Beta-trees: Multivariate histograms with confidence statements

Qian Zhao

University of Massachusetts, Amherst

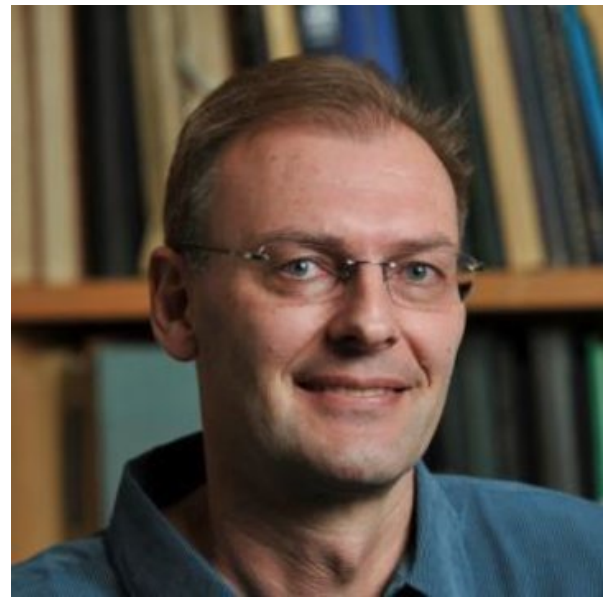
WNAR 2024

Beta-trees histogram

Multivariate histogram

+

Confidence interval of density in each region



A univariate histogram

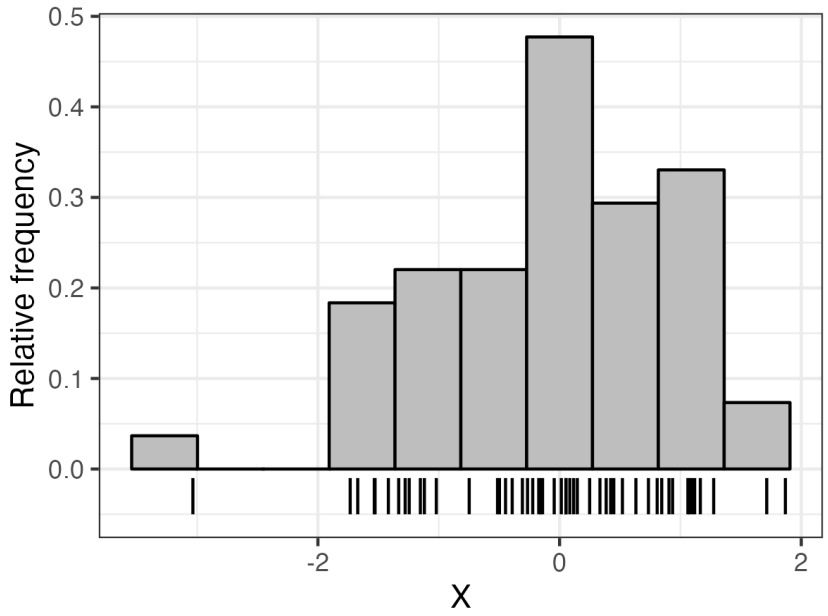


Figure: Histogram of 50 $\mathcal{N}(0, 1)$ obs.

A univariate histogram

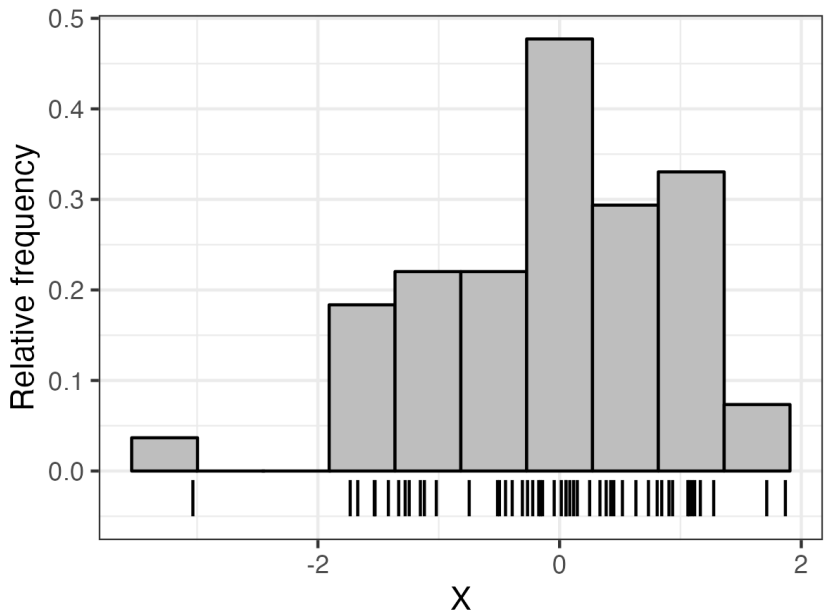


Figure: Histogram of 50 $\mathcal{N}(0, 1)$ obs.

We use histograms to

1. Summarize data
2. Visualize data

A univariate histogram

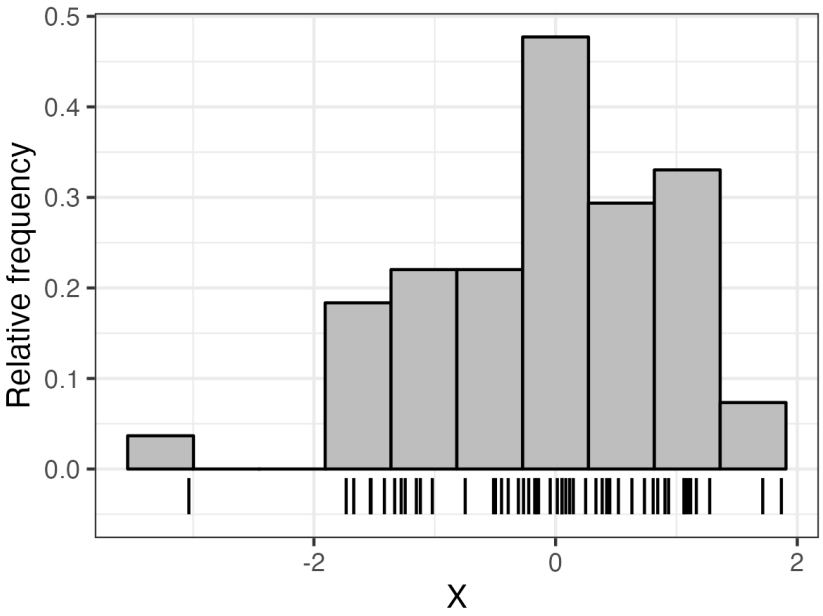


Figure: Histogram of 50 $\mathcal{N}(0, 1)$ obs.

We use histograms to

1. Summarize data
2. Visualize data
3. Estimate density

Optimal bin width to minimize Asymptotic Mean Integrated Squared Error (AMISE) is $h^\star = [6/R(f')]^{1/3} n^{-1/3}$ ($R(\cdot)$ is the ℓ_2 norm) (Freedman & Diaconis, 1981)

A multivariate histogram

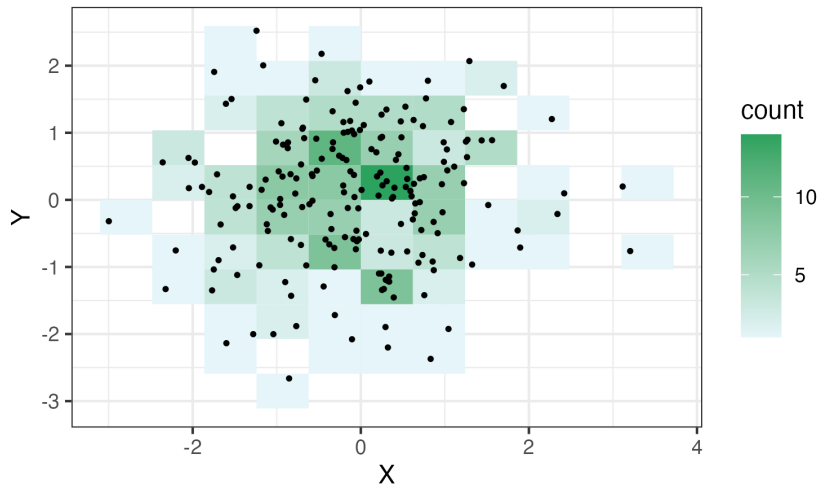


Figure: Two-dimensional histogram of 200 $\mathcal{N}(0, I)$ obs.

A multivariate histogram

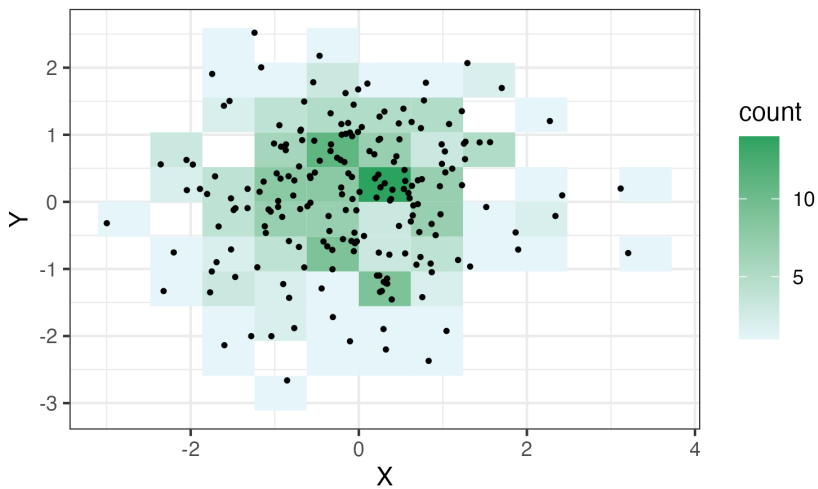


Figure: Two-dimensional histogram of 200 $\mathcal{N}(0, I)$ obs.

- If fix # bins in each dimension
  $\Longrightarrow$ #regions grows exponentially with d
- In higher dimensions, most bins would be empty ("Curse of dimensionality")

A multivariate histogram

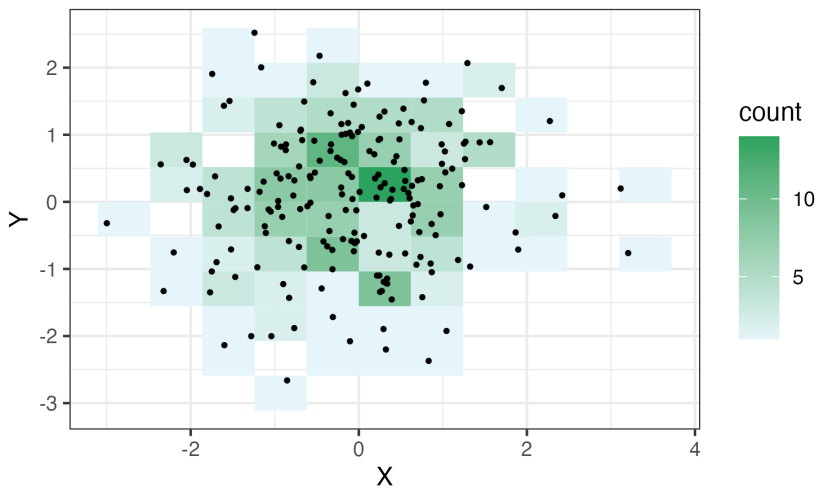


Figure: Two-dimensional histogram of 200 $\mathcal{N}(0, I)$ obs.

- If fix # bins in each dimension
  $\Longrightarrow$ #regions grows exponentially with d
- In higher dimensions, most bins would be empty ("Curse of dimensionality")

A multivariate histogram

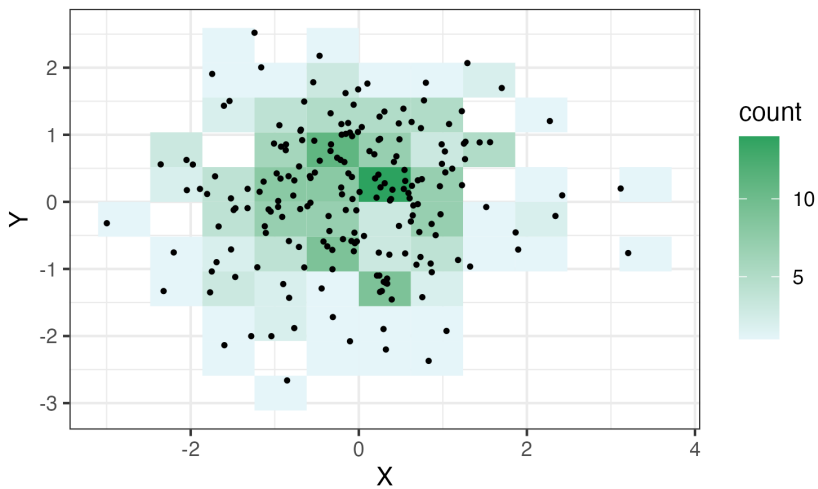
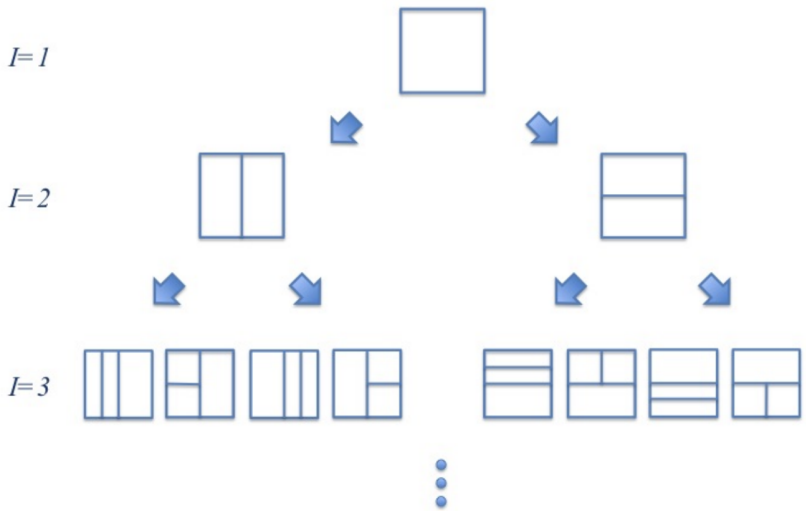


Figure: Two-dimensional histogram of 200 $\mathcal{N}(0, I)$ obs.

- If fix # bins in each dimension
  $\implies$ #regions grows exponentially with d
- In higher dimensions, most bins would be empty ("Curse of dimensionality")
- Optimal bin width is $O(n^{-1/(2+d)})$ in each dimension [1]

1. *Multivariate Density Estimation*, Scott, D. W. (2015)

Adaptive partitioning histograms



- At step I, partition each region into into d regions.
- Choose the partition that maximizes the likelihood (call the histogram "sieve MLE")
- If for some f_I supported on these partitions, $\rho(f, f_I) \leq AI^{-r}$, then the sieve MLE converges to f at rate $n^{-\frac{r}{2r+1}}(\log n)^{\frac{1}{2}+\frac{r}{2r+1}}$,

Multivariate density estimation via adaptive partitioning (i) Sieve MLE, by Liu and Wong (2014)

Roadmap

1. Some theory of order statistics
2. Constructing the beta-trees histogram
   - 2.1 Recursive partitioning
   - 2.2 Bottom-up merging
3. Application of the beta-trees histogram
   - 3.1 Data visualization
   - 3.2 Mode hunting
   - 3.3 Analyzing flow cytometry data

Univariate order statistic

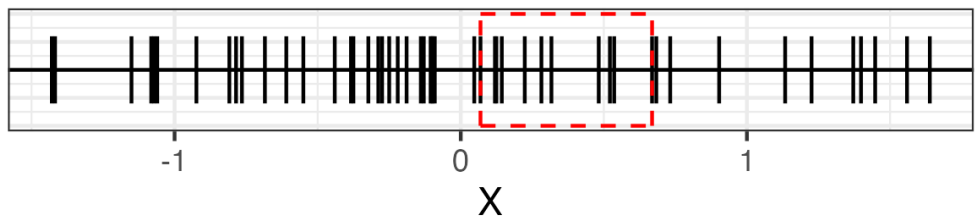
- $X_1, \ldots, X_n \overset{iid}{\sim} F$
- Order statistic $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$

$$F((X_{(i)}, X_{(j)})) \sim \text{Beta}(j - i, n + 1 - (j - i))$$

Univariate order statistic

- $X_1, \ldots, X_n \overset{iid}{\sim} F$
- Order statistic $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$

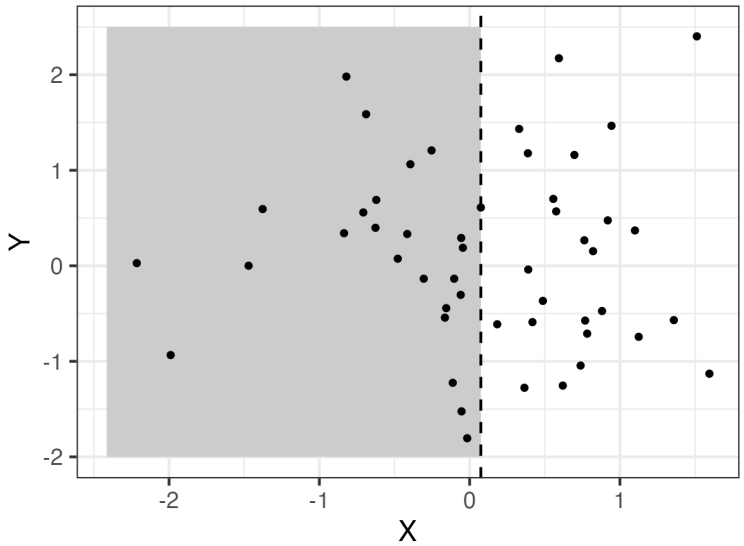
$$F((X_{(i)}, X_{(j)})) \sim \text{Beta}(j - i, n + 1 - (j - i))$$



E.g. $X_1, \ldots, X_{50} \overset{iid}{\sim} \mathcal{N}(0, 1)$, then $\Phi((X_{(30)}, X_{(40)})) \sim \text{Beta}(10, 41)$.

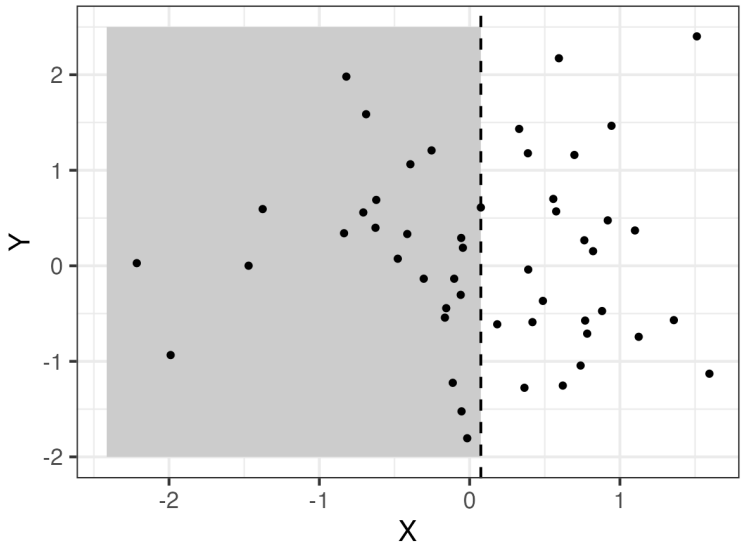
A 95% CI for $\Phi((X_{(30)}, X_{(40)}))$ is $[0.10, 0.31]$.

Multivariate scenario



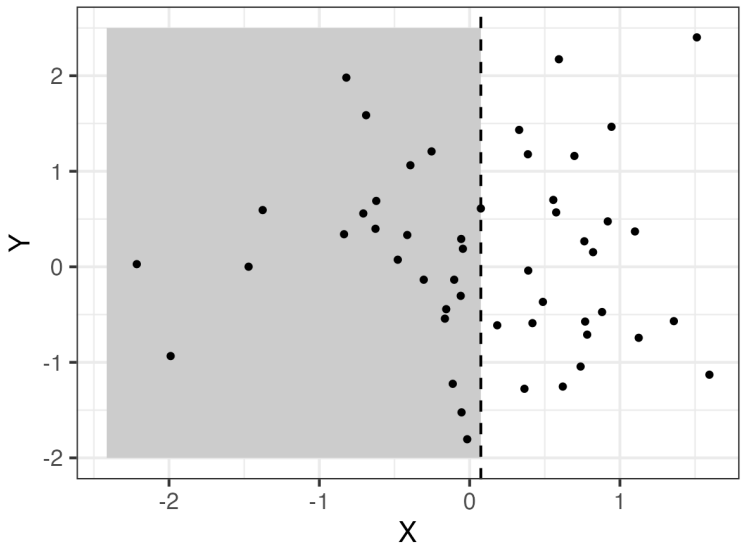
- $X_1, \ldots, X_{50} \overset{iid}{\sim} \mathcal{N}(0, I_2) := F$

Multivariate scenario



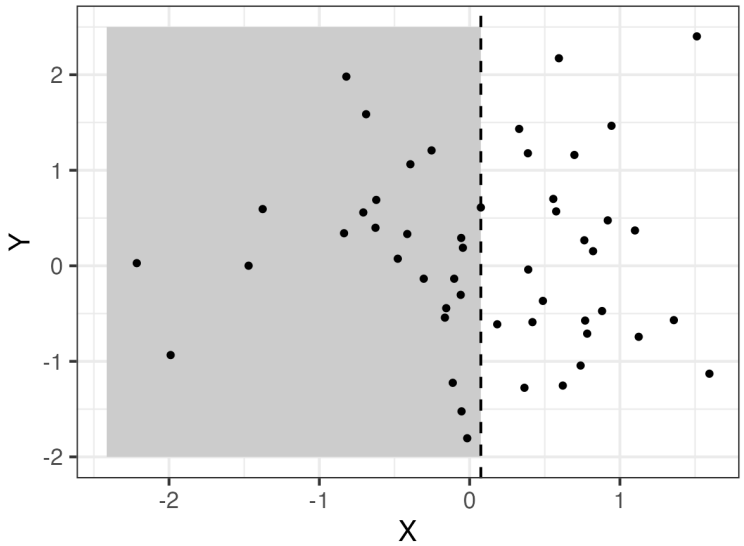
- $X_1, \ldots, X_{50} \overset{iid}{\sim} \mathcal{N}(0, I_2) := F$
- Order statistics along x-axis $X_{(1),1} \leq \ldots X_{(50),1}$

Multivariate scenario



- $X_1, \ldots, X_{50} \overset{iid}{\sim} \mathcal{N}(0, I_2) := F$
- Order statistics along x-axis $X_{(1),1} \leq \ldots X_{(50),1}$
- Let $R = \{x \in \mathbb{R}^2 : x_1 < X_{(25),1}\}$
- $F(R) \sim \text{Beta}(25, 26)$

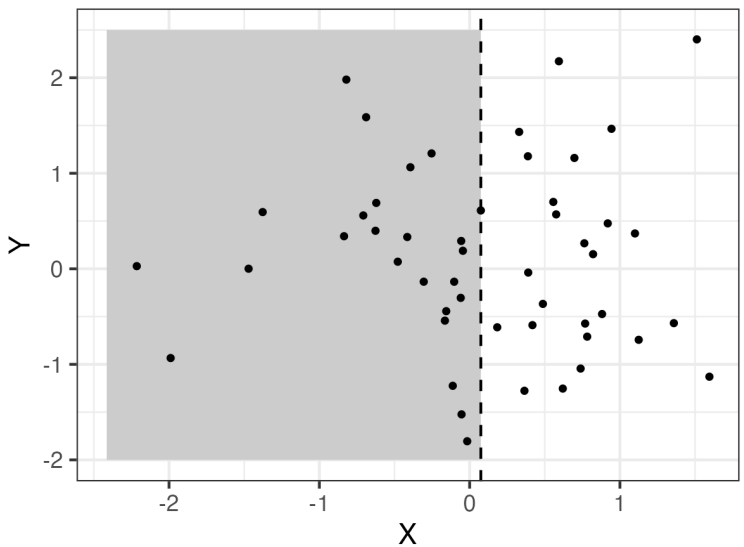
Multivariate scenario



- $X_1, \ldots, X_{50} \overset{iid}{\sim} \mathcal{N}(0, I_2) := F$
- Order statistics along x-axis $X_{(1),1} \leq \ldots X_{(50),1}$
- Let $R = \{x \in \mathbb{R}^2 : x_1 < X_{(25),1}\}$
- $F(R) \sim \text{Beta}(25, 26)$

This is our first split!

Multivariate case



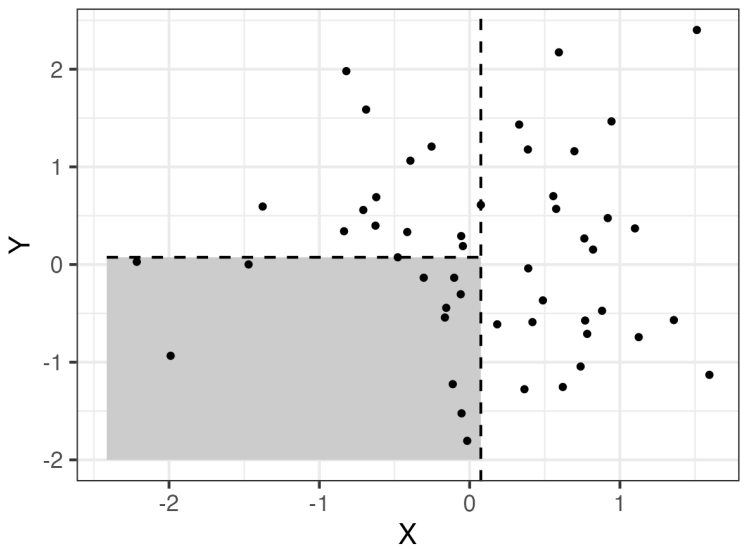
- $R = \{x \in \mathbb{R}^2 : x_1 < X_{(25),1}\}$
- Denote the obs. inside R as $\{Y_1, \ldots, Y_{24}\}$.

Lemma

Conditional on $X_{(25),1} = t$, $\{Y_1, \ldots, Y_{24}\}$ are i.i.d. from G,

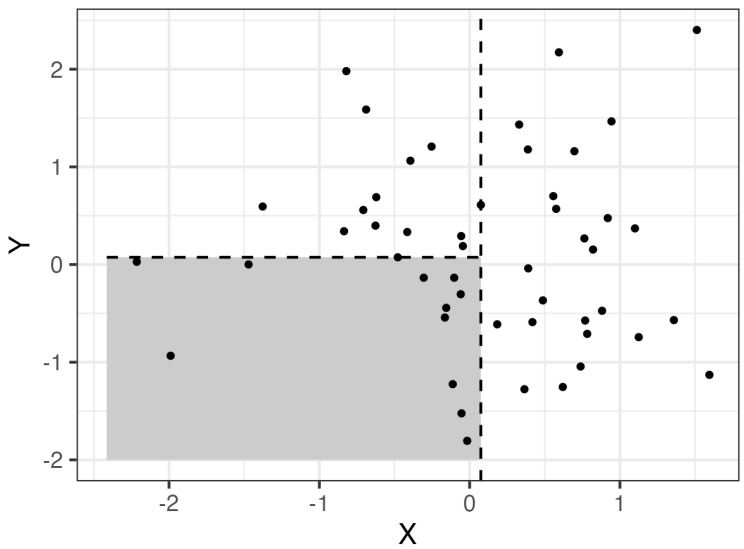
$$G(\cdot) = \frac{F(\cdot \cap R)}{F(R)}$$

Second split



- Let $S = \{x \in R : x_2 < Y_{(12),2}\}$

Second split

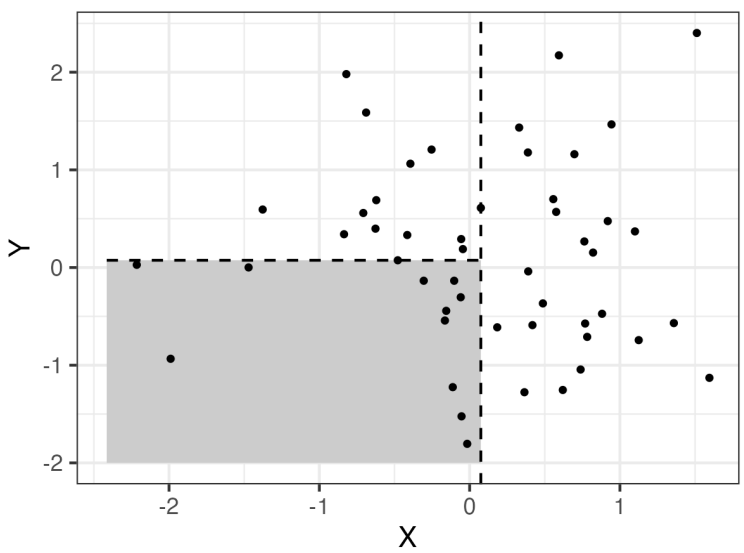


- Let $S = \{x \in R : x_2 < Y_{(12),2}\}$
- Conditional on the first split,

$$G(S) = \frac{F(S)}{F(R)} \sim \text{Beta}(12, 13)$$

and is independent of R.

Second split

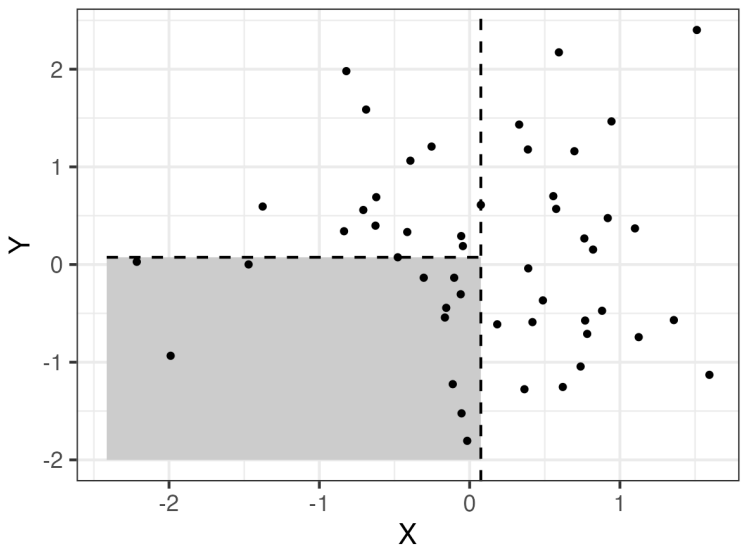


- Let $S = \{x \in R : x_2 < Y_{(12),2}\}$
- Conditional on the first split,

$$G(S) = \frac{F(S)}{F(R)} \sim \text{Beta}(12, 13)$$

  and is independent of R.
- $F(S) = G(S) \cdot F(R) \sim \text{Beta}(12, 39)$

Second split



This is our second split!

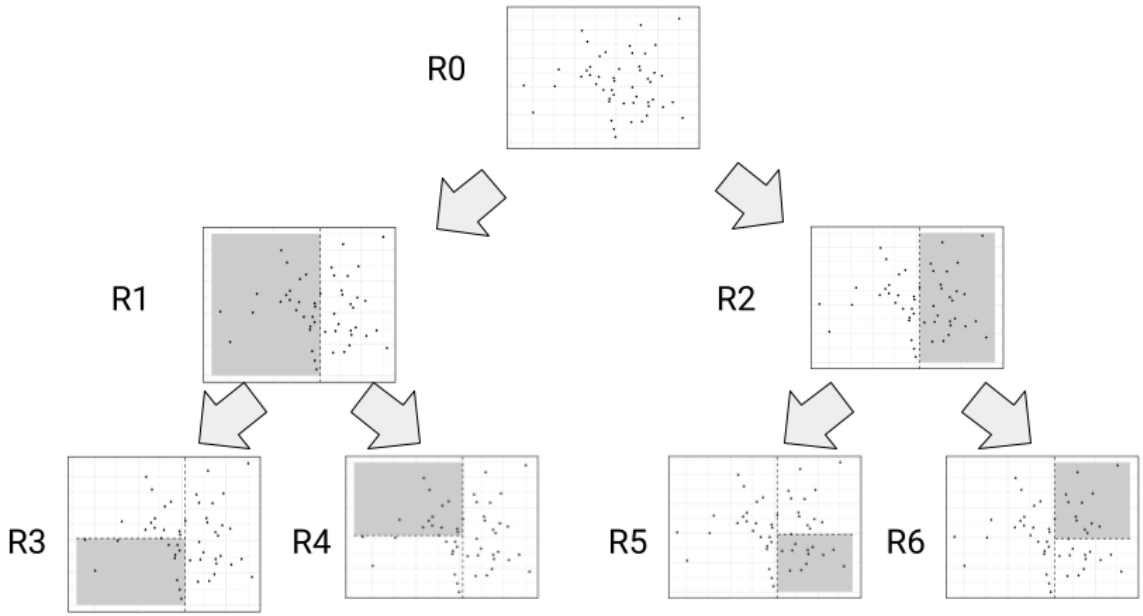
- Let $S = \{x \in R : x_2 < Y_{(12),2}\}$
- Conditional on the first split,

$$G(S) = \frac{F(S)}{F(R)} \sim \text{Beta}(12, 13)$$

and is independent of R.

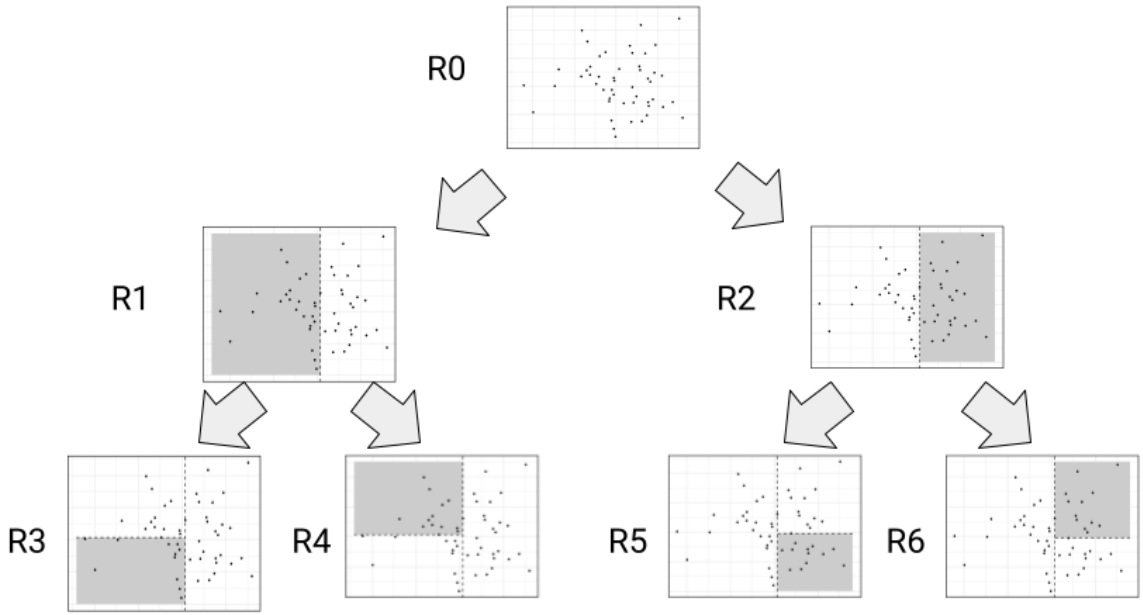
- $F(S) = G(S) \cdot F(R) \sim \text{Beta}(12, 39)$

Summary: Recursive partitioning (k-d tree)



1. Partition at the median along x-axis
2. For each region R_i,
   - 2.1 Pick a coordinate
   - 2.2 Partition along the median of obs. inside R_i.
3. Stop when # obs. inside is less than $4 \log n$.

Summary: Recursive partitioning (k-d tree)



Theorem

$F(R_i) \sim \text{Beta}(n_i + 1, n - n_i)$, n_i is the number of obs. inside R_i.

Summary: Recursive partitioning (k-d tree)

Theorem

$F(R_i) \sim \text{Beta}(n_i + 1, n - n_i)$, n_i is the number of obs. inside R_i.

Confidence intervals

Theorem

$F(R_i) \sim \text{Beta}(n_i + 1, n - n_i)$, n_i is the number of obs. inside R_i.

- We can compute $(1 - q_i)$ confidence interval for $F(R_i)$.

Confidence intervals

Theorem

$F(R_i) \sim \text{Beta}(n_i + 1, n - n_i)$, n_i is the number of obs. inside R_i.

- We can compute $(1 - q_i)$ confidence interval for $F(R_i)$.
- If $\sum_i q_i = \alpha$, then the CIs cover all regions **simultaneously** at level $(1 - \alpha)$.

Confidence intervals

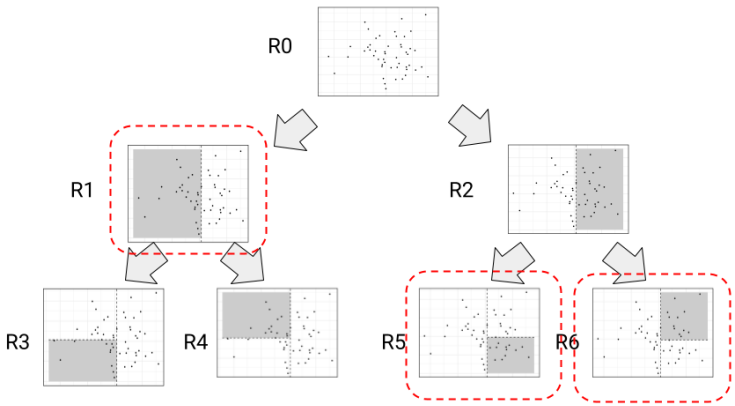
Theorem

$F(R_i) \sim \text{Beta}(n_i + 1, n - n_i)$, n_i is the number of obs. inside R_i.

- We can compute $(1 - q_i)$ confidence interval for $F(R_i)$.
- If $\sum_i q_i = \alpha$, then the CIs cover all regions **simultaneously** at level $(1 - \alpha)$.

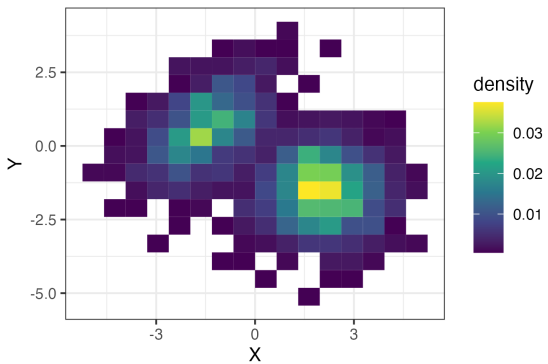
Which histogram should we choose?

Merging regions (bottom-up)

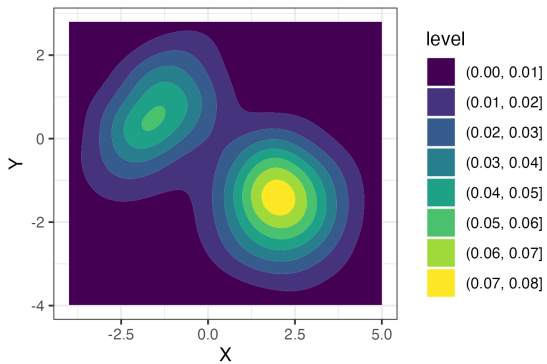


If $f(R_i)$ lies in the confidence intervals of both of its children, then pick R_i.

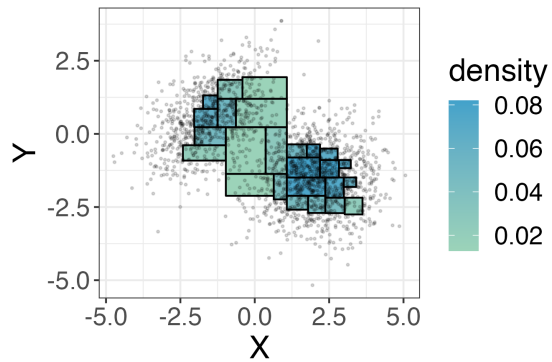
Data visualization: 2-d Gaussian mixture



(a) Fixed bin width (15 bins in each dimension)

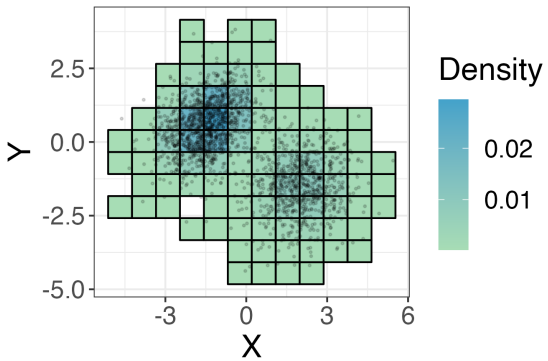


(b) Kernel density estimate (select bandwidth by cross validation)

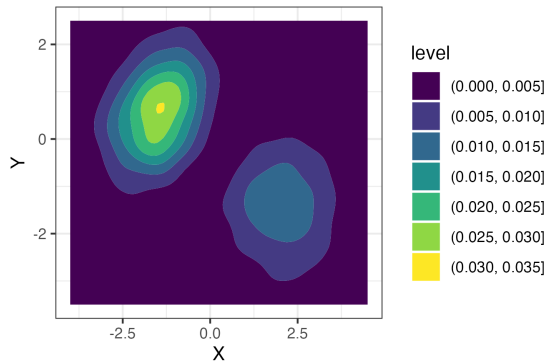


(c) Beta-trees histogram

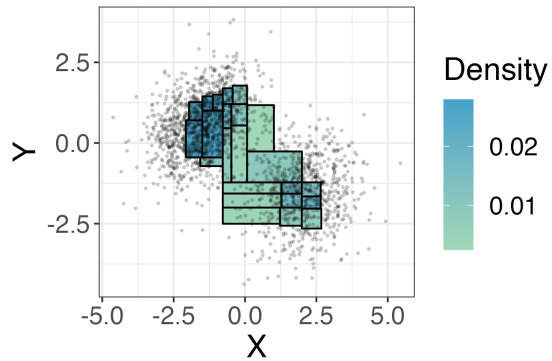
Data visualization: 3-d Gaussian mixture



(d) Fixed bin width (15 bins in each dimension; 834 non-empty regions)



(e) Kernel density estimate (select bandwidth by plug-in estimator)

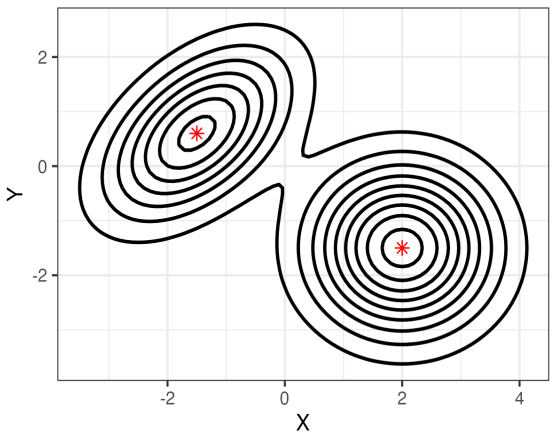


(f) Beta-trees histogram (only 125 regions)

We plot the density along $z = 1$ and obs. inside a slab of $0.8 \leq z \leq 1.2$.

Mode hunting

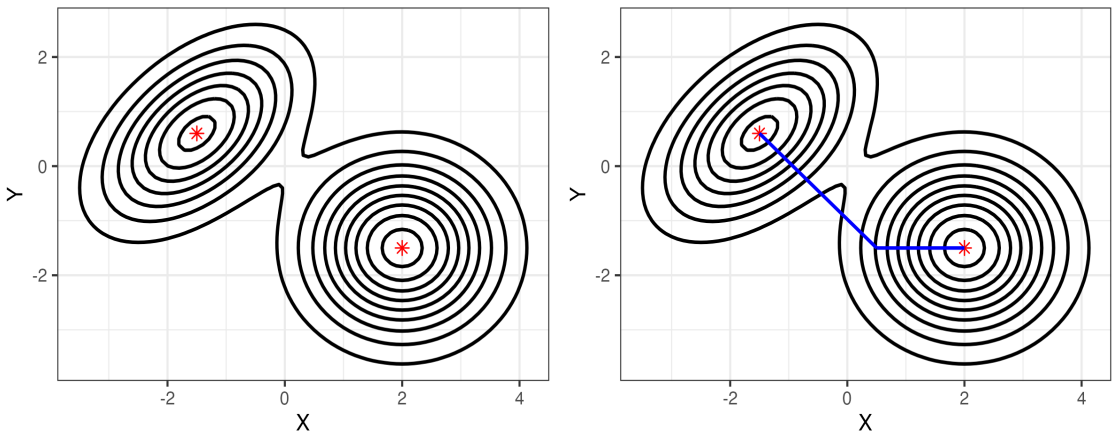
A mode is where density is a local maximum.



(g) Density contours

Mode hunting

A mode is where density is a local maximum.

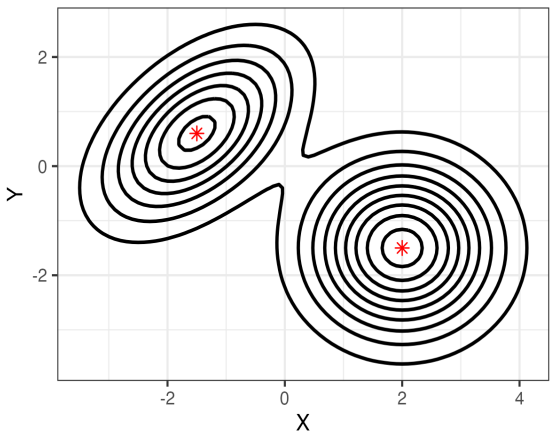


(j) Density contours

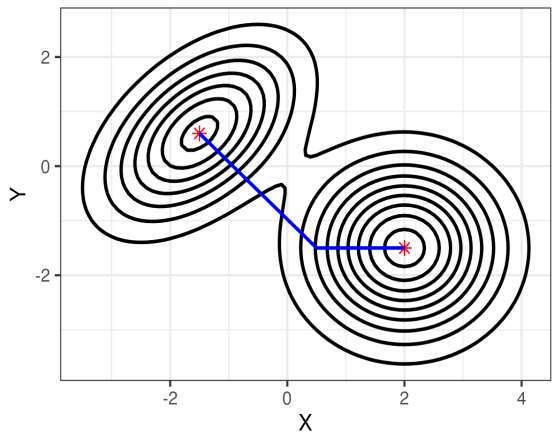
(k) A path between the two modes

Mode hunting

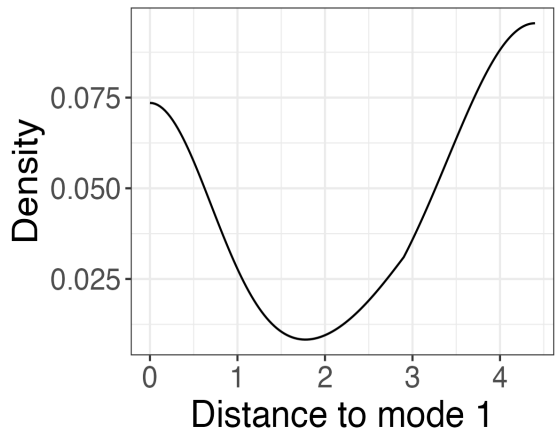
A mode is where density is a local maximum.



(m) Density contours

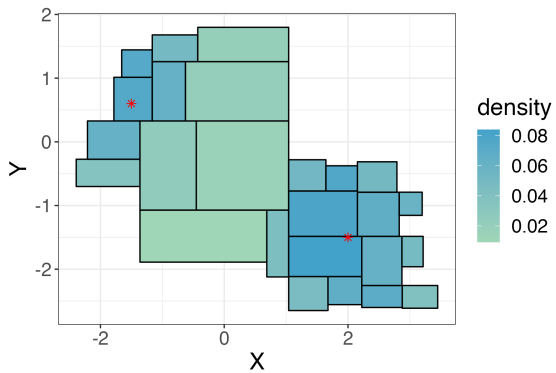


(n) A path between the two modes



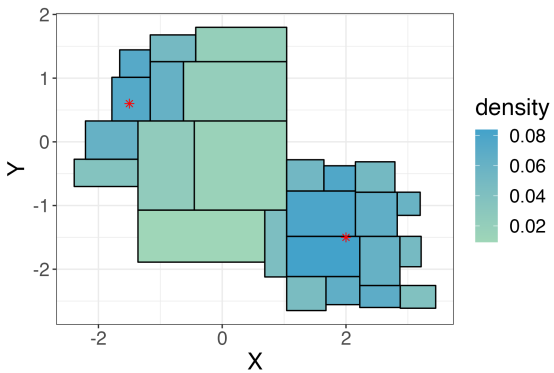
(o) Density along the path

Mode hunting using Beta-trees histogram

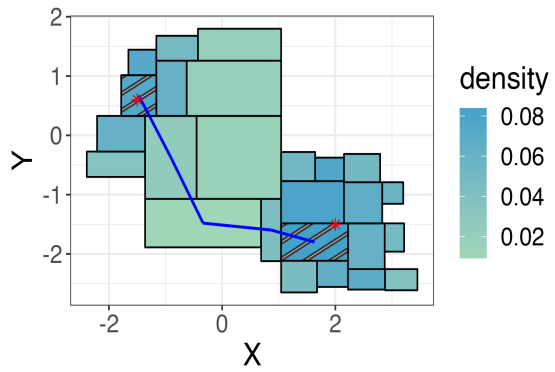


(p) Beta-trees histogram
($n = 2,000$, $\alpha = 0.1$)

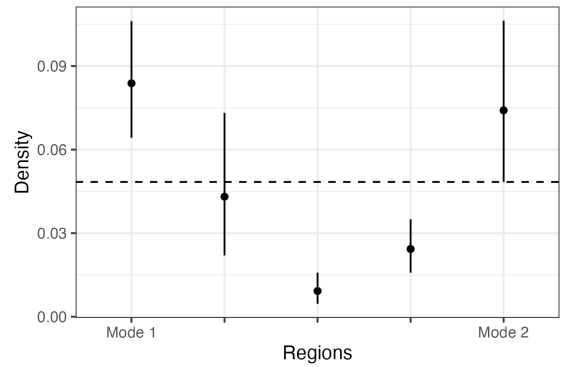
Mode hunting using Beta-trees histogram



(s) Beta-trees histogram ($n = 2,000$, $\alpha = 0.1$)



(t) A path between the two identified modes



(u) Confidence interval of the density along the path

If A and B are two distinct modes, then there should exist a point along any path connecting A and B whose density is lower than **both** A and B.

Application: Flow cytometry

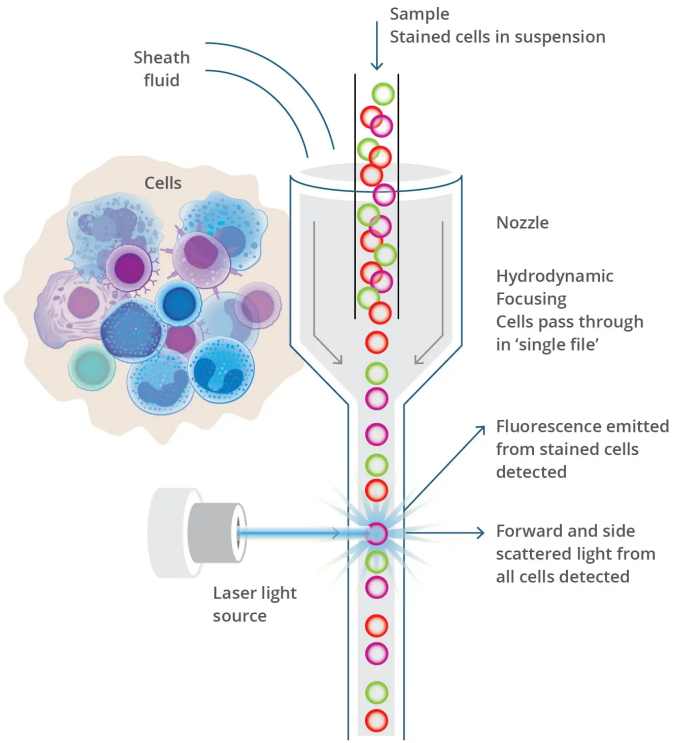


Figure: Illustration of a flow cytometer

Application: Flow cytometry

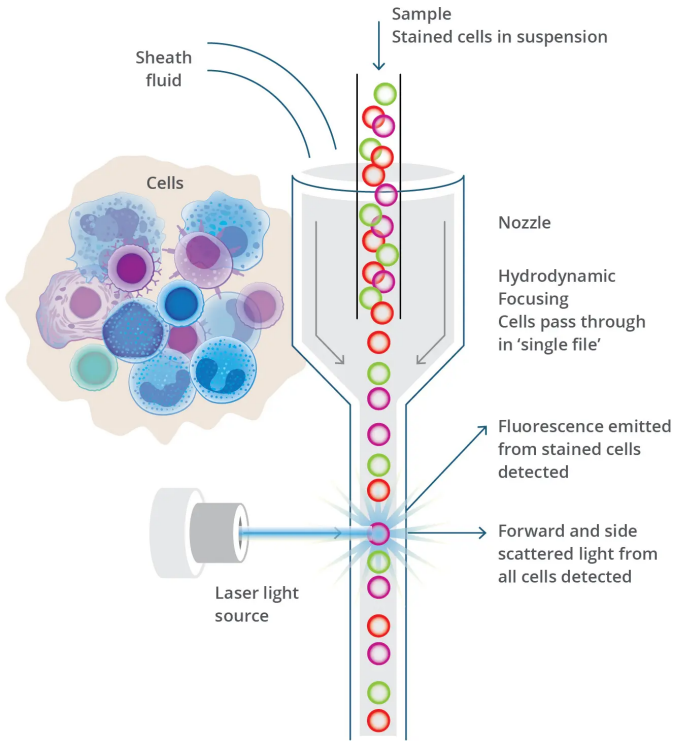
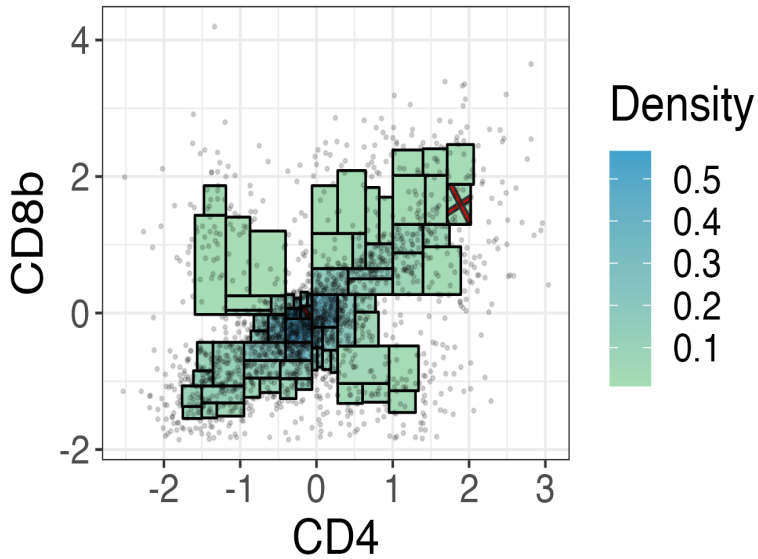


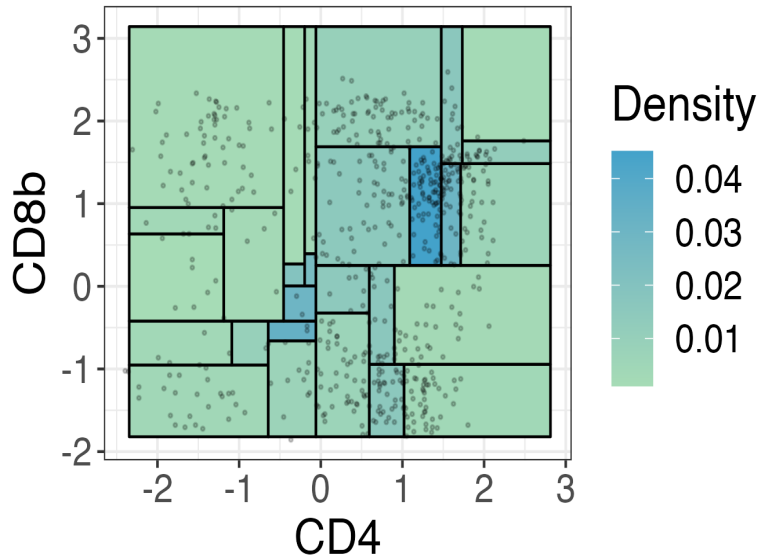
Figure: Illustration of a flow cytometer

- Data set `RvHD` from the R package `mclust` (Brinkman et al. 2007).
- Goal: Identify biomarkers associated with graft-versus-host disease (GvHD).
- 4 biomarkers.
- 9083 obs from a case patient and 6809 obs from a control patient.

Visualizing a flow cytometry data set



(a) A two-dimensional histogram. Two identified modes are indicated by red stripes.



(b) A slice of a 3-dim histogram of CD4, CD8b, CD3 at the slice CD3 = 1.

Researchers identified the CD3+ CD4+ CD8b+ population to be associated with GvHD (Brinkman et al. 2007).

Conclusion

- Beta-trees histogram
  - Automatically adapt to location of the obs.
  - $F(R_i) \sim \text{Beta}(n_i + 1, n - n_i) \implies$ simultaneous CI for every region.
  - Compact representation of data (merge regions based on a goodness-of-fit test)

Conclusion

- Beta-trees histogram
  - Automatically adapt to location of the obs.
  - $F(R_i) \sim \text{Beta}(n_i + 1, n - n_i) \implies$ simultaneous CI for every region.
  - Compact representation of data (merge regions based on a goodness-of-fit test)
- Using beta-trees histogram to identify modes in the distribution

Conclusion

- Beta-trees histogram
  - Automatically adapt to location of the obs.
  - $F(R_i) \sim \text{Beta}(n_i + 1, n - n_i) \implies$ simultaneous CI for every region.
  - Compact representation of data (merge regions based on a goodness-of-fit test)
- Using beta-trees histogram to identify modes in the distribution
- Future directions
  - Still cannot handle high-dimensions (each split reduces sample size by half) $\implies$ can we leverage information about the distribution?
  - Can we use Beta-trees histogram to identify *changes* in the distribution?

Thank you! Questions?

https://arxiv.org/abs/2308.00950

Mode hunting using the Beta-trees histogram (algorithm)

1. Order regions in decreasing order of density. Assign the region with highest density as a mode.
2. Iterate through every region R
   - 2.1 Iterate through M in the current list of modes
     - 2.1.1 Iterate through every path connecting R and M
     - 2.1.2 Is there a region whose CI intersects that of either R and M?
       If "yes", R is **not** a mode; continue to next region
       If "no", continue to check the next path.
   - 2.2 Add R to the list of modes.
3. Report the list of modes.