# Combining extreme value theory and machine learning for novelty detection

Luca Steyn

# INTRODUCTION

- Two topics:
  - ➢Extreme value theory
  - ➢Novelty detection
- A new idea for multivariate extreme value theory and multivariate anomaly detection
- Brings together research from Statistics and Computer Science

# What is novelty detection?

- Novelty detection is the process of identifying when new observations differ from what is expected as normal behaviour.
- Classification problem, *i.e.* normal or anomalous (positive or negative).
- Conventional classification algorithms fail to detect novel observations.
- Use a one-class classification approach $\Rightarrow$ threshold a distribution representing the normal state of the system. (Is this a bad thing?)
- Assumption: Novel observations are scarce and differ to some extent from the observations in the normal class.

# Methods to perform novelty detection

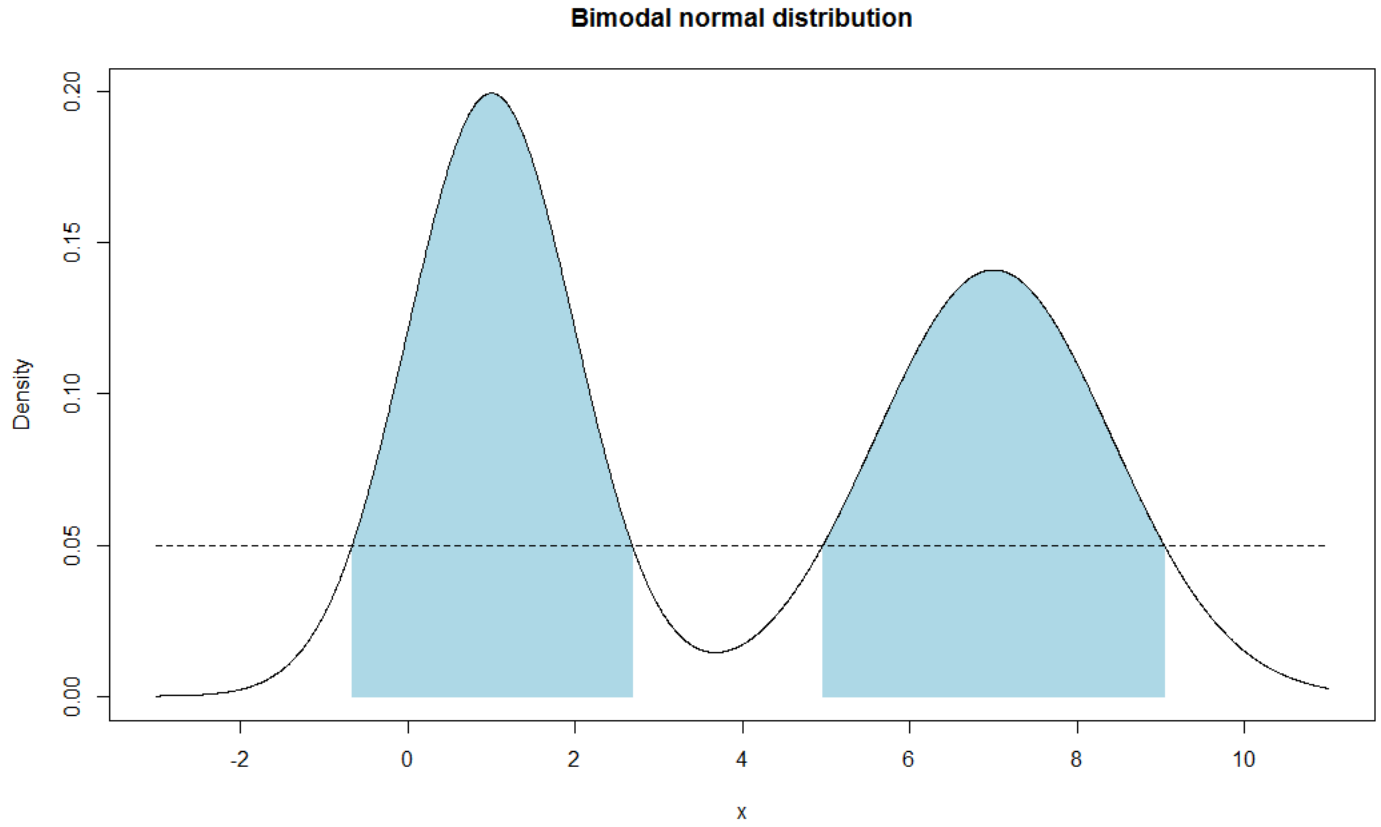
Many algorithms for novelty detection have been proposed. Broad approaches are:

- A distance-based approach
  - Modified KNN algorithm
- A domain-based approach
  - One-class support vector machines
- A reconstruction-based approach
  - Neural networks or PCA
- A probabilistic approach
  - Density estimation and thresholding

# A probabilistic approach

- Let $\underline{X} \in \mathbb{R}^p$ and denote the probability density function (pdf) by $f\left(\underline{x}\right) = \frac{d}{dx} F\left(\underline{x}\right)$.
- Choose a threshold $t$ such that $F_S\left(t\right) = \int_{\underline{x}: f\left(\underline{x}\right) \geq t} f\left(\underline{x}\right) d\underline{x}$ is large, *i.e.* $F_S\left(t\right) = 0.9$.
- Then, a new observation $\underline{x}^*$ is novel if $f\left(\underline{x}^*\right) < t$.

# A probabilistic approach



# A probabilistic approach

- If a new observation is below the threshold, how much certainty do we have that this observation is anomalous?
- Extreme value theory estimates a probability that an observation is anomalous.

# Extreme value theory: Fisher-Tippett theorem

- Let $\{X_1, X_2, X_3, \ldots\}$ be a sequence of independent and identically distributed (iid) random variables and let $M_n = \max\{X_i\}_{i=1}^{n}$. If sequences of constants $\{a_n\} > 0$ and $\{b_n\}$ exist such that $a_n^{-1}(M_n - b_n) \to G(x), n \to \infty$ , then $G(x)$ is necessarily the Generalized Extreme Value (GEV) distribution.

# Extreme value theory: Fisher-Tippett theorem

- The GEV distribution is given by

$$G_\gamma(x) = \begin{cases} \exp\left\{-\left(1+\gamma x\right)^{-1/\gamma}\right\}, & \gamma \neq 0,\ \left(1+\gamma x\right) > 0 \\ \exp\left\{-\exp\left\{-x\right\}\right\}, & \gamma = 0,\ x \in \mathbb{R} \end{cases}$$

- Move from a non-parametric to a parametric setting (in the limit).
- Three *types* of GEV distributions: Frechét-Pareto, Gumbel, (extremal) Weibull.
- Note: $\min\left(X\right) = -\max\left(-X\right)$.

# Extreme value theory: Pickands-Balkema-de Haan theorem

- The distribution $F$ is in the domain of attraction of the GEV distribution if and only if for some auxiliary function $b(\cdot)$ and for all $1+\gamma x > 0$ ,

$$\frac{1-F\left(y+b(y)x\right)}{1-F(y)} \to (1+\gamma x)^{-1/\gamma} \text{ as } y \to \infty$$

Furthermore,

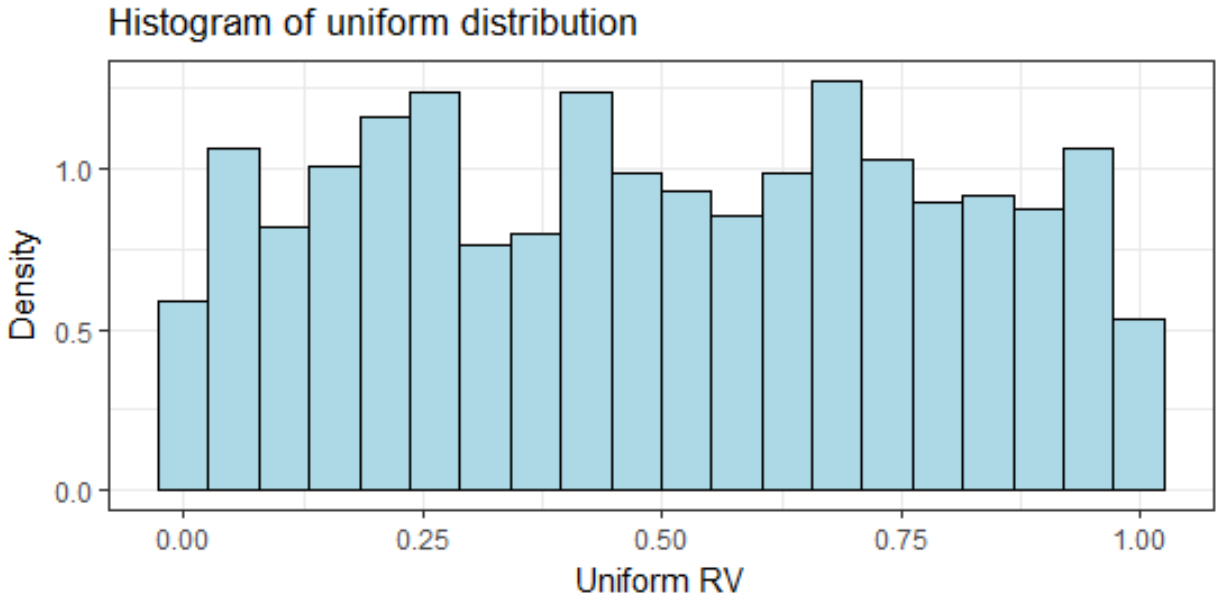
$$\frac{b\left(y+b(y)x\right)}{b(y)} \to u^{\gamma} = 1+\gamma x$$

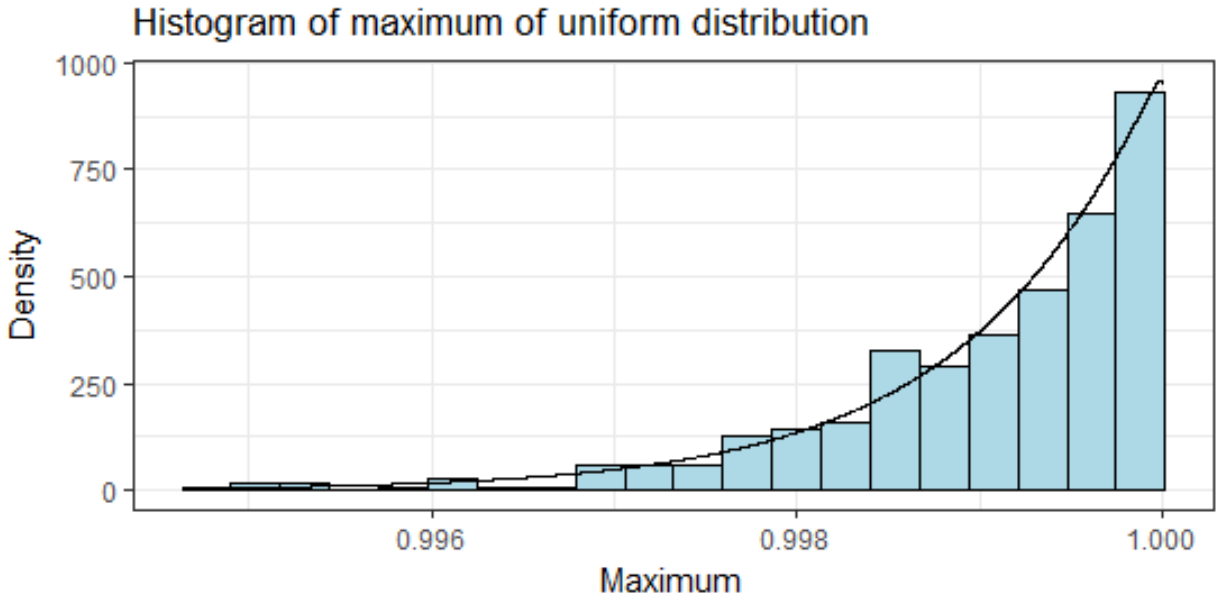
# Extreme value theory: Pickands-Balkema-de Haan theorem

- Essentially, this theorem states that there exists a high enough threshold $t$ such that the exceedances $Z = X - t$ are approximately generalised Pareto (GP) distributed. Hence, for a large threshold $t$,

$$P\left(Z > z \middle| X > t\right) \approx \left(1 + \gamma \frac{z}{b(t)}\right)^{-1/\gamma}$$

# Example: Uniform distribution

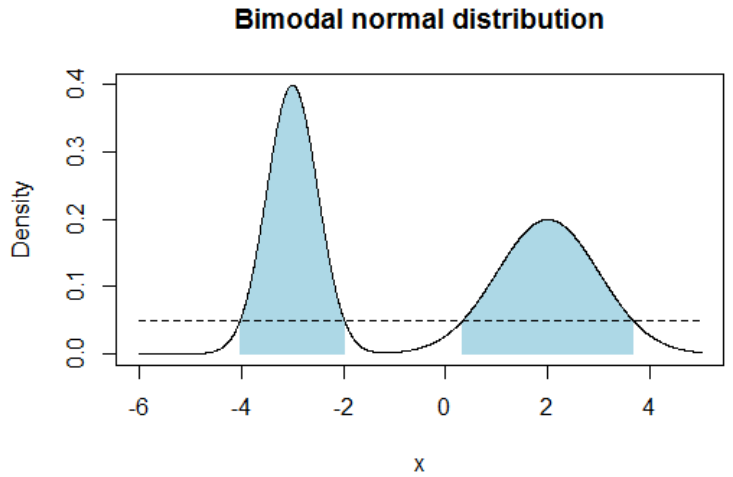


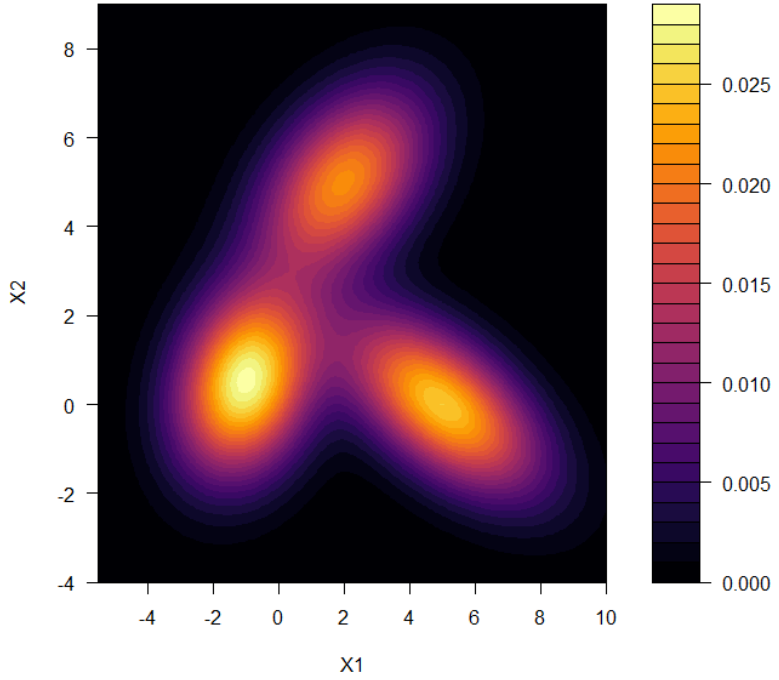



# Other problems with EVT

- The problem is multivariate
- The distribution under normal conditions is multimodal

Hence, one needs a method that transforms the data to overcome these issues.



# An approach based on minimum probability density

- Redefine extreme value theory in terms of minimum probability density.
- Let $E_n = \arg\min_{\underline{X}_i; i=1,\ldots,n} \left\{ f\left(\underline{X}_i\right) \right\}$ such that $f\left(E_n\right) = \min_i \left\{ f\left(\underline{X}_i\right) \right\} \equiv \min_i Y_i$
- Assume $\underline{X} \sim N\left(\underline{\mu}, \Sigma\right)$
- It can be shown that

$$P\left[ f\left(E_n\right) \le y \right] \approx 1 - \exp\left\{ -\left[ a_n^{-1} y \right] \right\} \equiv \text{ Weibull type GEV}$$

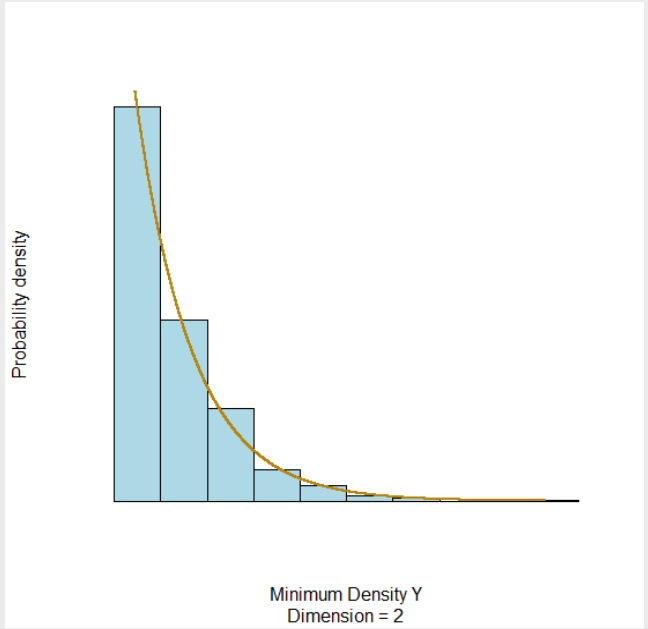
Furthermore, we can choose $a_n = G_d^{-1}\left(1/n\right)$ where $G_d\left(y\right)$ is the known distribution of $Y = f\left(\underline{X}\right)$.

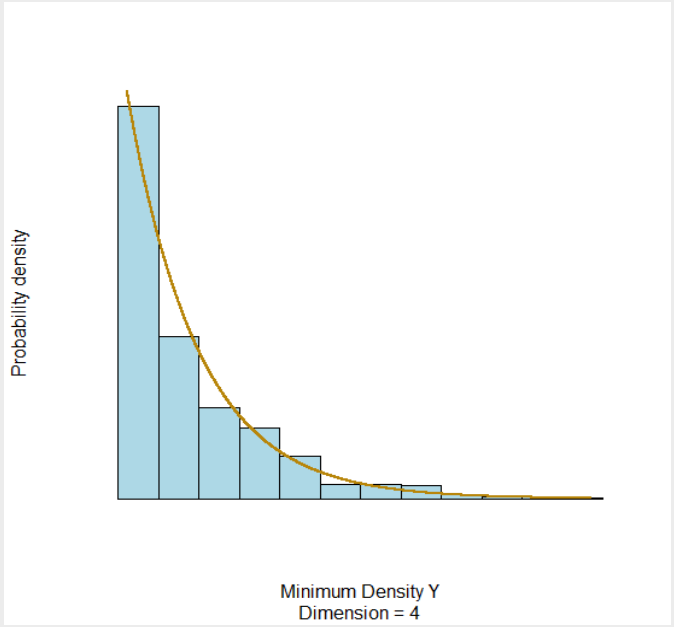
# An approach based on minimum probability density

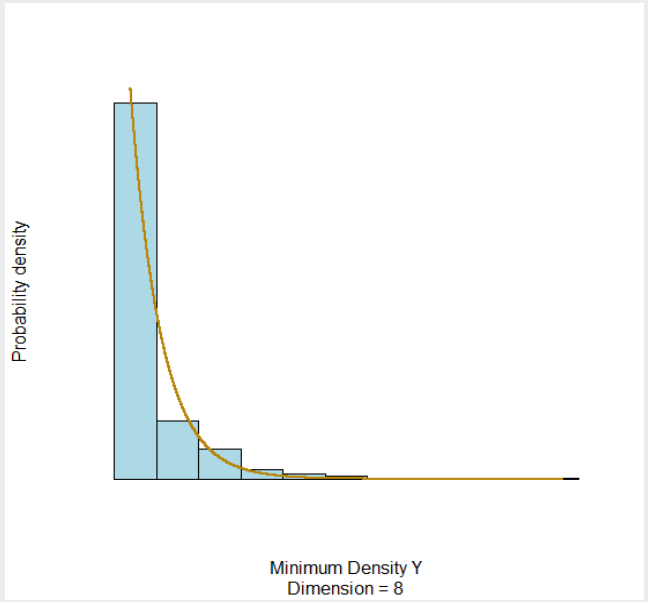
- Hence, the probability that a new observation $\underline{x}^*$ is novel is given by the probability that the density estimate at this observation $y^* = f\left(\underline{x}^*\right)$ is less than the distribution of minimum probability density, *i.e.*:

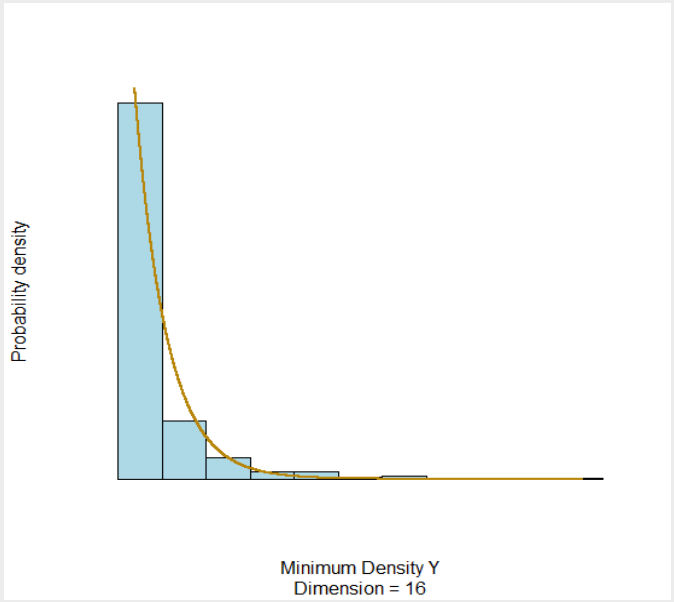
$$P\left(\underline{x}^* \text{ is novel}\right) = P\left(f\left(E_n\right) > y^*\right) \approx \exp\left\{-a_n^{-1} y^*\right\}$$

# An approach based on minimum probability density



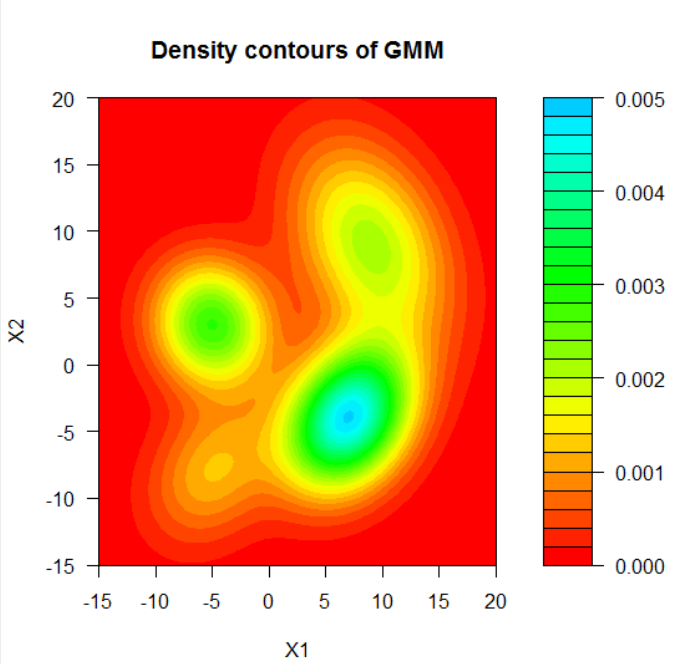



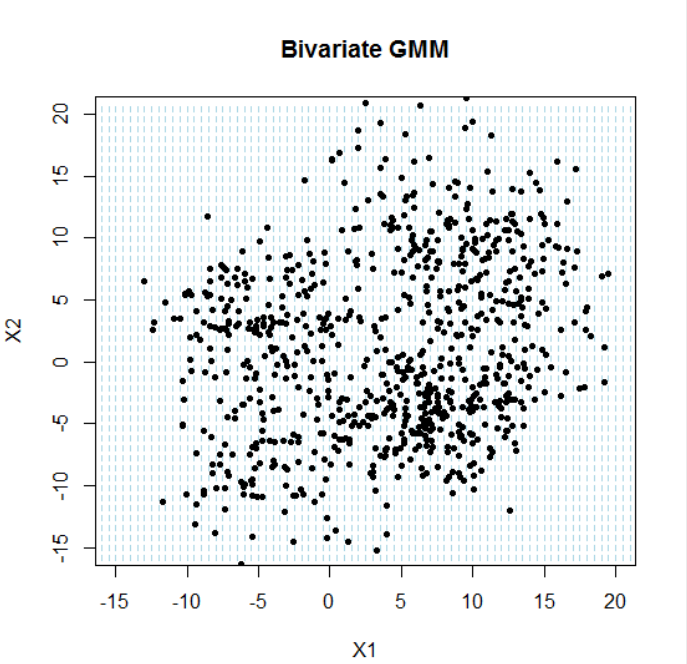



# An approach based on minimum probability density

Problem: Gaussian assumption is too strict.



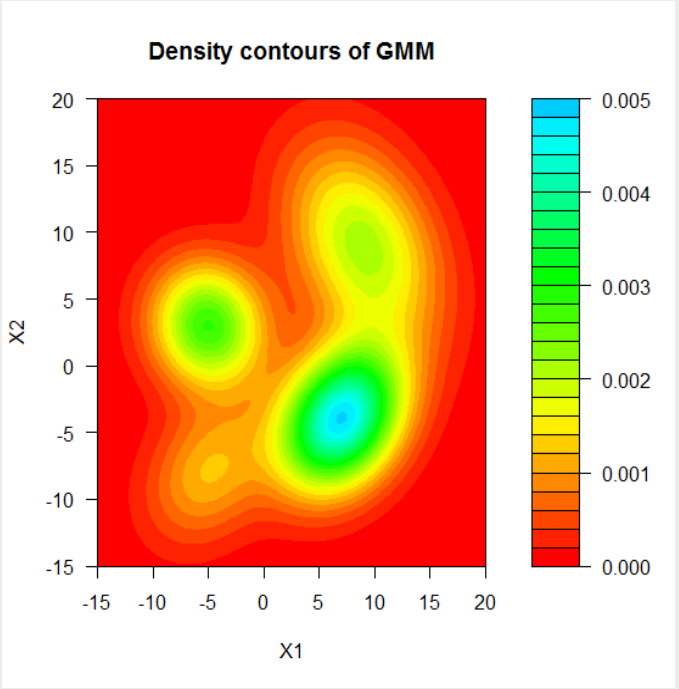



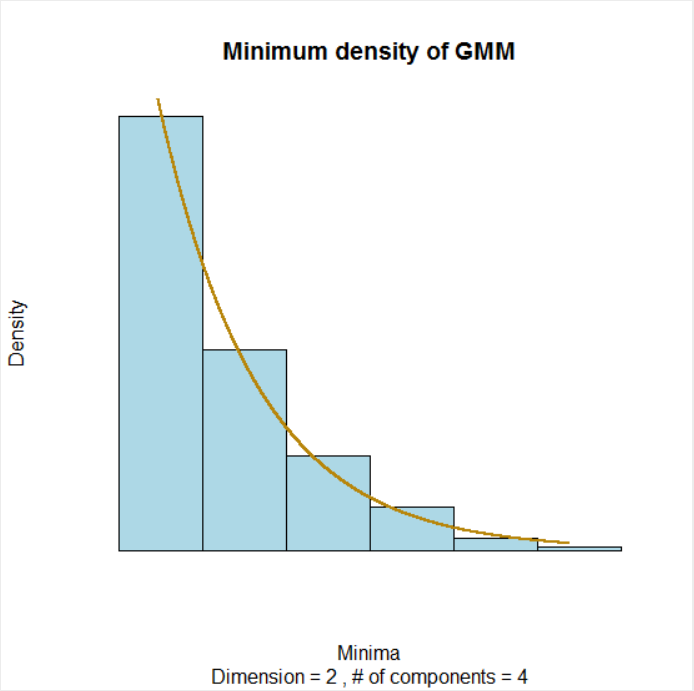
# An approach based on minimum probability density

- Gaussian assumption leads to analytical expression of parameter estimates.
- Minimum of GMM density bounded at zero.
- Hence, density of GMM is in domain of attraction of Weibull type GEV.
- However, parameters must be estimated via maximum likelihood.

# An approach based on minimum probability density

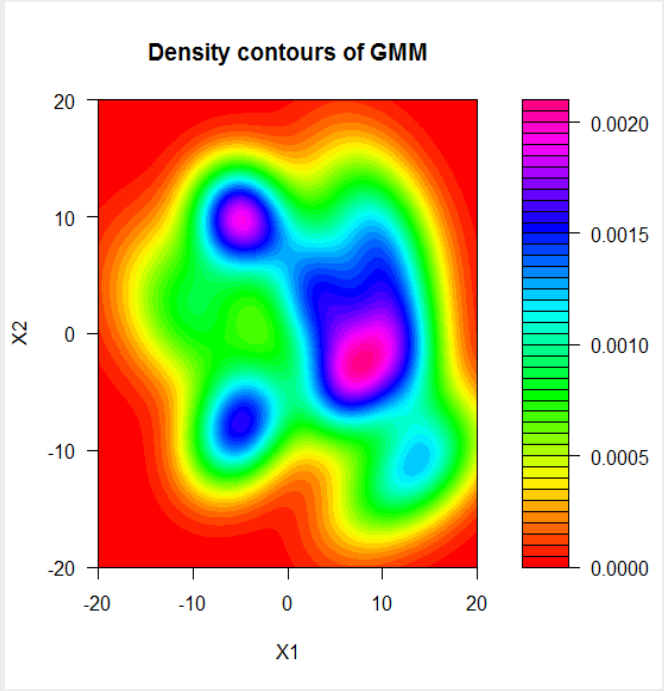
## Weibull density of GMM minimum density:

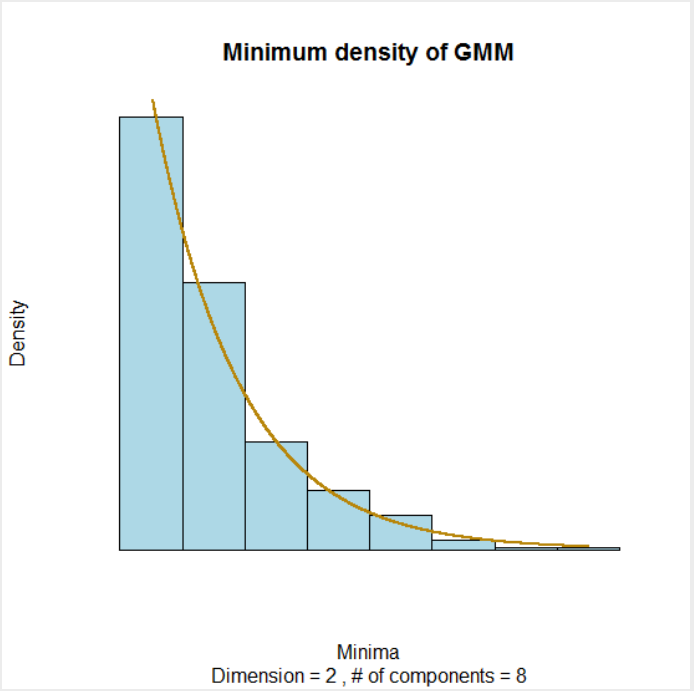


# An approach based on minimum probability density

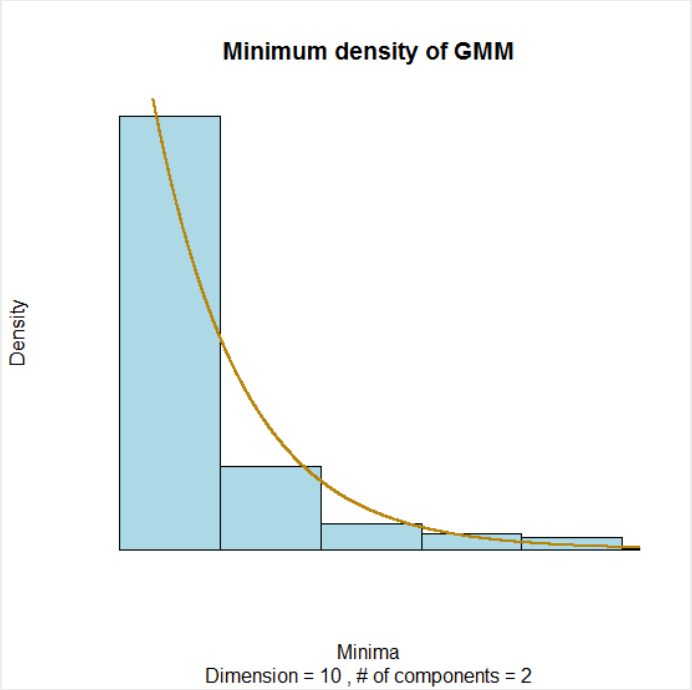
## Weibull density of GMM minimum density :

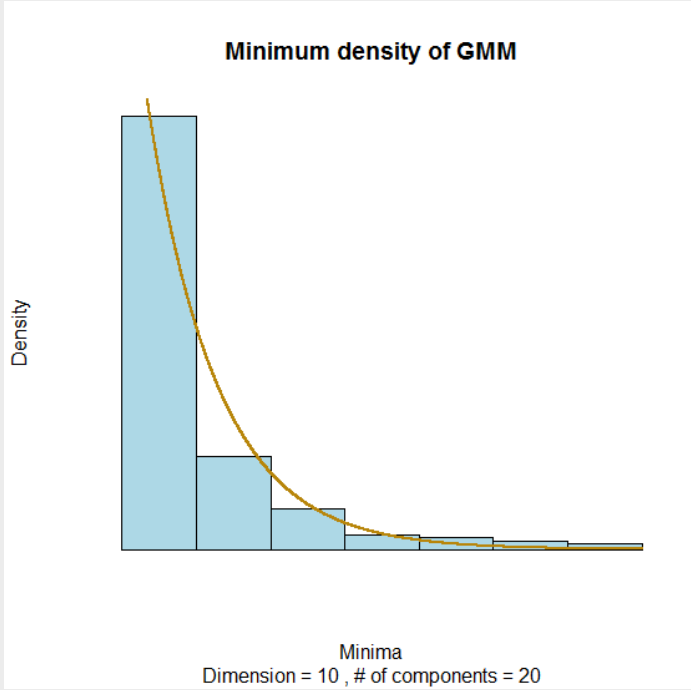


# An approach based on minimum probability density

## Weibull density of GMM minimum density :



# Banknote authentication example

- Dataset: Wavelet transform of banknotes – variables are variance, skewness, kurtosis and entropy of Wavelet transformed image.
- There are 600 real banknotes in the training data.
- There are 162 real and 610 forged banknotes in the test set.

# Banknote authentication example

- Select number of components in GMM with BIC criterion.
- Optimal was 5 Gaussian components.
- Estimate distribution of minimum density of real banknotes using Weibull GEV of minimum density.
- Use this distribution to determine probability of forged banknote on test set.

# Banknote authentication example

- Results:

| Predicted | Response | |
|---|---|---|
| | Real | Forged |
| Real | **162** | **1** |
| Forged | **0** | **609** |

- Clearly, the model does very well in detecting fake banknotes.
- However, very easy data.

# Supervised novelty detection and Open-set Recognition

- Open-set recognition: Perform classification under the assumption that not all classes are known at training.
- Use extreme value theory to detect new classes.
- Similar concepts used for supervised novelty detection.

# A new approach based on the GP distribution

- Problem: Testing set possibly contains classes not seen at training.
- Use a supervised model to classify known classes.
- Use extreme value theory to adjust predicted probabilities to account for other classes.
- Estimate the probability that an observation is from a new class not seen at training.

# A new approach based on the GP distribution

Consider a model that produces $P\left(Y=k\middle|\underline{x}\right)$ , $k=1,2,\ldots,K$

For each class:

1. Find the correctly classified training data $\underline{x}_{jk}=\underline{x}_j\,|\left(\hat{y}=k\right)$ , $j=1,\ldots n_k$
2. Let $\underline{\mu}_k=mean\left(\underline{x}_{jk}\right)$ and compute $d_{jk}=\left\|\underline{x}_{jk}-\underline{\mu}_k\right\|$
3. Fit a GP distribution to the exceedances $Z_{jk}=D_{jk}-t_k$ above a threshold $t_k$.

The probability that an observation $\underline{x}$ is *not novel* with respect to class *k* is:

$P\left(Z_k>z\,|\,D_k>t_k\right)$ where $Z=D-t_k$ and $D=\left\|\underline{X}-\underline{\mu}_k\right\|$.

Notice a per-class estimation strategy is followed.

# A new approach based on the GP distribution

Update probabilities:

We update each class probability with

$$
\begin{aligned}
& P^{new}\left(Y=k \middle| \underline{X}=\underline{x}^{*}\right) \\
&= P\left(\{Y=k\} \cap \{Z_k > z_k\} \middle| \underline{X}=\underline{x}^{*}\right) \\
&= P\left(Y=k \middle| \underline{X}=\underline{x}^{*}\right) \cdot P\left(Z_k > z_k \middle| Y=k,\ \underline{X}=\underline{x}^{*}\right) \\
&\approx P\left(Y=k \middle| \underline{X}=\underline{x}^{*}\right) \cdot \left(1+\gamma_k {z_k}/{\sigma_k}\right)^{-1/\gamma_k}
\end{aligned}
$$

The probability that an observation is from none of the classes is then

$$P\left(Y \text{ novel}\right) = 1 - \sum_k P^{new}\left(Y=k \mid \underline{X}=\underline{x}^{*}\right)$$

Classify as class with maximum probability.

# Handwritten digits example

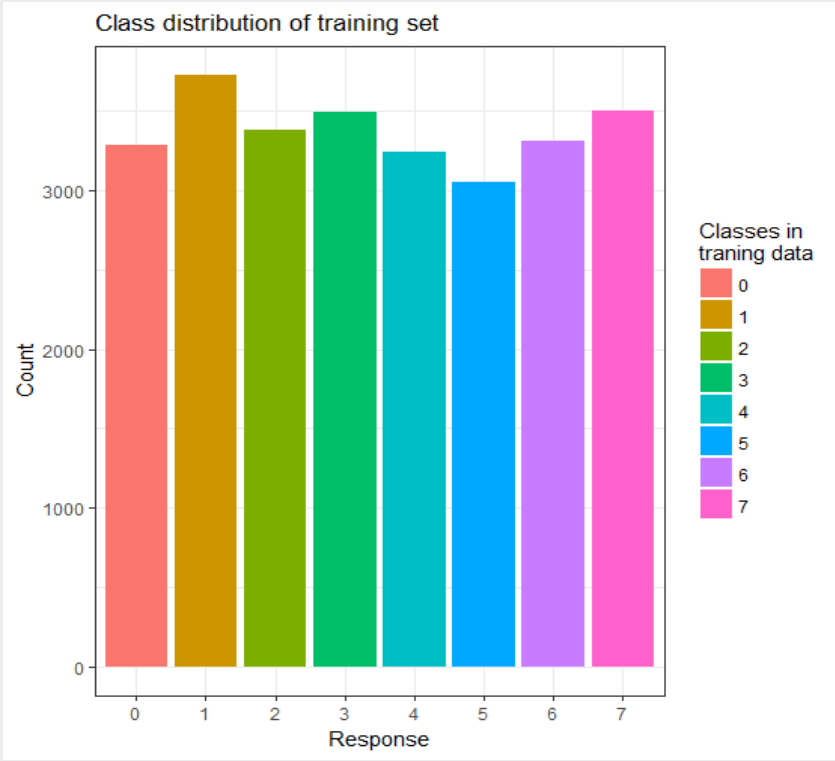
Approach:

- Images of handwritten digits downloaded from Kaggle.
- Use 0 to 7 as known classes in training data.
- Use 0 to 9 in testing data, *i.e.* 8 and 9 are new classes.
- Fit CNN on training data and find correctly classified training data.
- Extract activations in final hidden layer for each classes' correctly classified training data.
- Use these features to estimate probability that an observation is from a new class.

# Handwritten digits example

## Training data:

| Class | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| Observations | 3285 | 3728 | 3382 | 3496 | 3243 | 3054 | 3312 | 3501 |



# Handwritten digits example

CNN model:

- Two convolutional layers and four fully connected layers.
- Use ReLU activations on all hidden layers.
- SoftMax activation on output layer.
- Extract correctly classified training data on the final fully connected layer.
- Split data by class, *i.e.* 1 dataset of correctly classified training data for each class. Each dataset contains output of the ultimate hidden layer.

# Handwritten digits example

## Training results:

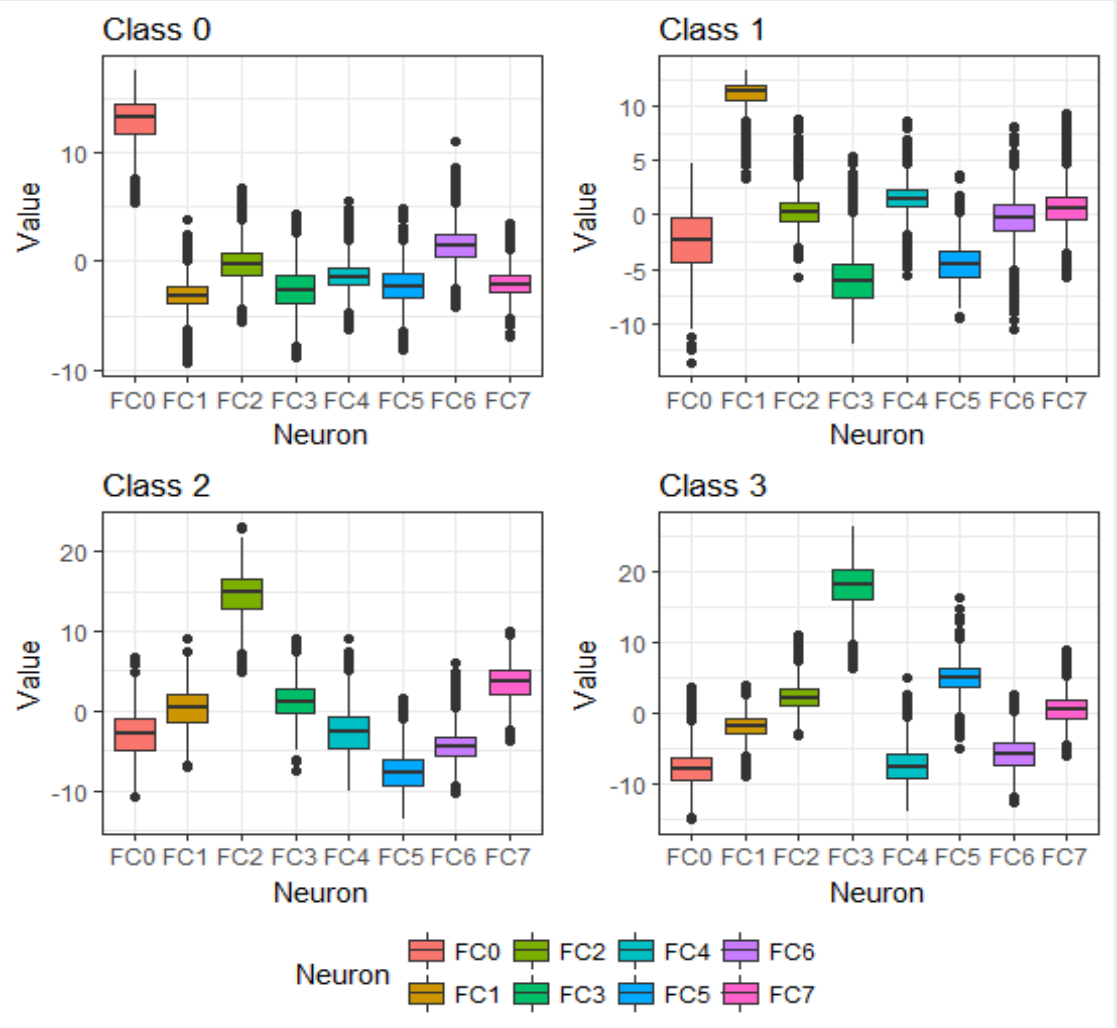
| | Response | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Prediction | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 0 | 3284 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 1 | 0 | 3711 | 0 | 0 | 1 | 0 | 0 | 2 |
| 2 | 0 | 2 | 3379 | 1 | 0 | 0 | 0 | 3 |
| 3 | 0 | 0 | 2 | 3493 | 0 | 0 | 0 | 0 |
| 4 | 0 | 2 | 0 | 0 | 3240 | 0 | 1 | 0 |
| 5 | 0 | 0 | 0 | 1 | 0 | 3046 | 0 | 0 |
| 6 | 1 | 3 | 0 | 0 | 1 | 8 | 3310 | 0 |
| 7 | 0 | 9 | 1 | 1 | 1 | 0 | 0 | 3496 |

## Misclassification error:

0,156%

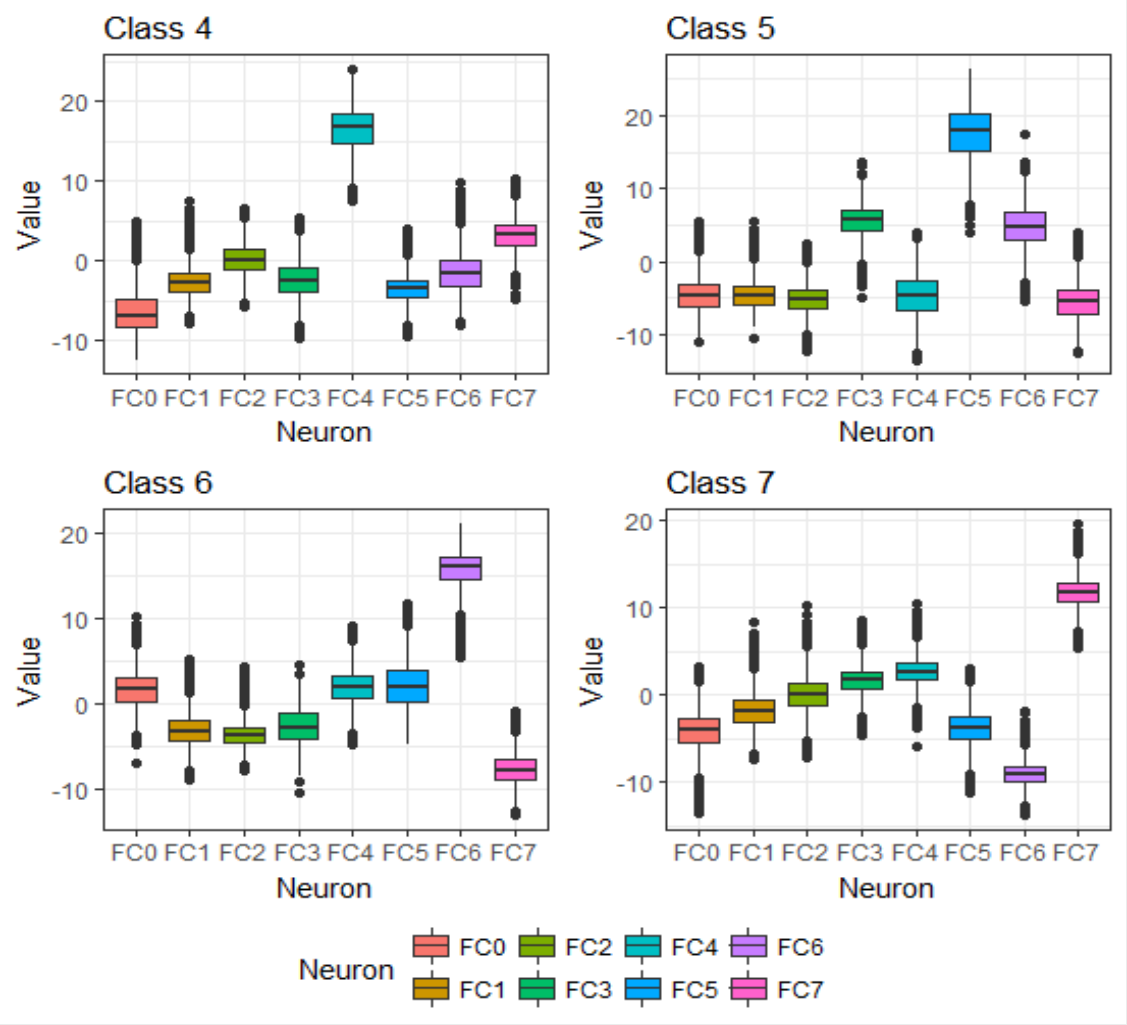
# Handwritten digits example

Training results:



# Handwritten digits example

Training results:



# Handwritten digits example

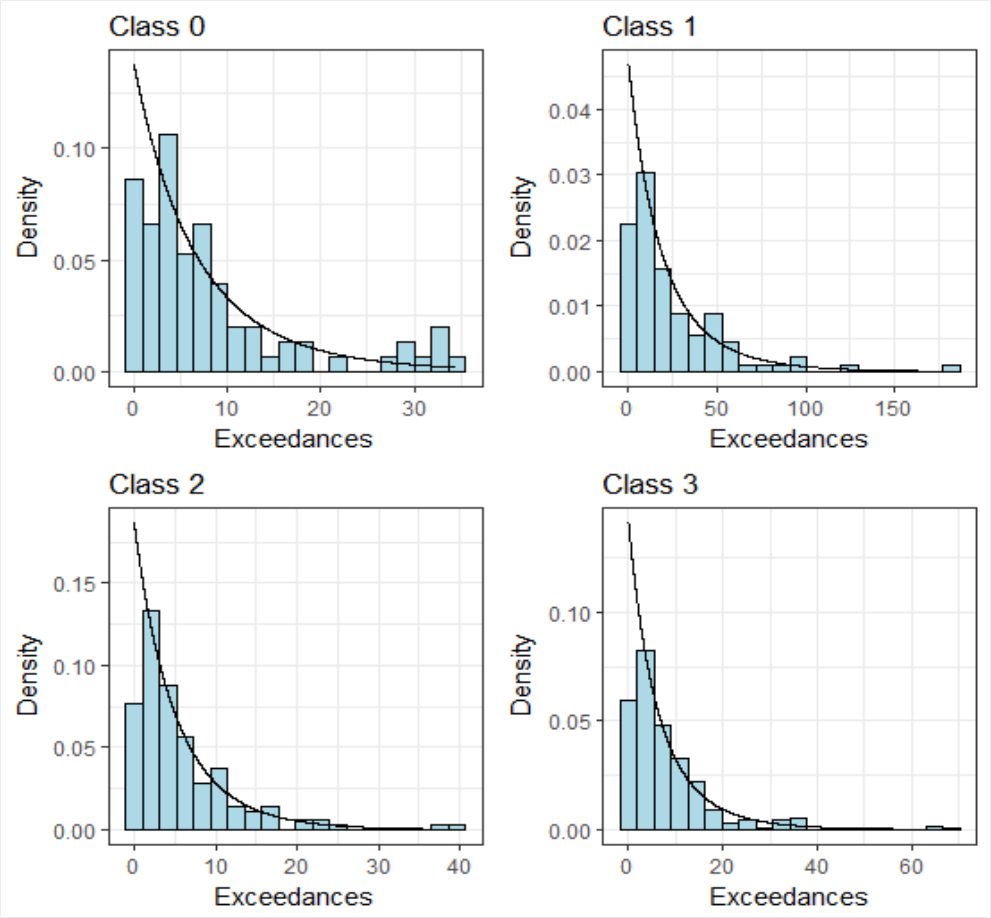
Estimate GP distribution for rescaling:

For each class:

- Use correctly classified training data.
- Find Mahalanobis distance for each observation.
- Select a high threshold.
- Estimate GP distribution of exceedances above the threshold.

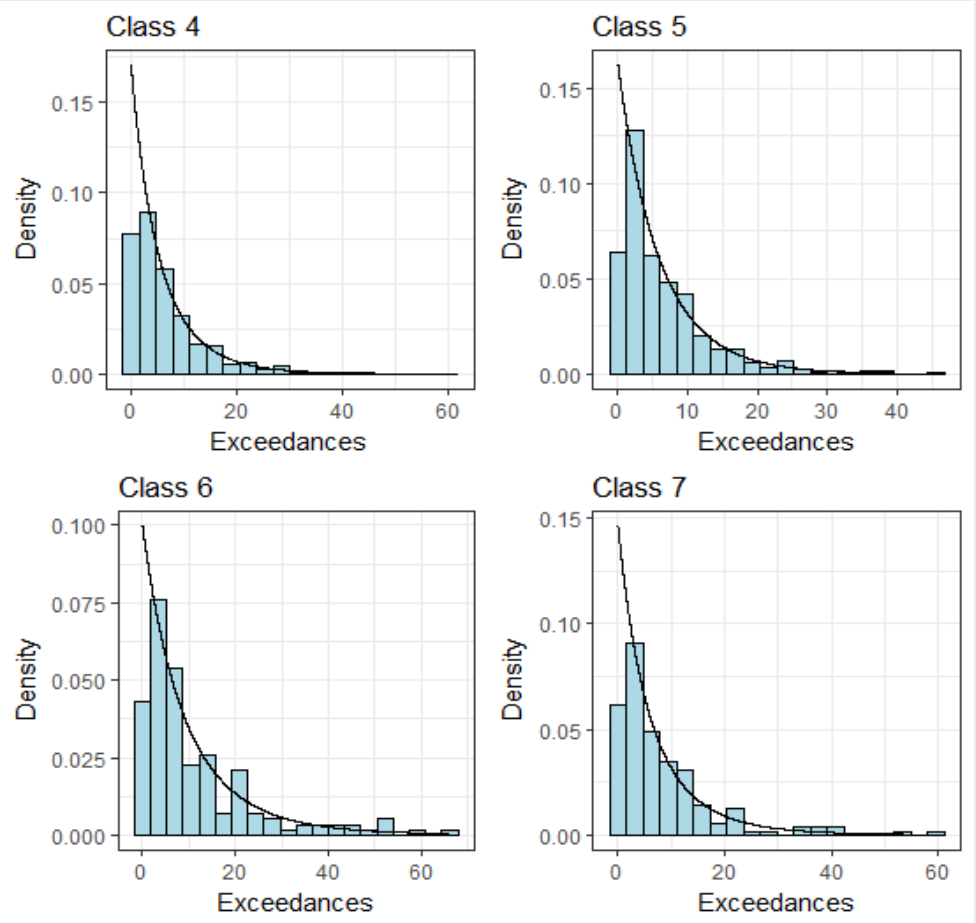
# Handwritten digits example

Estimate GP distribution for rescaling:



# Handwritten digits example

Estimate GP distribution for rescaling:



# Handwritten digits example

Rescale class probabilities of test set:

- Extract activations on final hidden layer of test data.
- Find model predictions.
- For the predicted class, use that GP distribution to rescale probability.
- Classify to class with maximum class probability.

# Handwritten digits example

## Results on testing set:

| | Response | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Prediction | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Unknown |
| 0 | 834 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 1 | 0 | 918 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| 2 | 0 | 0 | 760 | 0 | 0 | 0 | 0 | 0 | 8 |
| 3 | 0 | 0 | 0 | 794 | 0 | 0 | 0 | 0 | 23 |
| 4 | 0 | 0 | 0 | 0 | 706 | 0 | 0 | 0 | 19 |
| 5 | 0 | 0 | 0 | 0 | 0 | 682 | 1 | 0 | 14 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 | 791 | 0 | 8 |
| 7 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 869 | 22 |
| Unknown | 13 | 37 | 33 | 61 | 123 | 59 | 33 | 31 | 1548 |

## Test error:

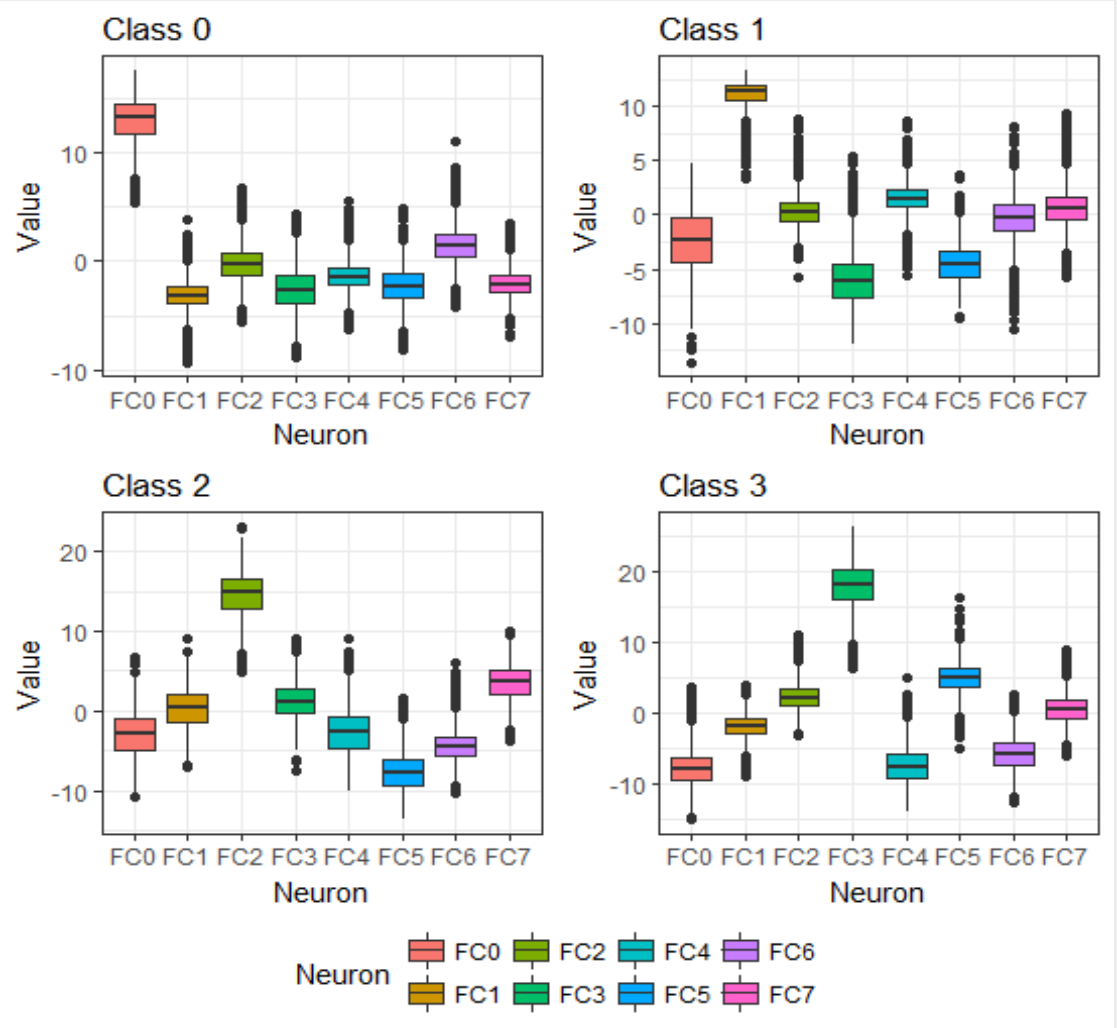
5.91%

## Test error without rescaling:

20.08%

# Handwritten digits example

Perhaps a better model:



# Handwritten digits example

Perhaps a better model:

- Each activation in the final layer is large when an observation is from the corresponding class.

For each class:

- Find output on corresponding node of correctly classified training data.
- Fit GP distribution on peaks below a small threshold.
- Use this probability to rescale.

# Handwritten digits example

## Perhaps a better model:

| | Response | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Prediction | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Unknown |
| 0 | 842 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 |
| 1 | 0 | 930 | 0 | 0 | 0 | 0 | 0 | 0 | 8 |
| 2 | 0 | 1 | 781 | 0 | 1 | 0 | 0 | 0 | 10 |
| 3 | 0 | 0 | 0 | 768 | 0 | 0 | 0 | 0 | 33 |
| 4 | 0 | 0 | 0 | 0 | 756 | 0 | 0 | 0 | 11 |
| 5 | 0 | 0 | 0 | 0 | 0 | 700 | 1 | 0 | 28 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 | 811 | 0 | 14 |
| 7 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 846 | 52 |
| Unknown | 5 | 25 | 13 | 87 | 71 | 41 | 13 | 54 | 1486 |

## Test error:

5.69%

# An idea for random forests

Question: Can a regression tree be used for density estimation?

If so, can we use this model to detect anomalous observations?

Main problem: Need a valid splitting criterion to determine optimal split recursively.

Criminisi, Shotton & Konukoglu (2011) proposed the information gain with the continuous entropy of a multivariate Gaussian.

# An idea for random forests

Consider splitting the root node into two decision nodes. Let the data in the root node be denoted by the set $S$ and let the left and right decision nodes be denoted by $S_L$ and $S_R$, respectively.

The information gain of this split (for the multivariate Gaussian) is then

$$I = \ln|\Sigma| - \frac{|S_L|}{|S|}\ln|\Sigma_L| - \frac{|S_R|}{|S|}\ln|\Sigma_R|.$$

This splitting criterion is used with recursive binary partitioning to grow a *density* tree.

# An idea for random forests

The density estimate is obtained from the Gaussian distribution in each terminal node.

Let the leaf reached by an input $x$ be denoted by $l(x)$.

Then, the probability density at the input $x$ is given by

$f(x) = \frac{\pi_{l(x)}}{K} \phi\left(x, \mu_{l(x)}, \Sigma_{l(x)}\right)$, where

$K$ is a normalising constant, $\pi_{l(x)}$ is the proportion of observations in that node and $\phi(\cdot)$ is the multivariate Gaussian density.

# An idea for random forests

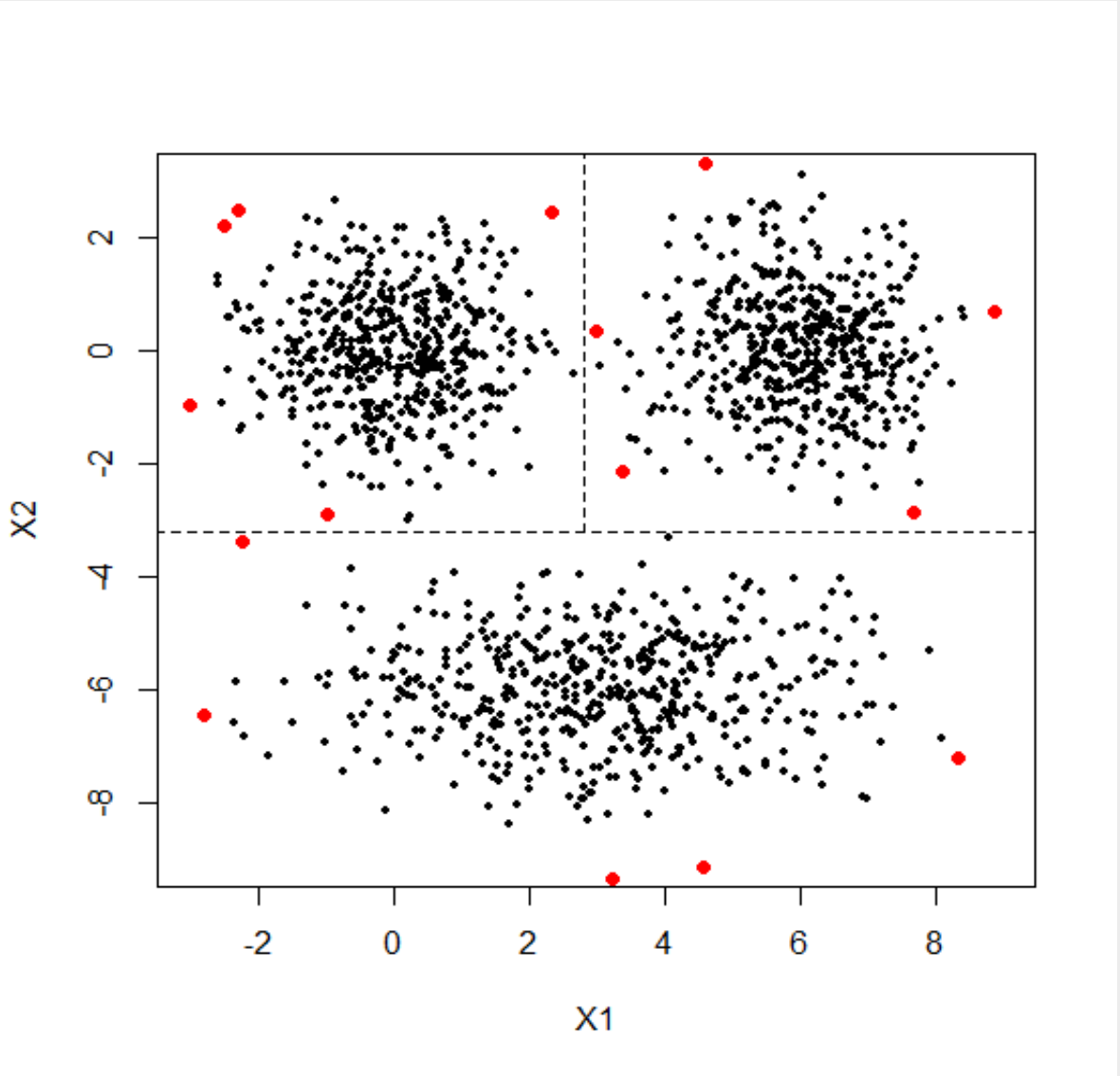
The normalising constant is given as

$$K = \sum_{l} \int_{\underline{x} \in l(\underline{x})} \pi_l \phi\left(\underline{x}, \underline{\mu}_l, \Sigma_l\right) d\underline{x}.$$

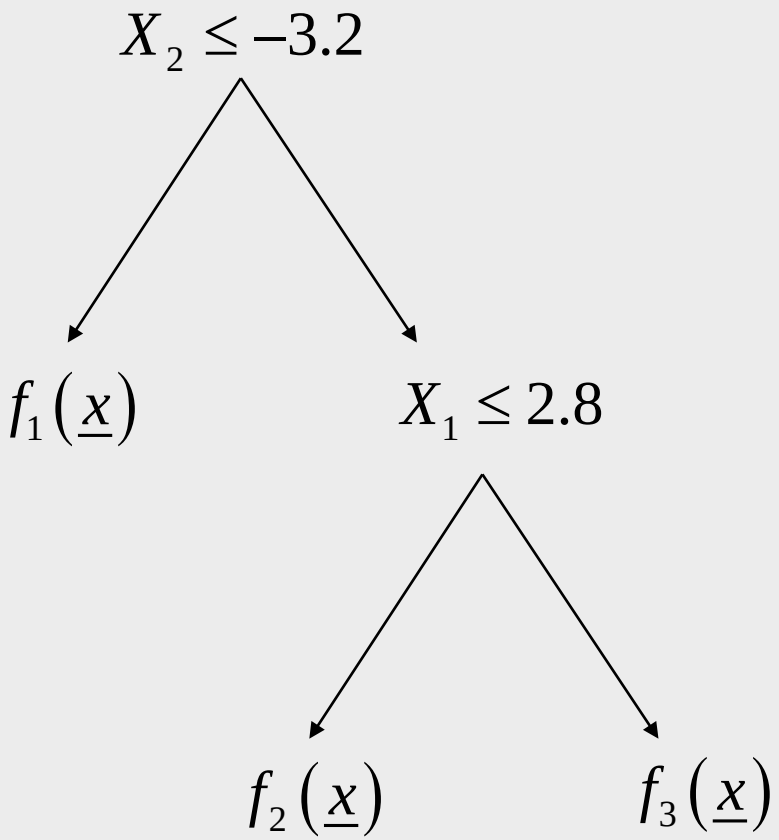
For each leaf, its density estimates are used to estimate the GP distribution associated with the peaks below a small threshold.

This distribution is then used to detect if a new observation is anomalous.

# An idea for random forests: An example



# An idea for random forests : An example

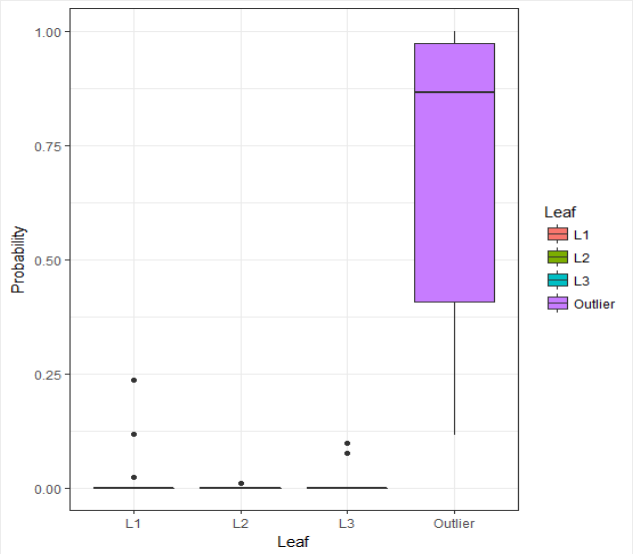


# An idea for random forests : An example

$$K = \sum_{l} \int_{\underline{x} \in l(\underline{x})} \pi_l \phi\left(\underline{x}, \underline{\mu}_l, \Sigma_l\right) d\underline{x} \approx 0.988$$

Consequently, the density in each region is

$$f(x) = \frac{\pi_{l(x)}}{0.988} \phi\left(x, \mu_{l(x)}, \Sigma_{l(x)}\right).$$



# Conclusion

- New idea for novelty detection.
- Unsupervised:
  - Estimate a density or similarity measure.
  - Perform EVT to detect anomalies.
- Supervised:
  - Estimate probabilities/scores for each well-sampled class.
  - Use EVT to rescale probabilities/scores to detect new classes.
- New connection between theoretical statistics and computer science.
- Thanks for listening!