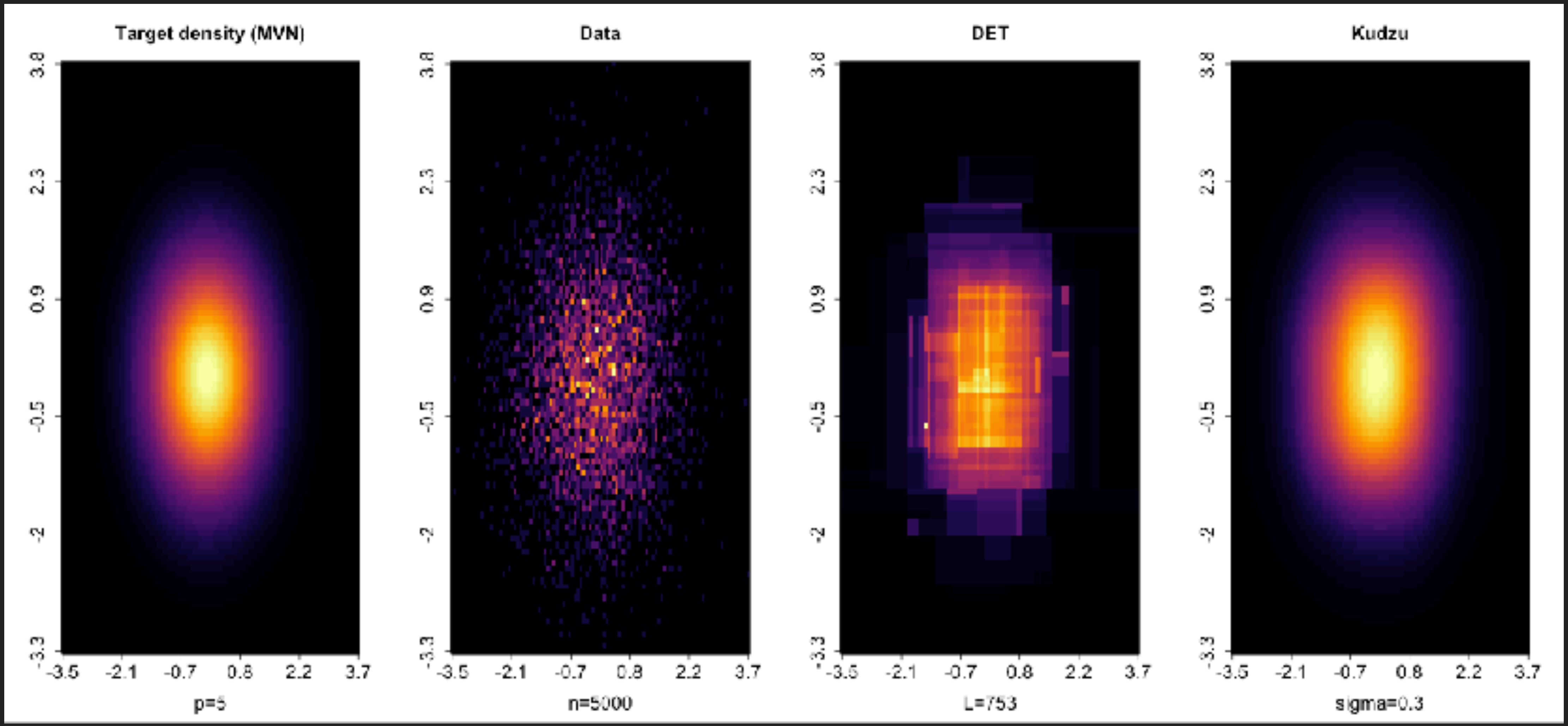


ROBERT GRANT — BAYESCAMP & KINGSTON UNIVERSITY, UK

SCALABLE HIGH-DIMENSIONAL NON-PARAMETRIC DENSITY ESTIMATION, WITH BAYESIAN APPLICATIONS

OVERVIEW

- Our goal: turn a dataset into a probability density function
- The PDF should be smooth
- The method should work for many variables / dimensions
- We use density estimation trees and smooth them
- We call this a kudzu density function
- This performs quite well

- Thanks: Gordon Hunter, Lucio Morettini

DENSITY ESTIMATION

- Typically, either parametric or not extensible to several dimensions (*p*)
- Non-parametric estimation:
  - kernels distort massively as $p>4$
  - newer methods are better, but computationally intensive (i.e. bad for the planet), and forbidding to most users

MOTIVATION

- Density estimation is useful for probabilistic prediction, visualisation, simulation, etc.
- My motivation was Bayesian updating:
  - Model fitted on data X_1, giving a posterior sample from $P(\theta \mid X_1)$.
  - Now data X_2 arrives. Re-analysis of (X_1, X_2) will take too long.
  - Use X_1's posterior $P(\theta \mid X_1)$ as the prior, and compute likelihood only on X_2. No guarantee of a convex, unimodal p[oste]rior, so we need a non-parametric method. *p* could be large.

CORE IDEA

- Fit a density estimation tree (DET; Ram & Gray 2011) and smooth it.
- This produces a kudzu density function, named after a vine (a.k.a. Japanese arrowroot) which grows rapidly over trees, smoothing out their shape.

DENSITY ESTIMATION TREES

- Proposed by Ram & Gray (2011) with little subsequent uptake.
- CART algorithm, but using integrated squared error (ISE), which controls over-fitting to some extent (and smoothing helps too).
- Trees scale well to high *n* and high *p* (compute time and accuracy).
- Terminal nodes of the tree are *L* "leaves". The tree is defined by two *L*-by-*p* matrices (top and bottom edges) and a *L*-vector of densities.
- ISE collapses to an extremely simple formula for DETs.

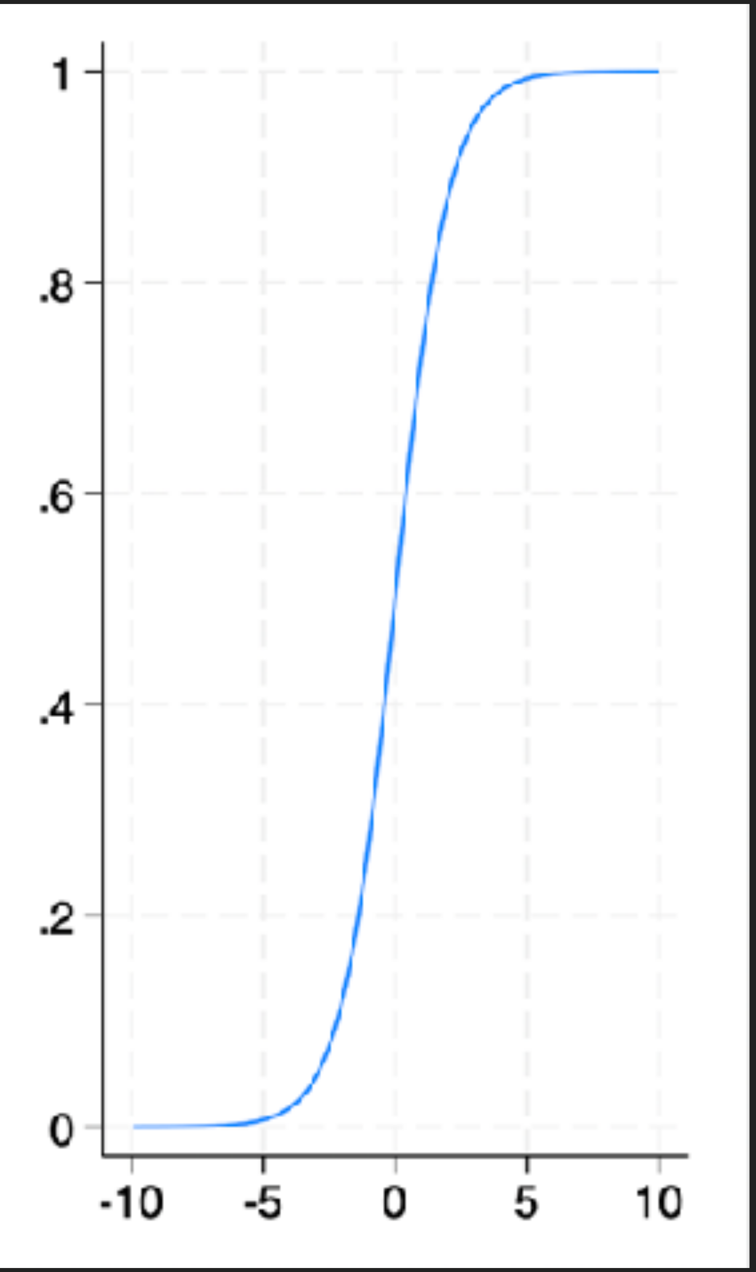
😱 (Technical aside: The volume of the leaf is important in DET, so the p variables/unknowns must jointly define a metric space: no ordinal or nominal variables, though integers are fine.)

KUDZU DENSITY FUNCTION (1)

$$\frac{h}{1+e^{-\frac{x-\mu}{\sigma}}}$$

- We replace each edge of each leaf with a smooth ramp (monotonic, two horizontal asymptotes).
- They are centred on the edges, and have bandwidth σ. Inverse logistic function is in the class of computationally minimal smooth ramps (one power series).

😱 (Technical aside: you can also conceive of it as a convolution of each leaf-dimension with the logistic PDF.)



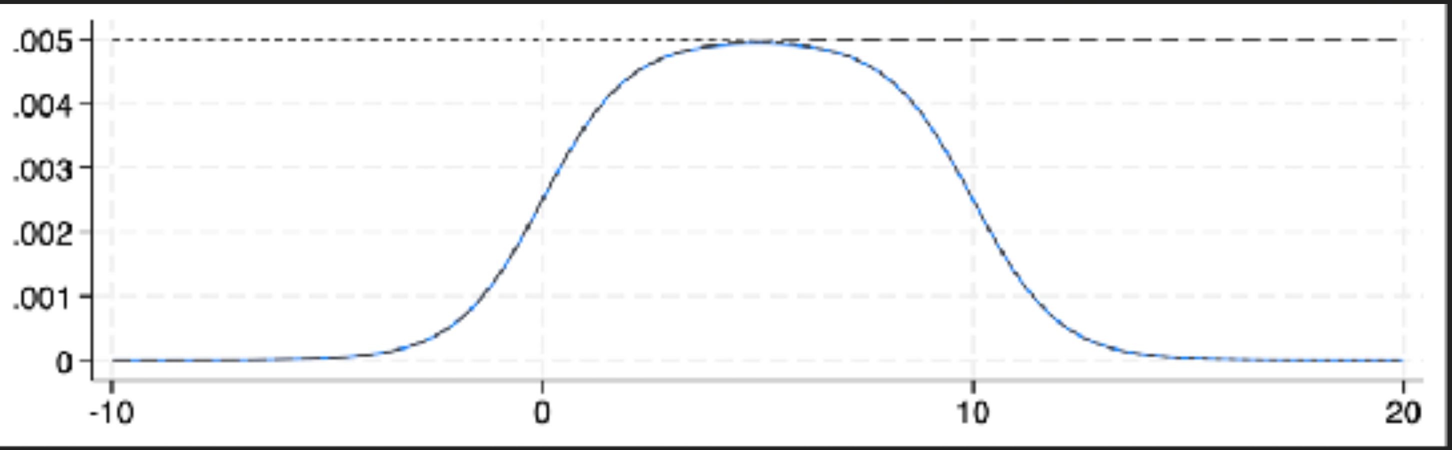
KUDZU DENSITY FUNCTION (2)

- Each dimension of each leaf is the product of the top and bottom ramps, and the predicted DET density

$$\hat{f}_{\text{kudzu}}(x_j|\ell) \propto \hat{f}_{\text{DET}}(x_j|\ell)\left(\frac{1}{1+e^{\frac{\mu_{b\ell j}-x_j}{\sigma}}}\right)\left(\frac{1}{1+e^{\frac{x_j-\mu_{t\ell j}}{\sigma}}}\right)$$

- Each leaf is the product of these *p* dimensions (which are orthogonal and independent)

$$\hat{f}_{\text{kudzu}}(\boldsymbol{x}|\ell) \propto \hat{f}_{\text{DET}}(\boldsymbol{x}|\ell)\prod_{j=1}^{p}\left(\frac{1}{1+e^{\frac{\mu_{b\ell j}-x_j}{\sigma}}}\right)\left(\frac{1}{1+e^{\frac{x_j-\mu_{t\ell j}}{\sigma}}}\right)$$



KUDZU DENSITY FUNCTION (3)

- The whole tree is the sum of the leaves

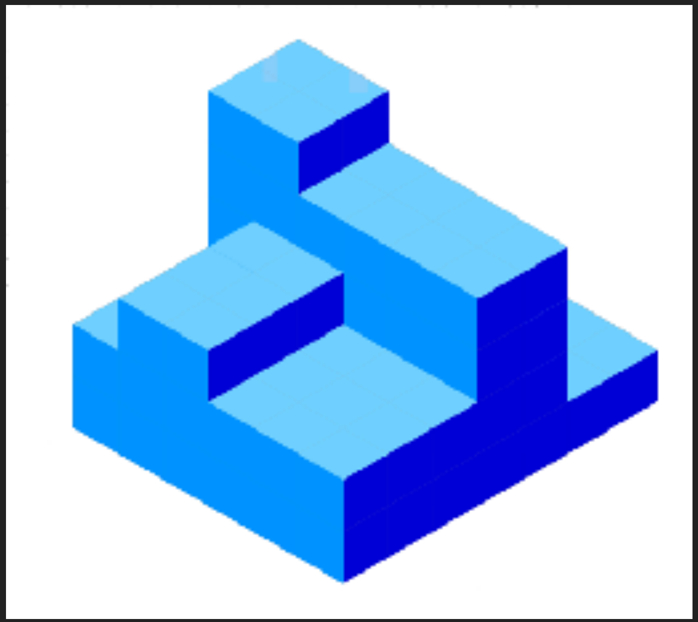
$$\hat{f}_{\text{kudzu}}(\boldsymbol{x}) \propto \sum_{\ell=1}^{L} \hat{f}_{\text{DET}}(\boldsymbol{x}|\ell) \prod_{j=1}^{p} \left(\frac{1}{1+e^{\frac{\mu_{b\ell j}-x_j}{\sigma}}} \right) \left(\frac{1}{1+e^{\frac{x_j-\mu_{t\ell j}}{\sigma}}} \right)$$

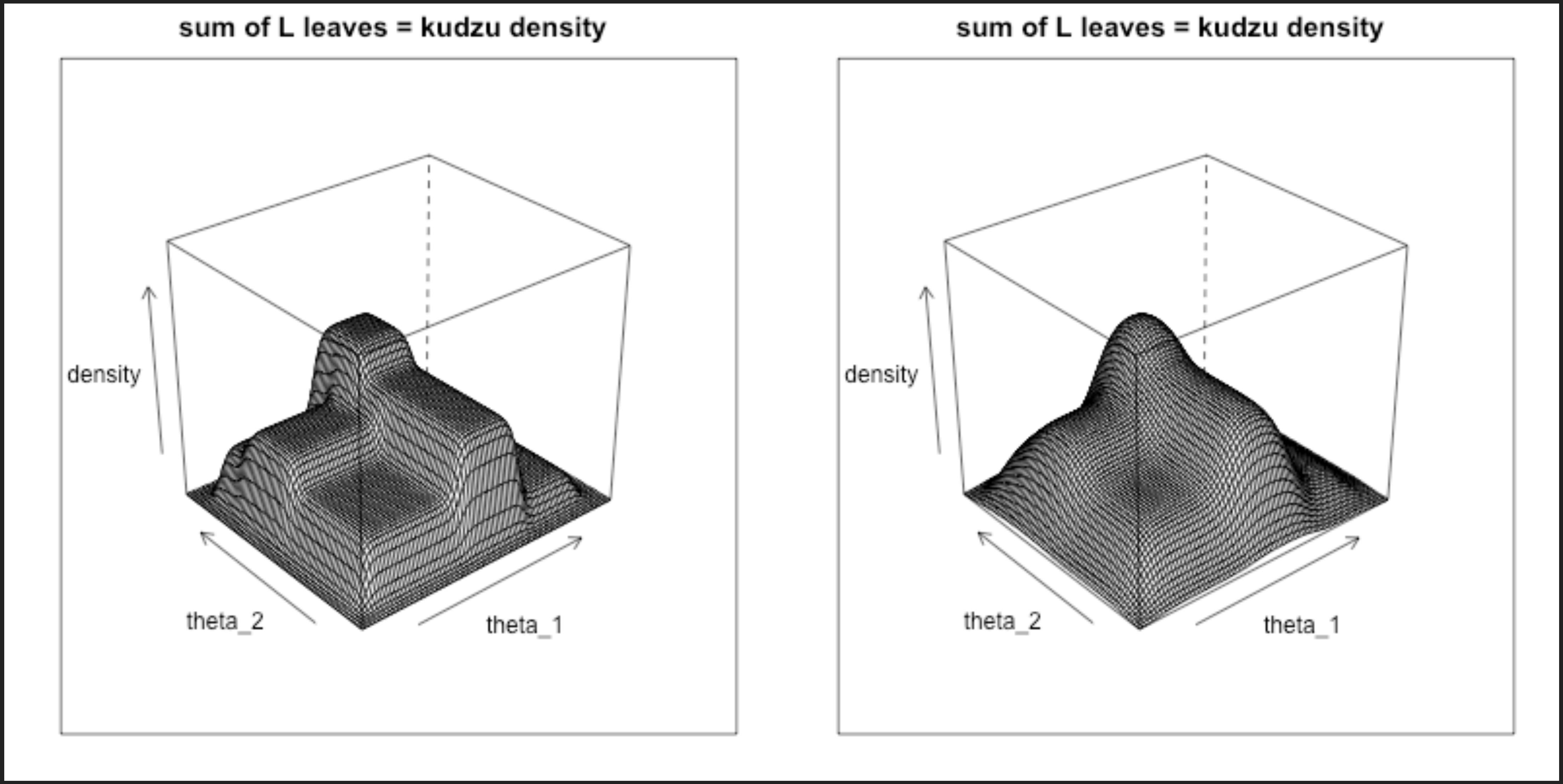
- But it might not integrate to one, so can optionally be normalised by dividing by the definite integral out to $\pm\phi\sigma$ (beyond which it is negligible).

$$\hat{f}_{\text{kudzu}}(\boldsymbol{x}) = \frac{\sum_{\ell=1}^{L} \hat{f}_{\text{DET}}(\boldsymbol{x}|\ell) \prod_{j=1}^{p} \left(\frac{1}{1+e^{\frac{\mu_{b\ell j}-x_j}{\sigma}}} \right) \left(\frac{1}{1+e^{\frac{x_j-\mu_{t\ell j}}{\sigma}}} \right)}{\sum_{\ell=1}^{L} \hat{f}_{\text{DET}}(\boldsymbol{x}|\ell) \prod_{j=1}^{p} \sigma \frac{e^{u_{\ell j}}}{e^{u_{\ell j}}-1} \ln\left(\frac{e^{u_{\ell j}+\phi}+e^{-(u_{\ell j}+\phi)}+2}{e^{\phi}+e^{-\phi}+2} \right)}$$

- When we integrate, we store all the leaf integrals and leaf-dimension integrals.

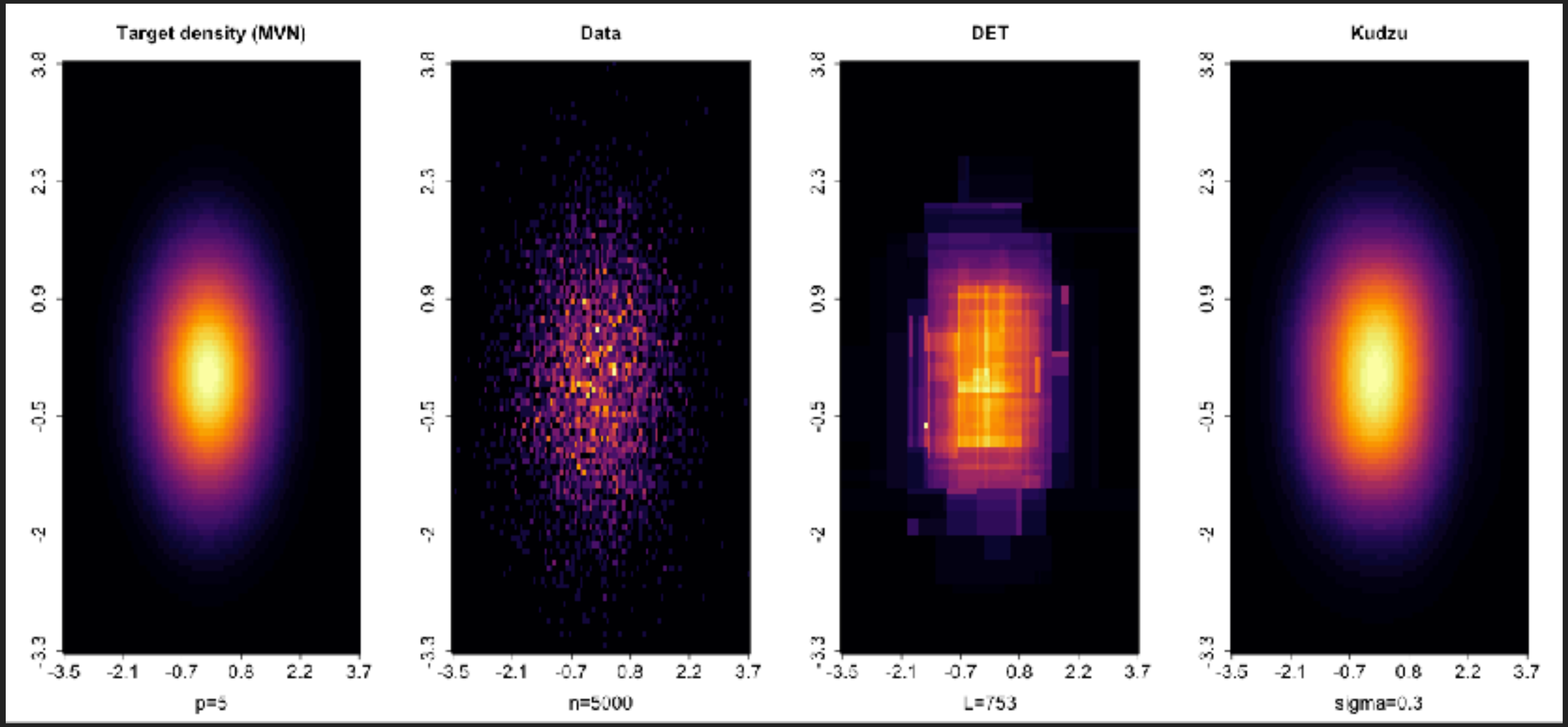
KUDZU DENSITY FUNCTION (4)





PERFORMANCE

- DET fit time is $O(L)$ and $O(p)$. Density evaluation is very fast. We only evaluate neighbouring leaves and use stored integral components for marginalisation.

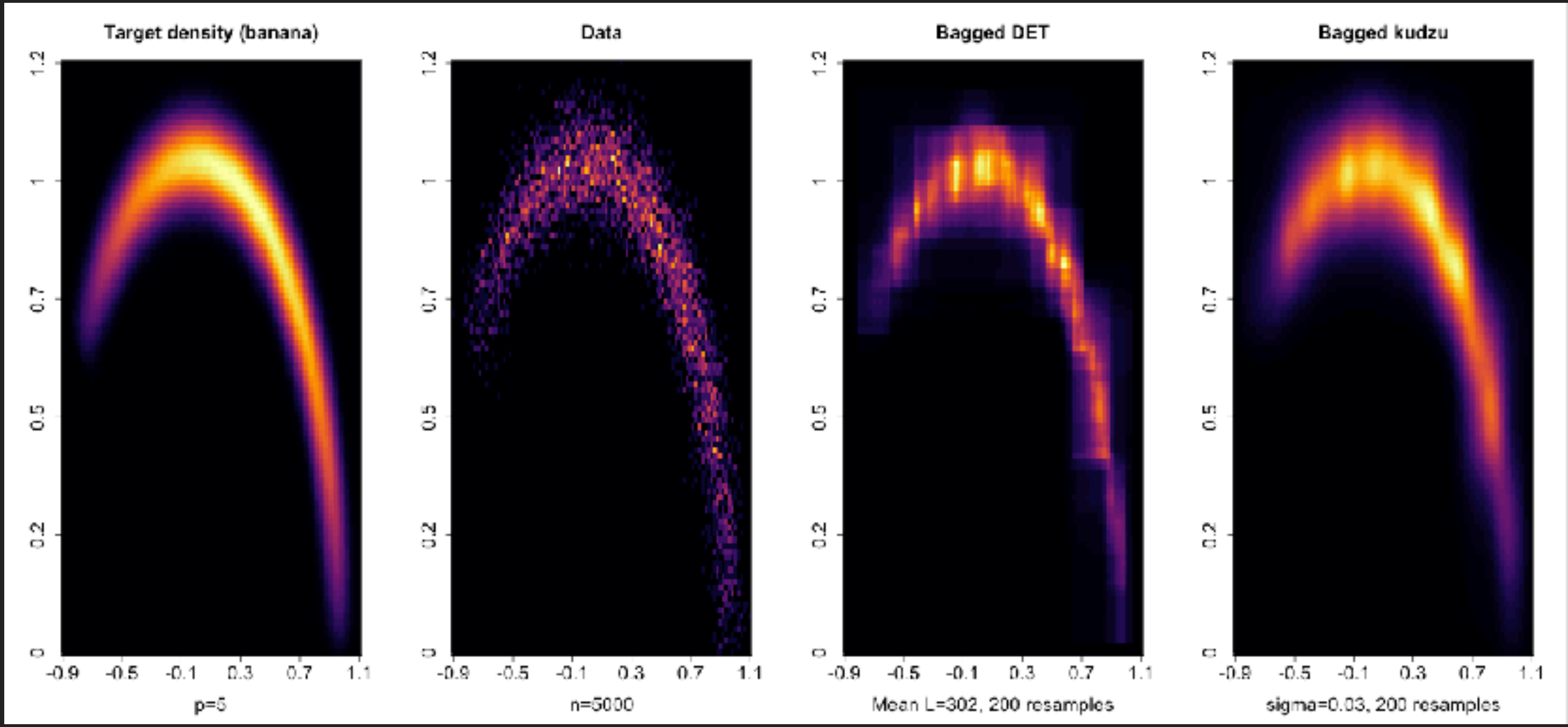


😱 NOT AS MANY DIMENSIONS AS YOU THOUGHT

- *p* dimensions will, in practice, be broken into:
  - those that are uncorrelated with anything else (just use univariate density)
  - mutually correlated blocks
    - blocks that can be linearly transformed to uncorrelated, convex distributions can be dealt with as univariate
    - multimodal distributions can be dealt with mode by mode
    - but those that are not convex require kudzu

ENSEMBLES

- Trees struggle with shapes that cannot line up with the axes.
- Ensembles of kudzu density functions are promising, and we have implemented bagging.



STATA / MATA IMPLEMENTATION

- Sharing for alpha testing by September: Mata functions and Stata .ado
- Tree command `kudzu_det` for density (ISE).

😱 (Technical aside: tree functions are extendable to classification and regression in future, or any loss function that is made out of sums of x^i.)

- Structs for DET and kudzu, return in r(), command to save to / load from .dta.
- Density evaluation at *p*-vector. Missing elements mean to marginalise that dimension. `kudzu_density, kudzufile(`*`filename`*`) at(`*`numlist`*`)`
- Very fast pseudo-RNG from kudzu density by reflection around edge. `rkudzu, kudzufile(`*`filename`*`) n(`*`#`*`)`
- Export BUGS/JAGS, Stan and bayesmh evaluator code for p[oste]rior.

POTENTIAL FUTURE WORK

- auto-setting σ
  - including fast approximations to kudzu ISE
- more ensembles

FIND OUT MORE

- Thank you for listening
- References:

  - P Ram & A Gray (2011). "Density estimation trees", KDD '11: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. pp. 627–635.
  - CART: see Breiman et al book (1984)
  - DW Scott (2015). "Multivariate density estimation: theory, practice, and visualization." Wiley.

- **robert@bayescamp.com**
- (By the way, I'm job hunting for 2025.)