Adaptive Histograms from a Randomized Queue that is Prioritized for Statistically Equivalent Blocks

Gloria Teng Jennifer Harlow Raazesh Sainudiin

Department of Mathematics and Statistics, University of Canterbury, New Zealand

August 19, 2010

Introduction

- Present *statistical regular sub-pavings* as an efficient, data-driven, multi-dimensional data-structure for non-parametric density estimation of massive data sets;
- Apply our methods to earthquakes in NZ, weather and aircraft trajectories over a busy US airport and samples simulated from challenging multi-dimensional densities, including Levy and Rosenbrock.

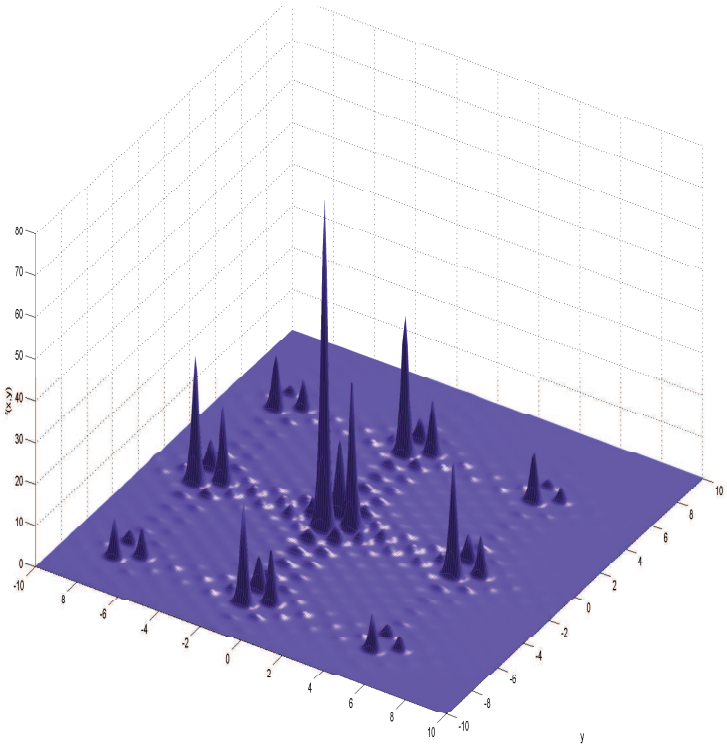
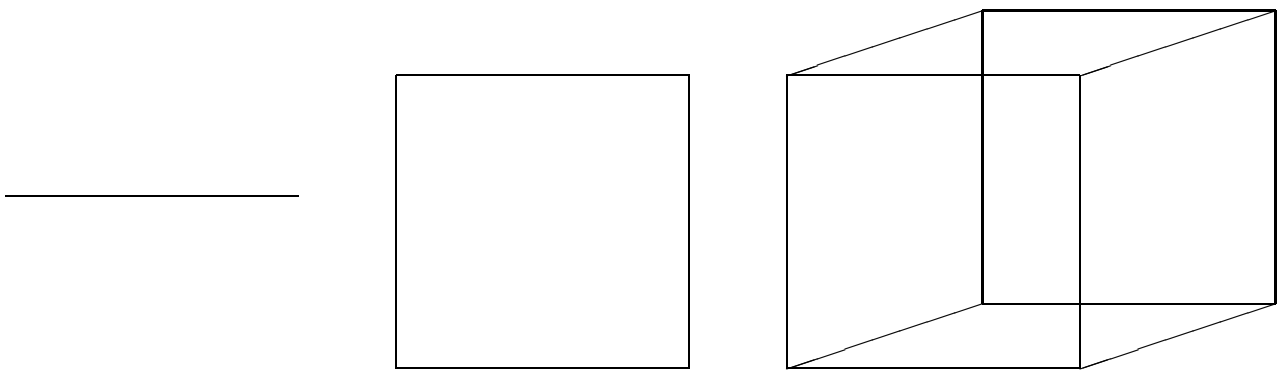


Figure: Shape of a Levy density with 700 modes.

Intervals and Boxes in $\mathbb{R}^d$

Intervals and *Boxes* as interval vectors:

$$\mathbf{x} = [\underline{x}_1, \overline{x}_1] \times [\underline{x}_2, \overline{x}_2] \times \ldots \times [\underline{x}_d, \overline{x}_d] \,, \, \underline{x}_i \leq \overline{x}_i \ .$$



1-dim. 2-dim. 3-dim.

Figure: Boxes in 1D, 2D, and 3D.

Binary Tree Representation

These boxes can also be represented by ordered binary trees.
An operation of bisection on a box is equivalent to performing the operation on its corresponding node in the tree, i.e.:

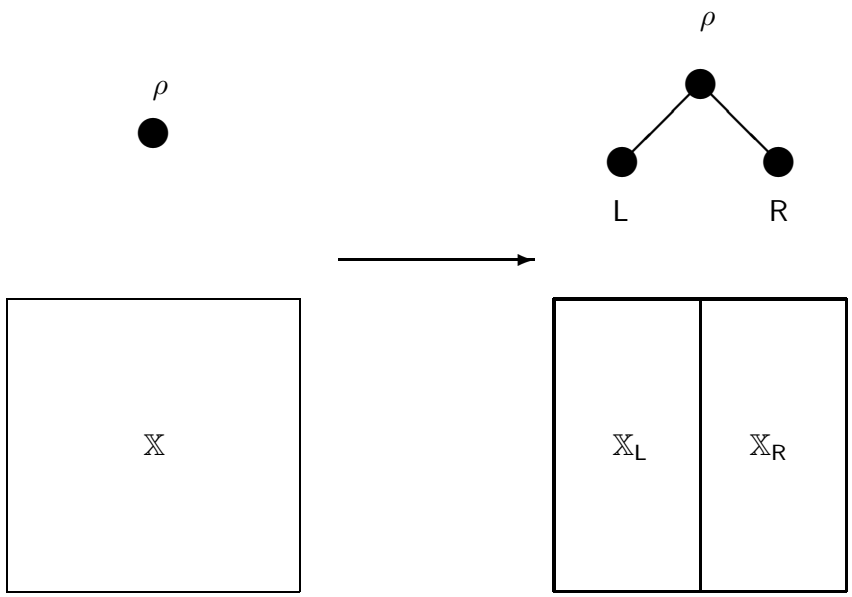
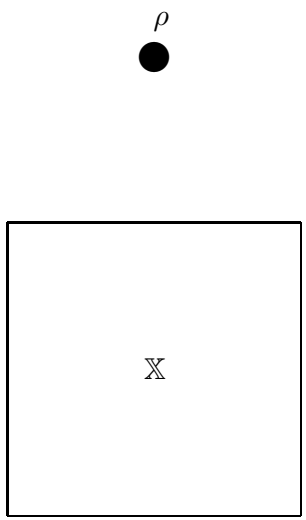


Figure: Bisecting a box or its equivalent node.

Regular Sub-pavings (RSPs) (Jaulin et. al., 2001)

- A sequence of bisections of boxes;
- Start from the root box;
- Along the first widest dimension.

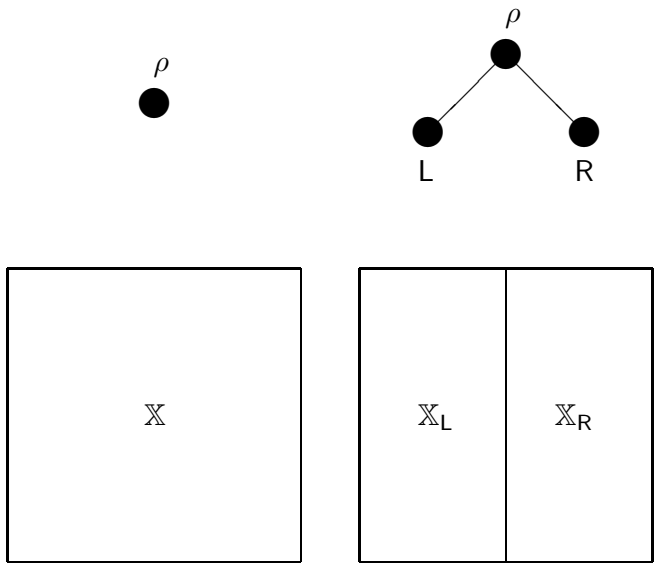
Figure: A sequence of bisections on root box $\mathbb{X}$ to produce a 4-leafed RSP s.



Regular Sub-pavings (RSPs) (Jaulin et. al., 2001)

- A sequence of bisections of boxes;
- Start from the root box;
- Along the first widest dimension.

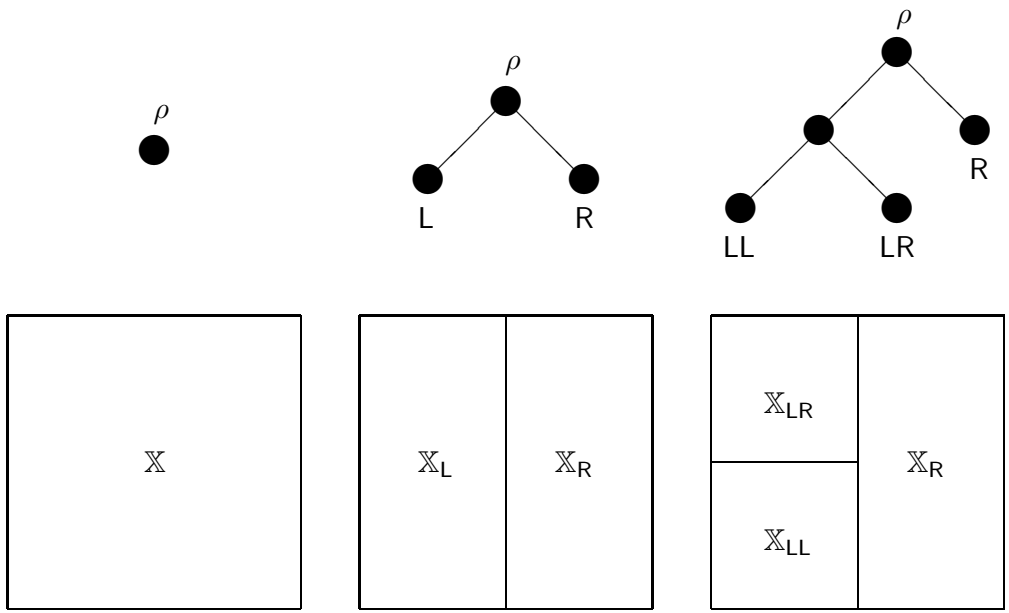
Figure: A sequence of bisections on root box $\mathbb{X}$ to produce a 4-leafed RSP s.



Regular Sub-pavings (RSPs) (Jaulin et. al., 2001)

- A sequence of bisections of boxes;
- Start from the root box;
- Along the first widest dimension.

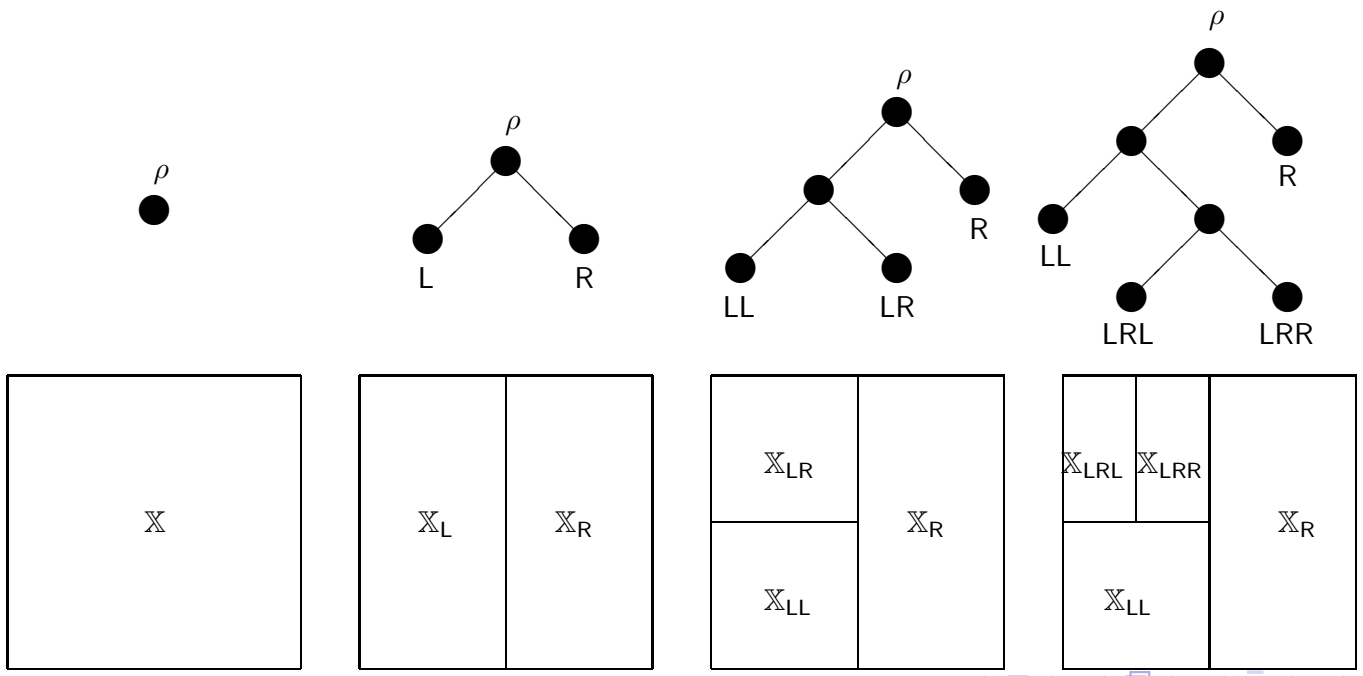
Figure: A sequence of bisections on root box $\mathbb{X}$ to produce a 4-leafed RSP s.



Regular Sub-pavings (RSPs) (Jaulin et. al., 2001)

- A sequence of bisections of boxes;
- Start from the root box;
- Along the first widest dimension.

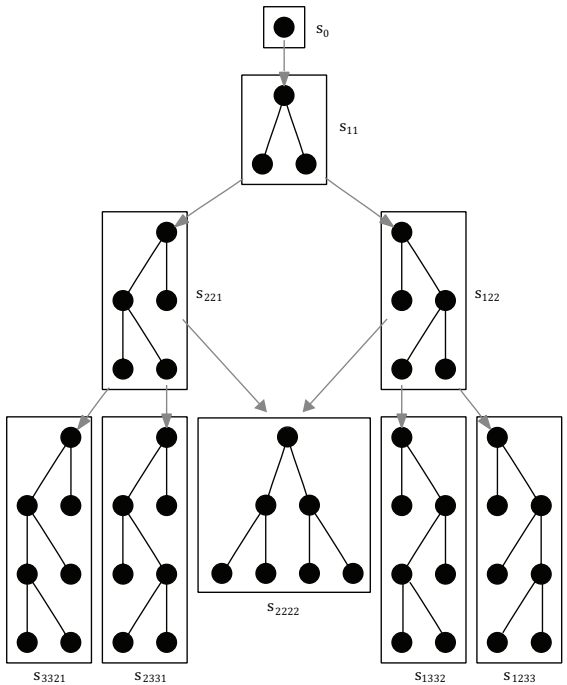
Figure: A sequence of bisections on root box $\mathbb{X}$ to produce a 4-leafed RSP s.



The Space of All Possible RSPs

The number of distinct RSP with i splits is equal to the Catalan number:

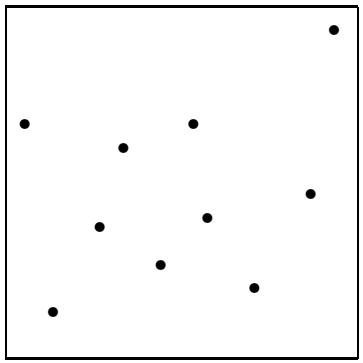
$$C_i = \frac{1}{i+1}\binom{2i}{i} = \frac{(2i)!}{(i+1)!(i!)} \ .$$



Statistical Regular Sub-pavings (SRSPs)

- Extended from the RSP;
- Caches recursively computable statistics at each box or node as data falls through;
- These statistics include:
  - the sample count;
  - the sample mean vector;
  - the sample variance-covariance matrix;
  - and the volume of the box.

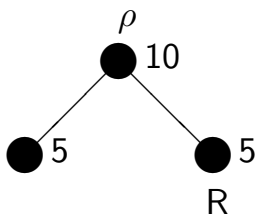
Figure: Caching the sample count in each node (or box).

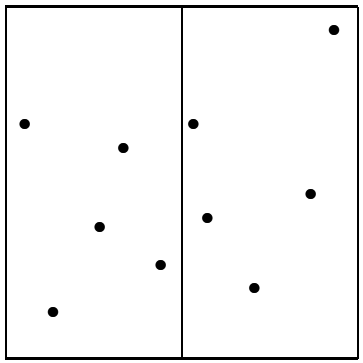


Statistical Regular Sub-pavings (SRSPs)

- Extended from the RSP;
- Caches recursively computable statistics at each box or node as data falls through;
- These statistics include:
  - the sample count;
  - the sample mean vector;
  - the sample variance-covariance matrix;
  - and the volume of the box.

Figure: Caching the sample count in each node (or box).

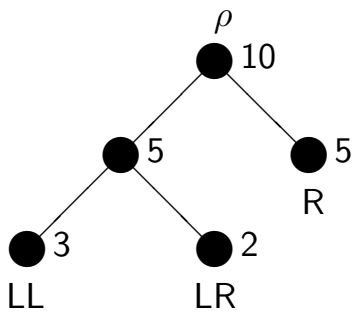


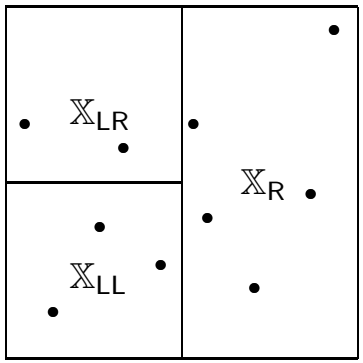


Statistical Regular Sub-pavings (SRSPs)

- Extended from the RSP;
- Caches recursively computable statistics at each box or node as data falls through;
- These statistics include:
  - the sample count;
  - the sample mean vector;
  - the sample variance-covariance matrix;
  - and the volume of the box.

Figure: Caching the sample count in each node (or box).





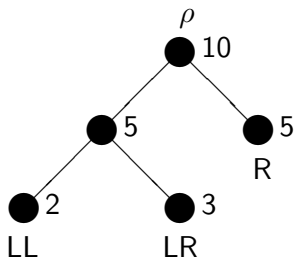
SRSPs as Adaptive Histograms

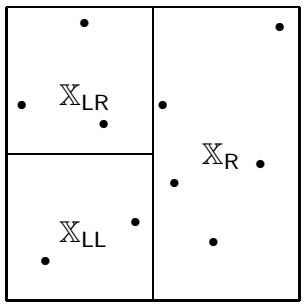
The histogram estimate of i.i.d. random variables $X_1, X_2, \ldots, X_n$ in $\mathbb{R}^d$ with density f is given by:

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{I_{X_i \in \mathbf{x}(x)}}{vol(\mathbf{x})}$$

$\mathbf{x}(x)$: the leaf box $\mathbf{x}$ that contains x $\quad$ $vol(\mathbf{x})$: volume of box $\mathbf{x}$

Figure: A SRSP as a histogram estimate.





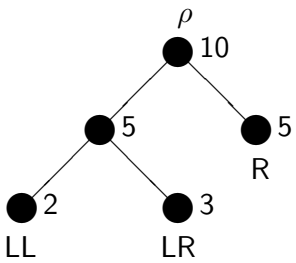
SRSPs as Adaptive Histograms

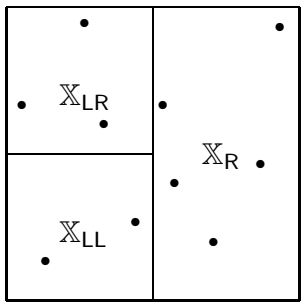
The histogram estimate of i.i.d. random variables $X_1, X_2, \ldots, X_n$ in $\mathbb{R}^d$ with density f is given by:

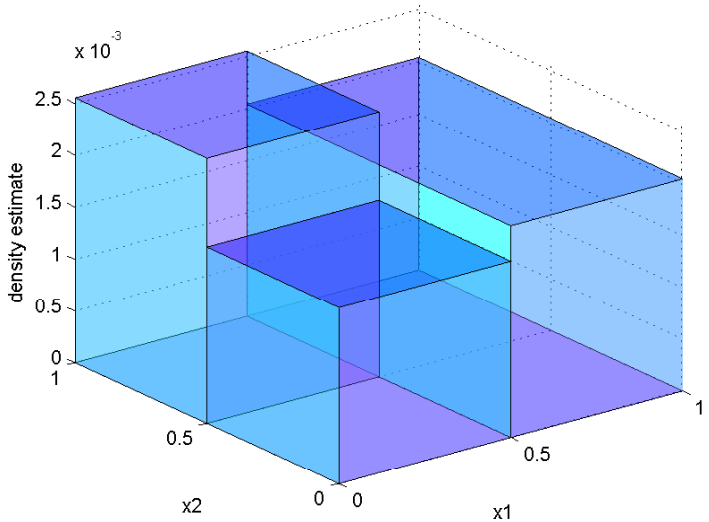
$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{I_{X_i \in \mathbf{x}(x)}}{vol(\mathbf{x})}$$

$\mathbf{x}(x)$: the leaf box $\mathbf{x}$ that contains x $vol(\mathbf{x})$: volume of box $\mathbf{x}$

Figure: A SRSP as a histogram estimate.



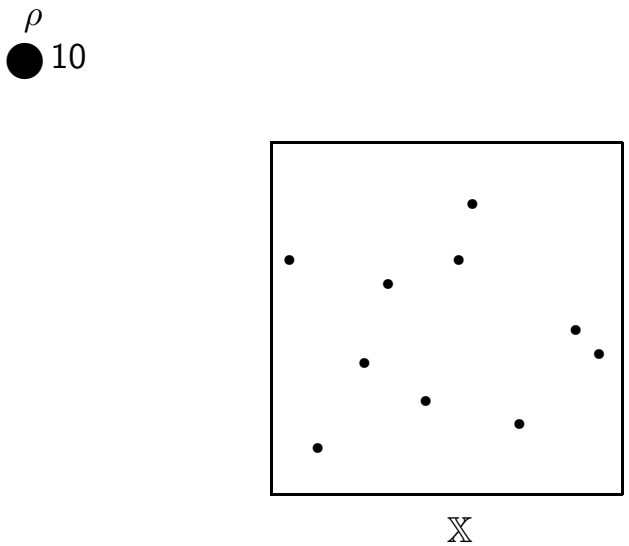




A Prioritized Queue based Algorithm

Algorithm `SplitMostCounts`

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

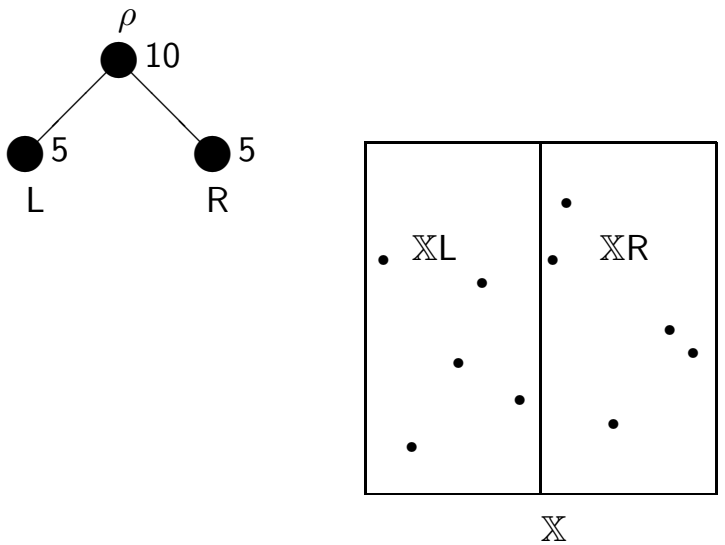


A Prioritized Queue based Algorithm

Algorithm `SplitMostCounts`

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

Split the root box.

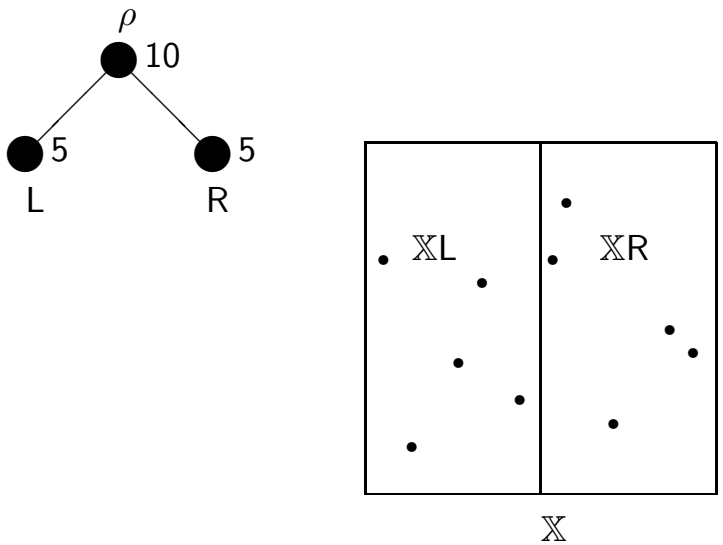


A Prioritized Queue based Algorithm

Algorithm `SplitMostCounts`

As data arrives, order the leaf boxes of the SRSP 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?

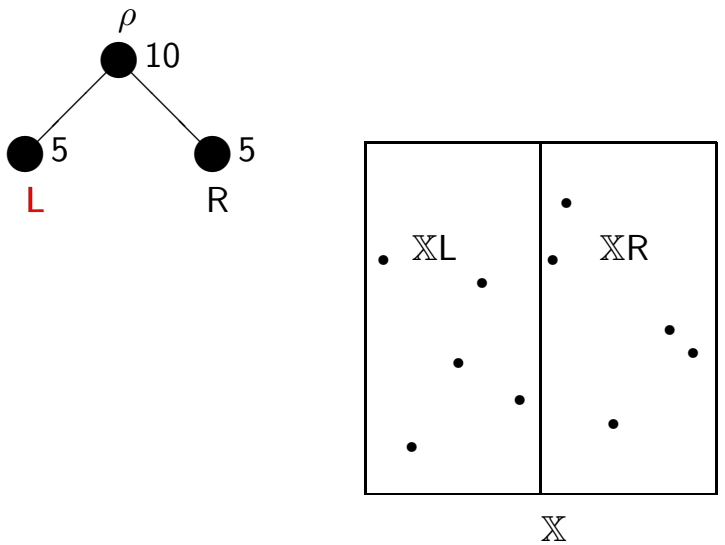


A Prioritized Queue based Algorithm

Algorithm `SplitMostCounts`

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

Break ties by picking these boxes at random for the next bisection.

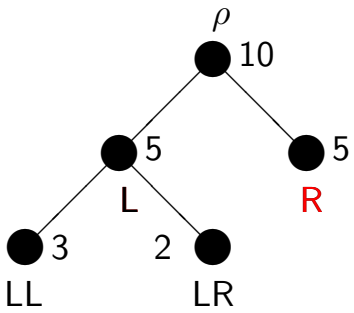


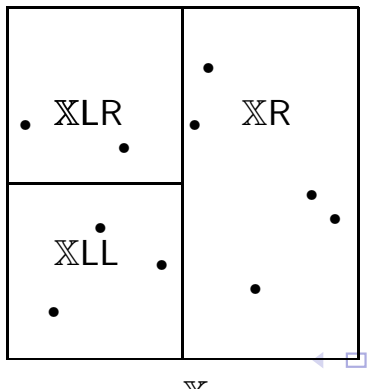
A Prioritized Queue based Algorithm

Algorithm `SplitMostCounts`

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

Keep bisecting till each box has less than or equal to k_n number of points (let $k_n = 3$ here).



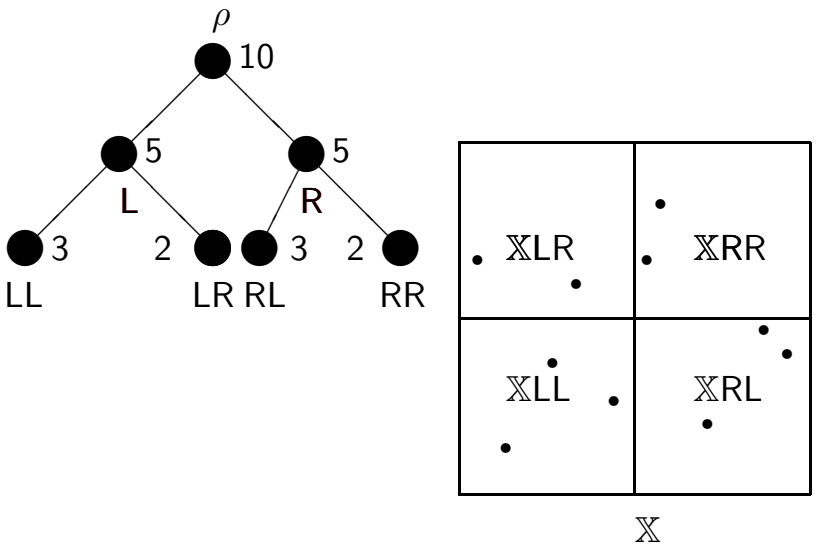


A Prioritized Queue based Algorithm

Algorithm `SplitMostCounts`

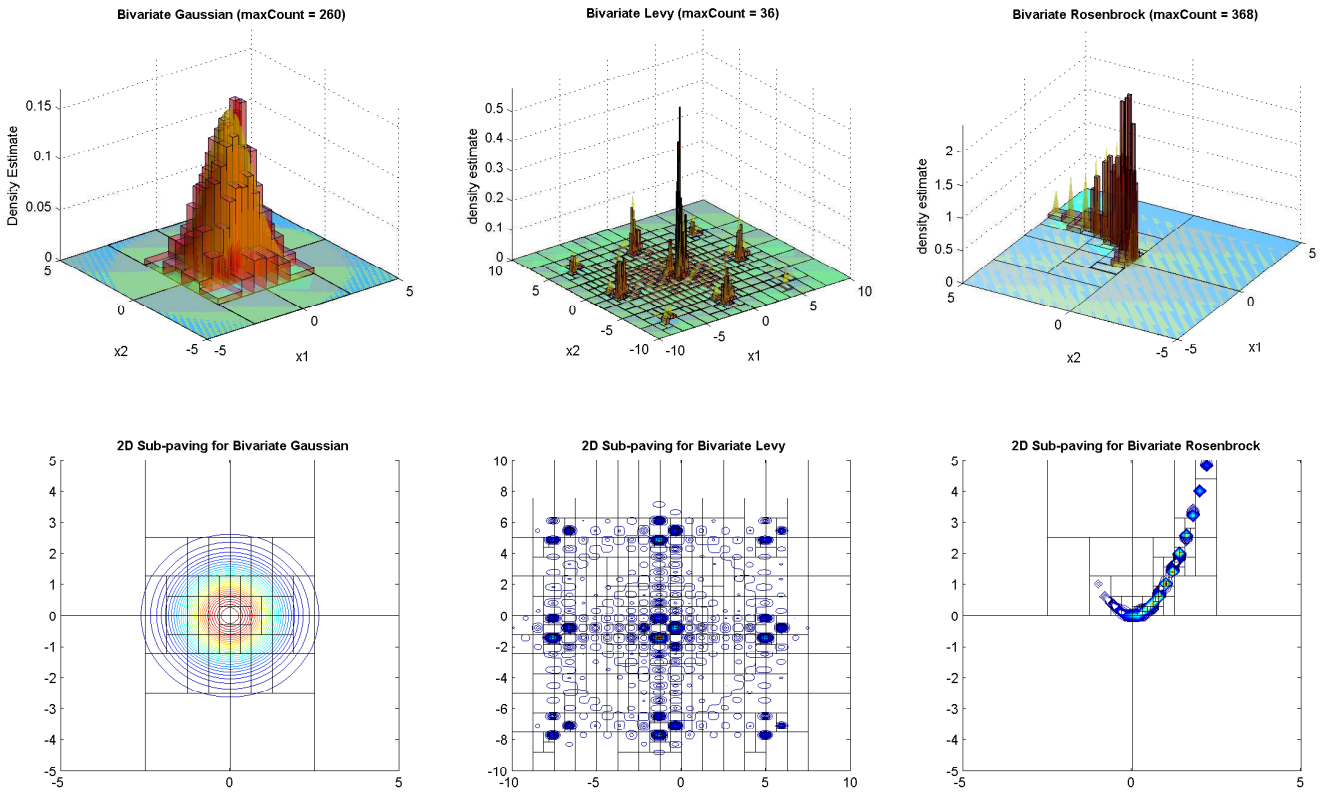
As data arrives, order the leaf boxes of the SRSP so that the leaf box with **the most number of points** will be chosen for the next bisection.

Final state



Some Examples

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



Choice of k_n

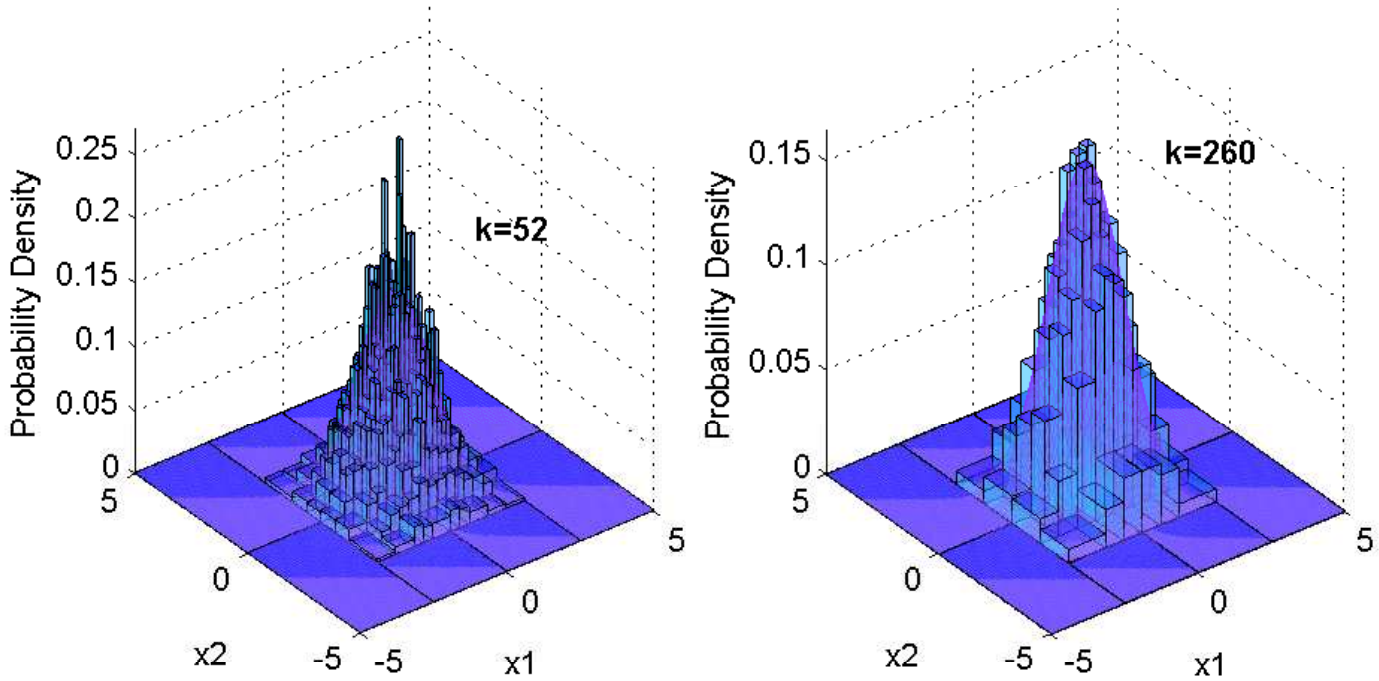
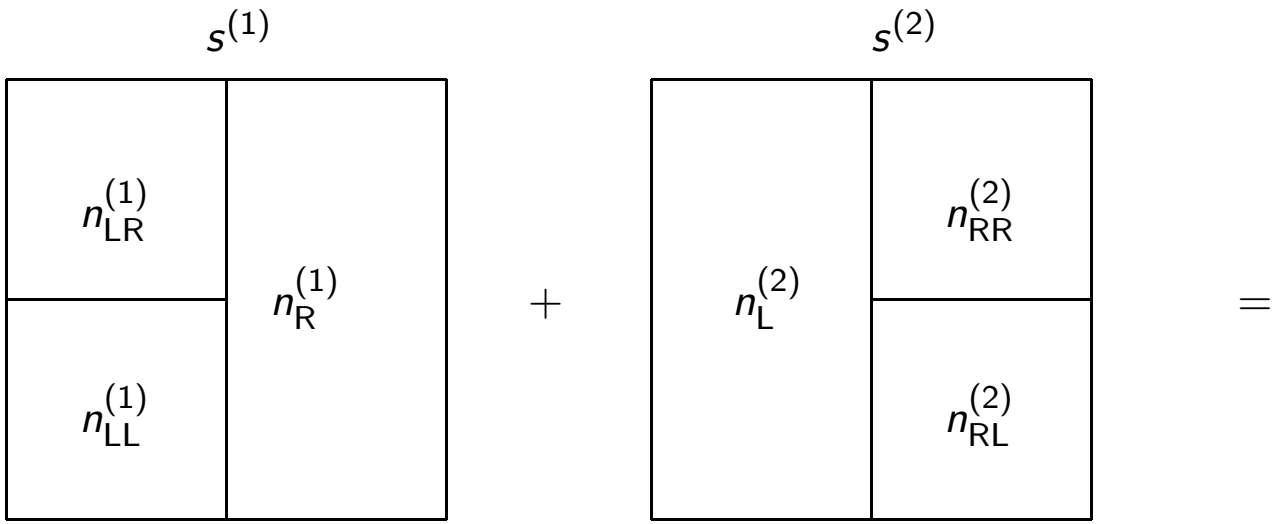


Figure: Two histogram density estimates for the standard bivariate gaussian density with different choices of k_n. The histogram is under-smoothed when k_n is relatively smaller than n and over-smoothed when k_n is relatively larger.

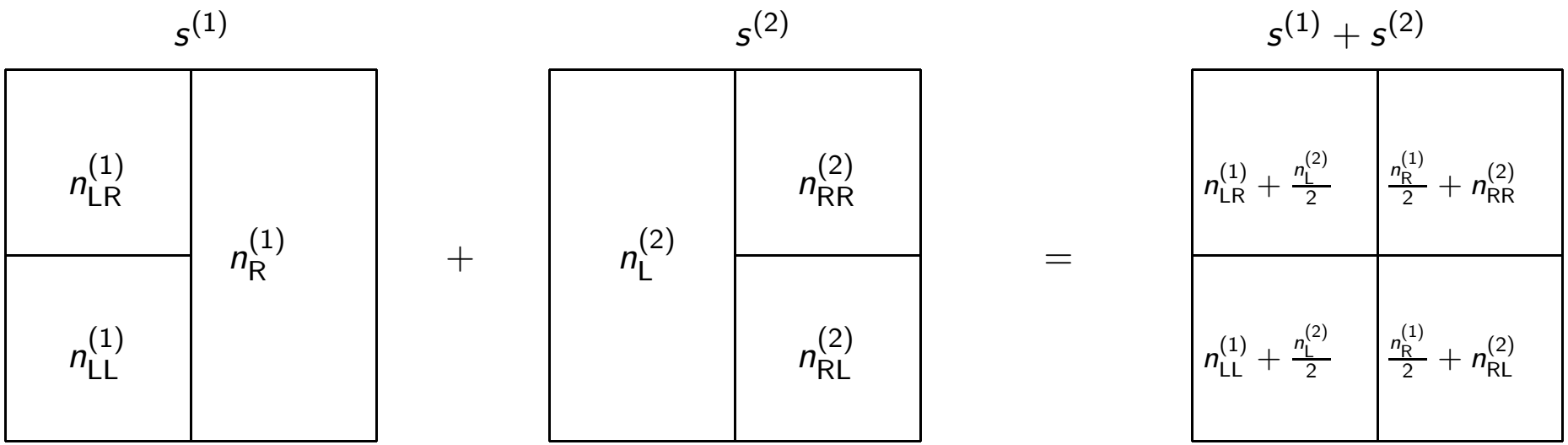
Adding and Averaging SRSPs

Perform a non-minimal union (or add sub-pavings) and adjust counts:



Adding and Averaging SRSPs

Perform a non-minimal union (or add sub-pavings) and adjust counts:



Adding and Averaging SRSPs

Adding m histogram density estimates

$$\sum_{i=1}^{m} \hat{f}^{(i)} = \hat{f}^{(1)} + \hat{f}^{(2)} + \hat{f}^{(3)} + \ldots + \hat{f}^{(m)}$$

$$= \left(\left(\left(\hat{f}^{(1)} + \hat{f}^{(2)} \right) + \hat{f}^{(3)} \right) + \ldots + \hat{f}^{(m)} \right) .$$

Adding and Averaging SRSPs

Adding m histogram density estimates

$$\sum_{i=1}^{m} \hat{f}^{(i)} = \hat{f}^{(1)} + \hat{f}^{(2)} + \hat{f}^{(3)} + \ldots + \hat{f}^{(m)}$$
$$= \left(\left(\left(\hat{f}^{(1)} + \hat{f}^{(2)}\right) + \hat{f}^{(3)}\right) + \ldots + \hat{f}^{(m)}\right) .$$

Averaging m histogram density estimate

$$\overline{\hat{f}} = \frac{1}{m} \sum_{i=1}^{m} \hat{f}^{(i)} .$$

An Example

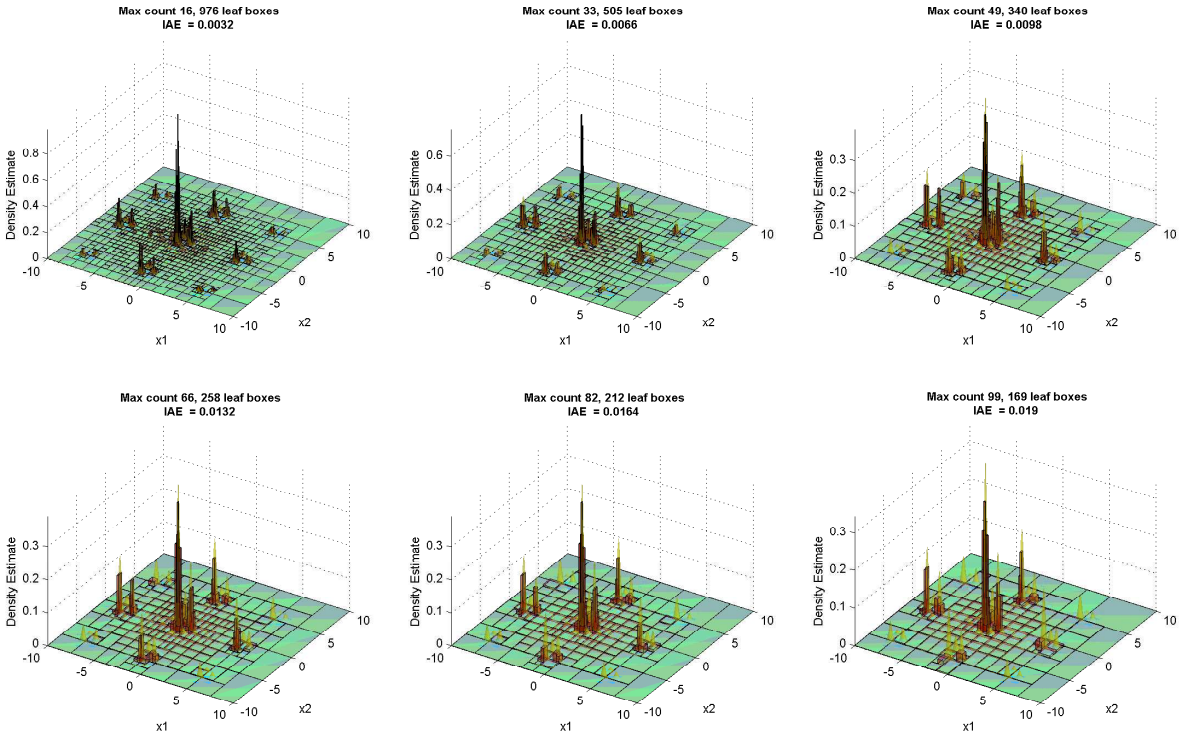


Figure: Histogram density estimates of the bivariate Levy using different values of k_n.

An Example

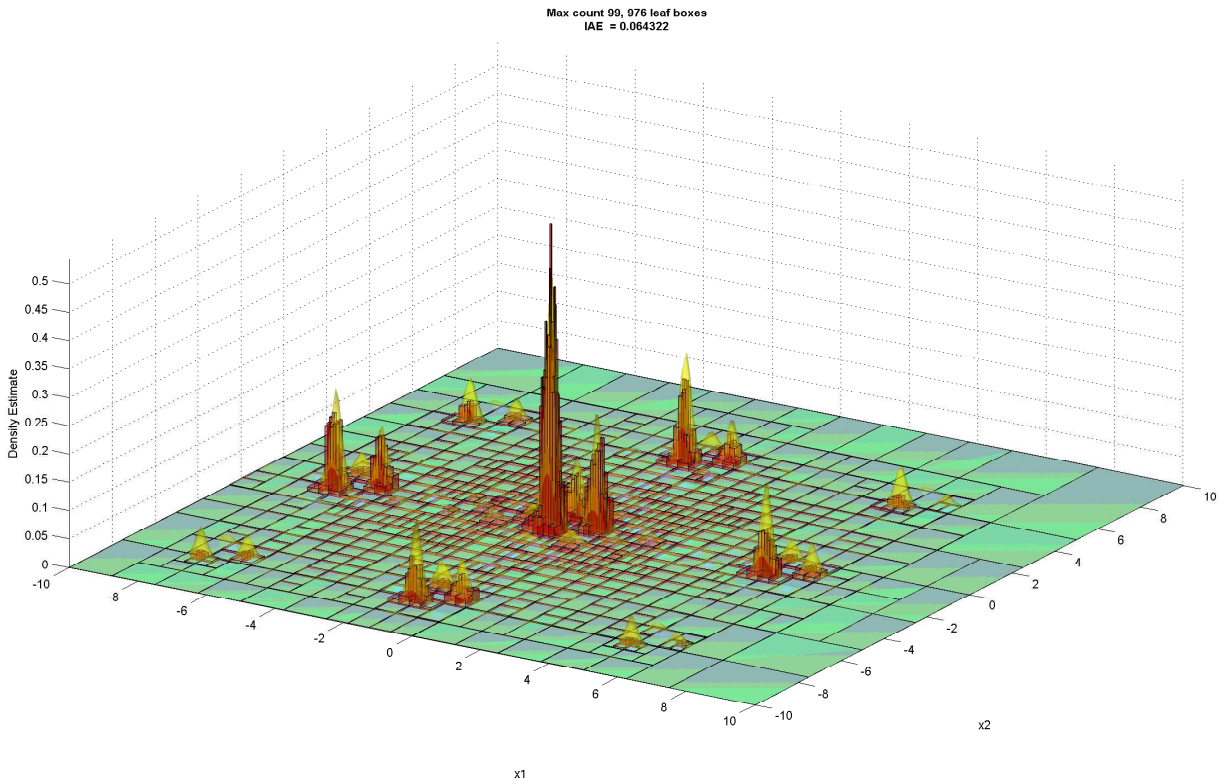


Figure: The averaged histogram density estimate.

An Example of Application

Example

Air Traffic Data (Link to SAGE server): interested in applying SRSPs to the analysis of thunderstorm effects on aggregated aircraft trajectories.

Conclusions

- We proposed an efficient, data-driven, multi-dimensional data-structure, **SRSPs**, for non-parametric density estimation of massive data sets;
- The SRSP can be represented by a binary tree and can either grow (through bisection of nodes) or be pruned (through merging nodes) adaptively;
- Arithmetic operations can be efficiently extended to these data structures, i.e. averaging histograms.

References

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

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

Sainudiin, R. and York, T. L. (2005). *An Auto-validating Rejection Sampler*. BSCB Dept. Technical Report BU-1661-M, Cornell University, Ithaca, New York.

Tucker, W. (2004). *Auto-validating numerical methods*. Lecture Notes, Uppsala University, Sweden.

Thank you!