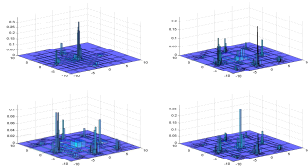


Statistical Regular Sub-pavings for Multivariate Density Estimation

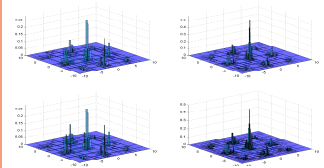
Gloria Teng, Jennifer Harlow, Dominic Lee and Raazesh Sainudiin

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



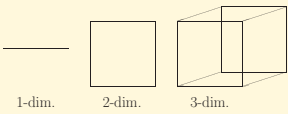
SUMMARY

- We propose an efficient, data-driven, multi-dimensional, metric data-structure that is sufficient for non-parametric density estimation of massive data sets.
- Arithmetic operations are extended to these data structures in order to efficiently obtain the average of two histograms.
- A non-parametric point-estimate of the density is obtained from the sample mean of a Markov chain whose stationary distribution is the posterior distribution over a class of histograms.



BOXES in \mathbb{R}^d

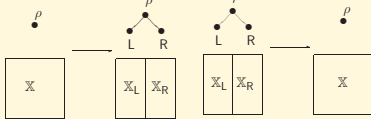
$$\mathbf{x} = [\underline{x}_1, \bar{x}_1] \times [\underline{x}_2, \bar{x}_2] \times \dots \times [\underline{x}_d, \bar{x}_d], \underline{x}_i \leq \bar{x}_i$$



These boxes can also be represented by ordered binary trees. An operation of bisection or reunion done on a box or a sibling pair of boxes is equivalent to performing the operation on its corresponding node in the tree, as shown below:

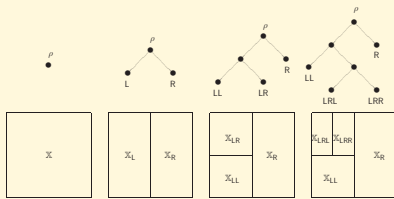
Bisecting a box

Reuniting two sub-boxes



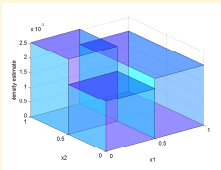
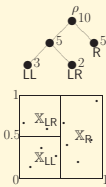
REGULAR SUB-PAVINGS (RSP)

A sequence of bisections of boxes along the first widest dimension, starting from the root box produces a set of partitions of \mathbb{R}^d known as *regular sub-pavings* (RSP) (Jaulin et al., 2001).



STATISTICAL REGULAR SUB-PAVINGS (SRSP)

A *statistical regular sub-paving* (SRSP) is extended from the RSP and is capable of caching 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.



A SRSP and its equivalent binary tree with sample counts. Whenever a box (or node) is bisected, the counts associated with each sub-box produced is then updated. We then get a histogram estimate of the data.

The histogram estimate of the data X_1, X_2, \dots, X_n is given by:

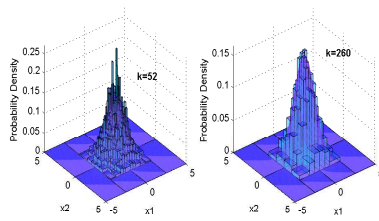
$$\hat{f}(\mathbf{x}; X_1, X_2, \dots, X_n) = \frac{1}{n} \sum_{\mathbf{x}} \frac{n_{\mathbf{x}}}{vol(\mathbf{x})}$$

$n_{\mathbf{x}}$: number of observations in box \mathbf{x}
 $vol(\mathbf{x})$: volume of box \mathbf{x}

A priority queue based algorithm is employed to obtain histogram estimates. The bisection process stops when each box has less than or equal to k_n number of points. The choice of k_n will determine the total number and volume of boxes produced and in turn, affects the bias and variance of our estimator.

PROBLEM

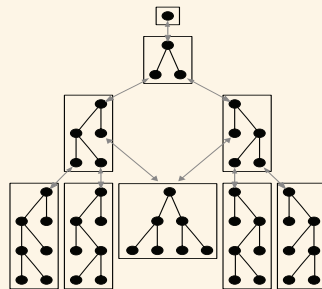
What is an 'optimal' k_n ?



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.

PROPOSED METHOD

Find a non-parametric point estimate of the density from the sample mean of a Markov chain whose stationary distribution is the posterior distribution over the space of all possible histograms over regular sub-pavings.



- The figure shows a state transition diagram of RSPs with 0, 1, 2 and 3 splits;
- This state space is denoted by $\mathcal{S}_{0,3}$;
- Number of j splits is the Catalan number: $C_j = \frac{1}{j+1} \binom{2j}{j}$;
- There is more than one way to reach a RSP by j splits;
- We are interested to use Markov chains on $\mathcal{S}_{0,\infty}$.

POSTERIOR DISTRIBUTION OVER HISTOGRAMS IN $\mathcal{S}_{0,\infty}$

Let s be a histogram with partition $\ell(s)$ given by the leaves of a RSP with j splits and $j+1$ leaves.

The likelihood function:

$$P(x_1, \dots, x_n | s) = \prod_{\mathbf{x} \in \ell(s)} \left(\frac{n_{\mathbf{x}}}{n \cdot vol(\mathbf{x})} \right)^{n_{\mathbf{x}}}$$

A possible prior probability:

$$P(s) \propto \frac{1}{C_j^2}$$

Resulting posterior distribution:

$$P(s | x_1, \dots, x_n) \propto P(x_1, \dots, x_n | s) \cdot P(s) = \prod_{\mathbf{x} \in \ell(s)} \left(\frac{n_{\mathbf{x}}}{n \cdot vol(\mathbf{x})} \right)^{n_{\mathbf{x}}} \frac{1}{C_j^2}$$

The Metropolis-Hastings algorithm is used to obtain samples of histogram states. We use a proposal where a bisection and a reunion are equally probable. Each leaf node has an equal probability of being bisected; and each parent node has an equal probability of having its sibling nodes reunited to itself.

REFERENCES

- Jaulin, L., Kieffer, M., Didrit, O. & Walter, E. (2001). *Applied interval analysis*. London: Springer-Verlag.
 Sainudiin, R. and York, T. L. (2005). *An Auto-validating Rejection Sampler*. BSCB Dept. Technical Report BU-1661-M, Cornell University, Ithaca, New York.

POSSIBLE PROPOSAL STATES

$$P(\text{Reunion into L}) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$$

$$P(\text{Bisect LR}) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$$



$$P(\text{Bisect LL}) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$$

$$P(\text{Bisect R}) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$$



Current state

Possible proposal states

AVERAGED HISTOGRAM

The MCMC algorithm gives us a sample of histogram states. We then get the average of these states to get the posterior mean of the stationary distribution of the distribution. We can perform arithmetic on SRSPs.

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

$$\begin{matrix} s^{(1)} \\ \begin{matrix} n_{LR}^{(1)} & n_{RR}^{(1)} \\ n_{LL}^{(1)} & n_{RR}^{(1)} \end{matrix} \end{matrix} + \begin{matrix} s^{(2)} \\ \begin{matrix} n_{RR}^{(2)} \\ n_{RL}^{(2)} \end{matrix} \end{matrix} = \begin{matrix} s^{(1)} + s^{(2)} \\ \begin{matrix} n_{LR}^{(1)} + \frac{n_{LL}^{(2)}}{2} + n_{RR}^{(2)} \\ n_{LL}^{(1)} + \frac{n_{LL}^{(2)}}{2} + n_{RL}^{(2)} \end{matrix} \end{matrix}$$

We can use this method to add m histogram density estimates $f^{(1)}, \dots, f^{(m)}$ as follows:

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

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

Averaging the adjusting counts over the number of histograms:

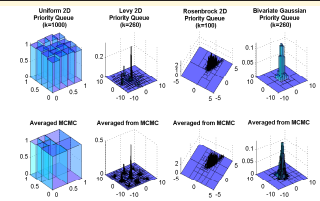
$$\frac{(s^{(1)} + s^{(2)})/2}{\begin{matrix} \frac{1}{2}(n_{LR}^{(1)} + \frac{n_{LL}^{(2)}}{2} + n_{RR}^{(2)}) \\ \frac{1}{2}(n_{LL}^{(1)} + \frac{n_{LL}^{(2)}}{2} + n_{RL}^{(2)}) \end{matrix}}$$

The histogram partition addition gives us a partition $\ell(s^{(1)} + \dots + s^{(m)})$ on which the averaged histogram is defined by

$$\bar{f} = \frac{1}{m} \sum_{i=1}^m f^{(i)}$$

SIMULATION RESULTS

Density	Time (s)	MIAE (root box)	MIAE (PQ)	Density	Time (s)	MIAE (root box)	MIAE (PQ)
U(0,1) 1D	5.2940	0.0132	0.0115	U(0,1) 2D	4.96	0.0125	0.0123
N(0,1) 1D	0.4857	0.0663	0.0651	Rosen. 10D	2.2900	NA	NA
Bivariate Gaussian	0.6206	0.2444	0.2702	U(0,1) 10D	14.775	0.0127	0.0119
Levy 2D	4.3200	0.4187	0.3272	U(0,1) 100D	107.3963	0.0108	0.0116
Rosen. 2D	9.5672	0.3273	0.4307	U(0,1) 1000D	970.4471	0.0117	0.0108



Comparison between priority-queue driven histograms with some choice of k_n and the averaged histograms obtained from collations of sample states for various 2-dimensional distributions.

CONCLUSIONS

- SRSP is a sufficient statistical data-structure for density estimation;
- We can grow (by bisection) or prune (by reunion) the SRSP tree adaptively;
- Arithmetic can be efficiently extended to SRSPs, i.e. averaging histograms;
- Thus obtain the posterior mean over the space of multivariate histogram with partitions in $\mathcal{S}_{0,\infty}$;
- Investigations currently conducted on convergence issues of the MCMC.