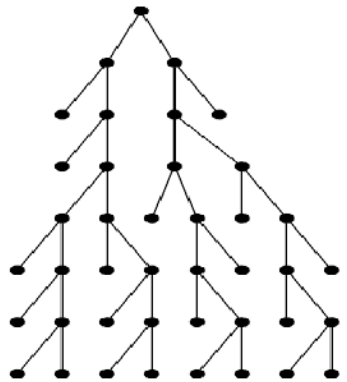

(a) SRP trajectory for aircraft position data.




(b) Shaded boxes in the SRP trajectory.



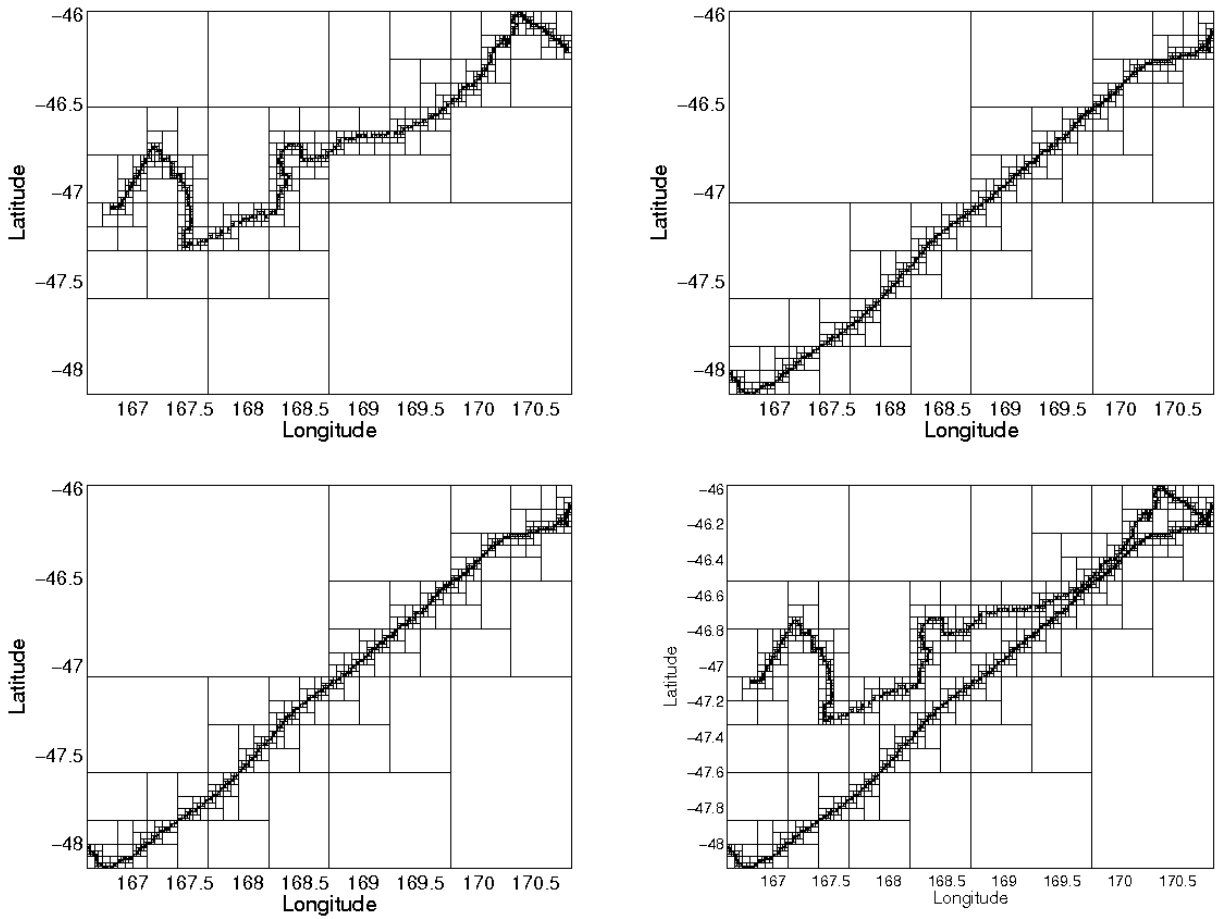

(c) Aircraft positions enclosed by boxes.



(d) The tree corresponding to (c).

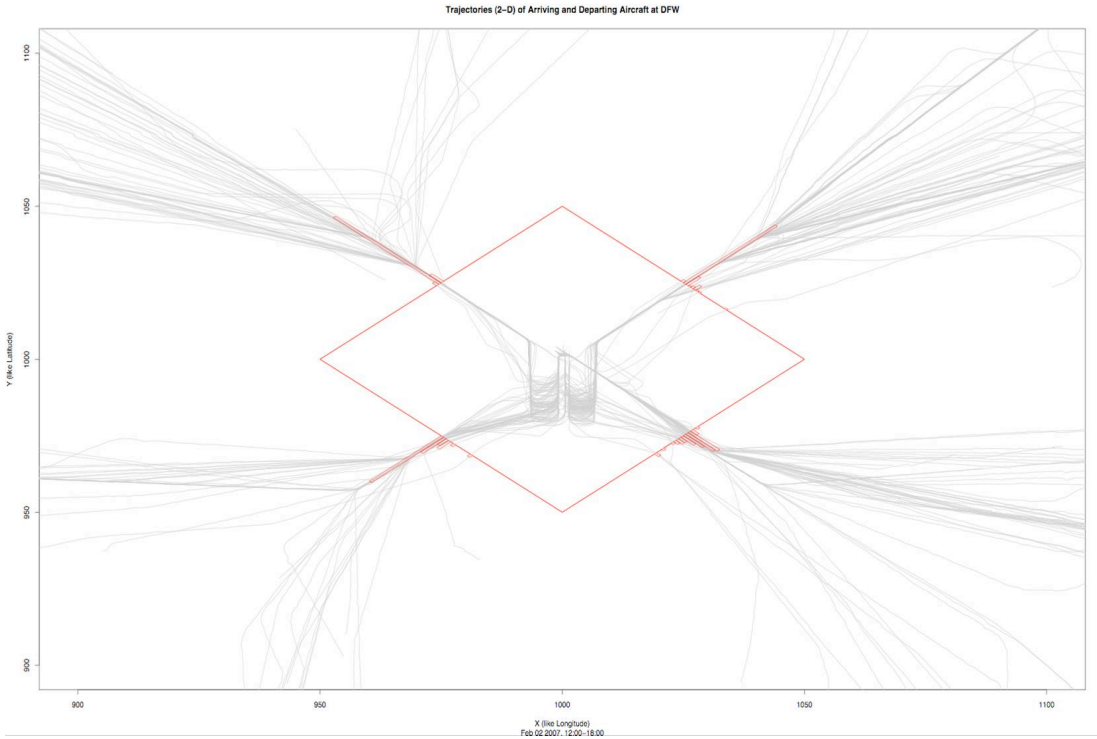
# Dynamic $\mathbb{Z}$-MRPs for 12GB-z (59-days of Air Traffic)

Three individual trajectories and their sum as $\mathbb{Z}$-MRPs



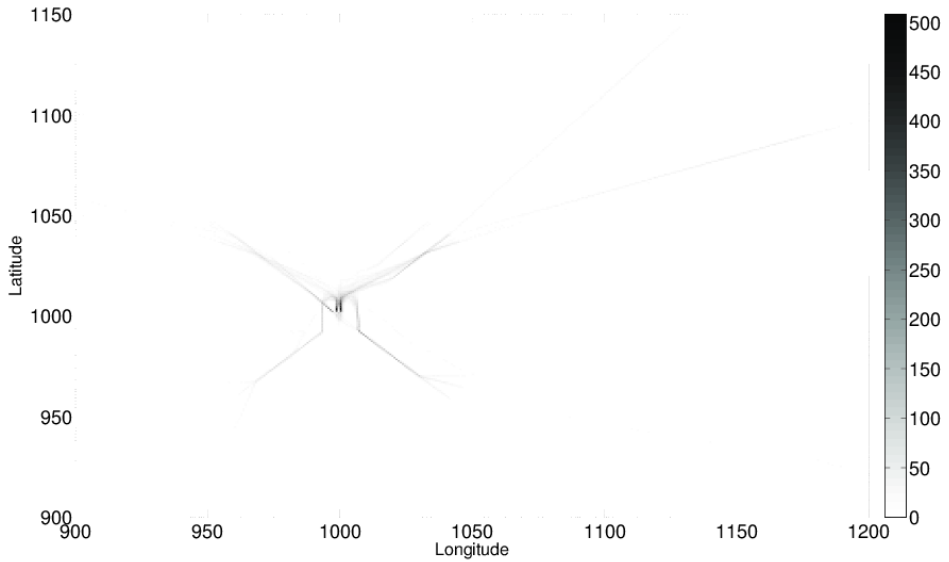
# Dynamic $\mathbb{Z}$-MRPs for 12GB-z (59-days of Air Traffic)

## On a Sunny Day over Atlanta, GA



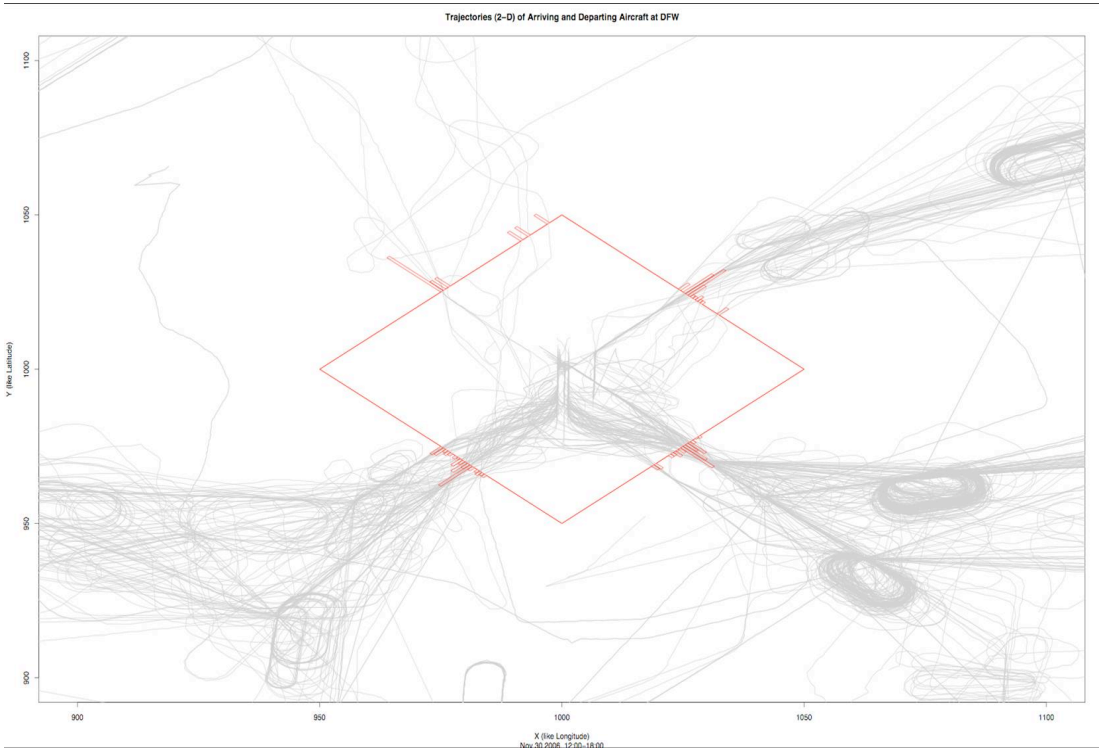
# Dynamic $\mathbb{Z}$-MRPs for 12GB-z (59-days of Air Traffic)

## $\mathbb{Z}$-MRP On this Sunny Day



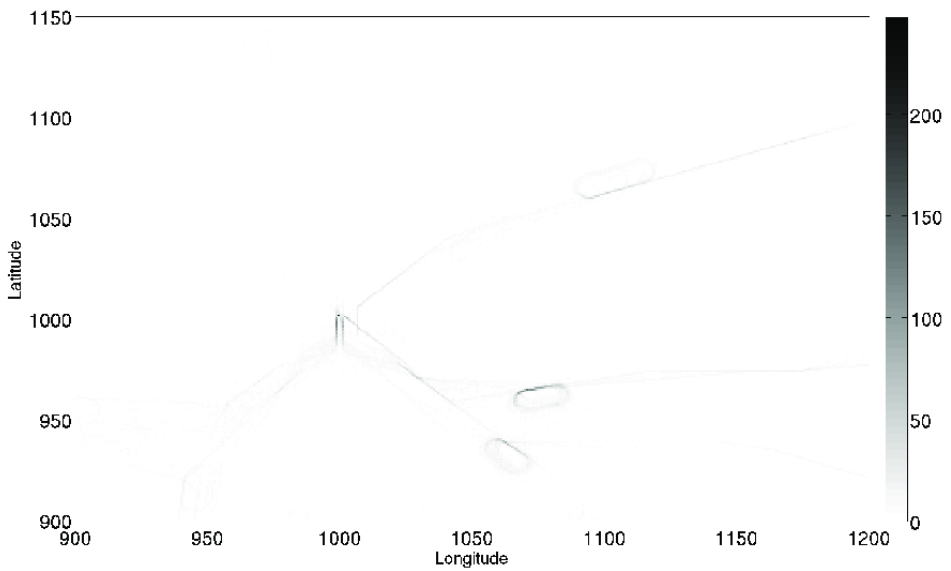
# Dynamic $\mathbb{Z}$-MRPs for 12GB-z (59-days of Air Traffic)

## On a Stormy Day over Altanta, GA



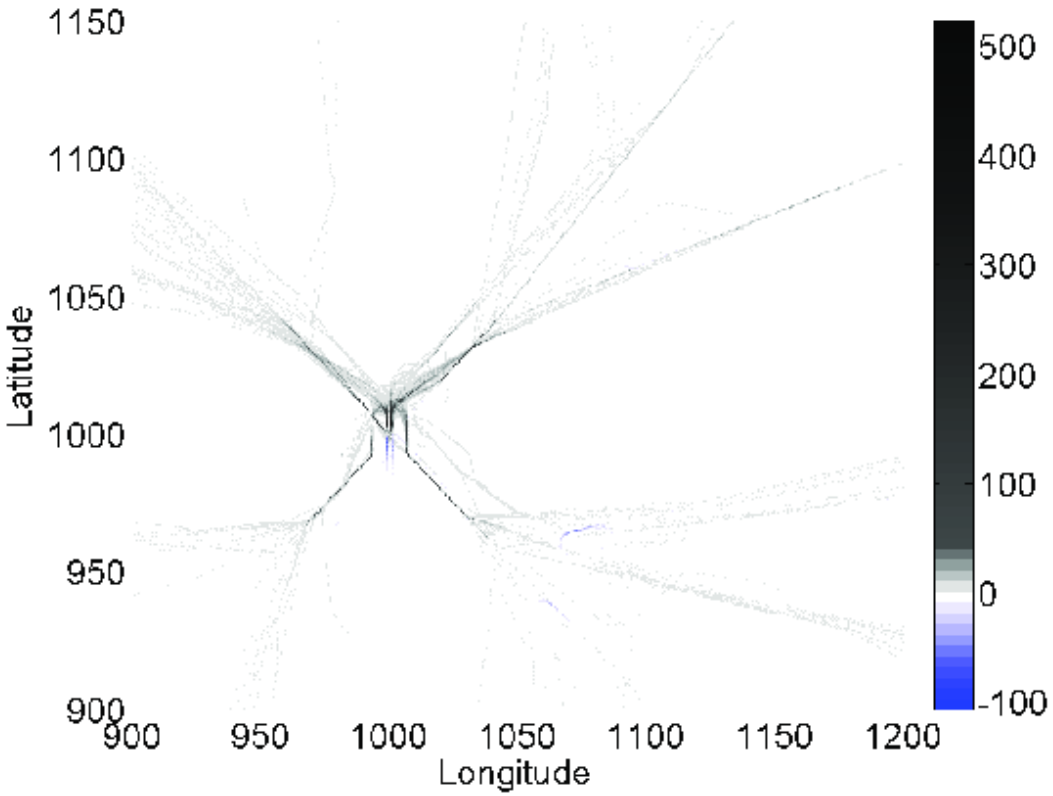
# Dynamic $\mathbb{Z}$-MRPs for 12GB-z (59-days of Air Traffic)

## $\mathbb{Z}$-MRP On this Stormy Day



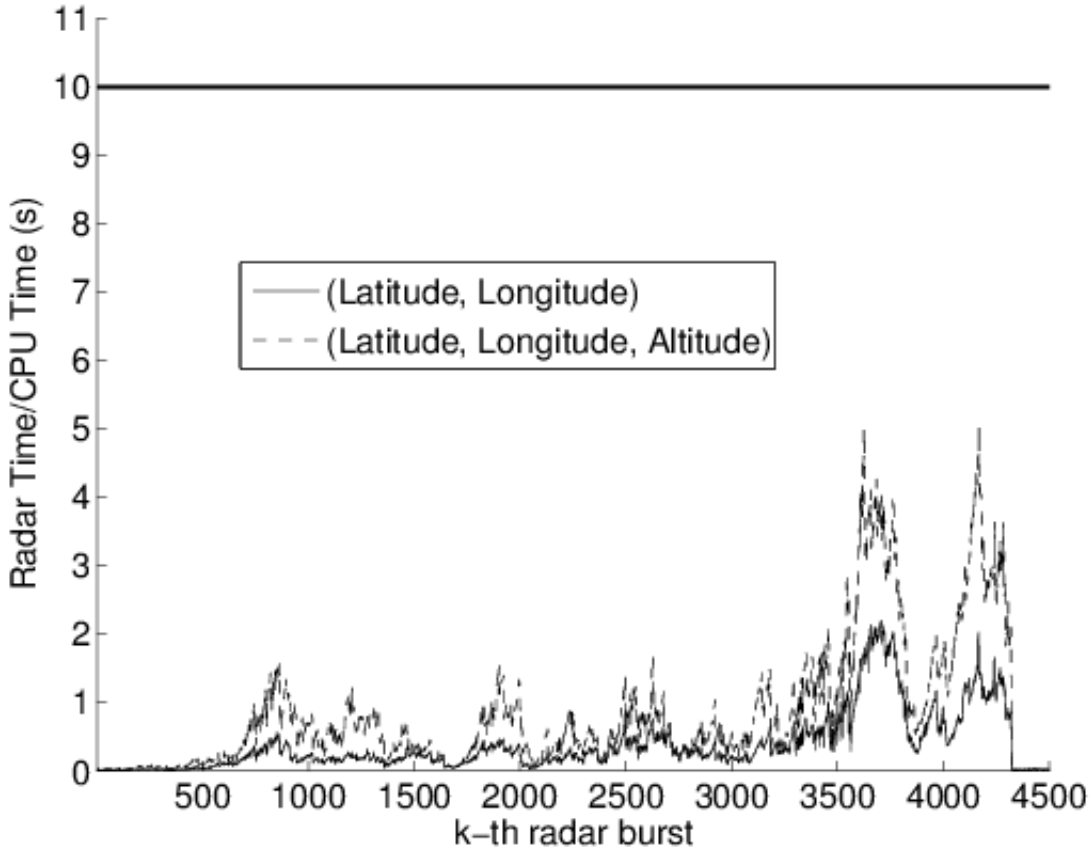
# Dynamic $\mathbb{Z}$-MRPs for 12GB-z (59-days of Air Traffic)

## $\mathbb{Z}$-MRP pattern for Sunny Day $-$ Stormy Day



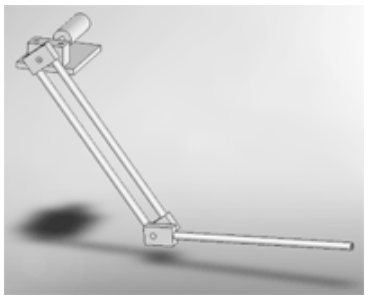
# Dynamic $\mathbb{Z}$-MRPs for 12GB-z (59-days of Air Traffic)

## $\mathbb{Z}$-MRP Dynamic Trees Can be Created in Real-time



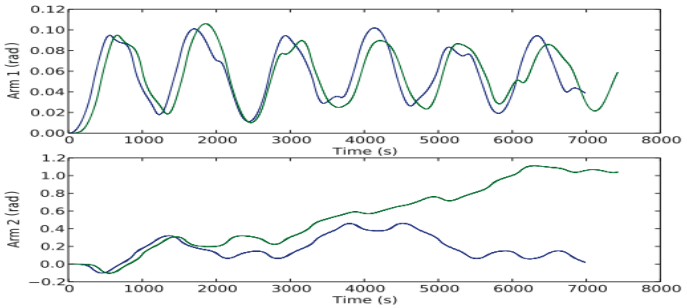
# Phenomenon: Damped Double Pendulum Trajectories

A: DP Schematic



B: Streaming DP data

C: Enclosures of two initially close trajectories



# ODE Model: Damped Double Pendulum Trajectories

## Parametric model

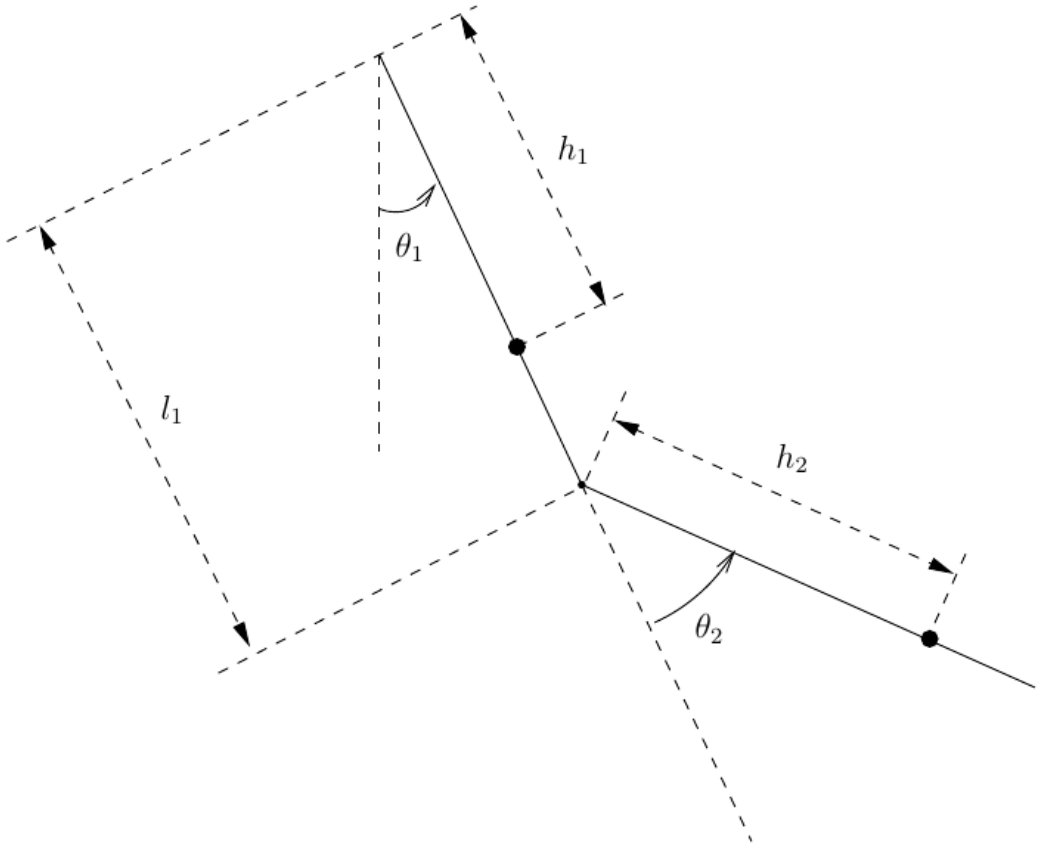



Figure 1: Scematic for the DP model parameters

# ODE Model: Damped Double Pendulum Trajectories

## Parametric model contd...

### 2 Parametric Model for Passive DP

Let the centre of mass of the inner arm is distance $h_1$ away from the main pivot that is located at the origin $(0, 0)$. Let the distance between pivots of the inner arm be $l_1$ and the centre of mass of the outer arm be at distance $l_2$ from the bottom pivot. We can measure the angular position $\theta_1$ and $\theta_2$ of the inner and outer arms, respectively, as shown in Figure 1. Thus, $(x_1, y_1)$ and $(x_2, y_2)$, the centres of mass of the inner and outer arms, respectively are given by:

$$x_1 = h_1 \sin(\theta_1), \qquad x_2 = l_1 \sin(\theta_1) + h_2 \sin(\theta_1 + \theta_2) \tag{1}$$

$$y_1 = -h_1 \cos(\theta_1), \quad y_2 = -\left(l_1 \cos(\theta_1) + h_2 \cos(\theta_1 + \theta_2)\right) \tag{2}$$

Let $m_1$ and $m_2$ be the masses of the outer and inner arms, respectively and $g$ be the acceleration due to gravity. Let $I_1$ be the moment of inertia of the inner arm about its pivot at the origin and $I_2$ be that of the outer arm about its centre of mass at $(x_2, y_2)$.

The kinetic energy $T_1$ of the inner arm and that $T_2$ of the outer arm are

$$T_1 = \frac{1}{2} I_1 \dot{\theta}_1^2 \ , \tag{3}$$

$$T_2 = \frac{1}{2} I_2 \dot{\theta}_2^2 + \frac{1}{2} m_2 \left(\dot{x}_2^2 + \dot{y}_2^2\right) \ . \tag{4}$$

# ODE Model: Damped Double Pendulum Trajectories

## Parametric model contd...

The potential energies of the inner and outer arms, respectively, are

$$U_1 = m_1 g y_1 = -m_1 g h_1 \cos\theta_1 \ , \tag{5}$$
$$U_2 = m_2 g y_2 = -m_2 g \left(l_1 \cos(\theta_1) + h_2 \cos(\theta_1 + \theta_2)\right) \ . \tag{6}$$

The Lagrangian $L$ is

$$\begin{aligned} L &= T_1 + T_2 - U_1 - U_2 \\ &= \frac{1}{2} I_1 \dot{\theta}_1^2 + \frac{1}{2} I_2 \dot{\theta}_2^2 + \frac{1}{2} m_2 \left( l_1^2 \dot{\theta}_1^2 + h_2^2 \left(\dot{\theta}_1 + \dot{\theta}_2\right)^2 + 2 l_1 h_2 \dot{\theta}_1 (\dot{\theta}_1 + \dot{\theta}_2) \cos(\theta_2) \right) \\ &\quad + m_1 g h_1 \cos(\theta_1) + m_2 g \left(l_1 \cos(\theta_1) + h_2 \cos(\theta_1 + \theta_2)\right) \ . \end{aligned} \tag{7}$$

The dissipation $K$ due to friction in each arm's bearings as well as others such as air resistance are modeled using two parameters as

$$K = k_1 \mathrm{abs}(\dot{\theta}_1) + k_2 \mathrm{abs}(\dot{\theta}_2) \ . \tag{8}$$

We can finally get the equations of motion from

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}_i}\right) - \frac{\partial L}{\partial \theta_i} + \frac{\partial K}{\partial \dot{\theta}_i} = 0, \quad \text{for } i = 1, 2 \ .$$

We have a total of ten parameters in this model and they are

$$I_1, I_2, m_1, m_2, l_1, h_1, h_2, g, k_1, k_2 \ .$$

# ODE Model: Damped Double Pendulum Trajectories

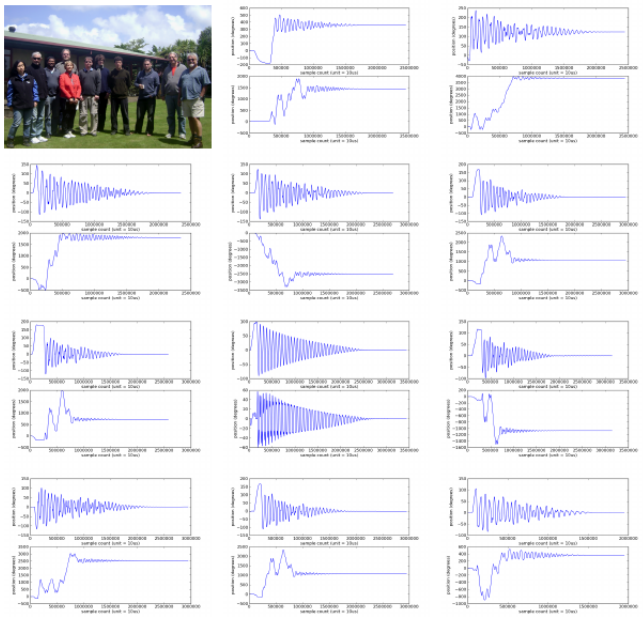
Finite versus infinite dimensional Models:

- Finite dimensional models can be *rigorously exhaused*
  - double-hooked bobs suspended by springs in each arm
  - add permanent magnets into each arm
  - attach to flimsier docking station via vice-grips
  - to rigorously exhaust parametric models we need computer-aided proofs in dynamics (work in progress)
- **Here**, we want to take a **nonparametric empirical process approach** ("infinite dimensional models from data with *universal performance guarantees*") to estimate the density of trajectories from multiple independent experiments directly

# ODE Model: Damped Double Pendulum Trajectories

## Trajectory signatures of ConstruMath South 2012 Participants

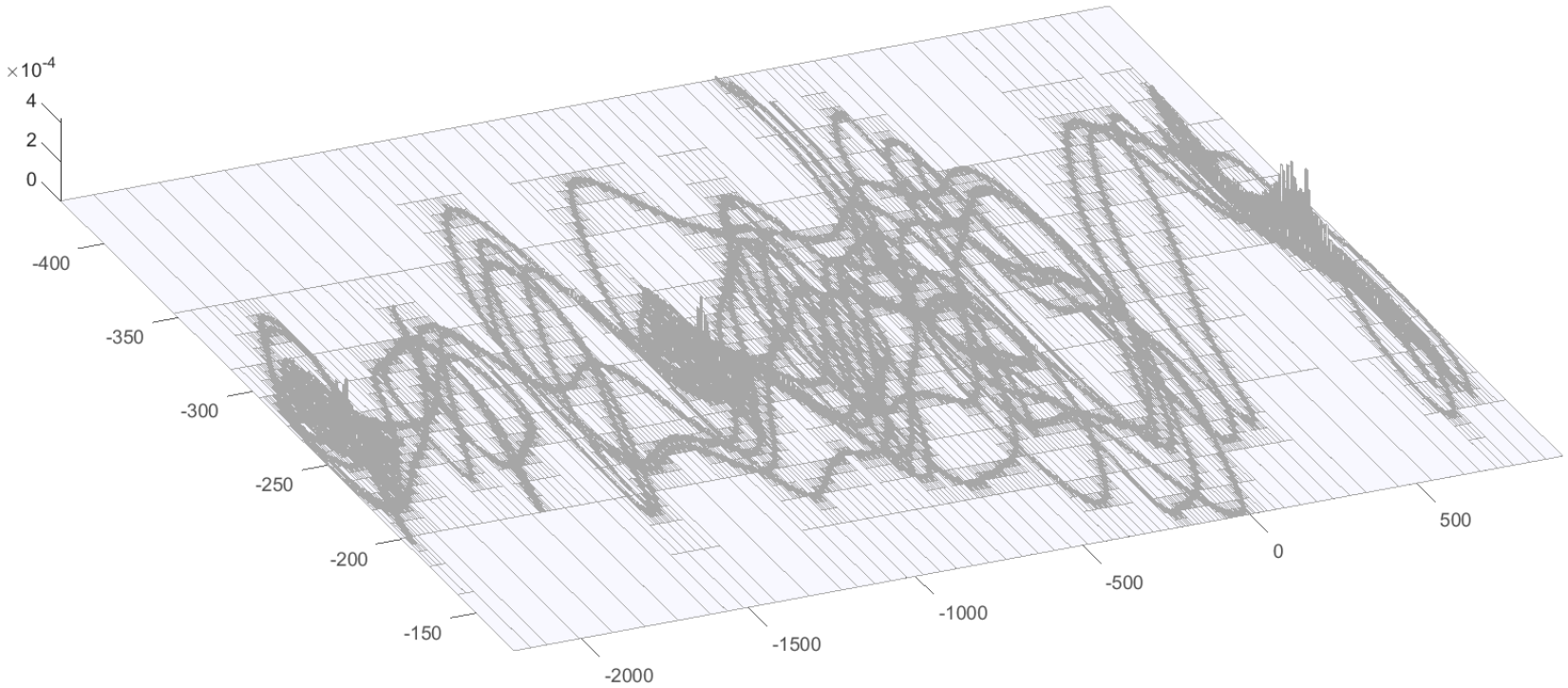
**Participants of**
***ConstruMath South 2012:*** *Applications of Non-Classical Logic*
**and Their Double Pendulum Release Signatures**



The participants from left to right in the first image are: Ruriko Yoshida, Cris Calude, Raazesh Sainudiin, Maarten McKubre-Jordens, Elena Calude, Nicholas Duncan, James Dent, Ty Baen, Zach Weber, Ed Mares and Bruce Burdick. Each participant released a mechatronically measurable double pendulum. The remaining eleven images (from left to right and row by row) show the positions of each arm of the double pendulum through time upon release by each participant in the above list order.

# ODE Model: Damped Double Pendulum Trajectories

Estimate of the Angular position of each arm from 3 trajectories



Section 2

# Non-parametric Density Estimation

# Massive Metric Data Streams – Introduction

- A massive metric data stream is:

$$\ldots, X_{-3}, X_{-2}, X_{-1}, X_0, \mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \mathbf{X_n}, X_{n+1}, \ldots \sim f, \quad X_i \in \mathbb{R}^d.$$

# Massive Metric Data Streams – Introduction

- A massive metric data stream is:

$$\ldots, X_{-3}, X_{-2}, X_{-1}, X_0, \mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \mathbf{X_n}, X_{n+1}, \ldots \sim f, \quad X_i \in \mathbb{R}^d.$$

- Large Effective Dimension:
  $1 \leq d \leq 1000$ (unstructured $f$), $1 \leq d \leq 10$ (highly structured $f$)

# Massive Metric Data Streams – Introduction

- A massive metric data stream is:

$$\ldots, X_{-3}, X_{-2}, X_{-1}, X_0, \mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \mathbf{X_n}, X_{n+1}, \ldots \sim f, \quad X_i \in \mathbb{R}^d.$$

- Large Effective Dimension:
  $1 \leq d \leq 1000$ (unstructured $f$), $1 \leq d \leq 10$ (highly structured $f$)
- Sample size for a single "burst": $10^4 \leq n \leq 10^{10}$

# Massive Metric Data Streams – Introduction

- A massive metric data stream is:

$$\ldots, X_{-3}, X_{-2}, X_{-1}, X_0, \mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \mathbf{X_n}, X_{n+1}, \ldots \sim f, \quad X_i \in \mathbb{R}^d.$$

- Large Effective Dimension:
  $1 \leq d \leq 1000$ (unstructured $f$), $1 \leq d \leq 10$ (highly structured $f$)
- Sample size for a single "burst": $10^4 \leq n \leq 10^{10}$
- Most estimators of $f$ grind to a halt on such data streams

# Massive Metric Data Streams – Introduction

- A massive metric data stream is:

$$\ldots, X_{-3}, X_{-2}, X_{-1}, X_0, \mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \mathbf{X_n}, X_{n+1}, \ldots \sim f, \quad X_i \in \mathbb{R}^d.$$

- Large Effective Dimension:
  $1 \leq d \leq 1000$ (unstructured $f$), $1 \leq d \leq 10$ (highly structured $f$)
- Sample size for a single "burst": $10^4 \leq n \leq 10^{10}$
- Most estimators of $f$ grind to a halt on such data streams
- Need a multi-dimensional metric data-structure that is:

# Massive Metric Data Streams – Introduction

- A massive metric data stream is:

$$\ldots, X_{-3}, X_{-2}, X_{-1}, X_0, \mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \mathbf{X_n}, X_{n+1}, \ldots \sim f, \quad X_i \in \mathbb{R}^d.$$

- Large Effective Dimension:
  $1 \leq d \leq 1000$ (unstructured $f$), $1 \leq d \leq 10$ (highly structured $f$)
- Sample size for a single "burst": $10^4 \leq n \leq 10^{10}$
- Most estimators of $f$ grind to a halt on such data streams
- Need a multi-dimensional metric data-structure that is:
  1. Computationally Efficient (do before the next radar sweep)

# Massive Metric Data Streams – Introduction

- A massive metric data stream is:

$$\ldots, X_{-3}, X_{-2}, X_{-1}, X_0, \mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \mathbf{X_n}, X_{n+1}, \ldots \sim f, \quad X_i \in \mathbb{R}^d.$$

- Large Effective Dimension:
  $1 \leq d \leq 1000$ (unstructured $f$), $1 \leq d \leq 10$ (highly structured $f$)
- Sample size for a single "burst": $10^4 \leq n \leq 10^{10}$
- Most estimators of $f$ grind to a halt on such data streams
- Need a multi-dimensional metric data-structure that is:
  1. Computationally Efficient (do before the next radar sweep)
  2. Statistically Consistent, i.e., $\int \mathrm{abs}(f_n - f) d\lambda \to 0$ as $n \to \infty$

# Massive Metric Data Streams – Introduction

- A massive metric data stream is:

$$\ldots, X_{-3}, X_{-2}, X_{-1}, X_0, \mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \mathbf{X_n}, X_{n+1}, \ldots \sim f, \quad X_i \in \mathbb{R}^d.$$

- Large Effective Dimension:
  $1 \leq d \leq 1000$ (unstructured $f$), $1 \leq d \leq 10$ (highly structured $f$)
- Sample size for a single "burst": $10^4 \leq n \leq 10^{10}$
- Most estimators of $f$ grind to a halt on such data streams
- Need a multi-dimensional metric data-structure that is:
  1. Computationally Efficient (do before the next radar sweep)
  2. Statistically Consistent, i.e., $\int \mathrm{abs}(f_n - f) d\lambda \to 0$ as $n \to \infty$
  3. Data-adaptive and Non-parametric (learn from data with minimal assumption)

# Massive Metric Data Streams – Introduction

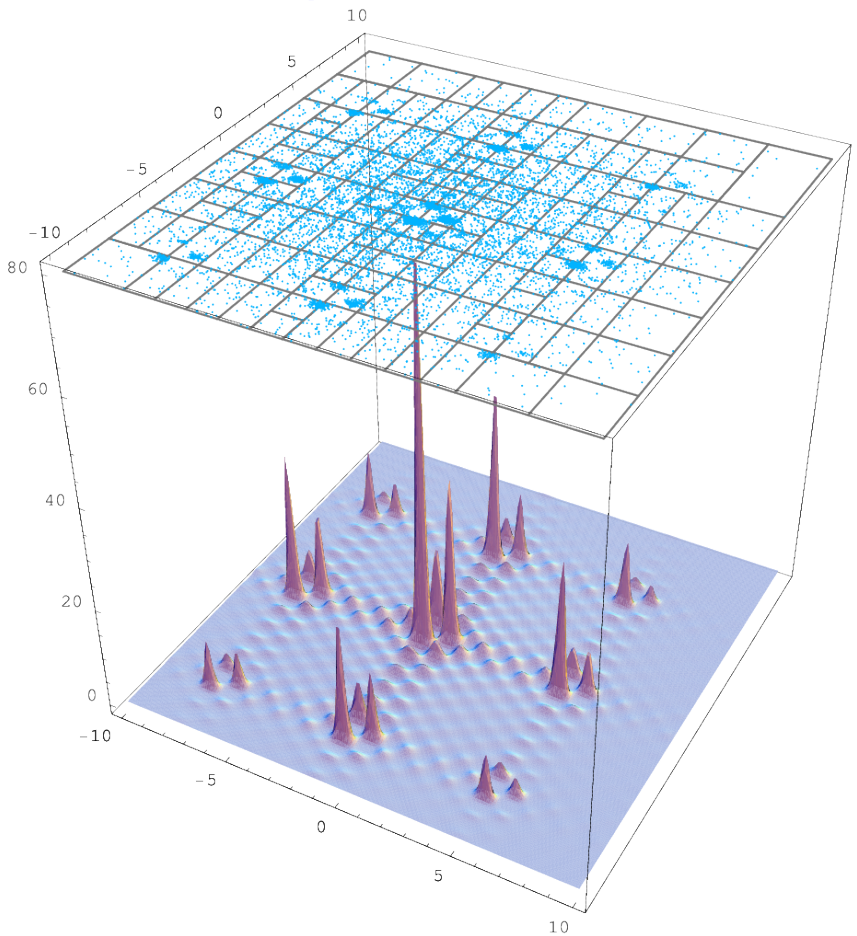
- A massive metric data stream is:

$$\ldots, X_{-3}, X_{-2}, X_{-1}, X_0, \mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \mathbf{X_n}, X_{n+1}, \ldots \sim f, \quad X_i \in \mathbb{R}^d.$$

- Large Effective Dimension:
  $1 \leq d \leq 1000$ (unstructured $f$), $1 \leq d \leq 10$ (highly structured $f$)
- Sample size for a single "burst": $10^4 \leq n \leq 10^{10}$
- Most estimators of $f$ grind to a halt on such data streams
- Need a multi-dimensional metric data-structure that is:
  1. Computationally Efficient (do before the next radar sweep)
  2. Statistically Consistent, i.e., $\int \text{abs}(f_n - f) d\lambda \to 0$ as $n \to \infty$
  3. Data-adaptive and Non-parametric (learn from data with minimal assumption)
  4. with Universal Performance Guarantees (error bounds for $\int \text{abs}(f_n - f) d\lambda$ as a function of data $x_1, \ldots, x_n$, with $n < \infty$
     — NEED to account for the combinatorial geometric complexity of observable *events*)

# Non-parametric Density Estimation – Problem

Take $X_1, X_2, \ldots, X_n$ IID samples from unknown density $f$

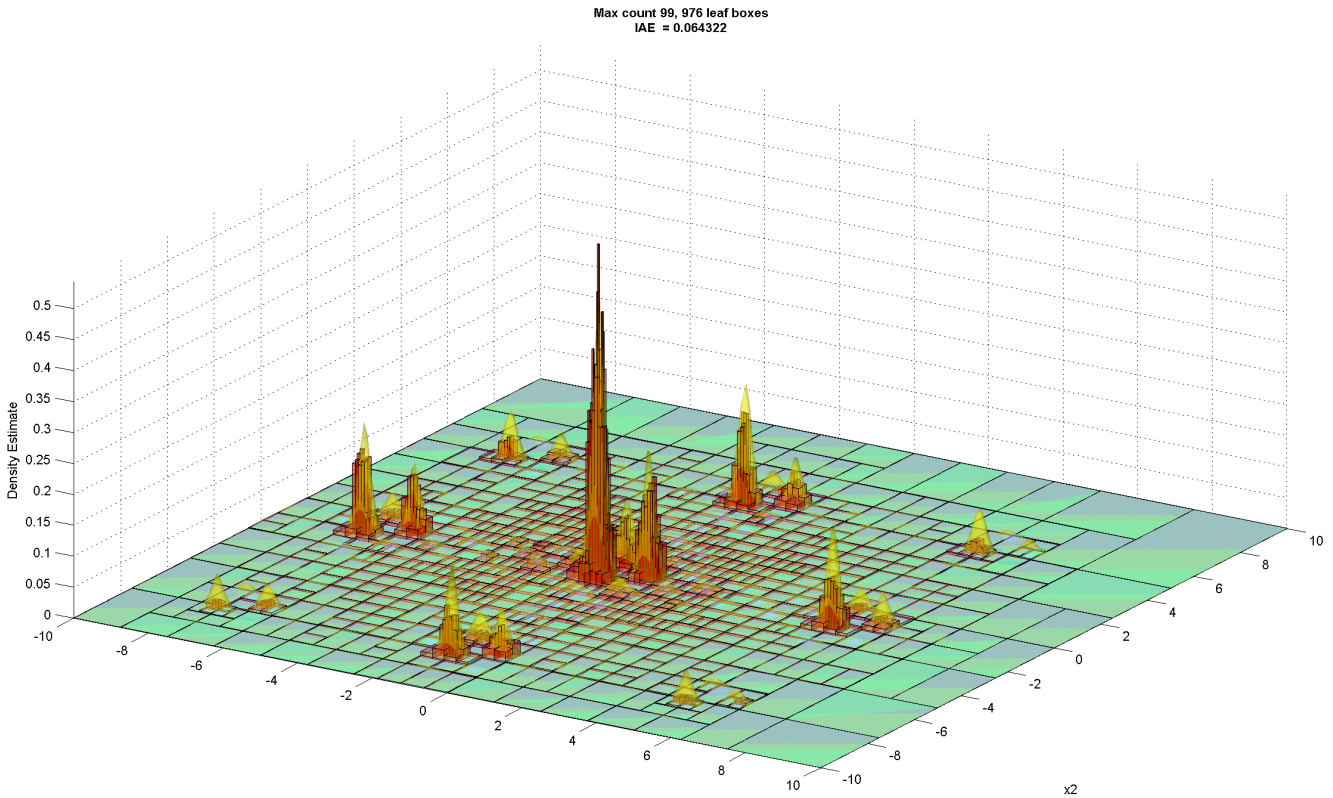



S & York, An auto-validating trans-dimensional universal rejection sampler for locally Lipschitz arithmetical expressions, Reliable Computing, 2013

# Non-parametric Density Estimation – Problem

and give a consistent estimator $f_n$ of $f$, i.e., $f_n : \left(\mathbb{R}^d\right)^n \times \mathbb{R}^d \to \mathbb{R}$

1. such that $f_n$ is imbued with arithmetic and
2. gives **universal performance guarantees** — $f \in L_1$



# The Smoothing Problem – controlling the data-dependent partitioning scheme

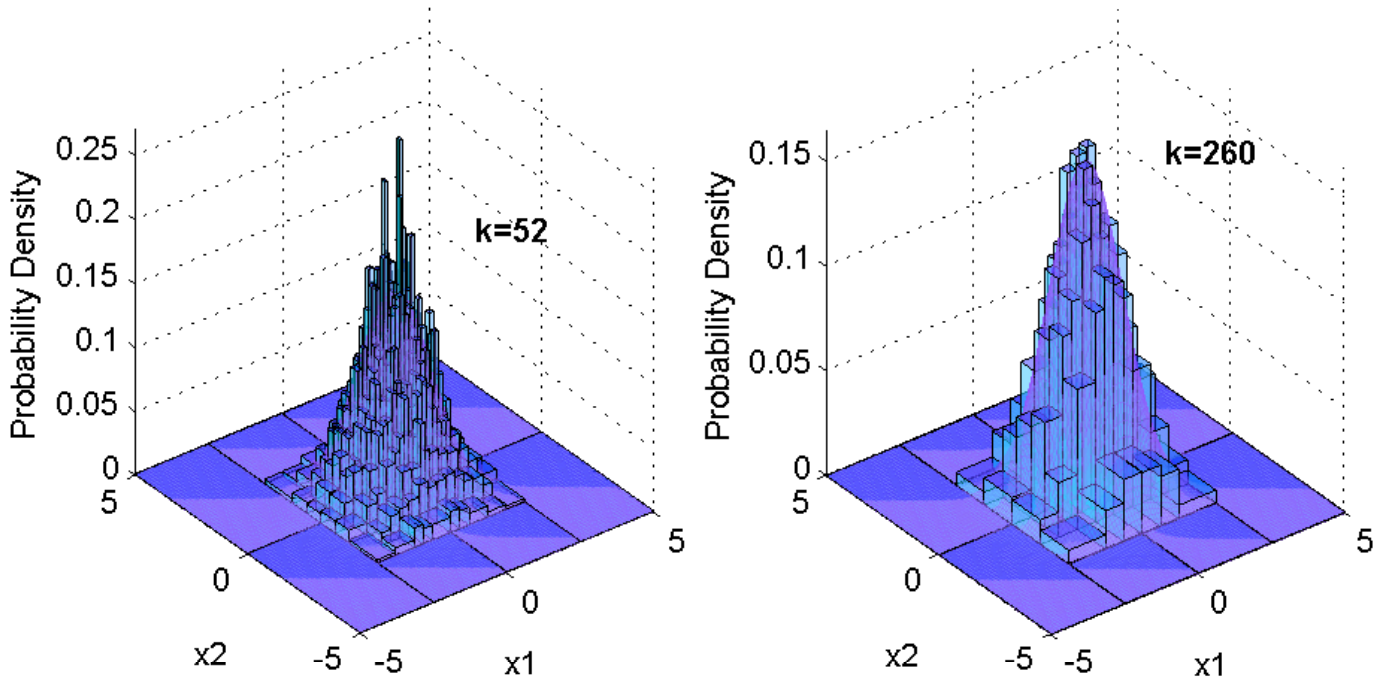



Figure: Two histogram density estimates for the standard bivariate gaussian density with different choices of partitions. The histogram is under-smoothed (left) and over-smoothed (right).

Section 3

# Arithmetic and Algebra of Plane Binary Trees - 2014 Primer

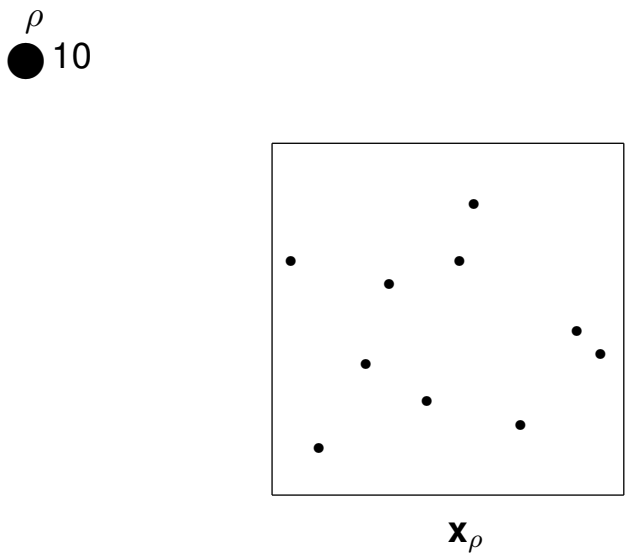
Section 4

# Randomized Priority Queue Markov chain

# A Prioritized Queue based Algorithm (for $L_1$ Consistent Initialization)

## Algorithm `SplitMostCounts`

As data arrives, order the leaf boxes of the SRP so that the leaf box with **the most number of points** will be chosen for the next bisection.

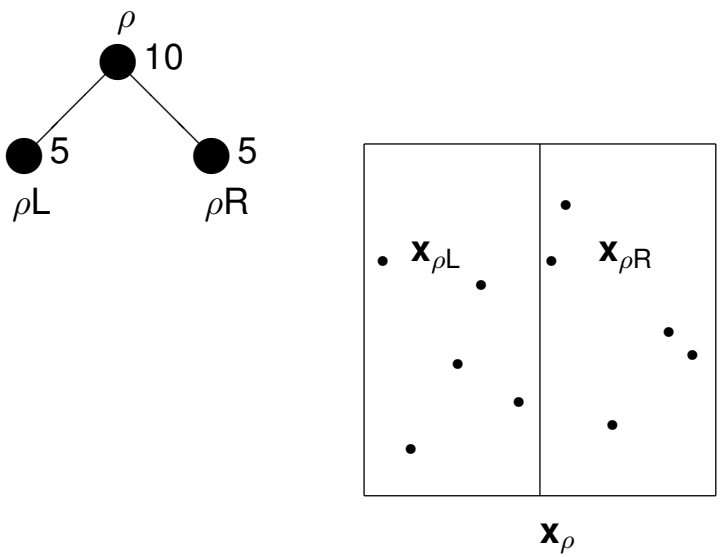


# A Prioritized Queue based Algorithm (for $L_1$ Consistent Initialization)

## Algorithm `SplitMostCounts`

As data arrives, order the leaf boxes of the SRP so that the leaf box with **the most number of points** will be chosen for the next bisection.

Split the root box.

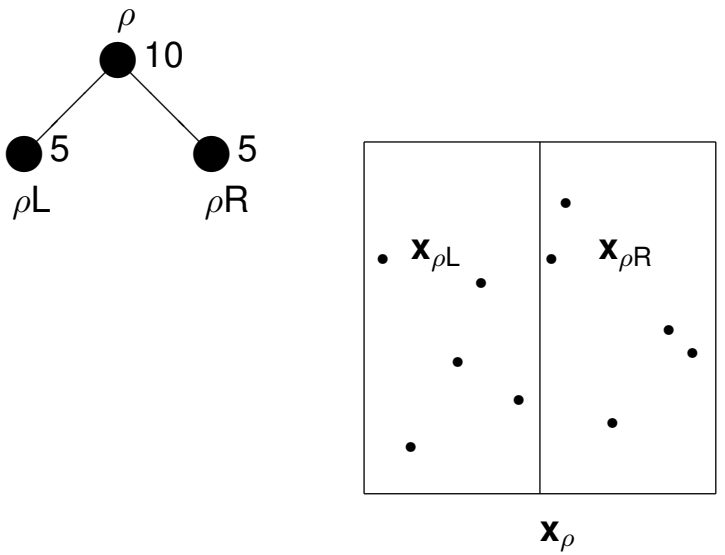


# A Prioritized Queue based Algorithm (for $L_1$ Consistent Initialization)

## Algorithm `SplitMostCounts`

As data arrives, order the leaf boxes of the SRP so that the leaf box with **the most number of points** will be chosen for the next bisection.

Two or more boxes with the most number of points?

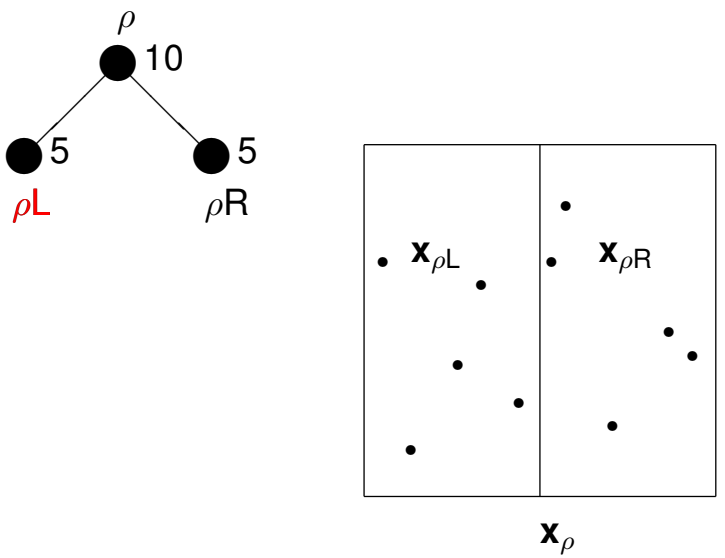


# A Prioritized Queue based Algorithm (for $L_1$ Consistent Initialization)

## Algorithm `SplitMostCounts`

As data arrives, order the leaf boxes of the SRP so that the leaf box with **the most number of points** will be chosen for the next bisection.

Break such ties by randomising the next bisection.

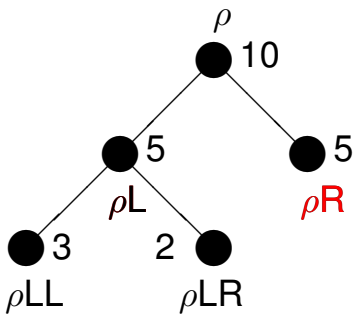


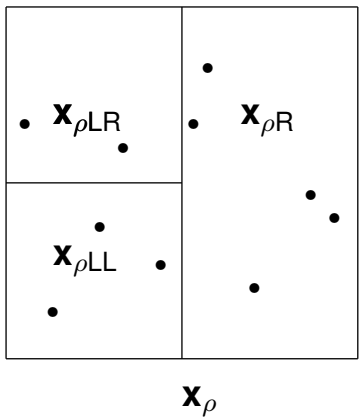
# A Prioritized Queue based Algorithm (for $L_1$ Consistent Initialization)

## Algorithm `SplitMostCounts`

As data arrives, order the leaf boxes of the SRP so that the leaf box with **the most number of points** will be chosen for the next bisection.

Bisect until each box has $\leq \overline{k}_n$ points (let $\overline{k}_n = 3$ here).



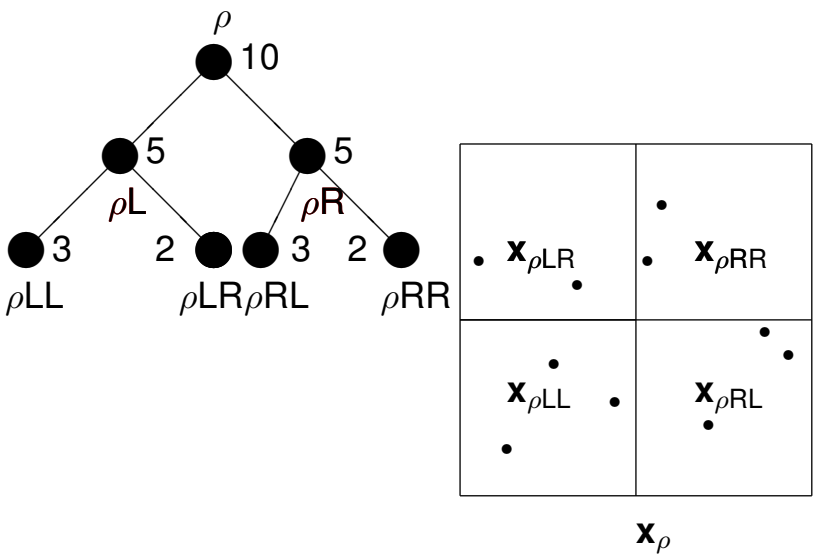



# A Prioritized Queue based Algorithm (for $L_1$ Consistent Initialization)

## Algorithm `SplitMostCounts`

As data arrives, order the leaf boxes of the SRP so that the leaf box with **the most number of points** will be chosen for the next bisection.

Final state



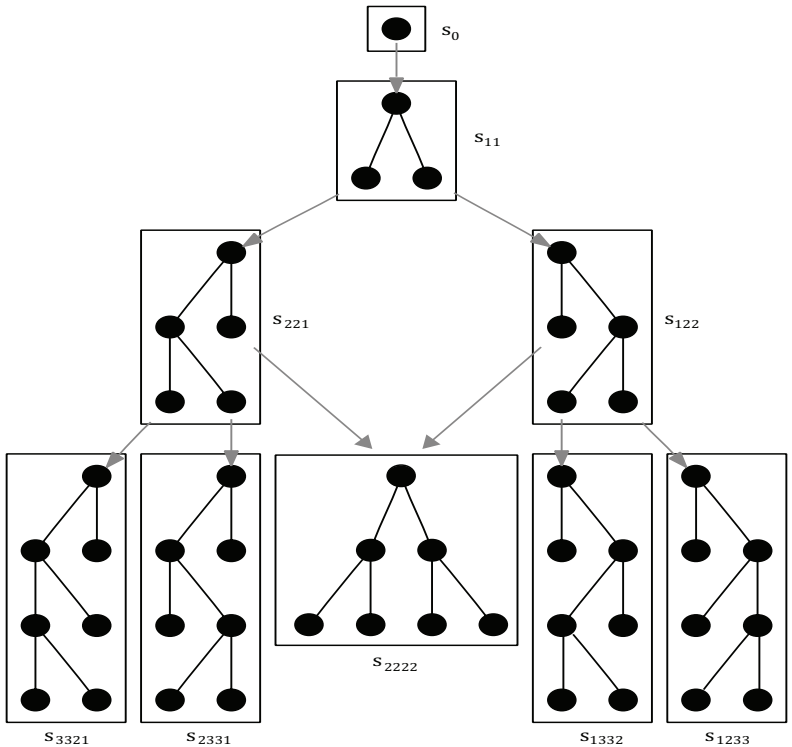
# The `SplitMostCounts` Algorithm

**Input:** (i) data: $x_1, \ldots, x_n \subseteq \mathbb{R}^d$; (ii) root box: $\mathbf{x}_\rho$;
(iii) SEB max: $\overline{k}_n$; (iv) maximum partition size: $\overline{m}_n$.
**Output:** histogram estimate $f_{n,s}$

initialize $i \leftarrow 1$; $s \leftarrow \mathbf{x}_\rho$;
repeat until
$\#\mathbf{x}_{\rho v} \leq \overline{k}_n$ for each $\mathbf{x}_{\rho v} \in \ell(s)$ and $i \leq \overline{m}_n$ // $\ell(s) = \{$`leaf boxes`$\}$
$\mathbf{x}_{\rho v} \leftarrow \text{Uniform}(\hat{\ell}(s))$ // `randomized PQ on leaf boxes`
$s \leftarrow$ `bisect`$(s, \mathbf{x}_{\rho v})$ // `bisect leaf box` $\mathbf{x}_{\rho v}$ `of` $s$
recursively update counts in $s$;
$i \leftarrow i + 1$;
return $f_{n,s}$

# Transition Diagram of Randomized PQ Markov chain

Let $\mathbb{S}_i$ be the set of all RPs of $\mathbf{x}_\rho$ made of $i$ splits and for $i, j \in \mathbb{N}$ with $i \leq j$, let $\mathbb{S}_{i:j}$ be the set of RPs with $k$ splits, $i \leq k \leq j$.




All possible RP partitions in $\mathbb{S}_{0:4}$.

# $L_1$-Consistency of `SplitMostCounts` Markov chain

## Theorem (S & Teng, 2012)

*Let $X_1, X_2, \ldots$ be independent and identical random vectors in $\mathbb{R}^d$ whose common distribution $\mu$ has a non-atomic density $f$, i.e., $f \ll \lambda^d$. Let $\{S_n(i)\}_{i=0}^{J}$ on $\mathbb{S}_{0:\infty}$ be the Markov chain formed using `SplitMostCounts` with terminal state $\dot{s}$ and histogram estimate $f_{n,\dot{s}}$ over the collection of partitions $\mathcal{L}_n$.*

# $L_1$-Consistency of `SplitMostCounts` Markov chain

## Theorem (S & Teng, 2012)

*Let $X_1, X_2, \ldots$ be independent and identical random vectors in $\mathbb{R}^d$ whose common distribution $\mu$ has a non-atomic density $f$, i.e., $f \ll \lambda^d$. Let $\{S_n(i)\}_{i=0}^{J}$ on $\mathbb{S}_{0:\infty}$ be the Markov chain formed using `SplitMostCounts` with terminal state $\dot{s}$ and histogram estimate $f_{n,\dot{s}}$ over the collection of partitions $\mathcal{L}_n$.*
*Then, as $n \to \infty$, if $\overline{k}_n \to \infty$, $\overline{k}_n/n \to 0$, $\overline{m}_n \geq n/\overline{k}_n$, and $\overline{m}_n/n \to 0$ then the density estimate $f_{n,\dot{s}}$ is strongly consistent in $L_1$, i.e.*

$$\int |f(x) - f_{n,\dot{s}}(x)| dx \to 0 \text{ with probability } 1.$$

# Proof Sketch

We will assume that $\overline{k}_n \to \infty$, $n^{-1}\overline{k}_n \to 0$, $\overline{m}_n \geq n/\overline{k}_n$, and $\overline{m}_n/n \to 0$, as $n \to \infty$, and show that the three conditions:

(a) $n^{-1} m(\mathcal{L}_n) \to 0$,
(b) $n^{-1} \log \Delta_n^*(\mathcal{L}_n) \to 0$, and
(c) $\mu(x : \mathrm{diam}(\mathbf{x}(x)) > \gamma) \to 0$ with probability 1 for every $\gamma > 0$,

are satisfied. Then by Theorem 1 of Lugosi & Nobel (Ann. Stats., 1996) our $f_{n,\dot{s}}$ is strongly consistent in $L_1$.

# Proof Sketch

We will assume that $\overline{k}_n \to \infty$, $n^{-1}\overline{k}_n \to 0$, $\overline{m}_n \geq n/\overline{k}_n$, and $\overline{m}_n/n \to 0$, as $n \to \infty$, and show that the three conditions:

(a) $n^{-1} m(\mathcal{L}_n) \to 0$,
(b) $n^{-1} \log \Delta_n^*(\mathcal{L}_n) \to 0$, and
(c) $\mu(x : \mathrm{diam}(\mathbf{x}(x)) > \gamma) \to 0$ with probability 1 for every $\gamma > 0$,

are satisfied. Then by Theorem 1 of Lugosi & Nobel (Ann. Stats., 1996) our $f_{n,\dot{s}}$ is strongly consistent in $L_1$.
These conditions mean:

(a) sub-linear growth of the number of leaf boxes
(b) sub-exponential growth of a combinatorial complexity measure of the growth of the partition
(c) shrinking leaf boxes in the partition

# (a) Sub-linear Growth of the Number of Leaf Boxes

Let $\{S_n(i)\}_{i=0}^{J}$ on $\mathbb{S}_{0:\infty}$ be the Markov chain formed using `SplitMostCounts`. The Markov chain terminates at some state $\dot{s}$ with partition $\ell(\dot{s})$.

# (a) Sub-linear Growth of the Number of Leaf Boxes

Let $\{S_n(i)\}_{i=0}^{J}$ on $\mathbb{S}_{0:\infty}$ be the Markov chain formed using `SplitMostCounts`. The Markov chain terminates at some state $\dot{s}$ with partition $\ell(\dot{s})$. Associated with the Markov Chain is a fixed, non-random collection of partitions

$$\mathcal{L}_n := \{\ell(\dot{s}) : \dot{s} \in \mathbb{S}_{0:\infty}, P(S(J) = \dot{s}) > 0\} \subseteq \mathbb{S}_{0:\overline{m}_n - 1} \quad .$$

# (a) Sub-linear Growth of the Number of Leaf Boxes

Let $\{S_n(i)\}_{i=0}^{J}$ on $\mathbb{S}_{0:\infty}$ be the Markov chain formed using `SplitMostCounts`. The Markov chain terminates at some state $\dot{s}$ with partition $\ell(\dot{s})$. Associated with the Markov Chain is a fixed, non-random collection of partitions

$$\mathcal{L}_n := \{\ell(\dot{s}) : \dot{s} \in \mathbb{S}_{0:\infty}, P(S(J) = \dot{s}) > 0\} \subseteq \mathbb{S}_{0:\overline{m}_n - 1} \ .$$

The size of the largest partition $\ell(\dot{s})$ in $\mathcal{L}_n$ is given by

$$m(\mathcal{L}_n) := \sup_{\ell(\dot{s}) \in \mathcal{L}_n} |\ell(\dot{s})| \leq \overline{m}_n \ .$$

# (a) Sub-linear Growth of the Number of Leaf Boxes

Let $\{S_n(i)\}_{i=0}^{J}$ on $\mathbb{S}_{0:\infty}$ be the Markov chain formed using `SplitMostCounts`. The Markov chain terminates at some state $\dot{s}$ with partition $\ell(\dot{s})$. Associated with the Markov Chain is a fixed, non-random collection of partitions

$$\mathcal{L}_n := \{\ell(\dot{s}) : \dot{s} \in \mathbb{S}_{0:\infty}, P(S(J) = \dot{s}) > 0\} \subseteq \mathbb{S}_{0:\overline{m}_n - 1} \ .$$

The size of the largest partition $\ell(\dot{s})$ in $\mathcal{L}_n$ is given by

$$m(\mathcal{L}_n) := \sup_{\ell(\dot{s}) \in \mathcal{L}_n} |\ell(\dot{s})| \leq \overline{m}_n \ .$$

Thus, (a) is satisfied by assumption that $\overline{m}_n / n \to 0$.

# (b) Sub-exponential Growth of the Partition

The complexity of $\mathcal{L}_n$ will be measured by a combinatorial quantity similar to the growth function for classes of sets proposed by Vapnik and Chervonenkis (1971).

# (b) Sub-exponential Growth of the Partition

The complexity of $\mathcal{L}_n$ will be measured by a combinatorial quantity similar to the growth function for classes of sets proposed by Vapnik and Chervonenkis (1971). Fix $n$ points $x_1, \ldots, x_n \in \mathbb{R}^d$ and let $B = \{x_1, \ldots, x_n\} \in \left(\mathbb{R}^d\right)^n$.

# (b) Sub-exponential Growth of the Partition

The complexity of $\mathcal{L}_n$ will be measured by a combinatorial quantity similar to the growth function for classes of sets proposed by Vapnik and Chervonenkis (1971). Fix $n$ points $x_1, \ldots, x_n \in \mathbb{R}^d$ and let $B = \{x_1, \ldots, x_n\} \in \left(\mathbb{R}^d\right)^n$. Let $\Delta(\mathcal{L}_n, B)$ be the number of distinct partitions of the finite set $B$ that are induced by partitions $\ell(\dot{s}) \in \mathcal{L}_n$:

$$\Delta(\mathcal{L}_n, B) := \left| \left\{ \left\{ \mathbf{x}_v \cap B : \mathbf{x}_v \in \ell(\dot{s}) \right\} : \ell(\dot{s}) \in \mathcal{L}_n \right\} \right| \quad .$$

# (b) Sub-exponential Growth of the Partition

The complexity of $\mathcal{L}_n$ will be measured by a combinatorial quantity similar to the growth function for classes of sets proposed by Vapnik and Chervonenkis (1971). Fix $n$ points $x_1, \ldots, x_n \in \mathbb{R}^d$ and let $B = \{x_1, \ldots, x_n\} \in \left(\mathbb{R}^d\right)^n$. Let $\Delta(\mathcal{L}_n, B)$ be the number of distinct partitions of the finite set $B$ that are induced by partitions $\ell(\dot{s}) \in \mathcal{L}_n$:

$$\Delta(\mathcal{L}_n, B) := \left| \left\{ \left\{ \mathbf{x}_v \cap B : \mathbf{x}_v \in \ell(\dot{s}) \right\} : \ell(\dot{s}) \in \mathcal{L}_n \right\} \right| \quad .$$

Define the growth function of $\mathcal{L}_n$ as

$$\Delta^*(\mathcal{L}_n, B) := \max_{B \in \left(\mathbb{R}^d\right)^n} \Delta(\mathcal{L}_n, B)$$

# (b) Sub-exponential Growth of the Partition

The complexity of $\mathcal{L}_n$ will be measured by a combinatorial quantity similar to the growth function for classes of sets proposed by Vapnik and Chervonenkis (1971). Fix $n$ points $x_1, \ldots, x_n \in \mathbb{R}^d$ and let $B = \{x_1, \ldots, x_n\} \in \left(\mathbb{R}^d\right)^n$. Let $\Delta(\mathcal{L}_n, B)$ be the number of distinct partitions of the finite set $B$ that are induced by partitions $\ell(\dot{s}) \in \mathcal{L}_n$:

$$\Delta(\mathcal{L}_n, B) := \mid \{ \{\mathbf{x}_v \cap B : \mathbf{x}_v \in \ell(\dot{s})\} : \ell(\dot{s}) \in \mathcal{L}_n\} \mid \quad .$$

Define the growth function of $\mathcal{L}_n$ as

$$\Delta^*(\mathcal{L}_n, B) := \max_{B \in \left(\mathbb{R}^d\right)^n} \Delta(\mathcal{L}_n, B) \leq |\mathcal{L}_n|$$

# (b) Sub-exponential Growth of the Partition

The complexity of $\mathcal{L}_n$ will be measured by a combinatorial quantity similar to the growth function for classes of sets proposed by Vapnik and Chervonenkis (1971). Fix $n$ points $x_1, \ldots, x_n \in \mathbb{R}^d$ and let $B = \{x_1, \ldots, x_n\} \in \left(\mathbb{R}^d\right)^n$. Let $\Delta(\mathcal{L}_n, B)$ be the number of distinct partitions of the finite set $B$ that are induced by partitions $\ell(\dot{s}) \in \mathcal{L}_n$:

$$\Delta(\mathcal{L}_n, B) := \left| \left\{ \left\{ \mathbf{x}_v \cap B : \mathbf{x}_v \in \ell(\dot{s}) \right\} : \ell(\dot{s}) \in \mathcal{L}_n \right\} \right| \quad .$$

Define the growth function of $\mathcal{L}_n$ as

$$\Delta^*(\mathcal{L}_n, B) := \max_{B \in \left(\mathbb{R}^d\right)^n} \Delta(\mathcal{L}_n, B) \leq |\mathcal{L}_n| \leq \sum_{k=0}^{\overline{m}_n} C_k$$

# (b) Sub-exponential Growth of the Partition

The complexity of $\mathcal{L}_n$ will be measured by a combinatorial quantity similar to the growth function for classes of sets proposed by Vapnik and Chervonenkis (1971). Fix $n$ points $x_1, \ldots, x_n \in \mathbb{R}^d$ and let $B = \{x_1, \ldots, x_n\} \in \left(\mathbb{R}^d\right)^n$. Let $\Delta(\mathcal{L}_n, B)$ be the number of distinct partitions of the finite set $B$ that are induced by partitions $\ell(\dot{s}) \in \mathcal{L}_n$:

$$\Delta(\mathcal{L}_n, B) := \left| \left\{ \{\mathbf{x}_v \cap B : \mathbf{x}_v \in \ell(\dot{s})\} : \ell(\dot{s}) \in \mathcal{L}_n \right\} \right| \quad .$$

Define the growth function of $\mathcal{L}_n$ as

$$\Delta^*(\mathcal{L}_n, B) := \max_{B \in \left(\mathbb{R}^d\right)^n} \Delta(\mathcal{L}_n, B) \leq |\mathcal{L}_n| \leq \sum_{k=0}^{\overline{m}_n} C_k \approxeq \frac{4^{\overline{m}_n+1}}{3\overline{m}_n\sqrt{(\pi\overline{m}_n)}}$$

where $\approxeq$ is a known partial Catalan sum result (Mattarei, 2010).

# (b) Sub-exponential Growth of the Partition

The complexity of $\mathcal{L}_n$ will be measured by a combinatorial quantity similar to the growth function for classes of sets proposed by Vapnik and Chervonenkis (1971). Fix $n$ points $x_1, \ldots, x_n \in \mathbb{R}^d$ and let $B = \{x_1, \ldots, x_n\} \in \left(\mathbb{R}^d\right)^n$. Let $\Delta(\mathcal{L}_n, B)$ be the number of distinct partitions of the finite set $B$ that are induced by partitions $\ell(\dot{s}) \in \mathcal{L}_n$:

$$\Delta(\mathcal{L}_n, B) := \left| \left\{ \left\{ \mathbf{x}_v \cap B : \mathbf{x}_v \in \ell(\dot{s}) \right\} : \ell(\dot{s}) \in \mathcal{L}_n \right\} \right| \quad .$$

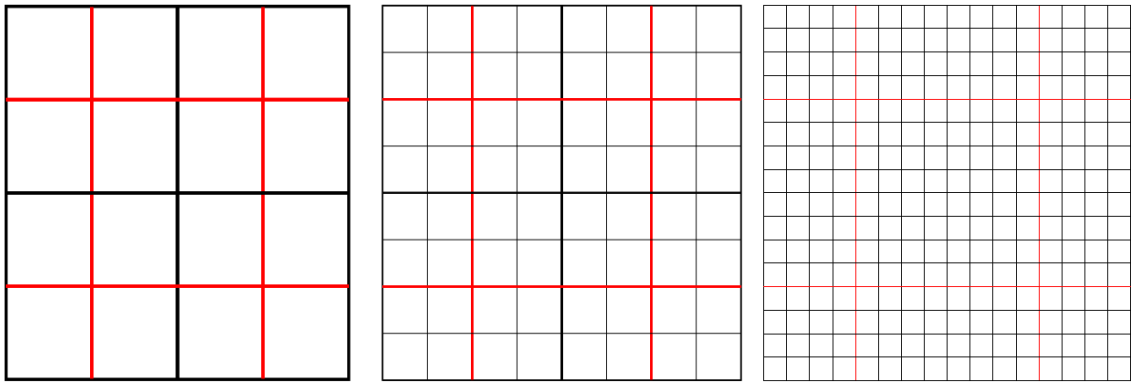
Define the growth function of $\mathcal{L}_n$ as

$$\Delta^*(\mathcal{L}_n, B) := \max_{B \in \left(\mathbb{R}^d\right)^n} \Delta(\mathcal{L}_n, B) \leq |\mathcal{L}_n| \leq \sum_{k=0}^{\overline{m}_n} C_k \approxeq \frac{4^{\overline{m}_n+1}}{3\overline{m}_n\sqrt{(\pi\overline{m}_n)}}$$

where $\approxeq$ is a known partial Catalan sum result (Mattarei, 2010). This ensures condition (b) is satisfied, i.e. $n^{-1} \log \Delta_n^*(\mathcal{L}_n) \to 0$.

# Shrinking Cells

$$\text{diam}(\mathbf{x}) = \sqrt{\sum_{i=1}^{d} (\overline{x}_i - \underline{x}_i)^2}, \mathbf{x} = [\underline{x}_1, \overline{x}_1] \times \ldots \times [\underline{x}_d, \overline{x}_d]$$



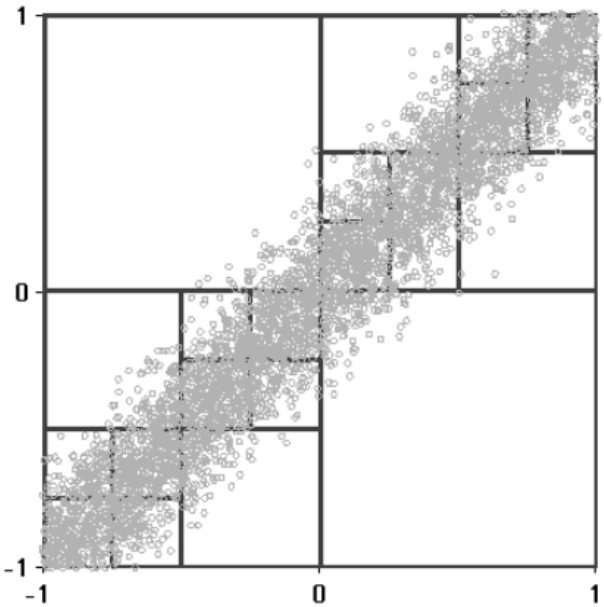
Basically find large enough box $[-M, +M]^d$ with almost all $\mu$ measure and diadically chop it to upper bound number of boxes with diameter $> \gamma$ and using VC $\neq$ to boxes in $\mathbb{R}^d$ to show:

(c) $\mu(x : \text{diam}(\mathbf{x}(x)) > \gamma) \to 0$ with probability 1 for every $\gamma > 0$.

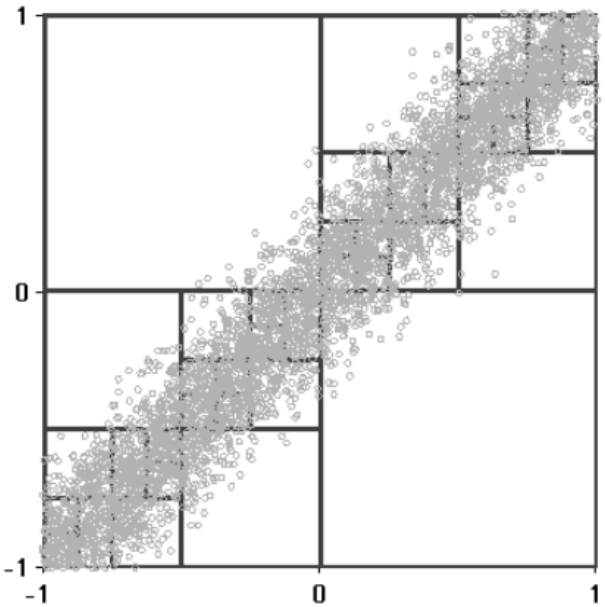
Q.E.D.

# Complementary PQ to "carve out" Support – A Trick

`SplitMostCounts` uses priority = $\mu_n(\mathbf{x}_{\rho v})$.



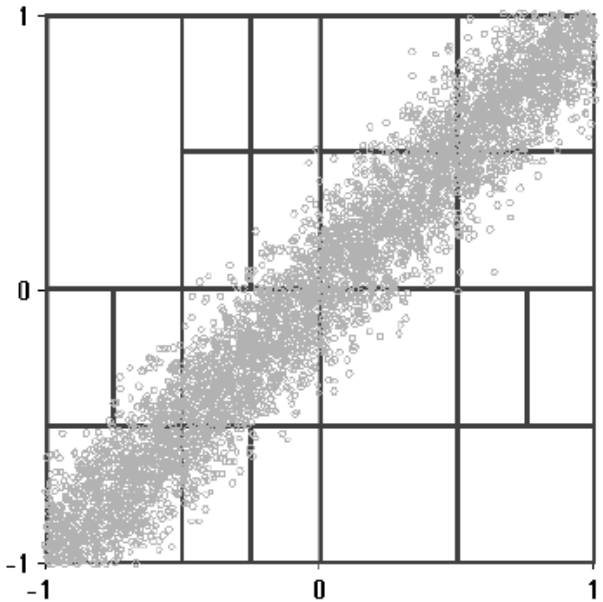

(a) 20 leaves.



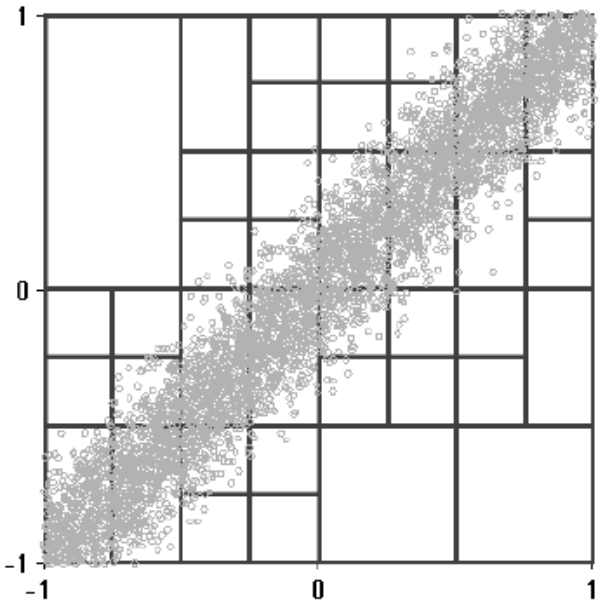

(b) 40 leaves.

# Complementary PQ to "carve out" Support – A Trick

`SupportCarver` uses priority = $(1 - \mu_n(\mathbf{x}_{\rho v}))\mathrm{vol}(\mathbf{x}_{\rho v})$.




(a) 20 leaves.

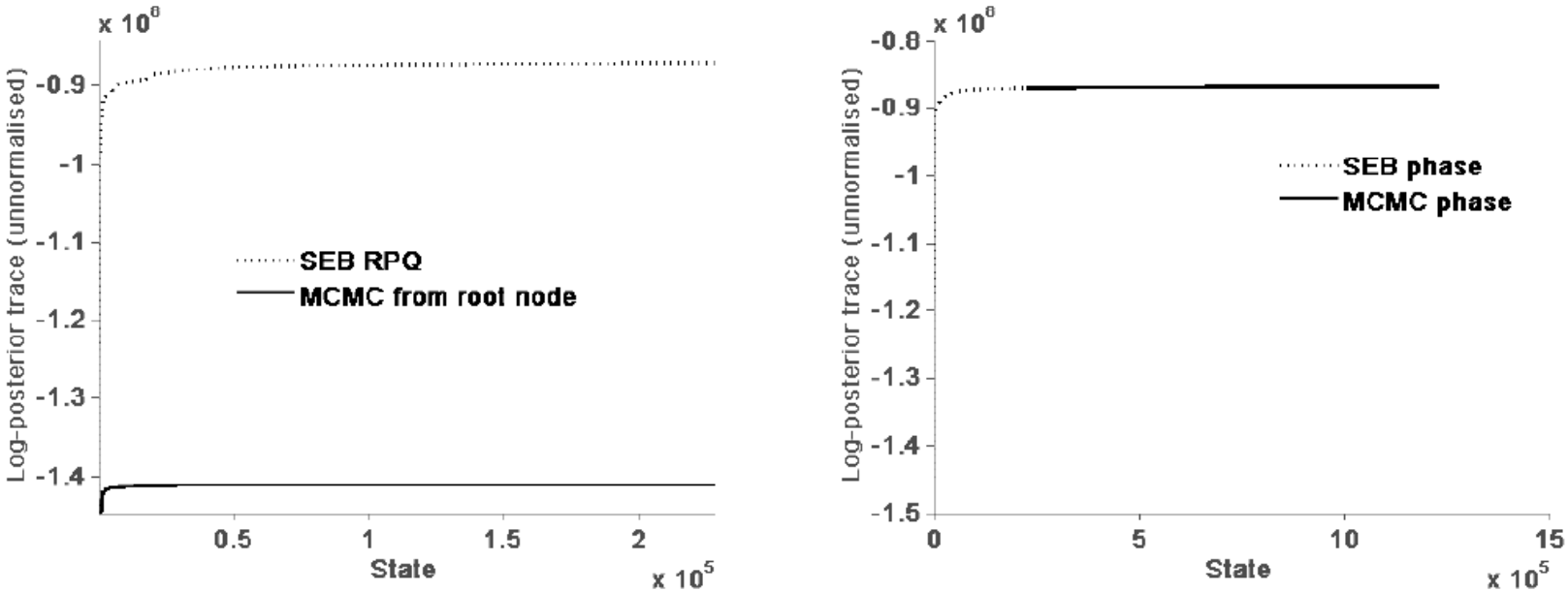



(b) 40 leaves.

Necessary to use `SupportCarver` for high-dimensional structured densities before using `SplitMostCounts`

# Ugly Revisited: $f \sim$ 6D Gaussian – Initialize MCMC by RPQ

Log-posterior traces of SEB RPQ Vs. MCMC started from root node



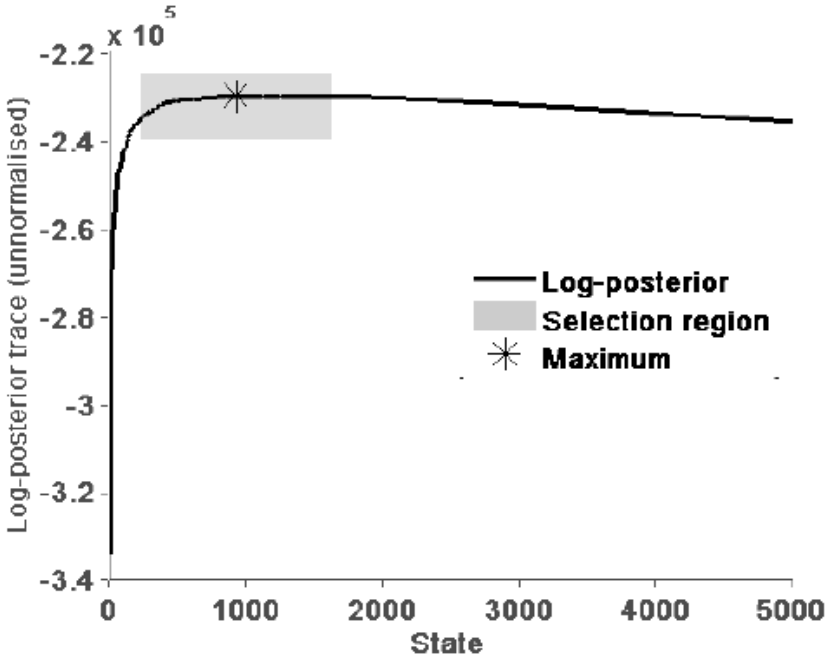

(a) Initial SEB phase compared with MCMC from root node.

(b) Combined log-posterior trace to $t = 1,000,000$.

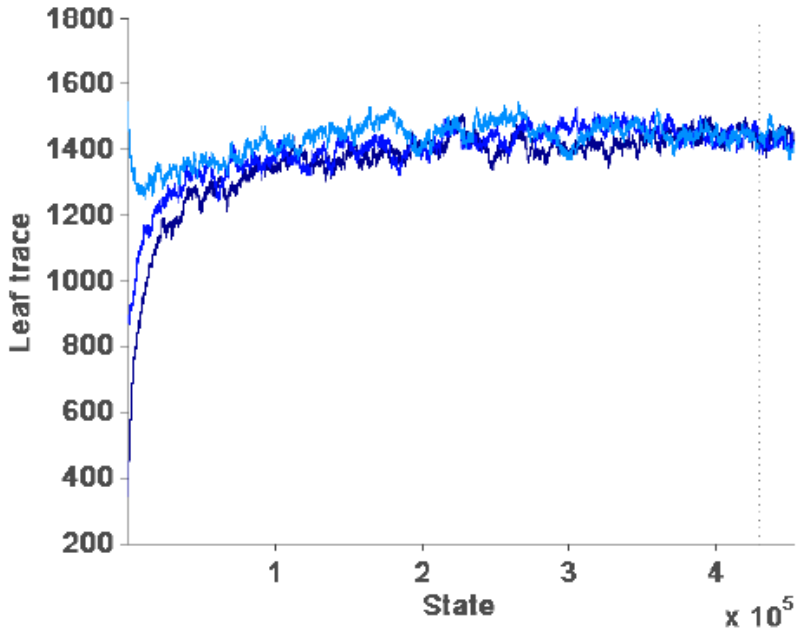
(data drawn from 6D Gaussian Density) – Initialize from highest log-posterior states visited by RPQ

# Ugly Revisited: $f \sim$ 6D Gaussian – Initialize MCMC by RPQ

Multiple MCMC chains started from high log-posterior region




(a) Selection region.

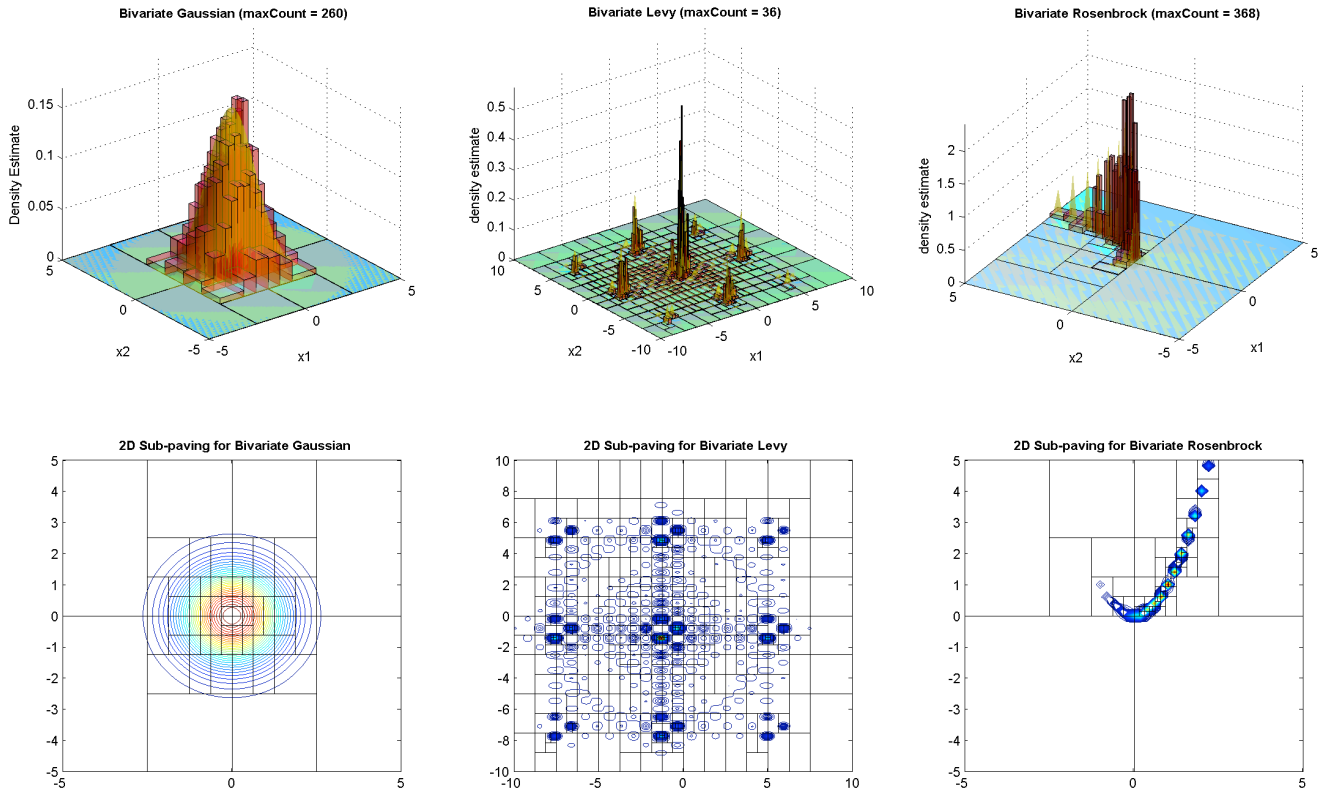



(b) Leaf trace for three chains.

(data drawn from mixture of two 3D Gaussian Densities)

# Some More Examples

Figure: Histogram density estimates their corresponding pavings for the bivariate Gaussian, Levy and Rosenbrock densities.



# Simulations for MCMC and `SplitMostCounts` PQ

MIAE (std. err.) for $n$ samples from approximated 1D-, 2D- and 10D-Gaussian densities, and 2D- and 10D-Rosenbrock densities ($L_1$-minimal Simple function approximation in $\mathbb{S}_\Lambda$).

| | | Standard Gaussian densities | | | Rosenbrock densities | |
|---|---|---|---|---|---|---|
| $\Lambda$ | n | 1D | 2D | 10D | 2D | 10D |
| $10^2$ | $10^2$ | 0.2665 (0.0415) | 0.4856 (0.0491) | 0.1192 (0.0662) | 0.5089 (0.0924) | 0.0323 (0.0511) |
| | $10^3$ | 0.1390 (0.0192) | 0.2558 (0.0127) | 0.0543 (0.0172) | 0.1712 (0.0224) | 0.0095 (0.0191) |
| | $10^4$ | 0.0620 (0.0047) | 0.0992 (0.0067) | 0.0382 (0.0036) | 0.0498 (0.0081) | 0.0025 (0.0050) |
| | $10^5$ | 0.0262 (0.0016) | 0.0279 (0.0019) | 0.0259 (0.0017) | 0.0143 (0.0025) | 0.0009 (0.0015) |
| | $10^6$ | 0.0099 (0.0008) | 0.0086 (0.0006) | 0.0073 (0.0009) | 0.0045 (0.0005) | 0.0004 (0.0005) |
| | $10^7$ | 0.0026 (0.0002) | 0.0027 (0.0003) | 0.0025 (0.0004) | 0.0017 (0.0010) | 0.0001 (0.0003) |
| $10^3$ | $10^2$ | 0.2946 (0.0678) | 0.6046 (0.1299) | 0.1702 (0.0907) | 1.0027 (0.0437) | 0.0323 (0.0511) |
| | $10^3$ | 0.1418 (0.0226) | 0.2973 (0.0174) | 0.0739 (0.0183) | 0.4747 (0.0191) | 0.0039 (0.0075) |
| | $10^4$ | 0.0648 (0.0052) | 0.1586 (0.0067) | 0.0555 (0.0045) | 0.2139 (0.0054) | 0.0013 (0.0028) |
| | $10^5$ | 0.0292 (0.0014) | 0.0768 (0.0016) | 0.0295 (0.0020) | 0.0789 (0.0023) | 0.0004 (0.0006) |
| | $10^6$ | 0.0136 (0.0006) | 0.0297 (0.0006) | 0.0108 (0.0005) | 0.0267 (0.0058) | 0.0001 (0.0002) |
| | $10^7$ | 0.0061 (0.0002) | 0.0091 (0.0003) | 0.0045 (0.0003) | 0.0082 (0.0011) | 0.0001 (0.0002) |
| $10^4$ | $10^2$ | 0.2864 (0.0487) | 0.5508 (0.0590) | 0.5210 (0.0799) | 1.1391 (0.0545) | 0.1941 (0.0820) |
| | $10^3$ | 0.1380 (0.0152) | 0.3301 (0.0120) | 0.2719 (0.0251) | 0.6018 (0.0139) | 0.0791 (0.0223) |
| | $10^4$ | 0.0664 (0.0062) | 0.1736 (0.0038) | 0.1157 (0.0047) | 0.3163 (0.0047) | 0.0391 (0.0041) |
| | $10^5$ | 0.0293 (0.0017) | 0.0957 (0.0014) | 0.0870 (0.0014) | 0.1691 (0.0053) | 0.0209 (0.0021) |
| | $10^6$ | 0.0138 (0.0005) | 0.0495 (0.0005) | 0.0788 (0.0009) | 0.0882 (0.0048) | 0.0123 (0.0012) |
| | $10^7$ | 0.0063 (0.0001) | 0.0244 (0.0008) | 0.0563 (0.0018) | 0.0479 (0.0057) | 0.0096 (0.0017) |

Section 5

# Minimum Distance Estimation

# Minimum Distance Estimation (MDE)

- Let $\theta$ be the current number of splits in `SplitMostCounts`.
- Let $f_{n,\theta}$ be the histogram estimate with corresponding SRP $s \in \mathbb{S}_\theta$, $\int f_{n,\theta} = 1$.
- Denote $\Theta$ as the set of the number of splits such that $\Theta := \{0, \ldots, \overline{m}_n - 1\}$ where $\overline{m}_n - 1$ is the maximum number of splits allowed.

# Minimum Distance Estimation (MDE)

- Let $\theta$ be the current number of splits in `SplitMostCounts`.
- Let $f_{n,\theta}$ be the histogram estimate with corresponding SRP $s \in \mathbb{S}_\theta$, $\int f_{n,\theta} = 1$.
- Denote $\Theta$ as the set of the number of splits such that $\Theta := \{0, \ldots, \overline{m}_n - 1\}$ where $\overline{m}_n - 1$ is the maximum number of splits allowed.

The goal is to select the optimal estimate amongst the $|\Theta|$ candidates, $f_{n,\theta}, \theta \in \Theta$ by using a hold-out method proposed by Devroye and Lugosi, 2004 for minimum distance estimation (MDE).

# Minimum Distance Estimation (MDE)

Let $0 < \varphi < 1/2$. Given $n$ points, use $n - \varphi n$ points as the training set and the remaining $\varphi n$ points as the validation set.

# Minimum Distance Estimation (MDE)

Let $0 < \varphi < 1/2$. Given $n$ points, use $n - \varphi n$ points as the training set and the remaining $\varphi n$ points as the validation set.

**Scheffé set**

For a pair $(\theta, \theta'), \theta, \theta' \in \Theta, \theta \neq \theta'$, the Scheffé set is

$$A_{\theta,\theta'} := A\left(f_{n-\varphi n,\theta}, f_{n-\varphi n,\theta'}\right) = \left\{x : f_{n-\varphi n,\theta}(x) > f_{n-\varphi n,\theta'}(x)\right\} \quad .$$

# Minimum Distance Estimation (MDE)

Let $0 < \varphi < 1/2$. Given $n$ points, use $n - \varphi n$ points as the training set and the remaining $\varphi n$ points as the validation set.

## Scheffé set

For a pair $(\theta, \theta'), \theta, \theta' \in \Theta, \theta \neq \theta'$, the Scheffé set is

$$A_{\theta,\theta'} := A\left(f_{n-\varphi n,\theta}, f_{n-\varphi n,\theta'}\right) = \left\{x : f_{n-\varphi n,\theta}(x) > f_{n-\varphi n,\theta'}(x)\right\} \quad .$$

## Yatracos class

The class of all sets of the form $A_{\theta,\theta'}$:

$$\mathcal{A}_\Theta := \left\{\left\{x : f_{n-\varphi n,\theta}(x) > f_{n-\varphi n,\theta'}(x)\right\} : \theta, \theta' \in \Theta, \theta \neq \theta'\right\} \quad .$$

# Minimum Distance Estimation (MDE)

**Minimum distance estimate**

The minimum distance estimate $f_{n-\varphi n,\theta^*}$ is the density estimate $f_{n-\varphi n,\theta}$ of smallest index $\theta^*$ that minimizes

$$\Delta_\theta = \sup_{A \in \mathcal{A}_\Theta} \left| \int_A f_{n-\varphi n,\theta}(A) - \mu_{\varphi n}(A) \right|$$

where $\mu_{\varphi n}$ is the empirical measure of the validation set $X_{n-\varphi n+1}, \ldots, X_n$.

# Minimum Distance Estimation (MDE)

Let $\Theta$ be the set of the number of splits such that $\Theta := \{0, \ldots, \overline{m}_n - 1\}$ where $\overline{m}_n - 1$ is the maximum number of splits allowed.
Every time a split happens during `SplitMostCounts`,

- Obtain the Yatracos class for the current split $\theta$;

# Minimum Distance Estimation (MDE)

Let $\Theta$ be the set of the number of splits such that $\Theta := \{0, \ldots, \overline{m}_n - 1\}$ where $\overline{m}_n - 1$ is the maximum number of splits allowed.
Every time a split happens during `SplitMostCounts`,

- Obtain the Yatracos class for the current split $\theta$;
- Compute $\Delta_\theta = \sup_{A \in \mathcal{A}_\Theta} \left| \int_A f_{n-\varphi n,\theta}(A) - \mu_{\varphi n}(A) \right|$ for each $\theta \in \Theta$.

# Minimum Distance Estimation (MDE)

Let $\Theta$ be the set of the number of splits such that $\Theta := \{0, \ldots, \overline{m}_n - 1\}$ where $\overline{m}_n - 1$ is the maximum number of splits allowed.
Every time a split happens during `SplitMostCounts`,

- Obtain the Yatracos class for the current split $\theta$;
- Compute $\Delta_\theta = \sup_{A \in \mathcal{A}_\Theta} \left| \int_A f_{n-\varphi n,\theta}(A) - \mu_{\varphi n}(A) \right|$ for each $\theta \in \Theta$.

The candidate estimate that minimizes $\Delta_\theta$ is the minimum distance estimate.

# Setting up MDE

- Need to track statistics for both the training and validation data - recursively computable statistics for validation data;

# Setting up MDE

- Need to track statistics for both the training and validation data - recursively computable statistics for validation data;
- Get the Yatracos class for the current $\Theta$ - use the RP as a collator (via non-minimal union) to track and compare the histogram estimate at each leaf box of each candidate.

# Recursively Computable Statistics for Validation Data

- The training data $\mathcal{T} := \{x_1, \ldots, x_{n-\varphi n}\}$ drive the randomized priority queue `RPQ` to form an SRP $s$.
- The validation data $\mathcal{V} := \{x_{n-\varphi n+1}, \ldots, x_n\}$ trickle through $s$ and stay in the boxes of $s$ that contain the data.

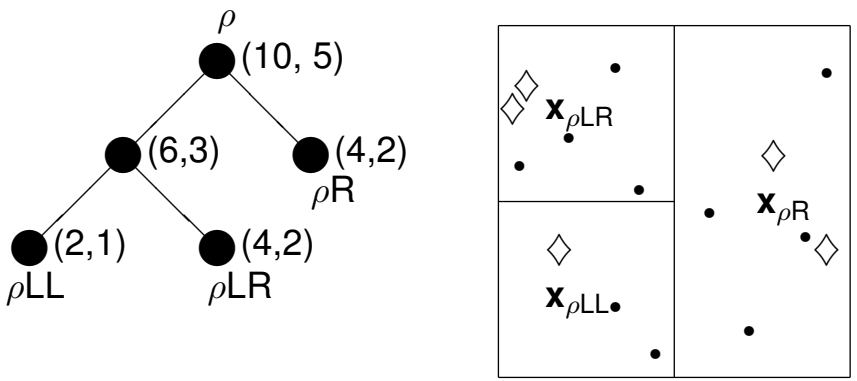



Figure: An SRP $s$ with training ($\bullet$) and validation data ($\diamond$) and their respective sample counts ($\#\mathbf{x}_{\rho v}, \check{\#}\mathbf{x}_{\rho v}$) that are updated recursively as data fall through the nodes of $s$.

# Recursively Computable Statistics for Validation Data

MDE requires the histogram estimate from the training data and the empirical mass of the validation data:

Histogram estimate obtained from the set of training data

$$f_{n-\varphi n}(\rho v) = \frac{\#\mathbf{x}_{\rho v}}{n \cdot \mathrm{vol}\left(\mathbf{x}_{\rho v}\right)} \quad .$$

The empirical measure of the validation data

$$\mu_{\varphi n}(\mathbf{x}_{\rho v}) = \frac{\check{\#}\mathbf{x}_{\rho v}}{\varphi n} \quad .$$

# Recursively Computable Statistics for Validation Data

The training data and the empirical mass of the validation data can be tracked as $(f_{n-\varphi n,\theta}(\rho v), \mu_{\varphi n}(\mathbf{x}_{\rho v}))$ at each leaf node.

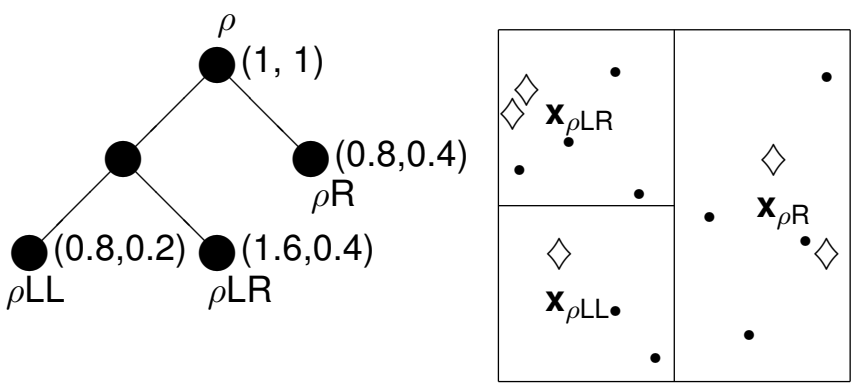



Figure: Tracking the histogram estimate and empirical mass at each node of an SRP $s$ built from a unit square.

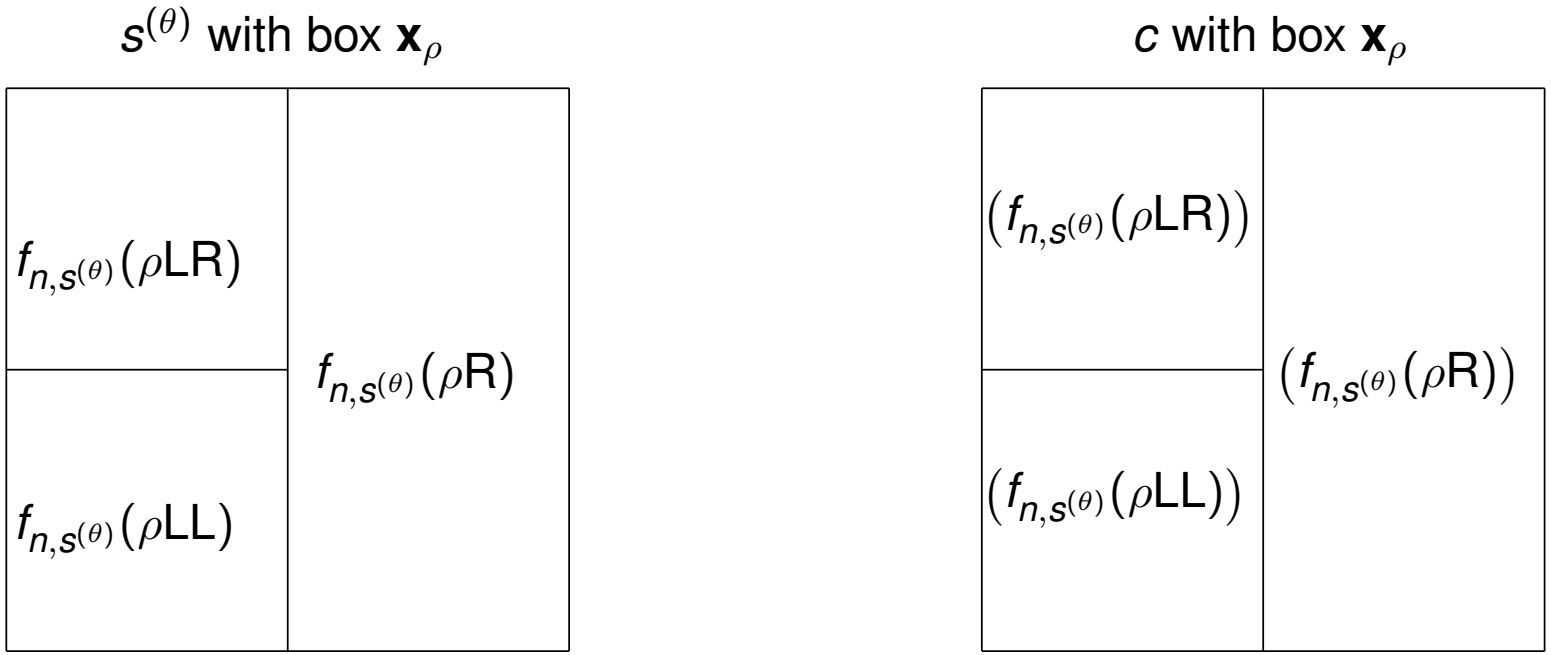
# RP as a Collator

- Need to compare the estimates between each SRP $s^{(i)}, 1 < i < \infty$.
- An efficient way to approach this is to collate these histograms onto an *collator regular paving* or *CRP* where the space of CRP is also $\mathbb{S}_{0:\infty}$.
- Consider two SRPs $s^{(\theta)}$ and $s^{(\theta')}$ that have the same mother box and for which the corresponding histogram estimates $f_{n,s^{(\theta)}}$ and $f_{n,s^{(\theta')}}$ are computed.
- By collating the two SRPs we get a CRP $c$ that stores $f_{n,s^{(\theta)}}$ and $f_{n,s^{(\theta')}}$ for each node $\rho v$ of $c$, such that each node $\rho v$ has a vector $\boldsymbol{f}_{n,c}(\rho v) := (f_{n,s^{(\theta)}}(\rho v), f_{n,s^{(\theta')}}(\rho v))$.

# RP as a Collator

Collating two SRPs $s^{(\theta)}$ and $s^{(\theta')}$ with the same root box $\mathbf{x}_\rho$:

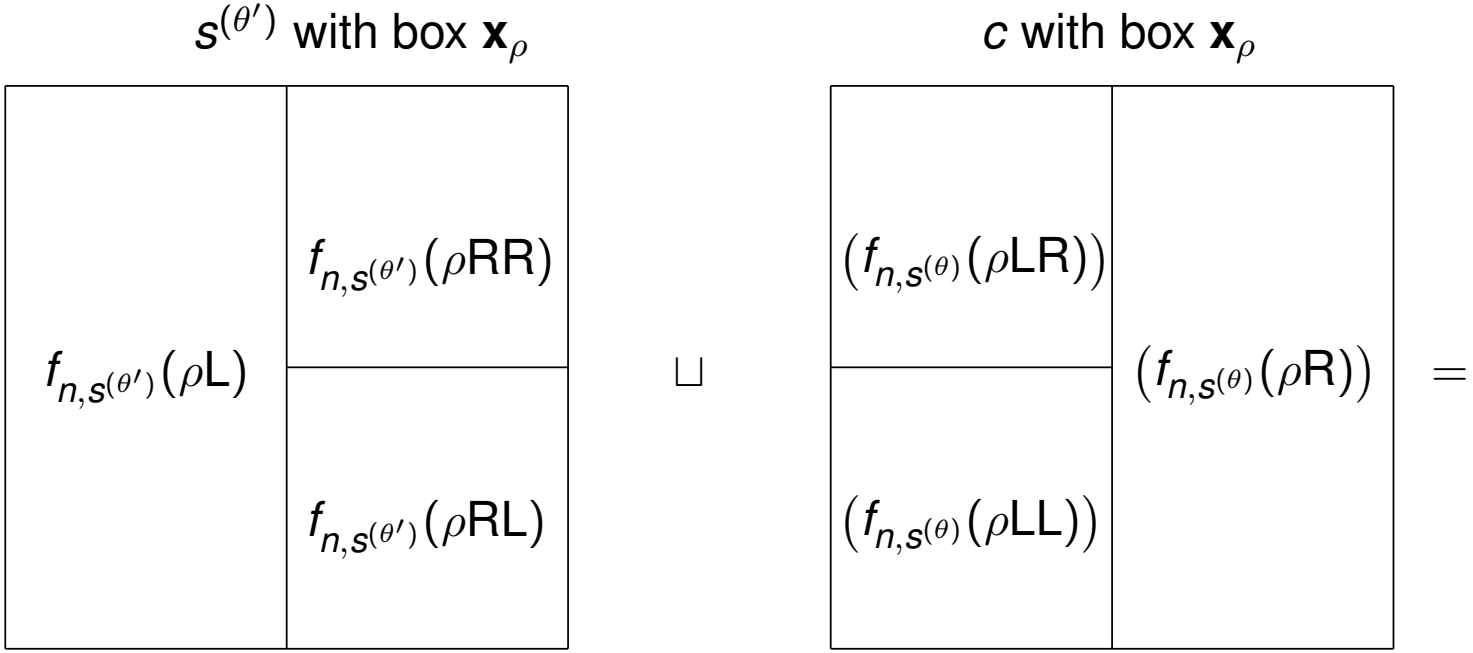
Figure: Make the SRP $s^{(\theta)}$ into a CRP $c$.



# RP as a Collator

Collating two SRPs $s^{(\theta)}$ and $s^{(\theta')}$ with the same root box $\mathbf{x}_\rho$:

Figure: Collate the SRP $s^{(\theta')}$ onto $c$.



# RP as a Collator

Collating two SRPs $s^{(\theta)}$ and $s^{(\theta')}$ with the same root box $\mathbf{x}_\rho$:

Figure: Collate the SRP $s^{(\theta')}$ onto $c$.
$c$ with box $\mathbf{x}_\rho$

| | |
|---|---|
| $\begin{pmatrix} f_{n,s^{(\theta)}}(\rho\mathrm{LR}) \\ f_{n,s^{(\theta')}}(\rho\mathrm{L}) \end{pmatrix}$ | $\begin{pmatrix} f_{n,s^{(\theta)}}(\rho\mathrm{R}) \\ f_{n,s^{(\theta')}}(\rho\mathrm{RR}) \end{pmatrix}$ |
| $\begin{pmatrix} f_{n,s^{(\theta)}}(\rho\mathrm{LL}) \\ f_{n,s^{(\theta')}}(\rho\mathrm{L}) \end{pmatrix}$ | $\begin{pmatrix} f_{n,s^{(\theta)}}(\rho\mathrm{R}) \\ f_{n,s^{(\theta')}}(\rho\mathrm{RL}) \end{pmatrix}$ |

# Minimum Distance Estimation (MDE): An Example

First candidate: $s_0$, $\theta = 0$.
Second candidate: $s_{1,1}$, $\theta = 1$.
So $\Theta = \Theta_1 = \{0, 1\}$.
Collate the two candidates:

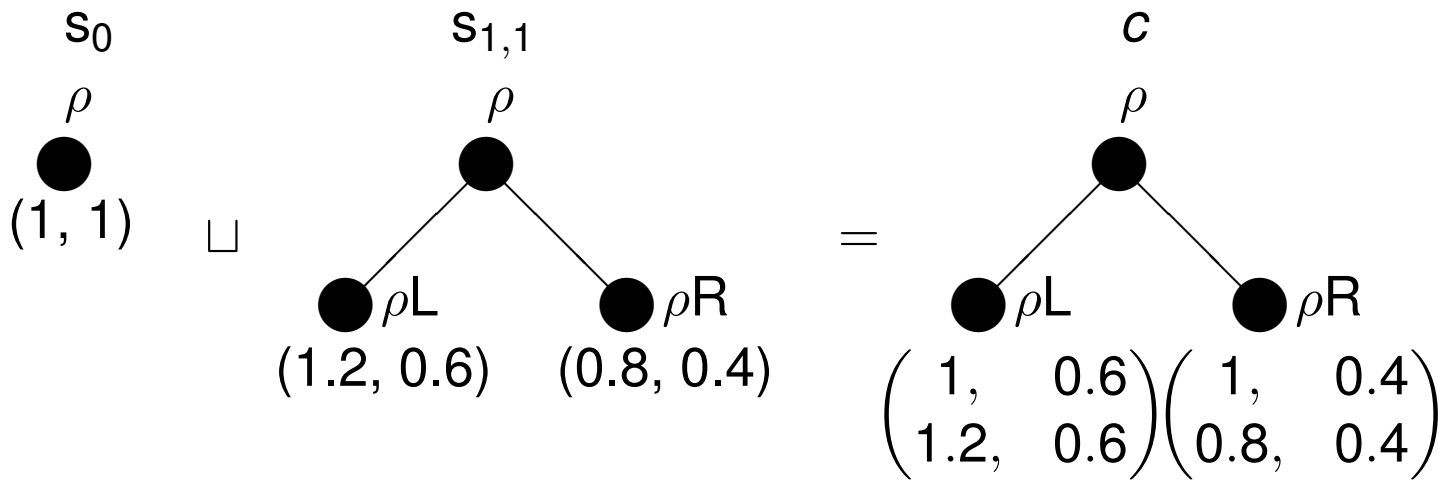
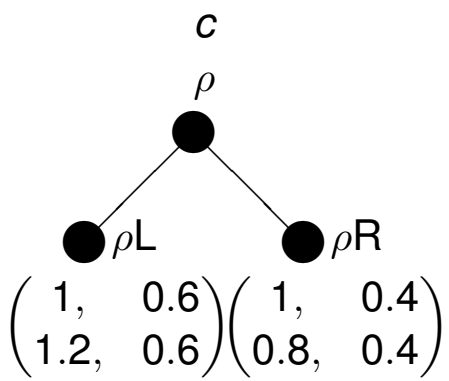



Figure: Collating $s_0$ and $s_{1,1}$.

# Minimum Distance Estimation (MDE): An Example

Compare $f_{n-\varphi n,\theta}, \theta \in \Theta$ for each leaf box $\mathbf{x}_{\rho v} \in \ell(c)$.




At $\mathbf{x}_{\rho \mathrm{L}}$,

$$f_{n-\varphi n,\theta=0}(\mathbf{x}_{\rho \mathrm{L}}) = 1 < f_{n-\varphi n,\theta=1}(\mathbf{x}_{\rho \mathrm{L}}) = 1.2 \ .$$

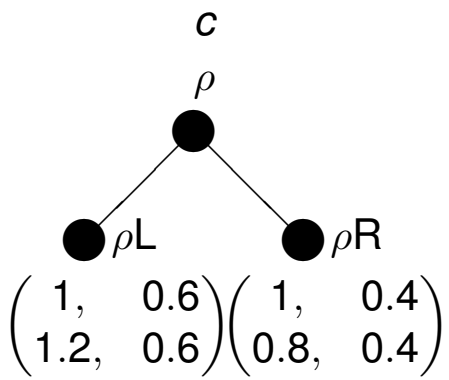
But,

$$f_{n-\varphi n,\theta=1}(\mathbf{x}_{\rho \mathrm{L}}) > f_{n-\varphi n,\theta=0}(\mathbf{x}_{\rho \mathrm{L}}) \ .$$

Thus $A_{\theta=1,\theta'=0} = \{\mathbf{x}_{\rho \mathrm{L}}\}$ and is in the Yatracos class $A_{\Theta_1}$.

# Minimum Distance Estimation (MDE): An Example

Compare $f_{n-\varphi n,\theta}, \theta \in \Theta$ for each leaf box $\mathbf{x}_{\rho v} \in \ell(c)$.



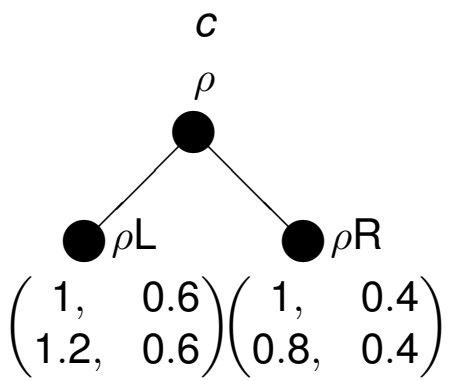

At $\mathbf{x}_{\rho \mathrm{R}}$,

$$f_{n-\varphi n,\theta=0}(\mathbf{x}_{\rho \mathrm{R}}) = 1 > f_{n-\varphi n,\theta=1}(\mathbf{x}_{\rho \mathrm{R}}) = 0.8 \ .$$

Thus $A_{\theta=0,\theta'=1} = \{\mathbf{x}_{\rho \mathrm{R}}\}$ and is in the Yatracos class $A_{\Theta_1}$.
Finally, we have $\mathcal{A}_{\Theta_1} = \{\mathbf{x}_{\rho \mathrm{L}}, \mathbf{x}_{\rho \mathrm{R}}\}$.

# Minimum Distance Estimation (MDE): An Example

Compare $f_{n-\varphi n,\theta}, \theta \in \Theta$ for each leaf box $\mathbf{x}_{\rho v} \in \ell(c)$.



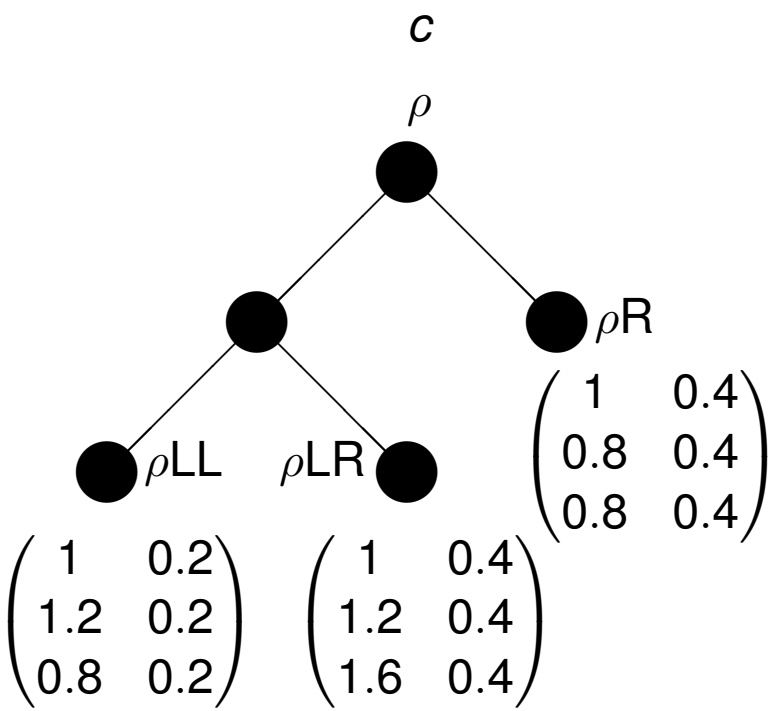

Can also express $\mathcal{A}_{\Theta_1}$ in a matrix as follows:

$$\mathcal{A}_{\Theta_1} = \begin{pmatrix} A_{0,0} & A_{0,1} \\ A_{1,0} & A_{1,1} \end{pmatrix} = \begin{pmatrix} \emptyset & \mathbf{x}_{\rho\mathrm{R}} \\ \mathbf{x}_{\rho\mathrm{L}} & \emptyset \end{pmatrix} \quad . \tag{1}$$

Note that the diagonal elements are all empty sets because there are no comparisons for the set $\{\theta, \theta'\}$ where $\theta = \theta'$.

# Minimum Distance Estimation (MDE): An Example

Make another split at node $\rho$L to produce the nodes $\rho$LL and $\rho$LR and get the SRP $s_{221}$. Perform collation to get the following collator $c$:



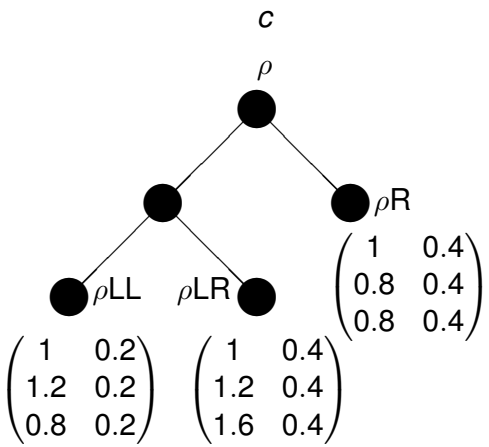
# Minimum Distance Estimation (MDE): An Example

Current $\Theta$ is now $\Theta = \{0, 1, 2\}$. To update the Yatracos class:

- Denote $\mathbf{x}^*_{\rho v}$ as the leaf box that is being split currently.
- Since only one leaf box is split every time, instead of comparing the estimates $f_{n-\varphi n,\theta}(\mathbf{x}_{\rho v})$ at each leaf box, we need only compare the estimates of its sub-boxes $\{\mathbf{x}^*_{\rho \mathrm{L}}, \mathbf{x}^*_{\rho \mathrm{R}}\}$ to update the Yatracos class.
- Thus we need not check the estimates at all the leaf boxes.

# Minimum Distance Estimation (MDE): An Example

Here $\mathbf{x}^*_{\rho v} = \mathbf{x}_{\rho \mathrm{L}}$. Only need to compare the estimates for each $\theta$ at the leaf boxes $\mathbf{x}_{\rho \mathrm{LL}}$ and $\mathbf{x}_{\rho \mathrm{LR}}$.



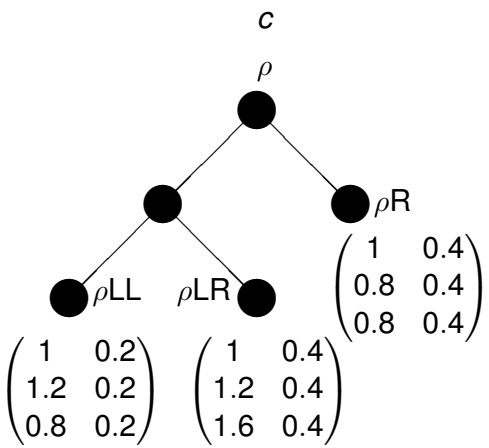

Compare the estimates of the pair $(0, 2)$ at $\mathbf{x}_{\rho \mathrm{LL}}$:

$$f_{n-\varphi n, \theta=0}(\mathbf{x}_{\rho \mathrm{LL}}) = 1 > f_{n-\varphi n, \theta'=2}(\mathbf{x}_{\rho \mathrm{LL}}) = 0.8 \ .$$

Thus $\mathbf{x}_{\rho \mathrm{LL}}$ will be taken into the set $A_{0,2}$.

# Minimum Distance Estimation (MDE): An Example

Here $\mathbf{x}^*_{\rho V} = \mathbf{x}_{\rho \mathrm{L}}$. Only need to compare the estimates for each $\theta$ at the leaf boxes $\mathbf{x}_{\rho \mathrm{LL}}$ and $\mathbf{x}_{\rho \mathrm{LR}}$.




From $A_{0,1}$ of Equation 1, we also know that $f_{n-\varphi n,\theta=0}(\mathbf{x}_{\rho \mathrm{R}})$ is larger than $f_{n-\varphi n,\theta=2}(\mathbf{x}_{\rho \mathrm{R}})$. This will also be true for the pair $(0, 2)$. Thus the set $A_{0,2}$ is $\mathbf{x}_{\rho \mathrm{R}} \cup \mathbf{x}_{\rho \mathrm{LL}}$.

# Minimum Distance Estimation (MDE): An Example

- If $A_{0,1}$ was a box that was split, no unions will be taken with this box since it is no longer a leaf box.
- In general, at some leaf box $\mathbf{x}_{\rho V}$ for which its estimates are being compared, for any pair $(\theta, \theta'), \theta \neq \theta'$, if $A_{\theta,\theta'} \neq \{\mathbf{x}_{\rho V^*}\}$, where $\mathbf{x}_{\rho V^*}$ was the box being split, we will take the union of $\mathbf{x}_{\rho V}$ with the elements of the set $A_{\theta-1,\theta'}$.
- Besides, since the sub-boxes will have the same estimate as its parent box, it will be redundant to make comparisons for the pair $\{0, 1\}$ .
- Therefore instead of doing $\binom{3}{2}$ comparisons, we now only need $\binom{2}{1}$ comparisons, i.e. comparing the estimates at the pairs $\{0, 2\}$ and $\{1, 2\}$.

# Minimum Distance Estimation (MDE): An Example

Continue the comparisons for the other sub-box $\mathbf{x}_{\rho \mathrm{LR}}$ and for all remaining comparison pairs.
The final Yatracos class $\mathcal{A}_{\Theta_2}$ is:

$$\mathcal{A}_{\Theta_2} = \begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} \\ A_{1,0} & A_{1,1} & A_{1,2} \\ A_{2,0} & A_{2,1} & A_{2,2} \end{pmatrix} = \begin{pmatrix} \emptyset & \mathbf{x}_{\rho \mathrm{R}} & \mathbf{x}_{\rho \mathrm{R}} \cup \mathbf{x}_{\rho \mathrm{LL}} \\ \mathbf{x}_{\rho \mathrm{L}} & \emptyset & \mathbf{x}_{\rho \mathrm{LL}} \\ \mathbf{x}_{\rho \mathrm{LR}} & \mathbf{x}_{\rho \mathrm{LR}} & \emptyset \end{pmatrix} .$$

# Minimum Distance Estimation (MDE): An Example

The corresponding $\Delta_\theta$ values for $\mathcal{A}_{\Theta_1}$ and $\mathcal{A}_{\Theta_2}$:

| | $\Delta_\theta = \int_{\mathbf{x}_{\rho v}} f_{n-\varphi n} - \mu_{\varphi n}(\mathbf{x}_{\rho v})$ | |
|---|---|---|
| $\theta$ | $\Theta_1 = \{0, 1\}$ | $\Theta_2 = \{0, 1, 2\}$ |
| $\theta = 0$ | 0.1 | 0.15 |
| $\theta = 1$ | 0 | 0.1 |
| $\theta = 2$ | - | **0** |

Table: Table of $\Delta_\theta$ values.

- Take $\mathcal{A}_{\Theta_2}$ to be the final Yatracos class;
- Observe the column associated with $\Theta_2$;
- The minimum distance estimate $f_{n-\varphi n, \theta^*}$ is the estimate at $\theta = 2$ since $\Delta_{\theta=2}$ is the minimum over all $\theta \in \Theta_2$.

Section 6

# Conclusions and References

# Conclusions

- Statistical Regular Paving (SRP) is a stat. sufficient data-adaptive structure for density estimation

# Conclusions

- Statistical Regular Paving (SRP) is a stat. sufficient data-adaptive structure for density estimation
- Arithmetic is efficiently extended through $\mathbb{R}$-MRPs

# Conclusions

- Statistical Regular Paving (SRP) is a stat. sufficient data-adaptive structure for density estimation
- Arithmetic is efficiently extended through $\mathbb{R}$-MRPs
- Combining PQ-based ($L_1$-consistent) initialization + Bayesian MCMC is powerful

# Conclusions

- Statistical Regular Paving (SRP) is a stat. sufficient data-adaptive structure for density estimation
- Arithmetic is efficiently extended through $\mathbb{R}$-MRPs
- Combining PQ-based ($L_1$-consistent) initialization + Bayesian MCMC is powerful
- Further decisions can be made with appropriate $\mathbb{R}$-MRP *arithmetic* (regression, anomaly detection, etc.)

# Conclusions

- Statistical Regular Paving (SRP) is a stat. sufficient data-adaptive structure for density estimation
- Arithmetic is efficiently extended through $\mathbb{R}$-MRPs
- Combining PQ-based ($L_1$-consistent) initialization + Bayesian MCMC is powerful
- Further decisions can be made with appropriate $\mathbb{R}$-MRP *arithmetic* (regression, anomaly detection, etc.)
- MDE over Yatracoss Classes of SRP histograms ($L_1$ School's Universal Performance Guarantees)

# Conclusions

- Statistical Regular Paving (SRP) is a stat. sufficient data-adaptive structure for density estimation
- Arithmetic is efficiently extended through $\mathbb{R}$-MRPs
- Combining PQ-based ($L_1$-consistent) initialization + Bayesian MCMC is powerful
- Further decisions can be made with appropriate $\mathbb{R}$-MRP *arithmetic* (regression, anomaly detection, etc.)
- MDE over Yatracoss Classes of SRP histograms ($L_1$ School's Universal Performance Guarantees)
- Future 1: Arithmetic over hyper-plane binary split trees (to account for linear transformation invariance)

# References

Devroye, L., and Lugosi, G. (2001). *Combinatorial methods in density estim.*, Springer.

Lugosi, G. and Nobel, A. (1996). Consistency of data-driven histogram methods for density estimation and classification. *Ann. Stat.* **24** 687–706.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *Ann. Stat.*. **20**, 1222–1235.

Jaulin, L., Kieffer, M., Didrit, O. & Walter, E. (2001). *Applied interval analysis*, Springer-Verlag.

Gray, A. G. & Moore, A. W. (2003). Nonparametric density estimation: toward computational tractability. In *Proc. of the SIAM Intl. Conf. on Data Mining*.

Harlow, J., Sainudiin, R. & Tucker, W. (2012). Mapped Regular Pavings, *Reliable Computing*, **16**, 252–282.

Teng G., Kuhn, K. and Sainudiin, R. (2012). Statistical regular pavings to analyze massive data of aircraft trajectories, *J. of Aerospace Comp. Inf. Commun.*, **9**:1, 14–25.

Sainudiin, R., Teng G., Harlow, J., and Lee D. (2013). Posterior expectation of regularly paved random histograms, *ACM Trans. on Model. Comp. Simul.*, **23**: 1, Article 6, 20 pp.

Sainudiin, R. & York, T. (2013). An auto-validating trans-dimensional universal rejection sampler for locally Lipschitz arithmetical expressions, *Reliable Computing*, **18**, 15–54.

# Acknowledgements

- RS's external consulting revenues from the New Zealand Ministry of Tourism
- WT's Swedish Research Council Grant 2008-7510 that enabled RS's visits to Uppsala in 2006, 2009, 2012
- Erskine grant from University of Canterbury that enabled WT's visit to Christchurch in 2011 & 2014
- University of Canterbury MSc Scholarship to JH.
- *Correctness by Construction*, 7th Framework Prog. of the EU, Marie Curie Actions-People, International Research Staff Exchange Scheme (IRSES), 2014 - 2017 (counter-part funds from Royal Soc. of NZ)

Thank you!

# KDE (diagonal badwidth) Vs. SRP MCMC

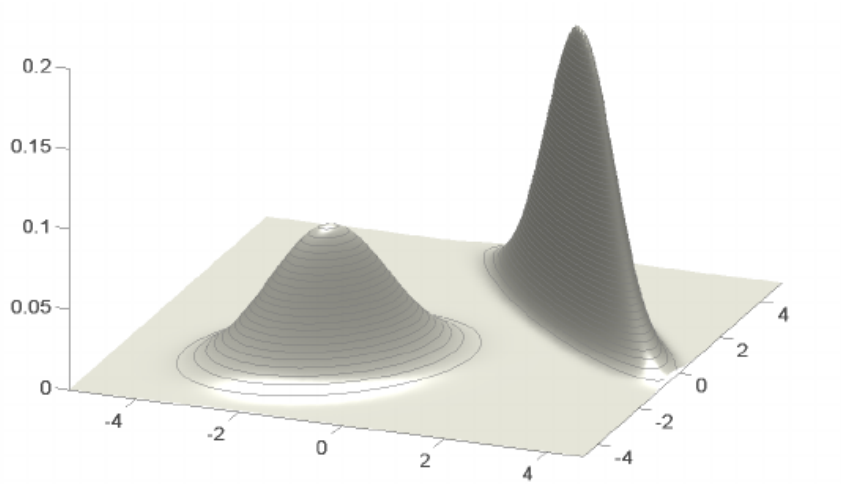



Figure B.2: Density II, $d = 2$.

When $d = 2$ Density II is a mixture of two bivariate Normal densities and is the same as Density A studied in Zhang et al. (2006):

$$\mu_a = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad \Sigma_a = \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix}, \quad \mu_b = \begin{pmatrix} -1.5 \\ -1.5 \end{pmatrix}, \quad \Sigma_b = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}.$$

# KDE (diagonal badwidth) Vs. SRP MCMC

Density II is a mixture of two multivariate Normal densities for $x \in \mathbb{R}^d$. Density II has high correlation between data coordinates and high bimodality:

$$f_{II}(x \,|\, \mu_a, \Sigma_a, \mu_b, \Sigma_b) = \frac{1}{2}\varphi(x \,|\, \mu_a, \Sigma_a) + \frac{1}{2}\varphi(x \,|\, \mu_b, \Sigma_b),$$

where $\varphi(x \,|\, \mu, \Sigma)$ is the multivariate Normal density with mean $\mu \in \mathbb{R}^d$ and $d \times d$ variance-covariance matrix $\Sigma$, and

$$\mu_a = \begin{pmatrix} 2.0 \\ \vdots \\ 2.0 \end{pmatrix}, \quad \Sigma_a = \begin{pmatrix} \sigma_a(x_1, x_1) & \sigma_a(x_1, x_2) & \cdots & \sigma_a(x_1, x_d) \\ \sigma_a(x_2, x_1) & \sigma_a(x_2, x_2) & \cdots & \sigma_a(x_2, x_d) \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_a(x_d, x_1) & \sigma_a(x_d, x_2) & \cdots & \sigma_a(x_d, x_d) \end{pmatrix},$$

$$\mu_b = \begin{pmatrix} -1.5 \\ \vdots \\ -1.5 \end{pmatrix}, \quad \Sigma_b = \begin{pmatrix} \sigma_b(x_1, x_1) & \sigma_b(x_1, x_2) & \cdots & \sigma_b(x_1, x_d) \\ \sigma_b(x_2, x_1) & \sigma_b(x_2, x_2) & \cdots & \sigma_b(x_2, x_d) \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_b(x_d, x_1) & \sigma_b(x_d, x_2) & \cdots & \sigma_b(x_d, x_d) \end{pmatrix},$$

and

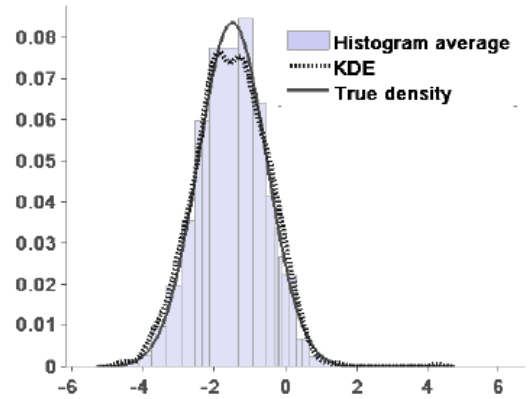
$$\sigma_a(x_i, x_j) = \begin{cases} 1 & \text{if } i = j, \\ -0.9^{|i-j|} & \text{if } i \neq j, \end{cases}, \quad \sigma_b(x_i, x_j) = \begin{cases} 1 & \text{if } i = j, \\ 0.3^{|i-j|} & \text{if } i \neq j, \end{cases}.$$

# KDE (diagonal badwidth) Vs. SRP MCMC

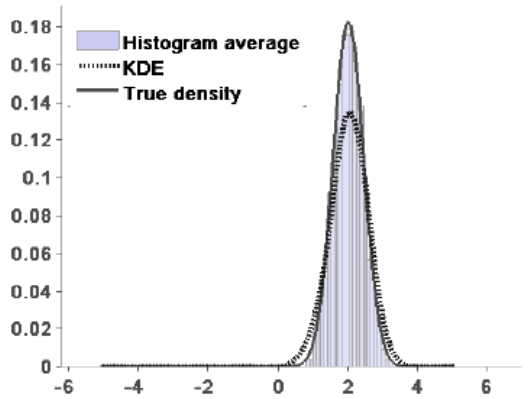
Table 7.2: Estimated errors for KDE and averaged SRP histogram RMRP.

| | $\hat{d}_{KL}$ | $\hat{L}_1$ error | Time (s) | | Leaves | |
|---|---|---|---|---|---|---|
| | | | min. | max. | min. | max. |
| 2-$d$ | | | | | | |
| KDE ($n_K = 2,000$) | 0.04 | 0.20 | 5,000 | 7,200 | $n/a$ | |
| Averaged RMRP histogram | | | | | | |
| $n = 10,000$ | 0.06 | 0.22 | 2 | 13 | 811 | 902 |
| $n = 50,000$ | 0.03 | 0.15 | 15 | 2,168 | 1,546 | 1,719 |
| 3-$d$ | | | | | | |
| KDE ($n_K = 2,000$) | 0.13 | 0.35 | 5,600 | 7,200 | $n/a$ | |
| Averaged RMRP histogram | | | | | | |
| $n = 10,000$ | 0.24 | 0.41 | 21 | 451 | 1,573 | 1,718 |
| $n = 50,000$ | 0.12 | 0.30 | 295 | 27,832 | 3,507 | 3,783 |
| 4-$d$ | | | | | | |
| KDE ($n_K = 2,000$) | 0.25 | 0.51 | 7,200 | 8,050 | $n/a$ | |
| Averaged RMRP histogram | | | | | | |
| $n = 50,000$ | 0.32 | 0.47 | 2,524 | 53,190 | 6,241 | 6,570 |
| $n = 100,000$ | 0.25 | 0.42 | 10,382 | 82,684 | 9,431 | 9,775 |
| 5-$d$ | | | | | | |
| KDE ($n_K = 2,000$) | 0.41 | 0.66 | 7,350 | 8,880 | $n/a$ | |
| Averaged RMRP histogram | | | | | | |
| $n = 50,000$ | 0.65 | 0.67 | 28,841 | 277,071 | 9,342 | 9,803 |
| $n = 100,000$ | 0.53 | 0.60 | 24,244 | 399,016 | 15,160 | 15,563 |

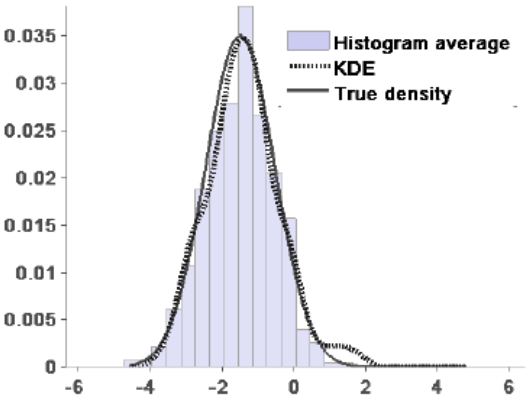
# KDE (diagonal badwidth) Vs. SRP MCMC



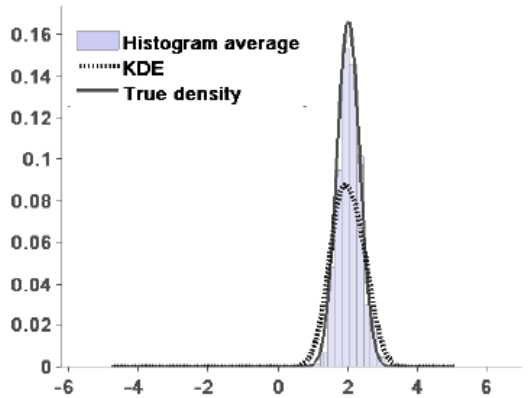

(a) $d = 2$, $x_2 = -1.5$, $n = 50,000$.




(b) $d = 2$, $x_1 = 2.0$, $n = 50,000$.




(c) $d = 3$, $x_2 = x_3 = -1.5$, $n = 50,000$.




(d) $d = 3$, $x_1 = x_3 = 2.0$, $n = 50,000$.

Figure 7.5: Density II, KDE and averaged RMRP histogram slice.